ENCYCLOPEDIA OF

MEDICAL DEVICES AND INSTRUMENTATION

Second Edition

VOLUME 5

Nanoparticles – Radiotherapy Accessories

ENCYCLOPEDIA OF MEDICAL DEVICES AND INSTRUMENTATION, SECOND EDITION

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

Editorial Board
David Beebe
University of Wisconsin–Madison

Jerry M. Calkins
University of Arizona College of Medicine

Michael R. Neuman
Michigan Technological University

Joon B. Park
University of Iowa

Edward S. Sternick
Tufts–New England Medical Center

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 5

Nanoparticles – Radiotherapy Accessories

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

Copyright © 2006 by John Wiley & Sons, Inc. All rights reserved.

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

Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222, Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at http://www.wiley.com/go/permission.

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-13 978-0-4070-04070-6 (v. 5 : cloth)
ISBN-10 0-471-04070-X (v. 5 : cloth)

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

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

E555 2006]
R856.A3E53 2006
610.2803–dc22
2005028946

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTRIBUTOR LIST

Abdel Hady, Mazen, *McMaster University, Hamilton, Ontario Canada*, Bladder Dysfunction, Neurostimulation of

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

Abreu, Beatriz C., *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

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

Ali, Abbas, *University of Illinois, at Urbana-Champaign*, Bioinformatics

Ali, Müftü, *School of Dental Medicine, Boston, Massachusetts*, Tooth and Jaw, Biomechanics of

Alperin, Noam, *University of Illinois at Chicago, Chicago, Illinois*, Hydrocephalus, Tools for Diagnosis and Treatment of

Anson, Denis, *College Misericordia, Dallas, Pennsylvania*, Environmental Control

Arena, John C., *VA Medical Center and Medical College of Georgia*, Biofeedback

Ariel, Gideon, *Ariel Dynamics, Canyon, California*, Biomechanics of Exercise Fitness

Armstrong, Steve, *University of Iowa, Iowa City, Iowa*, Biomaterials for Dentistry

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

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

Ayres, Virginia M., *Michigan State University, East Lansing, Michigan*, Microscopy, Scanning Tunneling

Azangwe, G., Ligament and Tendon, Properties of

Back, Lloyd H., *California Institute of Technology, Pasadena, California*, Coronary Angioplasty and Guidewire Diagnostics

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

Bandyopadhyay, Amit, *Washington State University, Pullman, Washington*, Orthopedic Devices, Materials and Design for

Banerjee, Rupak K., *University of Cincinnati, Cincinnati, Ohio*, Coronary Angioplasty and Guidewire Diagnostics

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

Barker, Steven J., *University of Arizona, Tucson, Arizona*, Oxygen Monitoring

Barth, Rolf F., *The Ohio State University, Columbus, Ohio*, Boron Neutron Capture Therapy

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

Belforte, Guido, *Politecnico di Torino – Department of Mechanics*, Laryngeal Prosthetic Devices

Benkeser, Paul, *Georgia Institute of Technology, Atlanta, Georgia*, Biomedical Engineering Education

Bennett, James R., *University of Iowa, Iowa City, Iowa*, Digital Angiography

Bersano-Begey, Tommaso, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

Biggs, Peter J., *Harvard Medical School, Boston, Massachusetts*, Radiotherapy, Intraoperative

Biyani, Ashok, *University of Toledo, and Medical College of Ohio, Toledo, Ohio*, Human Spine, Biomechanics of

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

Blue, Thomas E., *The Ohio State University, Columbus, Ohio*, Boron Neutron Capture Therapy

Blumsack, Judith T., *Disorders Auburn University, Auburn, Alabama*, Audiometry

Bogan, Richard K., *University of South Carolina, Columbia, South Carolina*, Sleep Laboratory

Bokros, Jack C., *Medical Carbon Research Institute, Austin, Texas*, Biomaterials, Carbon

Bongioannini, Guido, *ENT Division Mauriziano Hospital, Torino, Italy*, Laryngeal Prosthetic Devices

Borah, Joshua, *Applied Science Laboratories, Bedford, Massachusetts*, Eye Movement, Measurement Techniques for

Borden, Mark, *Director of Biomaterials Research, Irvine, California*, Biomaterials, Absorbable

Borton, Bettie B., *Auburn University Montgomery, Montgomery, Alabama*, Audiometry

Borton, Thomas E., *Auburn University Montgomery, Montgomery, Alabama*, Audiometry

Bose Susmita,, *Washington State University, Pullman, Washington*, Orthopedic Devices, Materials and Design for

Bova, Frank J., *M. D. Anderson Cancer Center Orlando, Orlando, FL*, Radiosurgery, Stereotactic

Brenner, David J., *Columbia University Medical Center, New York, New York*, Computed Tomography Screening

Brewer, John M., *University of Georgia*, Electrophoresis

Brian, L. Davis, *Lerner Research Institute, The Cleveland Clinic Foundation, Cleveland, Ohio*, Skin, Biomechanics of

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

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

Brozik, Susan M., *Sandia National Laboratories, Albuquerque, New Mexico*, Microbial Detection Systems

Bruner, Joseph P., *Vanderbilt University Medical Center, Nashville, Tennessee*, Intrauterine Surgical Techniques

Brunswig Newring, Kirk A., *University of Nevada, Reno, Nevada*, Sexual Instrumentatio n

Bruyant, Philippe P., *University of Massachusetts, North Worcester, Massachusetts*, Nuclear Medicine, Computers in

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

Calkins, Jerry M., *Defense Research Technologies, Inc., Rockville, Maryland*, Medical Gas Analyzers

Cannon, Mark, *Northwestern University, Chicago, Illinois*, Resin-Based Composites

Cappelleri, Joseph C., *Pfizer Inc., Groton, Connecticut*, Quality-of-Life Measures, Clinical Significance of

Cardoso, Jorge, *University of Madeira, Funchal, Portugal*, Office Automation Systems

Carello, Massimiliana, *Politecnicodi Torino – Department of Mechanics*, Laryngeal Prosthetic Devices

Caskey, Thomas C., *Cogene Biotech Ventures, Houston, Texas*, Polymerase Chain Reaction

Ceccio, Steven, *University of Michigan, Ann Arbor, Michigan*, Heart Valve Prostheses, In Vitro Flow Dynamics of

Chan, Jackie K., *Columbia University, New York, New York*, Photography, Medical

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

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

Chavez, Eliana, *University of Pittsburgh, Pittsburgh, Pennsylvania*, Mobility Aids

Chen, Henry, *Stanford University, Palo Alto, California*, Exercise Stress Testing

Chen, Jiande, *University of Texas Medical Branch, Galveston, Texas*, Electrogastrogram

Chen, Yan, *Lerner Research Institute, The Cleveland Clinic Foundation, Cleveland, Ohio*, Skin, Biomechanics of

Cheyne, Douglas, *Hospital for Sick Children Research Institute*, Biomagnetism

Chui, Chen-Shou, *Memorial Sloan-Kettering Cancer Center, New York, New York*, Radiation Therapy Treatment Planning, Monte Carlo Calculations in

Claxton, Nathan S., *The Florida State University, Tallahassee, Florida*, Microscopy, Confocal

Coderre, Jeffrey A., *Massachus etts Institute of Technology, Cambridge, Massachusetts*, Boron Neutron Capture Therapy

Collins, Beth, *University of Mississippi Medical Center, Jackson, Mississippi*, Hyperbaric Medicine

Collins, Diane, *University of Pittsburgh, Pittsburgh, Pennsylvania*, Mobility Aids

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

Cook, Albert, *University of Alberta, Edmonton, Alberta, Canada*, Communication Devices

Cooper, Rory, *University of Pittsburgh, Pittsburgh, Pennsylvania*, Mobility Aids

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

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

Craig, Leonard, *Feinberg School of Medicine of Northwestern University, Chicago, Illinois*, Ventilators, Acute Medical Care

Cress, Cynthia J., *University of Nebraska, Lincoln, Nebraska*, Communicative Disorders, Computer Applications for

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

Cunningham, John R., *Camrose, Alberta, Canada*, Cobalt 60 Units for Radiotherapy

D'Alessandro, David, *Montefiore Medical Center, Bronx, New York*, Heart–Lung Machines

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

Dadsetan, Mahrokh, *Mayo Clinic, College of Medicine, Rochester, Minnesota*, Microscopy, Electron

Daley, Michael L., *The University of Memphis, Memphis, Tennessee*, Monitoring, Intracranial Pressure

Dan, Loyd, *Linköping University, Linköping, Sweden*, Thermocouples

Das, Rupak, *University of Wisconsin, Madison, Wisconsin*, Brachytherapy, High Dosage Rate

Dattawadkar, Amruta M., *University of Wisconsin, Madison, Madison, Wisconsin*, Ocular Fundus Reflectometry

Davidson, Michael W., *The Florida State University, Tallahassee, Florida*, Microscopy, Confocal

De Luca, Carlo, *Boston University, Boston, Massachusetts*, Electromyography

De Salles, Antonio A.F., *UCLA Medical School, Los Angeles, California*, Stereotactic Surgery

Decau, Sabin, *University of Maryland, School of Medicine*, Shock, Treatment of

Dechow, Paul C., *A & M University Health Science Center, Dallas, Texas*, Strain Gages

Delbeke, Jean, *Catholique University of Louvain, Brussels, Belgium*, Visual Prostheses

Dell'Osso, Louis F., *Case Western Reserve University, Cleveland, Ohio*, Ocular Motility Recording and Nystagmus

Delorme, Arnaud, *University of San Diego, La Jolla, California*, Statistical Methods

Demenkoff, John, *Mayo Clinic, Scottsdale, Arizona*, Pulmonary Physiology

Demir, Semahat S., *The University of Memphis and The University of Tennessee Health Science Center, Memphis, Tennessee*, Electrophysiology

Demling, Robert H., *Harvard Medical School*, Skin Substitute for Burns, Bioactive

Dennis, Michael J., *Medical University of Ohio, Toledo, Ohio*, Computed Tomography

DeSanti, Leslie, *Harvard Medical School*, Skin Substitute for Burns, Bioactive

Deutsch, Steven, *Pennsylvania State University, University Park, Pennsylvania*, Flowmeters

Devineni, Trishul, *Conemaugh Health System*, Biofeedback

Di Bella Edward, V.R., *University of Utah*, Tracer Kinetics

Diakides, Nicholas A., *Advanced Concepts Analysis, Inc., Falls Church, Virginia*, Thermography

Dolan, Patricia L., *Sandia National Laboratories, Albuquerque, New Mexico*, Microbial Detection Systems

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

Douglas, Wilson R., *Children's Hospital of Philadelphia, Philadelphia, Pennsylvania*, Intrauterine Surgical Techniques

Draper, Crissa, *University of Nevada, Reno, Nevada*, Sexual Instrumentation

Drzewiecki, Tadeusz M., *Defense Research Technologies, Inc., Rockville, Maryland*, Medical Gas Analyzers

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

Dyro, Joseph F., *Setauket, New York*, Safety Program, Hospital

Dyson, Mary, *Herts, United Kingdom*, Heat and Cold, Therapeutic

Eckerle, Joseph S., *SRI International, Menlo Park, California*, Tonometry, Arterial

Edwards, Benjamin, *University of Wisconsin-Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

Edwards, Thayne L., *University of Washington, Seattle, Washington*, Chromatography

Eklund, Anders, *University of Illinois at Chicago, Chicago, Illinois*, Hydrocephalus, Tools for Diagnosis and Treatment of

El Solh, Ali A., *Erie County Medical Center, Buffalo, New York*, Sleep Studies, Computer Analysis of

Elmayergi, Nader, *McMaster University, Hamilton, Ontario, Canada*, Bladder Dysfunction, Neurostimulation of

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

Faddy, Steven C., *St. Vincents Hospital, Sydney, Darlinghurst, Australia*, Cardiac Output, Fick Technique for

Fahey, Frederic H., *Childrens Hospital Boston*, Computed Tomography, Single Photon Emission

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

Feldman, Jeffrey, *Childrens Hospital of Philadelphia, Philadelphia, Pennsylvania*, Anesthesia Machines

Fellers, Thomas J., *The Florida State University, Tallahassee, Florida*, Microscopy, Confocal

Ferrara, Lisa, *Cleveland Clinic Foundation, Cleveland, Ohio*, Human Spine, Biomechanics of

Ferrari, Mauro, *The Ohio State University, Columbus, Ohio*, Drug Delivery Systems

Fontaine, Arnold A., *Pennsylvania State University, University Park, Pennsylvania*, Flowmeters

Foust, Milton J., Jr, *Medical University of South Carolina Psychiatry and Behavioral Sciences, Charleston, South Carolina*, Electroconvulsive Therapy

Frasco, Peter, *Mayo Clinic Scottsdale, Scottsdale, Arizona*, Temperature Monitoring

Frazier, James, *Louisiana State University, Baton Rouge, Louisiana*, Ambulatory Monitoring

Freiesleben De Blasio, Birgitte, *University of Oslo, Oslo, Norway*, Impedance Spectroscopy

Fresta, Massimo, *University of Catanzaro Magna Græcia, Germaneto (CZ), Italy*, Drug Delivery Systems

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

Froehlicher, Victor, *VA Medical Center, Palo Alto, California*, Exercise Stress Testing

Fung, Edward K., *Columbia University, New York, New York*, Photography, Medical

Gage, Andrew A., *State University of New York at Buffalo, Buffalo, New York*, Cryosurgery

Gaglio, Paul J., *Columbia University College of Physicians and Surgeons*, Liver Transplantation

Gardner, Reed M., *LDS Hospital and Utah University, Salt Lake City, Utah*, Monitoring, Hemodynamic

Gejerman, Glen, *Hackensack University Medical, Hackensack, New Jersey*, Radiation Therapy, Quality Assurance in

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

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

Glasgow, Glenn P., *Loyola University of Chicago, Maywood, Illinois*, Radiation Protection Instrumentation

Glasgow, Glenn, *University of Wisconsin-Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

Goel, Vijay K., *University of Toledo, and Medical College of Ohio, Toledo, Ohio*, Human Spine, Biomechanics of

Goetsch, Steven J., *San Diego Gamma Knife Center, La Jolla, California*, Gamma Knife

Goldberg, Jay R., *Marquette University Milwaukee, Wisconsin*, Minimally Invasive Surgery

Goldberg, Zelenna, *Department of Radiation Oncology, Davis, California*, Ionizing Radiation, Biological Effects of

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

Gopas, Jacob, *Ben Gurion University of the Negev, Beer Sheva, Israel*, Monoclonal Antibodies

Gorgulho, Alessandra, *UCLA Medical School, Los Angeles, California*, Stereotactic Surgery

Gough, David A., *University of California, La Jolla, California*, Glucose Sensors

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

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

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

Grant, Walter III, *Baylor College of Medicine, Houston, Texas*, Radiation Therapy, Intensity Modulated

Grayden, Edward, *Mayo Health Center, Albertlea, Minnesota*, Cardiopulmonary Resuscitation

Green, Jordan R., *University of Nebraska, Lincoln, Nebraska*, Communicative Disorders, Computer Applications for

Haemmerich, Dieter, *Medical University of South Carolina, Charleston, South Carolina*, Tissue Ablation

Hamam, Habib, *Université de Moncton, Moncton New Brunswick, Canada*, Lenses, Intraocular

Hammond, Paul A., *University of Glasgow, Glasgow, United Kingdom*, Ion-Sensitive Field-Effect Transistors

Hanley, Joseph, *Hackensack University Medical, Hackensack, New Jersey*, Radiation Therapy, Quality Assurance in

Harley, Brendan A., *Massachusetts Institute of Technology*, Skin Tissue Engineering for Regeneration

Harper, Jason C., *Sandia National Laboratories, Albuquerque, New Mexico*, Microbial Detection Systems

Hasman, Arie, *Maastricht, The Netherlands*, Medical Education, Computers in

Hassouna, Magdy, *Toronto Western Hospital, Toronto, Canada*, Bladder Dysfunction, Neurostimulation of

Hayashi, Kozaburo, *Okayama University of Science, Okayama, Japan*, Arteries, Elastic Properties of

Hench, Larry L., *Imperial College London, London, United Kingdom*, Biomaterials: Bioceramics

Hettrick, Douglas A., *Sr. Principal Scientist Medtronic, Inc., Minneapolis, Minnesota*, Bioimpedance in Cardiovascular Medicine

Hirsch-Kuchma, Melissa, *University of Central Florida NanoScience Technology Center, Orlando, Florida*, Biosurface Engineering

Holder, Graham E., *Moorfields Eye Hospital, London, United Kingdom*, Electroretinography

Holmes, Timothy, *St. Agnes Cancer Center, Baltimore, Maryland*, Tomotherapy

Honeyman-Buck, Janice C., *University of Florida, Gainesville, Florida*, Radiology Information Systems

Hooper, Brett A., *Areté Associates, Arlington, Virginia*, Endoscopes

Horn, Bruce, *Kaiser Permanente, Los Angeles, California*, X-Rays Production of

Horner, Patricia I., *Biomedical Engineering Society Landover, Maryland*, Medical Engineering Societies and Organizations

Horowitz, Paul M., *University of Texas, San Antonio, Texas*, Fluorescence Measurements

Hou, Xiaolin, *Risø National Laboratory, Roskilde, Denmark*, Neutron Activation Analysis

Hovorka, Roman, *University of Cambridge, Cambridge, United Kingdom*, Pancreas, Artificial

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

Hunt, Alan J., *University of Michigan, Ann Arbor, Michigan*, Optical Tweezers

Hutten, Helmut, *University of Technology, Graz, Australia*, Impedance Plethysmography

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

Ibbott, Geoffrey S., *Anderson Cancer Center, Houston, Texas*, Radiation Dosimetry, Three-Dimensional

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

Isik, Can, *Syracuse University, Syracuse, New York*, Blood Pressure Measurement

James, Susan P., *Colorado State University, Fort Collins, Colorado*, Biomaterials: Polymers

Jensen, Winnie, *Aalborg University, Aalborg, Denmark*, Electroneurography

Jin, Chunming, *North Carolina State University, Raleigh, North Carolina*, Biomaterials, Corrosion and Wear of

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

Johnson, Arthur T., *University of Maryland College Park, Maryland*, Medical Engineering Societies and Organizations

Jones, Julian R., *Imperial College London, London, United Kingdom*, Biomaterials: Bioceramics

Joshi, Abhijeet, *Abbott Spine, Austin, Texas*, Spinal Implants

Jung, Ranu, *Arizona State University, Tempe, Arizona*, Functional Electrical Stimulation

Jurisson, Silvia S., *University of Missouri Columbia, Missouri*, Radionuclide Production and Radioactive Decay

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

Kamath, Celia C., *Mayo Clinic, Rochester, Minnesota*, Quality-of-Life Measures, Clinical Significance of

Kane, Mollie, *Madison, Wisconsin*, Contraceptive Devices

Katherine, Andriole P., *Harvard Medical School, Boston, Massachusetts*, Picture Archiving and Communication Systems

Katsaggelos, Aggelos K., *Northwestern University, Evanston, Illinois*, DNA Sequencing

Katz, J. Lawrence, *University of Missouri-Kansas City, Kansas City, Missouri*, Bone and Teeth, Properties of

Kesavan, Sunil, *Akebono Corporation, Farmington Hills, Michigan*, Linear Variable Differential Transformers

Khang, Gilson, *Chonbuk National University*, Biomaterials: Tissue Engineering and Scaffolds

Khaodhiar, Lalita, *Harvard Medical School, Boston, Massachusetts*, Cutaneous Blood Flow, Doppler Measurement of

Kim, Moon Suk, *Korea Research Institutes of Chemical Technology*, Biomaterials: Tissue Engineering and Scaffolds

Kim, Young Kon, *Inje University, Kimhae City, Korea*, Alloys, Shape Memory

Kindwall, Eric P., *St. Luke's Medical Center, Milwaukee, Wisconsin*, Hyperbaric Oxygenation

King, Michael A., *University of Massachusetts, North Worcester, Massachusetts*, Nuclear Medicine, Computers in

Klebe, Robert J., *University of Texas, San Antonio, Texas*, Fluorescence Measurements

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

Knoper, Steven R., *University of Arizona College of Medicine*, Ventilatory Monitoring

Kontaxakis, George, *Universidad Politécnica de Madrid, Madrid, Spain*, Positron Emission Tomography

Kottke-Marchant, Kandice, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Vascular Graft Prosthesis

Kripfgans, Oliver, *University of Michigan, Ann Arbor, Michigan*, Ultrasonic Imaging

Kulkarni, Amol D., *University of Wisconsin–Madison, Madison, Wisconsin*, Ocular Fundus Reflectometry, Visual Field Testing

Kumaradas, J. Carl, *Ryerson University, Toronto, Ontario, Canada*, Hyperthermia, Interstitial

Kunicka, Jolanta, *Bayer HealthCare LLC, Tarrytown, New York*, Differential Counts, Automated

Kwak, Kwanj Joo, *University of Miami Miller School of Medicine, Miami, Florida*, Microscopy, Scanning Force

Lakes, Roderic, *University of Wisconsin-Madison*, Bone and Teeth, Properties of

Lakkireddy, Dhanunjaya, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Hyperthermia, Ultrasonic

Larsen, Coby, *Case Western Reserve University, Cleveland, Ohio*, Vascular Graft Prosthesis

Laster, Brenda H., *Ben Gurion University of the Negev, Beer Sheva, Israel*, Monoclonal Antibodies

Latta, Loren, *University of Miami, Coral Gables, Florida*, Rehabilitation, Orthotics in

Leder, Ron S., *Universidad Nacional Autonoma de Mexico Mexico, Distrito Federal*, Continuous Positive Airway Pressure

Lee, Chin, *Harvard Medical School, Boston, Massachusetts*, Radiotherapy Treatment Planning, Optimization of; Hyperthermia, Interstitial

Lee, Hai Bang, *Korea Research Institutes of Chemical Technology*, Biomaterials: Tissue Engineering and Scaffolds

Lee, Sang Jin, *Korea Research Institutes of Chemical Technology*, Biomaterials: Tissue Engineering and Scaffolds

Lei, Liu, *Department of General Engineering, Urbana, Illinois*, Bioinformatics

Lei, Xing, *Stanford University, Stanford, California*, Radiation Dose Planning, Computer-Aided

Lewis, Matthew C., *Medical College of Wisconsin, Milwaukee, Wisconsin*, Hyperbaric Oxygenation

Li, Chaodi, *University of Notre Dame, Notre Dame, Indiana*, Bone Cement, Acrylic

Li, Jonathan G., *University of Florida, Gainesville, Florida*, Radiation Dose Planning, Computer-Aided

Li, Qiao, *University of Michigan, Ann Arbor, Michigan*, Immunotherapy

Li, Yanbin, *University of Arkansas, Fayetteville, Arkansas*, Piezoelectric Sensors

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

Ligas, James, *University of Connecticut, Farmington, Connecticut*, Respiratory Mechanics and Gas Exchange

Limoge, Aime, *The René Descartes University of Paris, Paris, France*, Electroanalgesia, Systemic

Lin, Pei-Jan Paul, *Beth Israel Deaconess Medical Center, Boston, Massachusets*, Mammography

Lin, Zhiyue, *University of Kansas Medical Center, Kansas City, Kansas*, Electrogastrogram

Lineaweaver, William C., *Unive rsity of Mississippi Medical Center, Jackson, Mississippi*, Hyperbaric Medicine

Lipping, Tarmo, *Tampere University of Technology, Pori, Finland*, Monitoring in Anesthesia

Liu, Xiaohua, *The University of Michigan, Ann Arbor, Michigan*, Polymeric Materials

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

Loeb, Robert, *University of Arizona, Tuscon, Arizona*, Anesthesia Machines

Lopes de Melo, Pedro, *State University of Rio de Janeiro, Térreo Salas, Maracanã, Thermistors*

Loudon, Robert G., Lung Sounds

Low, Daniel A., *Washington University School of Medicine, St. Louis, Missouri*, Radiation Therapy Simulator

Lu, Lichun, *Mayo Clinic, College of Medicine, Rochester, Minnesota*, Microscopy, Electron

Lu, Zheng Feng, *Columbia University, New York, New York*, Screen-Film Systems

Lyon, Andrew W., *University of Calgary, Calgary, Canada*, Flame Atomic Emission Spectrometry and Atomic Absorption Spectrometry

Lyon, Martha E., *University of Calgary, Calgary, Canada*, Flame Atomic Emission Spectrometry and Atomic Absorption Spectrometry

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

Macia, Narciso F., *Arizona State University at the Polytechnic Campus, Mesa, Arizona*, Pneumotachometers

Mackenzie, Colin F., *University of Maryland, School of Medicine*, Shock, Treatment of

Mackie, Thomas R., *University of Wisconsin, Madison, Wisconsin*, Tomotherapy

Madnani, Anju, *LSU Medical Centre, Shreveport, Louisiana*, Transcutaneous Electrical Nerve Stimulation (TENS)

Madnani, Sanjay, *LSU Medical Centre, Shreveport, Louisiana*, Transcutaneous Electrical Nerve Stimulation (TENS)

Madsen, Mark T., *University of Iowa, Iowa City, Iowa*, Anger Camera

Magnano, Mauro, *ENT Division Mauriziano Hospital, Torino, Italy*, Drug Delivery Systems

Mandel, Richard, *Boston University School of Medicine, Boston, Massachusetts*, Colorimetry

Manning, Keefe B., *Pennsylvania State University, University Park, Pennsylvania*, Flowmeters

Mao, Jeremy J., *University of Illinois at Chicago, Chicago, Illinois*, Cartilage and Meniscus, Properties of

Marcolongo, Michele, *Drexel University, Philadelphia, Pennsylvania*, Spinal Implants

Marek, Miroslav, *Georgia Institute of Technology, Atlanta, Georgia*, Biomaterials, Corrosion and Wear of

Marion, Nicholas W., *University of Illinois at Chicago, Chicago, Illinois*, Cartilage and Meniscus, Properties of

Masters, Kristyn S., *University of Wisconsin, Madison, Wisconsin*, Tissue Engineering

Maughan, Richard L., *Hospital of the University of Pennsylvania*, Neutron Beam Therapy

McAdams, Eric, *University of Ulster at Jordanstown, Newtownabbey, Ireland*, Bioelectrodes

McArthur, Sally L., *University of Sheffield, Sheffield, United Kingdom*, Biomaterials, Surface Properties of

McEwen, Malcom, *National Research Council of Canada, Ontario, Canada*, Radiation Dosimetry for Oncology

McGowan, Edward J., *E.J. McGowan & Associates*, Biofeedback

McGrath, Susan, *Dartmouth College, Hanover, New Hampshire*, Oxygen Analyzers

Meeks, Sanford L., *University of Florida, Gainesville, Florida*, Radiosurgery, Stereotactic

Melissa, Peter, *University of Central Florida NanoScience Technology Center, Orlando, Florida*, Biosurface Engineering

Mendelson, Yitzhak, *Worcester Polytechnic Institute*, Optical Sensors

Metzker, Michael L., *Baylor College of Medicine, Houston, Texas*, Polymerase Chain Reaction

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

Michler, Robert, *Montefiore Medical Center, Bronx, New York*, Heart–Lung Machines

Micic, Miodrag, *MP Biomedicals LLC, Irvine, California*, Microscopy and Spectroscopy, Near-Field

Miller, William, *University of Missouri Columbia, Missouri*, Radionuclide Production and Radioactive Decay

Mittra, Erik, *Stony Brook University, New York*, Bone Density Measurement

Modell, Mark, *Harvard Medical School, Boston, Massachusetts*, Fiber Optics in Medicine

More, Robert B., *RBMore Associates, Austin, Texas* Biomaterials Carbon

More, Robert, *Austin, Texas*, Heart Valves, Prosthetic

Morrow, Darren, *Royal Adelaide Hospital, Adelaide, Australia*, Intraaortic Balloon Pump

Mourtada, Firas, *MD Anderson Cancer Center, Houston, Texas*, Brachytherapy, Intravascular

Moy, Vincent T., *University of Miami, Miller School of Medicine, Miami, Florida*, Microscopy, Scanning Force

Müftü, Sinan, *Northeastern University, Boston, Massachusetts*, Tooth and Jaw, Biomechanics of

Murphy, Raymond L.H., Lung Sounds

Murphy, William L., *University of Wisconsin, Madison, Wisconsin*, Tissue Engineering

Murray, Alan, *Newcastle University Medical Physics, Newcastle upon Tyne, United Kingdom*, Pace makers

Mutic, Sasa, *Washington University School of Medicine, St. Louis, Missouri*, Radiation Therapy Simulator

Narayan, Roger J., *University of North Carolina, Chapel Hill, North Carolina*, Biomaterials, Corrosion and Wear of

Natale, Andrea, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Hyperthermia, Ultrasonic

Nazeran, Homer, *The University of Texas, El Paso, Texas*, Electrocardiography, Computers in

Neuman, Michael R., *Michigan Technological University, Houghton, Houghton, Michigan*, Fetal Monitoring, Neonatal Monitoring

Neuzil, Pavel, *Institute of Bioengineering and Nanotechnology, Singapore*, Immunologically Sensitive Field-Effect Transistors

Nickoloff, Edward L., *Columbia University, New York, New York*, X-Ray Quality Control Program

Niezgoda, Jeffrey A., *Medical College of Wisconsin, Milwaukee, Wisconsin*, Hyperbaric Oxygenation

Nishikawa, Robert M., *The University of Chicago, Chicago, Illinois*, Computer-Assisted Detection and Diagnosis

Nutter, Brian, *Texas Tech University, Lubbock, Texas*, Medical Records, Computers in

O'Donohue, William, *University of Nevada, Reno, Nevada*, Sexual Instrumentation

Orton, Colin, *Harper Hospital and Wayne State University, Detroit, Michigan*, Medical Physics Literature

Ozcelik, Selahattin, *Texas A&M University, Kingsville, Texas*, Drug Infusion Systems

Panitch, Alyssa, *Arizona State University, Tempe, Arizona*, Biomaterials: An Overview

Paolino, Donatella, *University of Catanzaro Magna Græcia, Germaneto (CZ), Italy*, Drug Delivery Systems

Papaioannou, George, *University of Wisconsin, Milwaukee, Wisconsin*, Joints, Biomechanics of

Park, Grace E., *Purdue University, West Lafayette, Indiana*, Porous Materials for Biological Applications

Parmenter, Brett A., *State University of New York at Buffalo, Buffalo, New York*, Sleep Studies, Computer Analysis of

Patel, Dimpi, *The Cleveland Clinic Foundation, Cleveland, Ohio*, Hyperthermia, Ultrasonic

Pearce, John, *The University of Texas, Austin, Texas*, Electrosurgical Unit (ESU)

Pelet, Serge, *Massachusetts Institute of Technology, Cambridge, Massachusetts*, Microscopy, Fluorescence

Periasamy, Ammasi, *University of Virginia, Charlottesville, Virginia*, Cellular Imaging

Persons, Barbara L., *University of Mississippi Medical Center, Jackson, Mississippi*, Hyperbaric Medicine

Piper, Ian, *The University of Memphis, Memphis, Tennessee*, Monitoring, Intracranial Pressure

Poletto, Christopher J., *National Institutes of Health*, Tactile Stimulation

Preminger, Glenn M., *Duke University Medical Center, Durham, North Carolina*, Lithotripsy

Prendergast, Patrick J., *Trinity College, Dublin, Ireland*, Orthopedics, Prosthesis Fixation for

Previte, Michael, *Massachusetts Institute of Technology, Cambridge, Massachusetts*, Microscopy, Fluorescence

Purdy, James A., *UC Davis Medical Center, Sacramento, California*, Radiotherapy Accessories

Qi, Hairong, *Advanced Concepts Analysis, Inc., Falls Church, Virginia*, Thermography

Qin, Yixian, *Stony Brook University, New York*, Bone Density Measurement

Quan, Stuart F., *University of Arizona, Tucson, Arizona*, Ventilatory Monitoring

Quiroga, Rodrigo Quian, *University of Leicester, Leicester, United Kingdom*, Evoked Potentials

Rahaghi, Farbod N., *University of California, La Jolla, California*, Glucose Sensors

Rahko, Peter S., *University of Wisconsin Medical School*, Echocardiography and Doppler Echocardiography

Ralph, Lieto, *University of Wisconsin–Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

Ramanathan, Lakshmi, *Mount Sinai Medical Center*, Analytical Methods, Automated

Rao, Satish S.C., *University of Iowa College of Medicine, Iowa City, Iowa*, Anorectal Manometry

Rapoport, David M., *NYU School of Medicine, New York, New York*, Continuous Positive Airway Pressure

Rebello, Keith J., *The Johns Hopkins University Applied Physics Lab, Laurel, Maryland*, Micro surgery

Reddy, Narender, *The University of Akron, Akron, Ohio*, Linear Variable Differential Transformers

Ren-Dih, Sheu, *Memorial Sloan-Kettering Cancer Center, New York, New York*, Radiation Therapy Treatment Planning, Monte Carlo Calculations in

Rengachary, Setti S., *Detroit, Michigan*, Human Spine, Biomechanics of

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

Ritchey, Eric R., *The Ohio State University, Columbus, Ohio*, Contact Lenses

Rivard, Mark J., *Tufts New England Medical Center, Boston, Massachusetts*, Imaging Devices

Robertson, J. David, *University of Missouri, Columbia, Missouri*, Radionuclide Production and Radioactive Decay

Roth, Bradley J., *Oakland University, Rochester, Michigan*, Defibrillators

Rowe-Horwege, R. Wanda, *University of Texas Medical School, Houston, Texas*, Hyperthermia, Systemic

Rumsey, John W., *University of Central Florida, Orlando, Florida*, Biosurface Engineering

Rutkowski, Gregory E., *University of Minnesota, Duluth, Minnesota*, Engineered Tissue

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

Samaras, Theodoros, *Aristotle University of Thessaloniki Department of Physics, Thessaloniki, Greece*, Thermometry

Sangole, Archana P., *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

Sarkozi, Laszlo, *Mount Sinai School of Medicine*, Analytical Methods, Automated

Schek, Henry III, *University of Michigan, Ann Arbor, Michigan*, Optical Tweezers

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

Schuckers, Stephanie A.C., *Clarkson University, Potsdam, New York*, Arrhythmia Analysis, Automated

Scope, Kenneth, *Northwestern University, Chicago, Illinois*, Ventilators, Acute Medical Care

Scott, Adzick N., *University of Pennsylvania, Philadelphia, Pennsylvania*, Intrauterine Surgical Techniques

Seal, Brandon L., *Arizona State University, Tempe, Arizona*, Biomaterials: An Overview

Seale, Gary, *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

Segers, Patrick, *Ghent University, Belgium*, Hemodynamics

Selim, Mostafa A., *Cleveland Metropolitan General Hospital, Palm Coast, Florida*, Colposcopy

Sethi, Anil, *Loyola University Medical Center, Maywood, Illinois*, X-Rays: Interaction with Matter

Severinghaus, John W., *University of California in San Francisco*, CO_2 Electrodes

Shalodi, Abdelwahab D., *Cleveland Metropolitan General Hospital, Palm Coast, Florida*, Colposcopy

Shanmugasundaram, Shobana, *New Jersey Institute of Technology, Newark, New Jersey*, Polymeric Materials

Shard, Alexander G., *University of Sheffield, Sheffield United Kingdom*, Biomaterials, Surface Properties of

Shen, Li-Jiuan, *National Taiwan University School of Pharmacy, Taipei, Taiwan*, Colorimetry

Shen, Wei-Chiang,*University of Southern California School of Pharmacy, Los Angeles, California*, Colorimetry

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

Sherman, David, *The Johns Hopkins University, Baltimore, Maryland*, Electroencephalography

Shi, Donglu, *University of Cincinnati, Cincinnati, Ohio*, Biomaterials, Testing and Structural Properties of

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

Siedband, Melvin P., *University of Wisconsin, Madison, Wisconsin*, Image Intensifiers and Fluoroscopy

Silberman, Howard, *University of Southern California, Los Angeles, California*, Nutrition, Parenteral

Silverman, Gordon, *Manhattan College*, Computers in the Biomedical Laboratory

Silvern, David A., *Medical Physics Unit, Rabin Medical Center, Petah Tikva, Israel*, Prostate Seed Implants

Sinha, Piyush, *The Ohio State University, Columbus, Ohio*, Drug Delivery Systems

Sinha, Abhijit Roy, *University of Cincinnati, Cincinnati, Ohio*, Coronary Angioplasty and Guidewire Diagnostics

Sinkjær, Thomas, *Aalborg University, Aalborg, Denmark*, Electroneurography

Sloan, Jeffrey A., *Mayo Clinic, Rochester, Minnesota*, Quality-of-Life Measures, Clinical Significance of

So, Peter T.C., *Massachusetts Institute of Technology, Cambridge, Massachusetts*, Microscopy, Fluorescence

Sobol, Wlad T., *University of Alabama at Birmingham Health System, Birmingham, Alabama*, Nuclear Magnetic Resonance Spectroscopy

Sood, Sandeep, *University of Illinois at Chicago, Chicago, Illinois*, Hydrocephalus, Tools for Diagnosis and Treatment of

Spector, Myron, *Brigham and Women's Hospital, Boston, Massachusetts*, Biocompatibility of Materials

Spelman, Francis A., *University of Washington*, Cochlear Prostheses

Srinivasan, Yeshwanth, *Texas Tech University, Lubbock, Texas*, Medical Records, Computers in

Sriram, Neelamegham, *University of Buffalo, Buffalo, New York*, Cell Counters, Blood

Starko, Kenton R., *Point Roberts, Washington*, Physiological Systems Modeling

Starkschall, George, *The University of Texas*, Radiotherapy, Three-Dimensional Conformal

Stavrev, Pavel, *Cross Cancer Institute, Edmonton, Alberta, Canada*, Radiotherapy Treatment Planning, Optimization of

Stenken, Julie A., *Rensselaer Polytechnic Institute, Troy, New York*, Microdialysis Sampling

Stiefel, Robert, *University of Maryland Medical Center, Baltimore, Maryland*, Equipment Acquisition

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

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

Su, Xiao-Li, *BioDetection Instruments LLC, Fayetteville, Arkansas*, Piezoelectric Sensors

Subhan, Arif, *Masterplan Technology Management, Chatsworth, California*, Equipment Maintenance, Biomedical

Sweeney, James D., *Arizona State University, Tempe, Arizona*, Functional Electrical Stimulation

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

Takayama, Shuichi, *University of Michigan, Ann Arbor, Michigan*, Microbioreactors

Tamul, Paul C., *Northwestern University, Chicago, Illinois*, Ventilators, Acute Medical Care

Tamura, Toshiyo, *Chiba University School of Engineering, Chiba, Japan*, Home Health Care Devices

Tang, Xiangyang, *GE Healthcare Technologies, Wankesha, Wisconsin*, Computed Tomography Simulators

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

Temple, Richard O., *Transitional Learning Center at Galveston, Galveston, Texas*, Rehabilitation, Computers in Cognitive

Ten, Stanley, *Salt Lake City, Utah*, Electroanalgesia, Systemic

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

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

Thierens, Hubert M.A., *University of Ghent, Ghent, Belgium*, Radiopharmaceutical Dosimetry

Thomadsen, Bruce, *University of Wisconsin–Madison, Madison, Wisconsin*, Codes and Regulations: Radiation

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

Togawa, Tatsuo, *Waseda University, Saitama, Japan*, Integrated Circuit Temperature Sensor

Tornai, Martin, *Duke University, Durham, North Carolina*, X-Ray Equipment Design

Tran-Son-Tay, Roger, *University of Florida, Gainesville, Florida*, Blood Rheology

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

ALLOYS, SHAPE MEMORY
AMBULATORY MONITORING
ANALYTICAL METHODS, AUTOMATED
ANESTHESIA MACHINES
ANESTHESIA, COMPUTERS IN
ANGER CAMERA
ANORECTAL MANOMETRY
ARRHYTHMIA ANALYSIS, AUTOMATED
ARTERIES, ELASTIC PROPERTIES OF
AUDIOMETRY
BIOCOMPATIBILITY OF MATERIALS
BIOELECTRODES
BIOFEEDBACK
BIOHEAT TRANSFER
BIOIMPEDANCE IN CARDIOVASCULAR MEDICINE
BIOINFORMATICS
BIOMAGNETISM
BIOMATERIALS, ABSORBABLE
BIOMATERIALS: AN OVERVIEW
BIOMATERIALS: BIOCERAMICS
BIOMATERIALS: CARBON
BIOMATERIALS, CORROSION AND WEAR OF
BIOMATERIALS FOR DENTISTRY
BIOMATERIALS: POLYMERS
BIOMATERIALS, SURFACE PROPERTIES OF
BIOMATERIALS, TESTING AND STRUCTURAL PROPERTIES OF
BIOMATERIALS: TISSUE ENGINEERING AND SCAFFOLDS
BIOMECHANICS OF EXERCISE FITNESS
BIOMEDICAL ENGINEERING EDUCATION
BIOSURFACE ENGINEERING
BIOTELEMETRY
BLADDER DYSFUNCTION, NEUROSTIMULATION OF
BLIND AND VISUALLY IMPAIRED, ASSISTIVE TECHNOLOGY FOR
BLOOD COLLECTION AND PROCESSING
BLOOD GAS MEASUREMENTS
BLOOD PRESSURE MEASUREMENT
BLOOD PRESSURE, AUTOMATIC CONTROL OF
BLOOD RHEOLOGY
BLOOD, ARTIFICIAL
BONE AND TEETH, PROPERTIES OF
BONE CEMENT, ACRYLIC
BONE DENSITY MEASUREMENT
BONE UNUNITED FRACTURE AND SPINAL FUSION, ELECTRICAL TREATMENT OF
BORON NEUTRON CAPTURE THERAPY
BRACHYTHERAPY, HIGH DOSAGE RATE
BRACHYTHERAPY, INTRAVASCULAR
CAPACITIVE MICROSENSORS FOR BIOMEDICAL APPLICATIONS
CARDIAC OUTPUT, FICK TECHNIQUE FOR
CARDIAC OUTPUT, INDICATOR DILUTION MEASUREMENT OF
CARDIAC OUTPUT, THERMODILUTION MEASUREMENT OF
CARDIOPULMONARY RESUSCITATION
CARTILAGE AND MENISCUS, PROPERTIES OF
CELL COUNTERS, BLOOD
CELLULAR IMAGING
CHROMATOGRAPHY
CO_2 ELECTRODES
COBALT 60 UNITS FOR RADIOTHERAPY
COCHLEAR PROSTHESES
CODES AND REGULATIONS: MEDICAL DEVICES
CODES AND REGULATIONS: RADIATION
COLORIMETRY
COLPOSCOPY
COMMUNICATION DEVICES
COMMUNICATIVE DISORDERS, COMPUTER APPLICATIONS FOR
COMPUTED TOMOGRAPHY
COMPUTED TOMOGRAPHY SCREENING
COMPUTED TOMOGRAPHY SIMULATORS
COMPUTED TOMOGRAPHY, SINGLE PHOTON EMISSION
COMPUTER-ASSISTED DETECTION AND DIAGNOSIS
COMPUTERS IN THE BIOMEDICAL LABORATORY
CONTACT LENSES
CONTINUOUS POSITIVE AIRWAY PRESSURE
CONTRACEPTIVE DEVICES
CORONARY ANGIOPLASTY AND GUIDEWIRE DIAGNOSTICS
CRYOSURGERY
CUTANEOUS BLOOD FLOW, DOPPLER MEASUREMENT OF
CYSTIC FIBROSIS SWEAT TEST
CYTOLOGY, AUTOMATED
DEFIBRILLATORS
DIFFERENTIAL COUNTS, AUTOMATED
DIGITAL ANGIOGRAPHY
DNA SEQUENCE
DRUG DELIVERY SYSTEMS
DRUG INFUSION SYSTEMS
ECHOCARDIOGRAPHY AND DOPPLER ECHOCARDIOGRAPHY
ELECTROANALGESIA, SYSTEMIC
ELECTROCARDIOGRAPHY, COMPUTERS IN
ELECTROCONVULSIVE THERAPY
ELECTROENCEPHALOGRAPHY
ELECTROGASTROGRAM
ELECTROMYOGRAPHY
ELECTRONEUROGRAPHY
ELECTROPHORESIS

ELECTROPHYSIOLOGY
ELECTRORETINOGRAPHY
ELECTROSURGICAL UNIT (ESU)
ENDOSCOPES
ENGINEERED TISSUE
ENVIRONMENTAL CONTROL
EQUIPMENT ACQUISITION
EQUIPMENT MAINTENANCE, BIOMEDICAL
ESOPHAGEAL MANOMETRY
EVOKED POTENTIALS
EXERCISE STRESS TESTING
EYE MOVEMENT, MEASUREMENT TECHNIQUES FOR
FETAL MONITORING
FIBER OPTICS IN MEDICINE
FLAME ATOMIC EMISSION SPECTROMETRY AND ATOMIC ABSORPTION SPECTROMETRY
FLOWMETERS
FLUORESCENCE MEASUREMENTS
FUNCTIONAL ELECTRICAL STIMULATION
GAMMA KNIFE
GAS AND VACUUM SYSTEMS, CENTRALLY PIPED MEDICAL
GASTROINTESTINAL HEMORRHAGE
GLUCOSE SENSORS
HEART VALVE PROSTHESES
HEART VALVE PROSTHESES, IN VITRO FLOW DYNAMICS OF
HEART VALVES, PROSTHETIC
HEART, ARTIFICIAL
HEART–LUNG MACHINES
HEAT AND COLD, THERAPEUTIC
HEMODYNAMICS
HIGH FREQUENCY VENTILATION
HIP JOINTS, ARTIFICIAL
HOME HEALTH CARE DEVICES
HUMAN FACTORS IN MEDICAL DEVICES
HUMAN SPINE, BIOMECHANICS OF
HYDROCEPHALUS, TOOLS FOR DIAGNOSIS AND TREATMENT OF
HYPERBARIC MEDICINE
HYPERBARIC OXYGENATION
HYPERTHERMIA, INTERSTITIAL
HYPERTHERMIA, SYSTEMIC
HYPERTHERMIA, ULTRASONIC
IMAGE INTENSIFIERS AND FLUOROSCOPY
IMAGING DEVICES
IMMUNOLOGICALLY SENSITIVE FIELD-EFFECT TRANSISTORS
IMMUNOTHERAPY
IMPEDANCE PLETHYSMOGRAPHY
IMPEDANCE SPECTROSCOPY
INCUBATORS, INFANT
INTEGRATED CIRCUIT TEMPERATURE SENSOR
INTRAAORTIC BALLOON PUMP
INTRAUTERINE SURGICAL TECHNIQUES
IONIZING RADIATION, BIOLOGICAL EFFECTS OF
ION-SENSITIVE FIELD-EFFECT TRANSISTORS
JOINTS, BIOMECHANICS OF
LARYNGEAL PROSTHETIC DEVICES
LENSES, INTRAOCULAR
LIGAMENT AND TENDON, PROPERTIES OF
LINEAR VARIABLE DIFFERENTIAL TRANSFORMERS
LITHOTRIPSY
LIVER TRANSPLANTATION
LUNG SOUNDS
MAGNETIC RESONANCE IMAGING
MAMMOGRAPHY
MEDICAL EDUCATION, COMPUTERS IN
MEDICAL ENGINEERING SOCIETIES AND ORGANIZATIONS
MEDICAL GAS ANALYZERS
MEDICAL PHYSICS LITERATURE
MEDICAL RECORDS, COMPUTERS IN
MICROARRAYS
MICROBIAL DETECTION SYSTEMS
MICROBIOREACTORS
MICRODIALYSIS SAMPLING
MICROFLUIDICS
MICROPOWER FOR MEDICAL APPLICATIONS
MICROSCOPY AND SPECTROSCOPY, NEAR-FIELD
MICROSCOPY, CONFOCAL
MICROSCOPY, ELECTRON
MICROSCOPY, FLUORESCENCE
MICROSCOPY, SCANNING FORCE
MICROSCOPY, SCANNING TUNNELING
MICROSURGERY
MINIMALLY INVASIVE SURGERY
MOBILITY AIDS
MONITORING IN ANESTHESIA
MONITORING, HEMODYNAMIC
MONITORING, INTRACRANIAL PRESSURE
MONITORING, UMBILICAL ARTERY AND VEIN
MONOCLONAL ANTIBODIES
NANOPARTICLES
NEONATAL MONITORING
NEUROLOGICAL MONITORS
NEUTRON ACTIVATION ANALYSIS
NEUTRON BEAM THERAPY
NONIONIZING RADIATION, BIOLOGICAL EFFECTS OF
NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY
NUCLEAR MEDICINE INSTRUMENTATION
NUCLEAR MEDICINE, COMPUTERS IN
NUTRITION, PARENTERAL
OCULAR FUNDUS REFLECTOMETRY
OCULAR MOTILITY RECORDING AND NYSTAGMUS
OFFICE AUTOMATION SYSTEMS
OPTICAL SENSORS
OPTICAL TWEEZERS
ORTHOPEDIC DEVICES, MATERIALS AND DESIGN FOR
ORTHOPEDICS, PROSTHESIS FIXATION FOR
OXYGEN ANALYZERS
OXYGEN MONITORING
PACEMAKERS
PANCREAS, ARTIFICIAL
PERIPHERAL VASCULAR NONINVASIVE MEASUREMENTS
PHANTOM MATERIALS IN RADIOLOGY
PHARMACOKINETICS AND PHARMACODYNAMICS
PHONOCARDIOGRAPHY
PHOTOGRAPHY, MEDICAL
PHYSIOLOGICAL SYSTEMS MODELING

PICTURE ARCHIVING AND COMMUNICATION SYSTEMS
PIEZOELECTRIC SENSORS
PNEUMOTACHOMETERS
POLYMERASE CHAIN REACTION
POLYMERIC MATERIALS
POROUS MATERIALS FOR BIOLOGICAL APPLICATIONS
POSITRON EMISSION TOMOGRAPHY
PROSTATE SEED IMPLANTS
PULMONARY PHYSIOLOGY
QUALITY-OF-LIFE MEASURES, CLINICAL SIGNIFICANCE OF
RADIATION DOSE PLANNING, COMPUTER-AIDED
RADIATION DOSIMETRY FOR ONCOLOGY
RADIATION DOSIMETRY, THREE-DIMENSIONAL
RADIATION PROTECTION INSTRUMENTATION
RADIATION THERAPY, INTENSITY MODULATED
RADIATION THERAPY SIMULATOR
RADIATION THERAPY TREATMENT PLANNING, MONTE CARLO CALCULATIONS IN
RADIATION THERAPY, QUALITY ASSURANCE IN
RADIOLOGY INFORMATION SYSTEMS
RADIONUCLIDE PRODUCTION AND RADIOACTIVE DECAY
RADIOPHARMACEUTICAL DOSIMETRY
RADIOSURGERY, STEREOTACTIC
RADIOTHERAPY ACCESSORIES
RADIOTHERAPY, HEAVY ION
RADIOTHERAPY, INTRAOPERATIVE
RADIOTHERAPY, THREE-DIMENSIONAL CONFORMAL
RADIOTHERAPY TREATMENT PLANNING, OPTIMIZATION OF
RECORDERS, GRAPHIC
REHABILITATION AND MUSCLE TESTING
REHABILITATION, COMPUTERS IN COGNITIVE
REHABILITATION, ORTHOTICS IN
RESIN-BASED COMPOSITES
RESPIRATORY MECHANICS AND GAS EXCHANGE
SAFETY PROGRAM, HOSPITAL
SCOLIOSIS, BIOMECHANICS OF
SCREEN-FILM SYSTEMS
SEXUAL INSTRUMENTATION
SHOCK, TREATMENT OF
SKIN SUBSTITUTE FOR BURNS, BIOACTIVE
SKIN TISSUE ENGINEERING FOR REGENERATION
SKIN, BIOMECHANICS OF
SLEEP LABORATORY
SLEEP STUDIES, COMPUTER ANALYSIS OF
SPINAL CORD STIMULATION
SPINAL IMPLANTS
STATISTICAL METHODS
STEREOTACTIC SURGERY
STERILIZATION OF BIOLOGIC SCAFFOLD MATERIALS
STRAIN GAGES
TACTILE STIMULATION
TELERADIOLOGY
TEMPERATURE MONITORING
THERMISTORS
THERMOCOUPLES
THERMOGRAPHY
THERMOMETRY
TISSUE ABLATION
TISSUE ENGINEERING
TOMOTHERAPY
TONOMETRY, ARTERIAL
TOOTH AND JAW, BIOMECHANICS OF
TRACER KINETICS
TRANSCUTANEOUS ELECTRICAL NERVE STIMULATION (TENS)
ULTRASONIC IMAGING
ULTRAVIOLET RADIATION IN MEDICINE
VASCULAR GRAFT PROSTHESIS
VENTILATORS, ACUTE MEDICAL CARE
VENTILATORY MONITORING
VISUAL FIELD TESTING
VISUAL PROSTHESES
X-RAY EQUIPMENT DESIGN
X-RAY QUALITY CONTROL PROGRAM
X-RAY THERAPY EQUIPMENT, LOW AND MEDIUM ENERGY
X-RAYS: INTERACTION WITH MATTER
X-RAYS, PRODUCTION OF

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)	$0.9144^{\dagger}$

N

NANOPARTICLES

O.V. Salata
University of Oxford
Oxford, United Kingdom

INTRODUCTION

Nanoparticle (From a Greek word “ο vάνοζ” meaning “dwarf”), “nano” is a prefix defining a billion times reduction (10^{-9}) in magnitude. Intuitively, a nanoparticle is an object that is a “nano” times smaller than just a particle. The generally accepted size range for the objects to be called “nanoparticles” is between 1 and 100 nm in at least in two dimensions. This broad definition can be narrowed by defining a nanoparticle of a certain material as a particle that is smaller than a critical size for this material (Fig. 1). The critical size is defined as a cross-point size for the transition from bulk to sized-dependent material property or properties, for example, solubility, melting point, magnetization, light absorption, and fluorescence. The actual value of the critical size depends on the strength of the interatomic or intermolecular forces. The weaker the forces, the weaker the overall interactions between the atoms in the particle. It is known from quantum mechanics that an atom or molecule can be characterized by a set of permitted electron energy levels. Those energy levels define physical and chemical properties of atoms and molecules. When two identical atoms are brought into proximity, they would experience both repulsing and attracting forces. An equilibrium distance might exist where the two forces are balanced. A new molecule containing two atoms will have a system of energy levels that is similar to that of the initial atoms, but each level will be split into two sublevels. The energy gap between the two closest sublevels in this molecule will now be smaller than the corresponding smallest energy gap in one atom. Each time a new atom is added, a new set of energy sublevels will be formed. The energy gap will be changing in value with the number of atoms used to build a nanoparticle, and because properties of the nanoparticle will depend on the energy gap, it will also depend on its size. Then millions of the same atoms form a bulk material; these energy sublevels are saturated and broadened into the energy bands that are insensitive to the changes in the number of atoms composing the solid body. Another component affecting the critical size of the nanoparticle is its surface-to-volume ratio. The smaller the size of the particle, the higher the proportion of the total number of atoms or molecules that are positioned at the surface. The surface atoms often have unsaturated or dangling bonds, leading to the high reactivity and catalytic ability of the nanoparticles. It also results in the restructuring and rearrangement of the crystal lattice near the surface. All of this affects the size-dependent properties of nanomaterials.

Nanoparticles can be made out of any material, including metals, ceramics, semiconductors, polymers, and biomolecules. They can possess a complex structure, which might contain a combination of different materials, or have a complex shape. Nanometer-sized objects that include nanoparticles fall in the realm of nanotechnology. Nanotechnology is developing in several directions: nanomaterials, nanodevices, and nanosystems. The nanomaterials/particles level is the most advanced currently, both in scientific knowledge and in commercial applications. Nanobiotechnology is a subfield of nanotechnology that deals with the applications of nanotechnology to biology. Understanding of biological processes on the nanoscale level is a strong driving force behind development of nanobiotechnology. Living organisms are built of cells that are typically more that 10 μm across. However, the cell-forming components are much smaller and are in the submicron size domain. Even smaller are the proteins with a typical size of just 5 nm, which is comparable with the dimensions of the smallest manmade nanoparticles. This simple size comparison gives an idea of using nanoparticles as very small probes (1) that would allow us to spy at the cellular machinery without introducing too much interference. Semiconductor nanoparticles, also known as “quantum dots” (2), show a strong dependence of their physical properties on their size. Just a decade ago, quantum dots were studied because of their size-dependent physical and chemical properties. One of the properties of the semiconductor nanoparticles that are changing with size is the color of their fluorescence, and now they are used as photostable fluorescent probes. As nanoparticles are rapidly taken up by all kinds of cells, they are also used in drug delivery. In pharmacology, the term “nanoparticles” specifically means polymer nanoparticles or, sometimes, submicron particles that carry a drug load (3). This term has been used in drug delivery for more than three decades. At about the same time, magnetic particles with submicron dimensions were employed for the first time to assist with cell separation. However, colloidal gold, which can be alternatively called “a dispersion of gold nanoparticles,” has been used in medicine for many decades if not centuries. Colloidal gold tinctures were used by alchemists to treat many illnesses. Colloidal gold was used as a contrast agent by the first optical microscopists as early as the 1600s. In the 1950s, work was started on the use of radioactive colloidal gold as a treatment for cancer (4). When functionalized with antibodies, gold nanoparticles are used to stain cellular organelles or membranes to create markers for the electron microscopy (5). Consequent decoration of the gold markers with silver assists in further signal magnification.

Out of a plethora of size-dependent physical properties available to someone who is interested in the practical side of nanomaterials, optical and magnetic effects are the most used for biological applications. Hybrid bionanomaterials

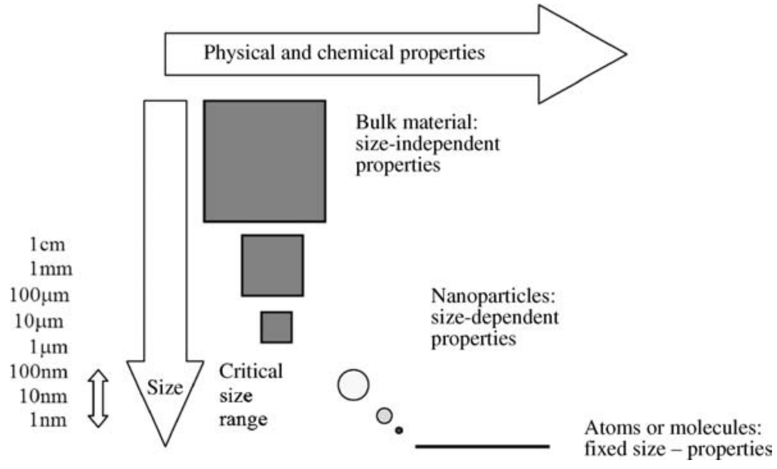

Physical and chemical properties
Bulk material: size-independent properties
1cm
1mm
100μm
10μm
1μm
100nm
10nm
1nm
Size
Critical size range
Nanoparticles: size-dependent properties
Atoms or molecules: fixed size – properties

Figure 1. Nanoparticles are a state of matter intermediate between the bulk (size-independent properties) and atomic or molecular (fixed properties) form of the same material with its physical and chemical properties (such as melting temperature, solubility, optical absorbance and fluorescence, magnetization, catalytic activity, and specific chemical reactivity) being dependent on the particle size. The critical size for the transition from the bulk to the nano-regime can vary from tenths to just a few nanometers. A 100 nm threshold is a convenient size to use for a generalized description of the expected bulk-nanoparticle transition.

can also be applied to build novel electronic, optoelectronics, and memory devices.

NANOPARTICLE FABRICATION

Two general ways are available to a manufacturer to produce nanoparticles (Fig. 2). The first way is to start with a bulk material and then break it into smaller pieces using mechanical, chemical, or another form of energy (top-down). An opposite approach is to synthesize the material from atomic or molecular species via physical condensation of atomized materials (energy released), or chemical reactions (exo- or endothermic), allowing the precursor particles to grow in size (bottom-up). Both approaches can be done in gas, liquid, or solid states, or under a vacuum. Both the top-down and the bottom-up processes can be happening during the formation of nanoparticles at the same time, for example, during mechano-chemical ball milling process.

The more detailed classification of the nanoparticle manufacturing techniques relies on the combination of the form of energy with the type of the controlled environment. Each technique has its advantages and disadvantages. Most manufacturers are interested in the ability to control particle size, particle shape, size distribution, particle composition, and degree of particle agglomeration. Absence of contaminants, processing residues or solvents, and sterility are often required in the case of biological and medical applications of nanomaterials. The scale-up of the production volume is also very important. Hence, the discussion of the nanoparticle production techniques is limited to some of those that are currently being pursued by the manufacturers.

Ball Milling

Ball milling is a process where large spheres of the milling media crush substantially smaller powder particles (6). Normally it is used to make fine powder blends or to reduce the degree of powder agglomeration. High-energy ball milling is a more energetic form capable of breaking ceramics into nanoparticles. It is used to create nano-structured composites, highly supersaturated solid solutions, and amorphous phases. The drawbacks of this technique include high energy consumption and poor control over particle sizes. A variation of high-energy ball milling is called mechano-chemical processing. Chemical reactions are mechanically activated during milling, forming nanoparticles via a bottom-up process from suitable precursors. A solid diluent's phase is used to separate the nanoparticles. In the pharmaceutical industry, wet ball milling is often used to produce nano-formulations of the drugs that are poorly soluble in their bulk form but acquire a much improved solubility when turned into nanoparticles.

Electro-Explosion

This process is used to generate 100 nm metal nanoparticles in the form of dry powders. Michael Faraday first observed it in 1773. It involves providing a very high current over a very short time through thin metallic wires, in either an inert or a reactive gas, such that extraordinary temperatures of 20,000 to 30,000 K are achieved. The wire is turned into plasma, contained, and compressed by the

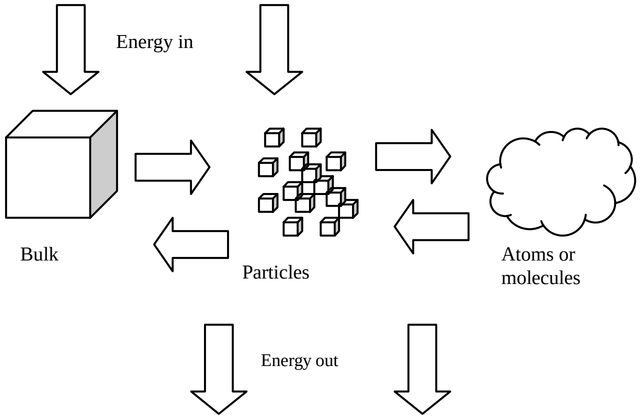


Figure 2. Two basic approaches to nanomaterials fabrication: top-down (shown here from left to the right) and bottom-up (from right to the left). Usually, energy in one of its forms (mechanical, thermal, etc.) is supplied to the bulk matter to create new surfaces. Chemical synthesis of nanomaterials from the atomic or molecular species can be either exothermic or endothermic. Condensation of atoms under vacuum results in cluster formation accompanied by the release of energy.

electromagnetic field. After that, the resistance of the wire becomes so high that the current terminates. The disappearing of the electromagnetic field makes the plasma expand very rapidly. The extremely fast cooling rate results in stabilization of otherwise meta-stable materials. The powders made by electro-explosion have greater purity and reactivity as compared with the ball-milled powders. This technique is used, for example, to produce high surface area nano-porous filtering media benefiting from the enhanced bactericidal properties of silver nanoparticles.

Laser Ablation

In laser ablation, pulsed light from an excimer laser is focused onto a solid target inside a vacuum chamber to supply thermal energy that would "boil off" a plume of energetic atoms of the target material. A substrate positioned to intercept the plume will receive a thin film deposit of the target material. This phenomenon was first observed with a ruby laser in the mid-1960s. Because this process then contaminated the films made with particles, little use was found for such "dirty" samples. The laser ablation method has the following advantages for the fabrication of nanomaterials: The fabrication parameters can be easily changed in a wide range; nanoparticles are naturally produced in the laser ablation plume so that the production rate is relatively high; and virtually all materials can be evaporated by laser ablation.

Colloidal Chemistry

Two general colloidal chemistry approaches are available to control the formation and growth of the nanoparticles. One is called an arrested precipitation; it depends on exhaustion of one of the reactants or on the introduction of the chemical that would block the reaction. Another method relies on a physical restriction of the volume available for the growth of the individual nanoparticles by using various templates (7).

Any material containing regular nanosized pores or voids can be used as a template to form nanoparticles. Examples of such templates include porous alumina, zeolites, di-block copolymers, dendrimers, proteins, and other molecules. The template does not have to be a 3D object. Artificial templates can be created on a plane surface or a gas–liquid interface by forming self-assembled monolayers. The template is usually removed by dissolving it in the solvent that is not affecting the formed nanoparticles. The main advantages of the colloidal chemistry techniques for the preparation of nanomaterials are low temperature of processing, versatility, and flexible rheology. They also offer unique opportunities for preparation of organic–inorganic hybrid materials. The most commonly used precursors for inorganic nanoparticles are oxides and alcoxides.

Aerosols

As an alternative to liquids, chemical reactions can be carried out in a gaseous media, resulting in the formation of nanoparticles aerosols (8). Aerosols can be defined as solid or liquid particles in a gas phase, where the particles can range from molecules up to 100 μm in size. Aerosol generation is driven by the pressure differential created with the help of compressed gases, vacuum, mechanical oscillations, or electrostatic forces acting on liquid. Aerosols were used in industrial manufacturing long before the basic science and engineering of the aerosols were understood. For example, carbon black particles used in pigments and reinforced car tires are produced by hydrocarbon combustion; titania pigment for use in paints and plastics is made by oxidation of titanium tetrachloride; fumed silica and titania formed from respective tetrachlorides by flame pyrolysis; and optical fibers are manufactured by a similar process. Aerosols are also widely used as a drug delivery technique.

Solvent Drying

This technique is frequently used to generate particles of soluble materials (9). Starting materials, for example, a drug and a stabilizing polymer, are dissolved in water-immiscible organic solvent, which is used to prepare an oil-in-water microemulsion. Water can be evaporated by heating under reduced pressure, leaving behind drug-loaded nanoparticles. Both nanospheres (uniform distribution of components) and nanocapsules (polymer encapsulated core) can be created with this method. A monomer can be used instead of the polymer, if the micelle polymerization step is possible. Solvent drying can be achieved via spray-drying step, where a homogeneous solution is fed to an aerosol generator, which produces uniformly sized droplets containing equal amounts of dissolved material. Solvent evaporation from the droplets under the right conditions would result in the formation of nanoparticles with a narrow size distribution.

Electro-Spinning

An emerging technology for the manufacture of thin polymer fibers is based on the principle of spinning dilute polymer solutions in a high-voltage electric field.

Electro-spinning is a process by which a suspended drop of polymer is charged with thousands of volts. At a characteristic voltage, the droplet forms a Taylor cone (the most stable shape with an apex angle of about 57°), and a fine jet of polymer releases from the surface in response to the tensile forces generated by the interaction of an applied electric field with the electrical charge carried by the jet. This produces a bundle of polymer fibers. The jet can be directed to a grounded surface and collected as a continuous web of fibers ranging in size from a few micrometers to less than 100 nm.

Self-Assembly (10)

The appropriate molecular building blocks can act as parts of a jigsaw puzzle that join in a perfect order without obvious driving force present. Various types of chemical bonding can be used to self-assemble nanoparticles. For example, electrostatic interaction between the oppositely charged polymers can be used to build multilayered nanocapsules, the difference in hydrophobicity between the different molecules in the mixture can lead to a formation of a 3D assembly, and proteins can be selected to self-assemble into virus like nanoparticles. Another example is a use of artificially created oligonucleotides that can be designed to assemble in a variety of shapes and forms.

Nanoparticle Surface Treatment (11)

“Bare” nanoparticles of the same material would rapidly agglomerate with each other, forming bulk material. Encapsulating nanoparticles after production helps to maintain particle size and particle size distribution by inhibiting particle growth that can be caused by evaporation/redeposition, dissolution/reprecipitation, or surface migration and/or flocculation/aggregation/agglomeration. Encapsulation quenches the particle’s reactivity and reduces degradation of either the particle or the matrix that surrounds it. The encapsulating coating may be functionalized to facilitate dispersion into organic or aqueous liquid systems. Surface treatment with functional groups enables direct interaction between nanoparticles and resins. Typical functional groups are as follows:

Acrylate
Epoxide
Amine
Vinyl
Isocyanate

The stability of the collected nanoparticle powders against agglomeration, sintering, and compositional changes can be ensured by collecting the nanoparticles in liquid suspension. For semiconductor particles, stabilization of the liquid suspension has been demonstrated by the addition of polar solvent; surfactant molecules have been used to stabilize the liquid suspension of metallic nanoparticles. Alternatively, inert silica encapsulation of nanoparticles by gas-phase reaction and by oxidation in colloidal solution has been shown to be effective for metallic nanoparticles.

When nanosized powders are dispersed in water, they aggregate due to attractive van der Waals forces. By altering the dispersing conditions, repulsive forces can be introduced between the particles to eliminate these aggregates. There are two ways of stabilizing nanoparticles. First, by adjusting the pH of the system, the nanoparticle surface charge can be manipulated such that an electrical double layer is generated around the particle. Overlap of two double layers on different nanoparticles causes repulsion and hence stabilization. The magnitude of this repulsive force can be measured via the zeta potential. The second method involves the adsorption of polymers onto the nanoparticles, such that the particles are physically prevented from coming close enough for the van der Waals attractive force to dominate; this is termed steric stabilization. The combination of two mechanisms is called electrosteric stabilization; it occurs when polyelectrolytes are adsorbed on the nanoparticle surface.

APPLICATIONS (12)

The fact that nanoparticles exist in the same size domain as proteins makes nanomaterials suitable for biotagging and payload delivery. However, size is just one of many characteristics of nanoparticles that it is rarely sufficient if one is to use nanoparticles as biological tags. To interact with a biological target, a biological or molecular coating or layer acting as a bioinorganic interface should be attached to the nanoparticle. Examples of biological coatings may include antibodies, biopolymers like collagen, or monolayers of small molecules that make the nanoparticles biocompatible. In addition, as optical detection techniques are widespread in biological research, nanoparticles should either fluoresce or change their optical properties. The approaches used in constructing nano-biomaterials are schematically presented below (Fig. 3). Nanoparticle usually forms the core of nano-biomaterial. It can be used as a convenient surface for molecular assembly and may be composed of inorganic or polymeric materials. It can also be in the form of a nano-vesicle, surrounded by a membrane or a layer. The shape is more often spherical, but cylindrical, plate-like, and other shapes are possible. The size and size distribution might be important in some cases, for example, if penetration through a pore structure of a cellular membrane is required. The size and size distribution are becoming extremely critical when quantum-sized effects are used to control material properties. A tight control of the average particle size and a narrow distribution of sizes allow creation of very efficient fluorescent probes that emit narrow light in a very wide range of wavelengths. This helps with creating biomarkers with many well-distinguished colors. The core might have several layers and be multifunctional. For example, combining magnetic and luminescent layers, one can both detect and manipulate the particles. The core particle is often protected by several monolayers of inert material, for example, silica. Organic molecules that are adsorbed or chemisorbed on the surface of the particle are also used for this purpose. The same layer might act as a biocompatible material. However, more often, an additional layer of linker molecules is required to proceed with further functionalization. This linear linker molecule has reactive

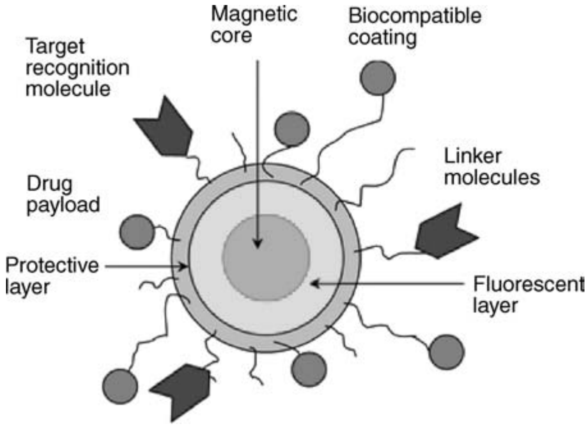


Figure 3. Schematic representation of an example of a biofunctionalized nanoparticle containing a magnetic core coated with a fluorescent layer, which, in turn, is coated by a thin protecting layer (e.g., silica). Linker molecules are attached to the protective layer at one end and to various functional molecules at the other.

groups at both ends. One group is aimed at attaching the linker to the nanoparticle surface, and the other is used to bind various moieties like biocompatibles (dextran), antibodies, and fluorophores, depending on the function required by the application.

Some current applications of nanomaterials to biology and medicine are listed as follows:

- Biological tags or labels: Mainly gold colloids are used for both electron and light microscopy; also, silver and silver-coated gold nanoparticles are used. A recent addition of semiconductor nanocrystals or quantum dots is finding increasing application as a substitute for the organic fluorophores. Greater photo-stability and the single-wavelength excitation option are among the quoted benefits. Badly designed quantum dots might disintegrate, releasing toxic components (e.g., cadmium or arsenic) that could be lethal at a cellular level. A use of porous silicon nanoparticles as fluorescent tags might be a safer option. Both quantum dots and colloidal gold were used recently for the detection of pathogens, proteins, and for the DNA sequencing. Colloidal gold and, more recently, quantum dots, have also been employed in phagokinetic studies.
- Drug delivery: Mainly use polymeric nanoparticles (13) because of their stability in biological fluids. Flexibility offered by a wide choice of polymers helps to control the rates of drug release and particle biodegradation. Polymeric nanoparticles can be used for all possible administration routes. Surface modifications allow creation of "stealth" as well as targeted drug carriers. Two major approaches are used in their preparation: from polymers (e.g., polyesters) or from monomers (e.g., alkylcyanoacrylate). Either solid nanospheres or liquid core nanocapsules can be produced. The fabrication technologies can be based on the solvent evaporation from oil-in-water microemulsions, created with help of surfactants, or by polymer precipitation caused by the addition of the nonsolvent. Another trend is to produce nanoparticles out of the poorly soluble drugs. Size reduction (nanosizing) (14) can significantly prolong drug bioavailability, increase its dissolution rate and maximum concentration, and dramatically shorten the onset of the drug action.
- MRI contrast enhancement: MRI detects the spatial distribution of the signals from water protons inside the body. These signals depend on the local amount of water and the proton relaxation times T1 and T2. The relaxation times can be affected by several factors; they can also be shortened by the presence of paramagnetic molecules or particles. The T1 shortening would result in the increased signal intensity, whereas the T2 shortening would lead to the opposite effect. These effects are nonlinear functions of the concentration of the contrasting agent.

  Superparamagnetic iron oxide (SPIO), whose average particle size is 50 nm, with dextrane or siloxane coating, is used as a tissue-specific contrast agent (Feridex, Endorem, GastroMARK, Lumirem, Resovist); ultra-small superparamagnetic oxide particles, 10 nm (Sinerem, Combidex), are used to distinguish between the metastatic and inflamed lymph nodes, and to identify arteriosclerosis plaque deposits. Intravenously injected SPIO would pass through the vascular endothelium into the interstitium. After that, the SPIO would be taken up by both healthy and inflamed lymph nodes. The uptake is followed by phagocytosis. Normal lymph nodes show a decrease in signal intensity on T2*- and T2-weighted MR images because of the effects of magnetic susceptibility and T2 shortening on the iron deposits that are a direct result of phagocytosis. However, metastatic lymph nodes are bad at phagocytosis, form no deposits, and do not show such reduction in the signal intensity. This effect can be used to distinguish between the healthy and the benign lymph nodes.
- Separation and purification of biomolecules and cells: Dynabeads are highly uniform in size and superparamagnetic; depending on the antibodies present on the surface, the beads can be applied to different tasks like separation of T-cells, detection of microbes and protozoa, HLA diagnostics for organ transplantation, and various *in vitro* diagnostics assays. They are also used for the isolation of DNA and proteins.
- Tissue engineering: Nanoparticles of hydroxyapatite are used to mimic the mineral particles occurring in the bone structure, whereas collagen is often replaced with a 3D porous scaffolding of a biodegradable polymer. This approach allows for the high mobility of the osteoblasts and, consequently, a uniform growth of the new bone. A similar strategy is used to promote the cellular growth on the surface of prosthetic implants.
- Tumor destruction via heating (hyperthermia): The nanoparticle approach is currently relying on the higher metabolic rates and enhanced blood supply to the tumors. As a result, the cancerous cells are likely to be enriched in a nanoparticulate matter, introduced in the blood circulation, or directly injected into the tumor. An external electromagnetic energy source is directed toward the tumor. The nanoparticles are designed to absorb the electromagnetic energy and convert it into localized heat, which would preferentially cause the apoptosis of the malignant cells. Localized heating might also result in the increased acidosis of the cancer cells. It is also been suggested that the high density of blood vessels in and around the growth stops them from expanding as efficiently as those in the healthy tissue, leading to a higher heat retention. The range of temperatures used is typically within 39 to 43 °C. Alternating magnetic fields can be used to heat up magnetic iron oxide nanoparticles concentrated inside the tumor tissue. More recently developed nanoshells rely on the illumination with a near-infrared laser. Nanoparticles can be designed to actively target the surface receptors on the malignant cells by coating the nanoparticles with the appropriate antibodies.

Recent Developments

Tissue Engineering. Natural bone surface often contains features that are about 100 nm across. If the surface of an artificial bone implant was left smooth, the body would try to reject it. So the smooth surface is likely to cause production of a fibrous tissue covering the surface of the implant. This layer reduces the bone-implant contact, which may result in loosening of the implant and further inflammation. It was demonstrated that by creating nanosized features on the surface of the hip or knee prosthesis, one could reduce the chances of rejection as well as stimulate the production of osteoblasts. The osteoblasts are the cells responsible for the growth of the bone matrix and are found on the advancing surface of the developing bone. The effect was demonstrated with polymeric, ceramic, and more recently, metal materials. More than 90% of the human bone cells from suspension adhered to the nanostructured metal surface, but only 50% did in the control sample. In the end, this finding would allow the design of a more durable and longer lasting hip or knee replacement and reduce the chances of the implant getting loose. Titanium is a well-known bone repairing material widely used in orthopedics and dentistry. It has a high fracture resistance, ductility, and weight-to-strength ratio. Unfortunately, it suffers from the lack of bioactivity, as it does not support cell adhesion and growth well. Apatite coatings are known to be bioactive and to bond to the bone. Hence, several techniques were used to produce an apatite coating on titanium. Those coatings suffer from thickness nonuniformity, poor adhesion, and low mechanical strength. In addition, a stable porous structure is required to support the nutrients' transport through the cell growth. It was shown that using a biomimetic approach — a slow growth of nanostructured apatite film from the simulated body fluid — resulted in the formation of a strongly adherent, uniform nanoporous layer. The layer was found to be built of 60 nm crystallites and possess a stable nanoporous structure and bioactivity. A real bone is a nanocomposite material, composed of hydroxyapatite crystallites in the organic matrix, which is mainly composed of collagen. Thanks to that, the bone is mechanically tough and plastic, so it can recover from mechanical damage. The actual nanoscale mechanism leading to this useful combination of properties is still debated. An artificial hybrid material was prepared from 15 to 18 nm ceramic nanoparticles and poly (methyl methacrylate) copolymer. Using the tribology approach, a viscoelastic behavior (healing) of the human teeth was demonstrated. An investigated hybrid material, deposited as a coating on the tooth surface, improved scratch resistance as well as possessed a healing behavior similar to that of the tooth.

Cancer Therapy. Photodynamic cancer therapy is based on the destruction of the cancer cells by laser-generated atomic oxygen, which is cytotoxic. A greater quantity of a special dye that is used to generate the atomic oxygen is taken in by the cancer cells when compared with a healthy tissue. Hence, only the cancer cells are destroyed and then exposed to a laser radiation. Unfortunately, the remaining dye molecules migrate to the skin and the eyes and make the patient very sensitive to the daylight exposure. This effect can last for up to 6 weeks. To avoid this side effect, the hydrophobic version of the dye molecule was enclosed inside a porous nanoparticle. The dye stayed trapped inside the Ormosil nanoparticle and did not spread to the other parts of the body. At the same time, its oxygen-generating ability has not been affected and the pore size of about 1 nm freely allowed for the oxygen to diffuse out.

Another recently suggested approach is to use gold-coated nanoshells. A surface plasmon resonance effect causes an intense size-dependent absorbance of the near-infrared light by the gold shells. This wavelength of light falls into the optical transparency window of the biological tissue, which can be used to detect cancerous growth up to a certain depth. Compact, powerful, and relatively inexpensive semiconductor lasers are readily available to generate light at this wavelength. The nanoshells are injected into a blood stream and rapidly taken up by the cancerous cells, as they possess a higher metabolism rate. Laser light is absorbed by the gold shells and converted into a local heating, which only kills the cancer cells and spares the healthy tissue. Surface-modified carbon nanotubes can also be used for the same purpose as they would absorb light in the infrared region and convert it into heat.

Multicolor Optical Coding for Biological Assays. The ever increasing research in proteomics and genomic generates an escalating number of sequence data and requires development of high throughput screening technologies. Realistically, various array technologies that are currently used in parallel analysis are likely to reach saturation when several array elements exceed several millions. A three-dimensional approach, based on optical "bar coding" of polymer particles in solution, is limited only by the number of unique tags one can reliably produce and detect. Single quantum dots of compound semiconductors were successfully used as a replacement of organic dyes in various bio-tagging applications. This idea has been taken one step further by combining differently sized and, hence, having different fluorescent colors quantum dots, and combining them in polymeric microbeads. A precise control of quantum dot ratios has been achieved. The selection of nanoparticles used in those experiments had 6 different colors as well as 10 intensities. It is enough to encode over one million combinations. The uniformity and reproducibility of beads was high, allowing for bead identification accuracies of 99.99%.

Manipulation of Cells and Biomolecules. Fictionalized magnetic nanoparticles have found many applications, including cell separation and probing; these and other applications are discussed in a recent review. Most of the magnetic particles studied so far are spherical, which somewhat limits the possibilities to make these nanoparticles multifunctional. Alternative cylindrically shaped nanoparticles can be created by employing metal electrodeposition into a nanoporous alumina template. Depending on the properties of the template, the nanocylinder radius can be selected in the range of 5 to 500 nm while their length can be as big as 60 μm. By sequentially depositing various thicknesses of different metals, the structure and the magnetic properties of individual cylinders can be tuned widely. As surface chemistry for functionalization of metal surfaces is well developed, different ligands can be selectively attached to different segments. For example, porphyrins with thiol or carboxyl linkers were simultaneously attached to the gold or nickel segments, respectively. Thus, it is possible to produce magnetic nanowires with spatially segregated fluorescent parts. In addition, because of the large aspect ratios, the residual magnetization of these nanowires can be high. Hence, the weaker magnetic field can be used to drive them. It has been shown that a self-assembly of magnetic nanowires in suspension can be controlled by weak external magnetic fields. This would potentially allow controlling cell assembly in different shapes and forms. Moreover, an external magnetic field can be combined with a lithographically defined magnetic pattern ("magnetic trapping").

Protein Detection. Proteins are the important part of the cell's language, machinery, and structure, and understanding their functionalities is extremely important for further progress in human well-being. Gold nanoparticles are widely used in immunohistochemistry to identify the protein–protein interaction. However, the multiple simultaneous detection capabilities of this technique are limited. Surface-enhanced Raman scattering spectroscopy is a well-established technique for detection and identification of single dye molecules. By combining both methods in a single nanoparticle probe, one can drastically improve the multiplexing capabilities of protein probes. The group of Prof. Mirkin has designed a sophisticated multifunctional probe that is built around a 13 nm gold nanoparticle. The nanoparticles are coated with hydrophilic oligonucleotides containing a Raman dye at one end and terminally capped with a small molecule recognition element (e.g., biotin). Moreover, this molecule is catalytically active and will be coated with silver in the solution of Ag(I) and hydroquinone. After the probe is attached to a small molecule or an antigen it is designed to detect, the substrate is exposed to silver and hydroquinone solution. Silver plating is happening close to the Raman dye, which allows for dye signature detection with a standard Raman microscope. Apart from being able to recognize small molecules, this probe can be modified to contain antibodies on the surface to recognize proteins. When tested in the protein array format against both small molecules and proteins, the probe has shown no cross-reactivity.

Commercial Exploration

Some companies involved in the development and commercialization of nanomaterials in biological and medical applications are listed below (Table 1). Most of the companies are small recent spinouts of various research institutions. Although not exhausting, this is a representative selection reflecting current industrial trends. Most companies are developing pharmaceutical applications, mainly for drug delivery. Several companies exploit quantum size effects in semiconductor nanocrystals for tagging biomolecules or use bioconjugated gold nanoparticles for labeling various cellular parts. Many companies are applying nanoceramic materials to tissue engineering and orthopedics. Most major and established pharmaceutical companies have internal research programs on drug delivery that are on formulations or dispersions containing components down to nanosizes. Colloidal silver is widely used in antimicrobial formulations and dressings. The high reactivity of titania nanoparticles, either on their own or then illuminated with UV light, is also used for bactericidal purposes in filters. Enhanced catalytic properties of surfaces of nano-ceramics or those of noble metals like platinum are used to destruct dangerous toxins and other hazardous organic materials.

Future Directions

As it stands, most commercial nanoparticle applications in medicine are geared toward drug delivery. In the biosciences, nanoparticles are replacing organic dyes in the applications that require high photo-stability as well as high multiplexing capabilities. There are some developments in directing and remotely controlling the functions of nanoprobes, for example, driving magnetic nanoparticles to the tumor and then making them either to release the drug load or just heating them to destroy the surrounding tissue. The major trend in further development of nanomaterials is to make them multifunctional and controllable by external signals or by local environment, thus essentially turning them into nanodevices.

HEALTH ISSUES (15)

It has been shown by several researchers that nanomaterials can enter the human body through several ports. Accidental or involuntary contact during production or use is most likely to happen via the lungs from where a rapid translocation through the blood stream is possible to other vital organs. On the cellular level, an ability to act as a gene vector has been demonstrated for nanoparticles. Carbon black nanoparticles have been implicated in interfering with cell signaling. There is work that demonstrates uses of DNA for the size separation of carbon nanotubes. The DNA strand just wraps around it if the tube diameter is right. Although excellent for separation purposes, it raises some concerns over the consequences of carbon nanotubes entering the human body.

Ports of Entry

Human skin, intestinal tract, and lungs are always in direct contact with the environment. Whereas skin acts as a barrier, lungs and intestinal tract also allow transport (passive and/or active) of various substances like water, nutrients, or oxygen. As a result, they are likely to be a first port of entry for the nanomaterials' journey into the human body. Our knowledge in this field mainly comes from drug delivery (pharmaceutical research) and toxicology (xenobiotics) studies.

Human skin functions as a strict barrier, and no essential elements are taken up through the skin (except radiation necessary to buildup vitamin D). The lungs exchange oxygen and carbon dioxide with the environment, and some water escapes with warm exhaled air. The intestinal tract

Table 1. Examples of Companies Commercializing Nanomaterials for Bio- and Medical Applications

Company	Applications	Technology
Advectus Life Sciences Inc.	Drug delivery	Polymeric nanoparticles engineered to carry anti-tumor drug across the blood-brain barrier
Alnis Biosciences, Inc.	Bio-pharmaceutical	Biodegradable polymeric nanoparticles for drug delivery
Argonide	Membrane filtration, implants	Nanoporous ceramic materials for endotoxin filtration, orthopedic and dental implants, DNA and protein separation
Biophan Technologies, Inc.	MRI shielding, nanomagnetic particles for guided drug delivery and release	Nanomagnetic/carbon composite materials to shield medical devices from RF fields
Capsulution NanoScience AG	Pharmaceutical coatings to improve solubility of drugs	Layer-by-layer poly-electrolyte coatings, 8–50 nm
Dynal Biotech (Invitrogen)	Cell/biomolecule separation	Superparamagnetic beads
Eiffel Technologies	Drug delivery	Reducing size of the drug particles to 50–100 nm
Evident Technologies	Luminescent biomarkers	Semiconductor quantum dots with amine or carboxyl groups on the surface, emission from 350 to 2500 nm
NanoBio Corporation	Pharmaceutical	Antimicrobial nano-emulsions
NanoCarrier Co., Ltd	Drug delivery	Micellar polymer nanoparticles for encapsulation of drugs, proteins, DNA
NanoPharm AG	Drug delivery	Polybutilcyanoacrylate nanoparticles are coated with drugs and then with surfactant, can go across the blood–brain barrier
Nanoprobes, Inc.	Gold nanoparticles for biological markers	Gold nanoparticles bioconjugates for TEM and/or fluorescent microscopy
Nanoshpere, Inc.	Gold biomarkers	DNA barcode attached to each nanoprobe for identification purposes, PCR is used to amplify the signal; also catalytic silver deposition to amplify the signal using surface plasmon resonance
NanoMed Pharmaceutical, Inc.	Drug delivery	Nanoparticles for drug delivery
PSiVida Ltd	Tissue engineering, implants, drugs and gene delivery, biofiltration	Exploiting material properties of nanostructured porous silicone
QuantumDot Corporation	Luminescent biomarkers	Bioconjugated semiconductor quantum dots

is in close contact with all of the materials taken up orally; here all nutrients (except gases) are exchanged between the body and the environment.

The histology of the environmental contact sides of these three organs is significantly different. The skin of an adult human is roughly 1.5 m^2; in area and is at most places covered with a relatively thick first barrier (10 μm), which is built of strongly keratinized dead cells. This first barrier is difficult to pass for ionic compounds as well as for water-soluble molecules.

The lung consists of two different parts: airways (transporting the air in and out the lungs) and alveoli (gas exchange areas). Our two lungs contain about 2,300 km of airways and 300 million alveoli. The surface area of the lungs is 140 m^2 in adults, as big as a tennis court. The airways are a relatively robust barrier, an active epithelium protected with a viscous layer of mucus. In the gas exchange area, the barrier between the alveolar wall and the capillaries is very thin. The air in the lumen of the alveoli is just 0.5 μm away from the blood flow. The large surface area of the alveoli and the intense air–blood contact in this region makes the alveoli less well protected against environmental damage when compared with airways.

The intestinal tract is a more complex barrier and exchange side; it is the most important portal for macromolecules to enter the body. From the stomach, only small molecules can diffuse through the epithelium. The epithelium of the small and large intestines is in close contact with ingested material so that nutrients can be used. A mixture of disaccharides, peptides, fatty acids, and monoglycerides generated by digestion in the small intestine are further transformed and taken in. The area of the gastrointestine tract (G_IT) is estimated as 200 m^2.

Lung. Most nanosized spherical solid materials will easily enter the lungs and reach the alveoli. These particles can be cleared from the lungs, as long as the clearance mechanisms are not affected by the particles or any other cause. Nanosized particles are more likely to hamper the clearance, resulting in a higher burden, possibly amplifying any possible chronic effects caused by these particles. It is also important to note that the specific particle surface area is probably a better indication for maximum tolerated exposure level than total mass. Inhaled nanofibers (diameter smaller than 100 nm) also can enter the alveoli. In addition, their clearing would depend on the length of the specific fiber. Recent publications on the pulmonary effects of carbon nanotubes confirm the intuitive fear that the nanosized fiber can induce a general nonspecific pulmonary response. Passage of solid material from the pulmonary epithelium to the circulation seems to be restricted to nanoparticles. The issue of particle translocation still needs to be clarified: both the trans-epithelial transport in the alveoli and the transport via nerve cells. Thus, the role of factors governing particle translocation such as the way of exposure, dose, size, surface chemistry, and time course should be investigated. For instance, it would be also very important to know how and to what extent lung inflammation modulates the extrapulmonary translocation of particles. Solid inhaled particles are a risk for those who suffer from cardiovascular disease. Experimental data indicate that probably many inhaled particles can affect cardiovascular parameters, via pulmonary inflammation. Nanosized particles, after passage in the circulation, can also play a direct role in, e.g., thrombogenesis.

Intestinal Tract. Already in 1926, it was recognized by Kumagai that particles could translocate from the lumen of the intestinal tract via aggregations of intestinal lymphatic tissue [Peirel's Patches (PP)] containing M-cells (specialized phagocytic enterocytes). Particulate uptake happens not only via the M-cells in the PP and the isolated follicles of the gut-associated lymphoid tissue, but also via the normal intestinal enterocytes. There have been several excellent reviews on the subject of intestinal uptake of particles. Uptake of inert particles has been shown to occur trans-cellular through normal enterocytes and PP via M-cells and, to a lesser extent, across paracellular pathways. Initially it was assumed that the PP did not discriminate strongly in the type and size of the absorb particles. Later it has been shown that modified characteristics, such as particle size, the surface charge of particles, attachment of ligands, or coating with surfactants, offers possibilities of site-specific targeting to different regions of the GIT including the PP.

The kinetics of particle translocation in the intestine depends on diffusion and accessibility through mucus, initial contact with enterocyte or M-cell, cellular trafficking, and post-translocation events. Cationic nanometer-sized particles became entrapped in the negatively charged mucus, whereas negatively charged nanoparticles can diffuse across this layer. The smaller the particle diameter, the faster they could permutate the mucus to reach the colonic enterocytes. Once in the sub-mucosal tissue, particles can enter both lymphatic and capillaries. Particles entering the lymphatic are probably important in the induction of secretory immune responses, whereas those that enter the capillaries become systemic and can reach different organs. It has been suggested that the disruption of the epithelial barrier function by apoptosis of enterocytes is a possible trigger mechanism for mucosal inflammation. The patho-physiological role of M-cells is unclear; for example, it has been found that in Crohn's disease, M-cells are lost from the epithelium. Diseases other than of gut origin, for example, diabetes, also have marked effects on the ability of the GIT to translocate particles. In general, the intestinal uptake of particles is understood better and studied in more detail than pulmonary and skin uptake. Because of this advantage, it is maybe possible to predict the behavior of some particles in the intestines.

Skin. Skin is an important barrier, protecting against insult from the environment. The skin is structured in three layers: the epidermis, the dermis, and the subcutaneous layer. The outer layer of the epidermis, the stratum corneum (SC), covers the entire outside of the body. In the SC we find only dead cells, which are strongly keratinized. For most chemicals, the SC is the rate-limiting barrier to percutaneous absorption (penetration). The skin of most mammalian species is covered with hair.

At the sites where hair follicles grow, the barrier capacity of the skin differs slightly from the "normal" stratified squamous epidermis. Most studies concerning penetration of materials into the skin have focused on whether drugs penetrate through the skin using different formulations containing chemicals and/or particulate materials as a vehicle. The main types of particulate materials commonly used in contact with skin are liposomes, solid poorly soluble materials such as TiO_2, and polymer particulates and submicron emulsion particles such as solid lipid nanoparticles. TiO_2 particles are often used in sunscreens to absorb UV light and therefore to protect skin against sunburn or genetic damage. It has been reported by Lademann et al. that micrometer-sized particles of TiO_2 get through the human stratum corneum and even into some hair follicles, including their deeper parts. However, the authors did not interpret this observation as penetration into living layers of the skin. In a recent review, it was stated that "very small titanium dioxide particles (e.g. 5–20 nm) penetrate into the skin and can interact with the immune system." Unfortunately, this has not been discussed any further.

Penetration of nonmetallic solid materials such as biodegradable poly(D,L-lactic-co-glycolic acid (PLGA)) microparticles, 1 to 10 μm with a mean diameter of 4.61 ± 0.8 μm, were studied after application on to porcine skin. The number of microparticles in the skin decreased with the depth (measured from the airside toward the subcutaneous layer). At 120 μm depth (where viable dermis is present), a relative high number of particles was found; at 400 μm (dermis), some microparticles were still observed. At a depth of 500 μm, no microparticles were found. In the skin of persons, who had an impaired lymphatic drainage of the lower legs, soil microparticles, frequently 0.4-0.5 μm but as larger particles of 25 μm diameter, were found in the dermis of the foot in a patient with endemic elephantiasis. The particles are observed to be in the phagosomes of macrophages or in the cytoplasm of other cells. The failure to conduct lymph to the node produces a permanent deposit of silica in the dermal tissues (a parallel is drawn with similar deposits in the lung in pneumoconiosis). This indicates that soil particles penetrate through (damaged) skin, most probably in every person, and normally are removed via the lymphatic system.

Liposomes penetrate the skin in a size-dependent manner. Microsized, and even submicron sized, liposomes do not easily penetrate into the viable epidermis, whereas liposomes with an average diameter of 272 nm can reach into the viable epidermis and some are found in the dermis. Smaller sized liposomes of 116 and 71 nm were found in higher concentration in the dermis. Emzaloid particles, a type of submicron emulsion particle such as liposomes and nonionic surfactant vesicles (niosomes), with a diameter of

50 nm to 1 μm, were detected in the epidermis in association with the cell membranes after application to human skin. The authors suggested that single molecules, which make up the particles, might penetrate the intercellular spaces and, at certain regions in the stratum corneum, can accumulate and reform into microspheres. In a subsequent experiment, it was shown that the used formulation allowed penetration of the spheres into melanoma cells, even to the nucleus.

From the limited literature on nanoparticles penetrating the skin, some conclusions can be drawn. First, penetration of the skin barrier is size dependent, and nanosized particles are more likely to enter more deeply into the skin than larger ones. Second, different types of particles are found in the deeper layers of the skin, and currently, it is impossible to predict the behavior of a particle in the skin. Third, materials, which can dissolve or leach from a particle (e.g., metals), or break into smaller parts (e.g., Emzaloid particles), can possibly penetrate into the skin.

Currently, there is no direct indication that particles, that had penetrated the skin also entered the systemic circulation. The observation that particles in the skin can be phagocytized by macrophages, Langerhan cells, or other cells is a possible road toward skin sensitization.

Summary of Health Risks

Particles in the nanosize range can certainly enter the human body via the lungs and the intestines; penetration via the skin is less evident. It is possible that some particles can penetrate deep into the dermis. The chances of penetration would depend on the size and surface properties of the particles and on the point of contact in the lung, intestines, or skin.

After penetration, the distribution of the particles in the body is a strong function of the surface characteristics of the particle. It seems that size can restrict the free movement of particles. The target organ-tissue or cell of a nanoparticle needs to be investigated, particularly in the case of potentially hazardous compounds. Before developing an *in vitro* test, it is essential to know the pharmacokinetic behavior of different types of nanoparticles; therefore, it would be important to compose a database of health risks associated with different nanoparticles.

Beside the study of the health effects of the nanomaterials, investigations should also take into consideration the presence of contaminates, such as metal catalysts present in nanotubes and their role in the observed health effects.

The increased risk of cardiopulmonary diseases requires specific measures to be taken for every newly produced or used nanoparticle. There is no universal "nanoparticle" to fit all cases; each nanomaterial should be treated individually when health risks are expected. The tests currently used to test the safety of materials should be applicable to identify hazardous nanoparticles.

BIBLIOGRAPHY

1. Alivisatos P. The use of nanocrystals in biological detection. Nature Biotechnol 2004;22:47–52.
2. Brus L. Electronic wave function in semiconductor clusters: Experiment and theory. J Phys Chem 1986;90:2555–2560.
3. Kreuter J. Nanoparticles. In: Kreuter J, editor. Colloidal drug delivery systems. New York: Marcel Dekker; 1994.
4. Mikheev NB. Radioactive colloidal solutions and suspensions for medical use. At Energy Rev 1976;14:3–36.
5. Horisberger M. Colloidal gold: a cytochemical marker for light and fluorescent microscopy and for transmission and scanning electron microscopy. Scan Electron Microsc 1981 (Pt 2):9–31.
6. Yavari AR. Mechanically prepared nanocrystalline materials. Mater T JIM 1995;36:228–239.
7. Huczko A. Template-based synthesis of nanomaterials. Appl Phys A-Mater 2000;70:365–376.
8. Gurav A, Kodas T, Pluym T, Xiong Y. Aerosol processing of materials. Aerosol Sci Technol 1993;19:411–452.
9. Jain RA. The manufacturing techniques of various drug loaded biodegradable poly(lactide-*co*-glycolide) (PLGA) devices. Biomaterials 2000;21:2475–2490.
10. Zhang SG. Emerging biological materials through molecular self-assembly. Biotechnol Adv 2002;20:321–339.
11. Caruso F. Nanoengineering of particle surfaces. Adv Mate 2001;13:11–22.
12. Salata OV. Applications of nanoparticles in biology and medicine. J Nanobiotechnol 2004;2:3–8.
13. Soppimatha KS, Aminabhavia TM, Kulkarnia AR, Rudzinski WE. Biodegradable polymeric nanoparticles as drug delivery devices. J Controlled Release 2001;70:1–20.
14. Merisko-Liversidge E, Liversidge GG, Cooper ER. Nanosizing: A formulation approach for poorly-water-soluble compounds. Eur J Pharm Sci 2003;18:113–120.
15. Hoet PH, Bruske-Hohlfeld I, Salata OV. Nanoparticles — known and unknown health risks. J Nanobiotechnol 2004;2:12–26.

FURTHER READING

General

Moriarty P. Nanostructured materials. *Rep Prog Phys* 2001;64: 297–381.

Schmid G, Baumle M, Geerkens M, Heim I, Osemann C, Sawitowski T. Current and future applications of nanoclusters. *Chem Soc Rev* 1999;28:179–185.

Fabrication of Nanoparticles

Bourgeat-Lami E. Organic-inorganic nanostructured colloids. *J Nanosci Nanotechnol* 2002;2:1–24.

Fendler JH. Colloid chemical approach to nanotechnology. *Korean J Chem Eng* 2001;18:1–13.

Gaffet E, Abdellaoui M, Malhourouxgaffet N. Formation of nanostructural materials induced by mechanical processings. *Mater T JIM* 1995;36:198–209.

Meier W. Polymer nanocapsules. *Chem Soc Rev* 2000;29:295–303.

Meisel D. Inorganic small colloidal particles. *Curr Opin Colloid In* 1997;2:188–191.

Shimomura M, Sawadaishi T. Bottom-up strategy of materials fabrication: a new trend in nanotechnology of soft materials. *Curr Opin Colloid In* 2001;6:11–16.

Tomalia DA, Wang ZG, Tirrell M. Experimental self-assembly: the many facets of self-assembly. *Curr Opin Colloid In* 1999;4:3–5.

Ullmann M, Friedlander SK, Schmidt-Ott A. Nanoparticle formation by laser ablation. *J Nanoparticle Res* 2002;4:499–509.

Yu SH. Hydrothermal/solvothermal processing of advanced ceramic materials. *J Ceram Soc Jpn* 2001;109:S65–S75.

Biological and Medical Applications

de la Isla A, Brostow W, Bujard B, Estevez M, Rodriguez JR, Vargas S, Castano VM. Nanohybrid scratch resistant coating for teeth and bone viscoelasticity manifested in tribology. *Mat Res Innovat* 2003;7:110–114.

Han M-Y, Gao X, Su JZ, Nie S-M. Quantum-dot-tagged microbeads for multiplexed optical coding of biomolecules. *Nature Biotechnol* 2001;19:631–635.

Loo C, Lowery A, Halas N, West J, Drezek R. Immunotargeted nanoshells for integrated cancer imaging and therapy. *Nano Lett* 2005;5:709–711. Ozkan M. Quantum dots and other nanoparticles: What can they offer to drug discovery? *DDT* 2004;9:1065–1071.

Ma J, Wong H, Kong LB, Peng KW. Biomimetic processing of nanocrystallite bioactive apatite coating on titanium. *Nanotechnology* 2003;14:619–623.

Molday RS, MacKenzie D. Immunospecific ferromagnetic iron dextran reagents for the labeling and magnetic separation of cells. *J Immunol Methods* 1982;52:353–367.

Pankhurst QA, Connolly J, Jones SK, Dobson J. Applications of magnetic nanoparticles in biomedicine. *J Phys D: Appl Phys* 2003;36:R167–R181.

Panyam J, Labhasetwar V. Biodegradable nanoparticles for drug and gene delivery to cells and tissue. *Adv Drug Del Rev* 2003;55: 329–347.

Parak WJ, Gerion D, Pellegrino T, Zanchet D, Micheel C, Williams SC, Boudreau R, Le Gros MA, Larabell CA, Alivisatos AP. Biological applications of colloidal nanocrystals. *Nanotechnology* 2003;14:R15–R27.

Pricea RL, Waidb MC, Haberstroha KM, Webster TJ. Selective bone cell adhesion on formulations containing carbon nanofibres. *Biomaterials* 2003;24:1877–1887.

Reich DH, Tanase M, Hultgren A, Bauer LA, Chen CS, Meyer GJ. Biological applications of multifunctional magnetic nanowires. *J Appl Phys* 2003;93:7275–7280.

Roy I, Ohulchanskyy TY, Pudavar HE, Bergey EJ, Oseroff AR, Morgan J, Dougherty TJ, Prasad PN. Ceramic-based nanoparticles entrapping water-insoluble photosensitizing anticancer drugs: A novel drug-carrier system for photodynamic therapy. *J Am Chem Soc* 2003;125:7860–7865.

Sinani VA, Koktysh DS, Yun BG, Matts RL, Pappas TC, Motamedi M, Thomas SN, Kotov NA. Collagen coating promotes biocompatibility of semiconductor nanoparticles in stratified LBL films. *Nano Lett* 2003;3:1177–1182.

Taton TA. Nanostructures as tailored biological probes. *Trends Biotechnol* 2002;20:277–279.

Weissleder R, Elizondo G, Wittenburg J, Rabito CA, Bengele HH, Josephson L. Ultrasmall superparamagnetic iron oxide:characterization of a new class of contrast agents for MR imaging. *Radiology* 1990;175:489–493.

Yoshida J, Kobayashi T. Intracellular hyperthermia for cancer using magnetite cationic liposomes. *J Magn Magn Mater* 1999;194:176–184.

Zhang Y, Kohler N, Zhang M. Surface modification of superparamagnetic magnetite nanoparticles and their intracellular uptake. *Biomaterials* 2002;23:1553–1561.

Commercial Exploration

Mazzola L. Commercializing nanotechnology. *Nature Biotechnol* 2003;21:1137–1143.

Paull R, Wolfe J, Hebert P, Sinkula M. Investing in nanotechnology. *Nature Biotechnol* 2003;21:1144–1147.

Health Risks

Anonymous. Nanotech is not so scary. *Nature* 2003;421:299.

Borm PJ. Particle toxicology: From coal mining to nanotechnology. *Inhal Toxicol* 2000;14:311–324.

Sanfeld A, Steinchen A. Does the size of small objects influence chemical reactivity in living systems? *C R Biol* 2003;326:141–147.

UK Royal Society and Royal Academy of Engineering. Nanoscience and nanotechnologies: opportunities and uncertainties. Final Report. (2004). Available: http://www.nanotec.org.uk/finalReport.htm.

See also DRUG DELIVERY SYSTEMS; MICROSCOPY, ELECTRON; TISSUE ENGINEERING.

NEAR-FIELD MICROSCOPY AND SPECTROSCOPY.

See MICROSCOPY AND SPECTROSCOPY, NEAR-FIELD.

NEONATAL MONITORING

MICHAEL R. NEUMAN
Michigan Technological
University
Houghton, Michigan

INTRODUCTION

The care of premature and newborn infants is quite different from other areas of clinical medicine. The infant represents a special patient with special problems not found in other patients. For this reason a subspecialty of pediatrics dealing with these patients has been established. Neonatology is concerned with newly born infants including those prematurely delivered and those delivered at term. The field generally covers infants through the first month of normal newborn life, and so for prematurely born infants this can be several additional months.

Neonatology includes special hospital care for infants who require it. In the case of the prematures, this involves specialized life-support systems, as well as special considerations for nutrition, thermal control, fluid and electrolyte therapy, pulmonary support, and elimination of products of metabolism. While the full-term infant generally only requires routine well-baby care, there are special cases that require intensive hospital care as well. These include treatment of infants of diabetic mothers, some infants delivered by cesarean section, infants with hemolytic diseases, infants who encounter respiratory distress, and other less common problems. Special hospital care is also necessary for infants requiring surgery. These infants are generally born with severe congenital anomalies that would be life threatening if not immediately repaired. These include anomalies of the gastrointestinal system, urinary tract, cardiovascular system, and nervous system. Pediatric surgery has developed to the point where many of these problems can be corrected, and the infant can grow and lead a normal life following the surgery.

The neonatal intensive care unit is a special nursery in major tertiary care hospitals that is devoted to the care of premature or other infants who require critical care. This unit is similar to its adult counterpart in that each patient is surrounded by equipment necessary for life support, diagnosis, and therapy. Often, as indicated in Fig. 1, the patient appears to be insignificant in the large array of equipment, but, of course, this is not the case. Nursing

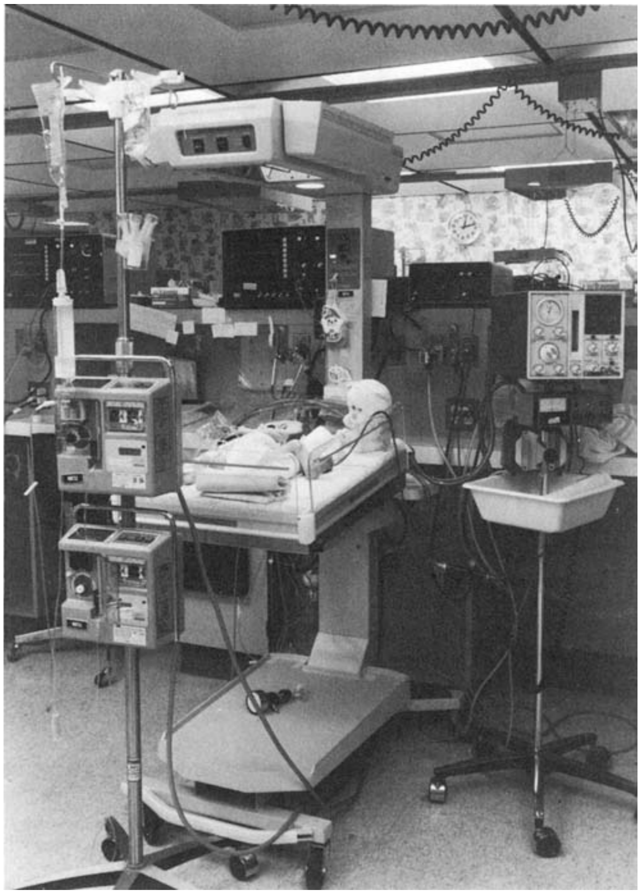


Figure 1. Typical infant station in a neonatal intensive care unit showing infant warmer, infusion pumps, ventilator, transcutaneous oxygen instrument, cardiopulmonary monitor, bilirubin lights, and other miscellaneous apparatus.

functions in the neonatal intensive care unit are very important. The patient/nurse ratio is small, and the nursing staff must be familiar with the equipment as well as special procedures and precautions in caring for these special patients.

Electronic monitoring of the infant plays an important role in neonatal intensive care. Not only does it allow the clinical caregivers to follow vital signs, such as pulse rate, temperature, blood pressure, and respiration rate, but other critical variables in the care of these special patients can be followed as well. These include blood gas tensions, acid–base balance, bilirubin, and glucose concentrations. Monitoring is especially important in fluid therapy for it can provide precise data for fluid control of these very small patients. Electronic monitoring, however, goes beyond just monitoring the patient and its physiologic functions. A good neonatal intensive care unit also monitors the functioning of life-support systems. These include incubators for maintaining an appropriate thermal environment, ventilators for providing respiratory support, and phototherapy units for the control of bilirubin.

Although electronic monitoring devices for just about all of the areas mentioned in the previous paragraph are used in adult intensive care medicine, their application in neonatology often represents a unique aspect of the technology. The infant should not be viewed as a miniature adult, but rather he/she is a unique physiologic entity. Although similar variables are measured to those measured in adults, they often must be measured in different ways. Frequently, sensors unique for infants must be applied because the sensors used for adults when interfaced to the infant might provide errors or change the variable being measured by their very presence. Size is an important aspect here. If one considers a sensor to be used on an infant and compares it to a sensor for the same variable on an adult, in most cases although the sensor for the infant is smaller than that for the adult, the ratio of sizes of the two sensors is quite different from the ratio of sizes of the different patients. Although sensors for use on infants are reduced in size, they are still quite large when compared to the size of the subject. This is especially true for premature infants and can result in the sensors actually interfering with the care of the patient.

There are also special problems related to the measurement of physiologic variables in infants resulting from the special physiology of newborns and especially premature newborns. One first must realize that a newborn has come to live in a new environment quite different from the uterus. In the case of premature infants, they are not ready for this major change in their lives, and special considerations need to be made to minimize the transitional trauma. In the case of the premature, some of the body systems are immature and not ready for life outside of the uterus. Two notable examples of this are the control of temperature and control of respiration. Both are obviously unnecessary in the uterus, but become crucial in extrauterine life. Instrumentation to assist these control systems or to detect when they are not functioning properly is essential in the care of many premature infants.

One also must realize in applying instrumentation systems for premature infants that the patient in many cases is much more fragile than an adult patient. Fluid and electrolyte balance has already been indicated as an important aspect of neonatal monitoring and control. When one considers some of the very small premature infants that are cared for in neonatal intensive care units today, this can be better appreciated. Infants between 500 and 1,000 g can be successfully cared for and nurtured until they are old enough and grow enough to go home with their parents. These very small babies, however, can easily run into problems if they receive either too much or too little fluid. Since feeding of these very small infants can be done by intravenous hyperalimentation, the possibility of a fluid overload is always present since it takes a certain amount of fluid to transport the nutritional requirements of the infant. Another example of the fragility of these very small patients is the simple problem of attaching devices to the infant's skin. In some infants, the skin is very sensitive and can easily become irritated by the attachment procedure or substance.

This article, looks more closely at electronic monitoring systems for neonatal intensive care and emphasize those aspects of these monitoring systems that differ from similar monitors for adult patients. The reader is encouraged to supplement information contained in the following paragraphs with other articles from this encyclopedia dealing with the sensors and instrumentation for similar monitoring in adults.

CARDIAC MONITORING

Cardiac monitoring involves the continuous assessment of heart function by electronic measurement of the electrocardiogram and determination of heart rate and rhythm from it by means of electronic signal processing. As such, cardiac monitors for neonatal use are very similar to those for use with adults. There are, however, two major differences. The sensors used with both types of monitors are biopotential electrodes, and in the case of infants the interface between the electrodes and the patients has more stringent requirements than in the adult case. Second, cardiac monitors designed for use with infants frequently are incorporated into cardiorespiratory monitors that include instrumentation for determining breathing rate and apnea as well as cardiac function.

The primary use of cardiac monitors for infants is in determining heart rate. These electronic devices are designed to indicate conditions of bradycardia (low heart rate) and tachycardia (high heart rate) by determining the heart rate from the electrocardiogram. In the case of infants with heart diseases, cardiac monitors are used to detect various arrhythmias as well.

Cardiac monitors for use with infants are organized similarly to their adult counterparts (see MONITORING, HEMODYNAMIC). There are some minor differences due to the fact that infant heart rates are higher than those of adults, and the Q–S interval of the infant electrocardiogram is less than it is in the adult. Thus, heart rate alarm circuits need to be able to respond to higher rates in the infant case than in the adult case. For example, it is not at all unusual to set the bradycardia alarm level at a rate of 90 or 100 $\text{beats}\cdot\text{min}^{-1}$ for an infant, which is well above the resting heart rate of a normal adult. Filtering circuits in the monitor for infants must be different from those of adult monitors for optimal noise reduction due to the different configuration of the neonatal electrocardiogram. Generally, bandpass filters used for isolating the QRS complex will have a higher center frequency than in the adult case.

Two types of cardiotachometer circuits can be used in cardiac monitors (1). The averaging cardiotachometer determines the mean number of heartbeats per predetermined interval to establish the heart rate. The mean R–R interval over a number, of heartbeats can also be used in average heart rate determination. In such systems the heart rate is calculated by averaging over from as few as three to as many as fifteen or more heartbeats. An instantaneous or beat-to-beat cardiotachometer determines the heart rate for each measured R–R interval. This type of cardiotachometer must be used when one is interested in beat-to-beat variability of the heart rate.

Biopotential electrodes for use with cardiac monitors for infants are usually scaled down versions of skin surface electrodes used for adult cardiac monitoring. As pointed out earlier, the scale factor does not correspond to the body size ratio between the neonate and an adult, and the smallest commercially available skin surface electrodes for neonates only approaches about one-fourth the size of those used in adults. For this reason, electrodes used with neonatal cardiac monitors and their method of attachment can cover a large portion of the neonatal thorax. This is especially true with the small premature infant and can interfere with direct observation of chest wall movements, an important diagnostic method. In addition to size, shape and flexibility of the electrode are important for biopotential electrodes in neonates. Stiff, flat electrode surfaces will not conform well to the curved, compliant surface of the infant. This means that optimal electrical contact is not always possible and it, therefore, becomes more difficult to hold electrodes in place. This problem is further complicated by the fact that the neonatal skin can be sensitive to the electrode adhesive. It is not at all unusual to find skin irritation and ulceration as a result of placement of biopotential electrodes on the infant. Such skin lesions are usually the result of the adhesive and the electrode attachment system, although the electrode itself can in some cases be the problem as well.

Since electrodes are relatively large on the small infant, an additional problem develops. The materials used in many electrode systems are X-ray opaque; hence, it is necessary to remove the electrodes when X rays are taken so that shadows do not appear in the resulting radiograph. Some biopotential electrodes especially developed for neonates have minimized this problem by utilizing special electrode structures that are translucent or transparent to X rays (2). These electrodes are based upon thin films of metals, usually silver, deposited upon polymer films or strips or various fabric materials. These films are sufficiently thin to allow X rays to penetrate with little absorption, and the plastic or polymer substrate is also X-ray transparent. Such electrodes have the advantage of increased flexibility, which helps them to remain in place for longer periods of time. In intensive care units, however, it is a good idea to change electrodes every 48 h to minimize the risk of infection.

Electrode lead wires and patient cables present special problems for cardiac monitors used with infants. Lead wires should be flexible so as not to apply forces to the electrodes that could cause them to become loose, but this increased flexibility makes it easier for them to become ensnarled with themselves and the infant. The potential for strangulation on older, active infants is always present. The connectors between the lead wires and the patient cable also present special problems. They must be capable of maintaining their connection with an active infant and provide a means of connection that will be unique for these components. The possibility of inadvertently connecting the lead wires, and hence the infant, to the power line must be eliminated (3).

RESPIRATORY MONITORING

Respiratory monitoring is the most frequently applied form of electronic monitoring in neonatology. In its most common application, it is used to identify periods of apnea and to set off an alarm when these periods exceed a predetermined limit. There are direct and indirect methods of sensing alveolar ventilation and breathing effort. The direct methods are those in which the sensor is coupled to the airway and measures the movement or other properties of the air transported into and out of the lungs. In the indirect methods, the sensor looks at variables related to air movement, but not at the air movement itself. Indirect methods involve no contact with the airway or the air being moved into or away from the lungs. Usually, indirect methods are noninvasive and can be mounted on or near the body surface. Some of the most frequently applied methods are described in the following paragraphs.

Direct Methods

Various direct methods of sensing breathing effort and ventilation have been in use in the pulmonary physiology and pülmonary function laboratories for many years. These involve the measurement of volume, flow, and composition of inspired and expired gasses. Table 1 lists some of the principal methods that have been used for the direct measurement of respiration in infants and neonates. Most of these methods are not appropriate for clinical monitoring, since they involve direct connection to the infant airway through the use of a mask over the mouth and nose or an endotrachial canula. In other cases a sensor must be located at the nasal-oral area for signal detection. These methods are, however, useful in some cases for diagnostic studies carried out for periods from several hours to overnight in the hospital setting.

Many of the methods listed in Table 1 are described in detail in the article on pulmonary function testing, and therefore are not repeated here. Others, however, have special application to neonatal monitoring and will be mentioned.

Pneumotachography. Clinicians and researchers involved in neonatal and infant care agree for the most part that the best measurement of ventilation can be obtained using the pneumotachograph (4). Although this involves direct connection to the airway and can add some dead space due to the plumbing, it, with an appropriate electronic integrator, provides good volume and flow measurements that can be used as a standard against which other direct or indirect methods can be calibrated and evaluated. Identical instrumentation as used for adults can be applied in the infant case, but it must be recognized that dead space due to the instrumentation represents a more important problem with the infant than it does with the adult. Lower flows and volumes as well as faster respiration rates will be encountered with infants than with adults.

Table 1. Direct Methods of Sensing Breathing and Ventilation

Method	Primary Sensed Variable
Pneumotachograph	Flow volume
Anemometer	Flow velocity
Expired air temperature	Temperature
Air turbulence sounds	Sound
Spirometer	Volume

Capnography. Special carbon dioxide sensors have been developed for measuring air expired from the lungs, and these are used as the basis of a direct respiration monitor (5). Expired air has a higher percentage of carbon dioxide than inspired air, and this can be sensed by placing an open-ended tube at the nose or mouth so that it samples the air entering and leaving the airway. The sampled gas is transported along the tube to an instrument that contains a rapidly responding carbon dioxide sensor. This is generally a sensor that detects the increased absorption of infrared (IR) radiation by carbon dioxide-containing gas. Thus, when a sample of expired gas reaches the sensor, an increase in carbon dioxide content is indicated, while a decrease in carbon dioxide is seen in samples of air about to be inspired. There is a delay in response of this instrument due to the time it takes the gas to be transported through the tube to the sensor; thus, it is important to have rapid passage through this tube to minimize this delay. This can present some problems since the tube must be thin and flexible and, therefore, offers a relatively high resistance to the flow of gas. While it is generally not necessary to have a quantitative measure of carbon dioxide for respiration monitoring, the system can be refined to the point where it can measure the carbon dioxide content of the end tidal expired air, which is the gas that actually was in the alveoli (6).

Temperature Sensor. Similar sensing systems based upon temperature variations have also been used to monitor respiration (7). These generally can be divided into two types: one that measures temperature differences between inspired and expired air and one that measures the cooling of a heated probe as inspired or expired air is transported past it. In both cases, the temperature sensor of choice is a small, low mass, and therefore fast responding, thermistor. In the first mode of operation, the thermistor changes its resistance proportionally to the change in temperature of the air drawn over it. This can then be electronically detected and processed to determine respiration rate. It is also possible to heat the thermistor by an electrical current. Some of this heat will be dissipated convectively by the air passing over the sensor. As the flow of air over the thermistor increases, more heat will be drawn from the thermistor, and it will cool to a lower temperature. Changes in the thermistor's temperature can be determined by measuring its electrical resistance. Thus, an electrically heated thermistor will cool during both inspiration and expiration, and it will become warmer in the interval between these two phases when air is not passing over it. This type of anemometer gives a respiration pattern that appears to be twice the breathing rate, whereas the unheated thermistor gives a pattern that is the same as the breathing rate. An important consideration in using the nasal thermistor for ventilation measurement is its placement in the flowing air. For young infants, the sensor package can be taped to the nose or face so that the thermistor itself is near the center of one nostril. Another technique is to place a structure containing two thermistors under the nose so that each thermistor is under one nostril and expired air flows over both thermistors.

Nasal temperature sensors, such as thermistors, have been used for monitoring ventilation in research studies and for making physiologic recordings in the hospital and in the laboratory (8). Their advantage is that the electronic circuit for processing the signal is relatively simple and inexpensive compared with other techniques. The major problem of the method is the placement of the thermistor on the infant and maintaining it in place. Thermistors can also become covered with mucus or condensed water, which can greatly reduce their response time. Most investigators who use this technique prefer the temperature sensing rather than the flow-detecting mode. The devices can also be used with radiotelemetry systems to eliminate the wire between the thermistor on the subject's face and the remainder of the monitoring apparatus (9).

Although thermistors have a high sensitivity and can be realized in a form with very low mass, they are fragile when in this low mass form and are relatively expensive components. Low mass, high surface area resistance temperature sensors can also be fabricated using thin- and thick-film temperature sensitive resistors.(10) These can either be fabricated from metal films with relatively high temperature coefficients of resistance or more sensitive films of thermistor materials. Single use disposable sensors have been produced for use in infant and adult sleep studies as shown in Figure 2.

Sound Measurement. Air passing over the end of an open tube generates sound by producing local turbulence. A miniature microphone at the other end of the tube can detect this sound, and the level of sound detected is roughly proportional to the turbulence and, hence, the air flowing past the open end. Nasal air flow can thus be detected by placing the open end of the tube in the stream of inspired or expired air at the nose by taping the tube to the infant's face in much the same way as was done for the carbon dioxide sensor mentioned previously (11). As with the thermistor anemometer, this technique can detect changes for both inspired and expired air and will give a pattern that appears to indicate double the actual respiration rate. The method has been demonstrated to give efficacious monitoring results, but can suffer from sensitivity to extraneous sounds other than the air passing the open ended tube. This can lead to incorrect detection of breaths.

Indirect Sensors of Ventilation

There are a wide variety of indirect sensors of ventilation that can be applied to monitoring in infants. Table 2 lists some of the principal examples of these various types of sensors and sensing systems, and those with aspects unique to neonatal monitoring will be described in the following paragraphs. The main advantage of the indirect methods of sensing ventilation is that attachment to the subject is easier than for the direct measurements and less likely to interfere with breathing patterns. Of the methods described in Table 2 and this section, the transthoracic electrical impedance method is the one used in most presently available respiration-apnea monitors for both hospital and home use. This, therefore, will be described in greatest detail in a separate section.

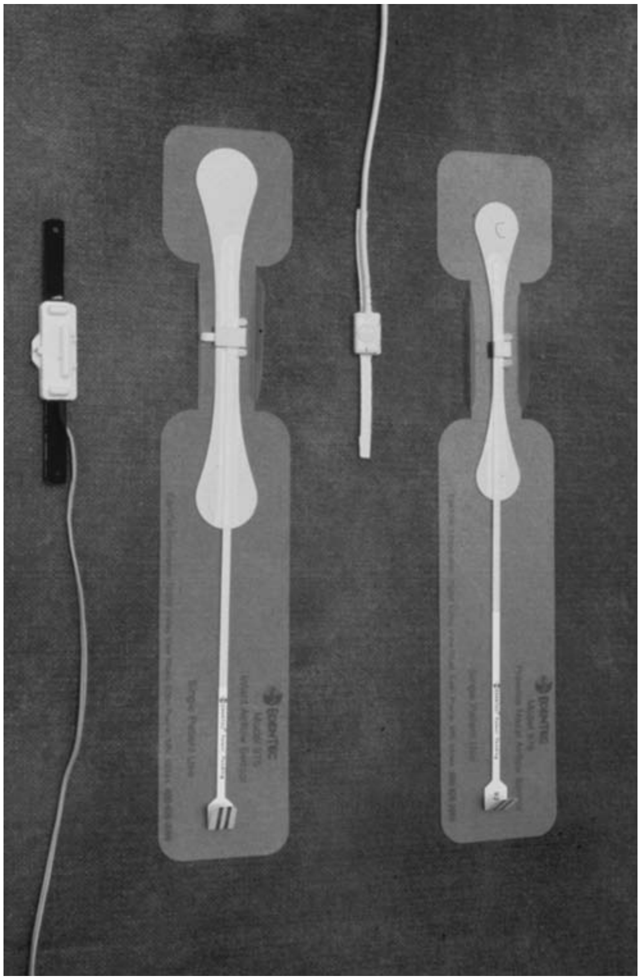


Figure 2. An example of commercially available thick-film nasal temperature sensors for measurement of breathing patterns. The small sensors on the illustration are conventional thermistor sensors.

The Whole-Body Plethysmograph. This method is used primarily in pulmonary function testing, and the reader is referred to the article on this subject for details on the method. Miniature whole-body plethysmographs have been designed for use with neonates and infants, but this

Table 2. Indirect Methods of Sensing Breathing and Ventilation

- Transthoracic electrical impedance
- Whole-body plethysmograph
- Contacting motion sensors
  - Strain gage
  - Air-filled capsule or vest
  - Magnetometer
  - Inductance respirometry
- Noncontacting motion sensors
  - Motion-sensing pad
  - Radiation reflection
  - Variable capacitance sensor
- Electromyography
- Breath sounds
- Intraesophageal pressure

application is strictly for purposes of research or diagnostic studies. The technique is not appropriate for routine clinical monitoring.

Contacting Motion Sensors. Breathing effort involves the movement of different parts of the body for pulmonary ventilation to occur. Sensors can be placed upon and attached to an infant to measure this motion. These contacting motion sensors pick up movements of the chest and/or abdomen, and there are several different types of sensors that fall within this category. These are described in the following sections.

Strain Gage Displacement Sensors. Strain gages measure small displacements or strain in an electrical conductor by measuring changes in its electrical resistance. Most strain gages used for general measurements are made of thin metal foils or wires and are useful for measuring only very small displacements due to their very low mechanical compliance. A special type of strain gage consisting of a compliant, thin-walled, rubber capillary tube filled with mercury was developed by Whitney as a limb plethysmograph (12). This compliant device can be placed on the chest or abdomen of an infant such that breathing movements cause it to stretch and contract without offering significant mechanical constraint to the breathing efforts of the infant. By taping the ends of such a strain gage at different points on the chest or abdomen such that the gage is slightly stretched, the changes in electrical resistance of the gage can then be used to monitor infant breathing movements. Simple electronic resistance measurement circuitry can be used for processing the signal.

This technique is used primarily in research and in some rare cases for in-hospital monitoring and recording of infant respiration patterns. Its limitations are related to use of a toxic substance that could escape from the sensor and put the infant at risk. In addition, the mercury column frequently becomes interrupted after several days of use, thereby limiting the sensor's reliability for infant monitoring. Nevertheless, workers who use this sensor for monitoring purposes are enthusiastic about its reliability in picking up high quality respiration patterns.

Air-Filled Capsule or Vest. Breathing efforts of an infant can also be determined for chest or abdominal movements by a sensor consisting of an air-filled compliant tube, disk, or entire vest attached around an infant. The tube and disk can be taped to the infant's chest or abdomen in a fashion similar to the strain gage, and the structures will be stretched or compressed by the infant's breathing movements. This causes the pressure of the air within to increase or decrease as a result of volume changes, and this pressure variation can be measured by coupling the sensor to a sensitive pressure transducer through a fine-gage flexible tube. The advantage of this system is that the sensors on the infant are simple and inexpensive and thus can be considered disposable devices. Since only air is contained within the sensors, they are not toxic and are much more reliable than the mercury strain gages. They can be produced as inexpensive disposable sensors.

Displacement Magnetometers. The magnetic field from a permanent magnet or an electromagnet decreases as one gets farther from the magnet. By placing such a magnet on an infant's chest or abdomen with a detector located on the back of the subject or underneath the infant, differences in separation between the magnet and the detector can be sensed as the infant breathes (13). It is important that such a system be designed so that it will only respond to breathing movements and will be insensitive to other movements of the infant. Unfortunately, this is not always the case, and sensors of this type can respond to infant limb movement as well as movements between the infant and the pad upon which it is placed.

Inductance Respirometry. The inductance of a loop of wire is proportional to the area enclosed by that loop. If a wire is incorporated in a compliant belt in a zigzag fashion so that the wire does not interfere with the stretching of the belt, such a belt can be wrapped around the chest or abdomen of an infant to form a loop. As the infant inhales or exhales the area enclosed by this loop will change, and so the inductance of the loop will also change. These changes can be measured by appropriate electronic circuits and used to indicate breathing efforts. Investigators have shown that the use of such a loop around the chest and the abdomen of an adult can, when appropriately calibrated, measure tidal volume as well as respiratory effort (14). Although the system is simple in concept, realizing it in practice can involve complicated and therefore costly electronic circuitry (15). Often as the subject moves to a new position, the calibration constant relating inductance and volume will change thereby making the instrument less quantitative, yet still allowing it to be suitable for qualitative measurements. Variations in tidal volume measurements using this technology have been reported by Brooks et al. (16) Since the instrument is sensitive to inductance changes in the wire loop, anything in the vicinity of the wire that affects its inductance also will affect the measurement. Thus, the instrument can also be sensitive to moving electrical conductors or other magnetic materials in the vicinity of the infant.

Noncontacting Motion Sensors. Sensors of infant breathing effort and pulmonary ventilation that detect breathing movements of the infant without direct patient contact fit in this category. These sensors can consist of devices that are placed under the infant or can sense movement of the infant by means of a remotely located sensor. A clinician, in effect, is an indirect motion sensor when he or she determines infant breathing patterns by watching movements of the chest and abdomen. Devices in this category have a special appeal for monitoring systems that are used outside of the hospital, such as instruments for use in the home. With many of the noncontacting sensors, the infant-sensor interface can be created by individuals who do not have specialized training. For example, the motion sensing pad discussed in the next paragraph is attached to the infant by simply placing the infant on top of it in a bassinet or crib.

Motion Sensing Pad. Movements of neonates and infants can be sensed by a flexible pad that responds to compression by producing an electrical signal when the infant is placed on top of the pad. There are two different forms of this sensor that can be used for motion detection. The first utilizes a piezoelectric polymer film, polyvinylidene fluoride, that has its surfaces metalized to form electrical contacts. Depending on the piezoelectric properties of the film, an electrical signal is produced between the metalized layers when the polymer is either compressed or flexed. In the former case, the polymer film and its metalized electrode need only to be packaged in an appropriate pad structure to be used, while in the latter case the package must be a little more complex with the polymer film positioned between two corrugated, flexible layers so that compression of the structure causes the piezoelectric polymer to be flexed (17). The second form of the pad uses an electret material to generate the electrical signal. The actual pad structure in this case is similar to that for the piezoelectric material.

The sensitive portion of the motion sensing pad structure is usually smaller than the overall size of the infant and is located under the infant's thoracic and/or lumbar regions. Infant breathing efforts result in periodic compression of the pad as the center of mass of the infant shifts cephalad and caudad with respiratory motion. This generates a periodic electrical signal related to the breathing effort.

The major limitation of the motion sensing pad is its sensitivity to movements other than those related to respiratory efforts of the infant. Other body movements can be picked up by the sensor, and the device can even respond to movements that are not associated with the infant at all, such as an adult walking near or bumping the bassinet or crib, a heavy truck, train, or subway passing nearby, or even earthquakes.

Radiation Reflection. Electromagnetic radiation in the microwave range (radar) or ultrasonic radiation (sonar) can be reflected from the surface of an infant. If this surface is moving, as, for example, would be the chest or abdominal wall during breathing efforts, the reflected radiation will be shifted in frequency according to the Doppler effect. In some cases the reflected signal's amplitude will be shifted as well as a result of this motion. These changes can be detected and used to sense breathing efforts without actually contacting the infant. The problem with these methods is that the movement of any surface that reflects the radiation will be detected. Body movements of the infant that are unrelated to respiratory movements can be detected and mistakenly identified as breathing effort, and even in some cases movement of objects in the vicinity of the infant, such as a sheet of paper shifting due to air currents, will also be detected as infant respiration. Thus, this type of monitor has the possibility of indicating apparent breathing activity during periods of apnea if moving objects other than the infant are within the range of the radiation sensor. This technique of noncontacting detection of breathing is not considered to be reliable enough for routine clinical use, and a commercial device based on this principle has been withdrawn from the market.

Variable Capacitance Displacement Sensor. A parallel plate capacitor can be fabricated so that an infant is placed between the parallel conducting planes. For example, such a capacitor could be formed in an incubator by having the base upon which the mattress and infant are placed serving as one plate of the capacitor and having the second plate just inside the top of the incubator (18). To maintain good clinical practice, this second plate should consist of a transparent conductor, such as an indium tin oxide film, so that it does not interfere with a clinician's ability to observe the patient. Since a major component of the infant's tissue is water, and water has a relatively high dielectric constant compared to air, movements of the infant will produce changes in capacitance between plates that can be detected by an electronic circuit. Such changes can be the result of breathing movements by the infant, but they also can result from other infant movement or movement of some other materials in the vicinity of the conducting plates. Therefore, for this system to be effective, adequate electrical shielding of the capacitor is essential. Thus, this indirect motion sensor suffers from some of the same problems as other sensors in this classification: the lack of specificity for breathing movements.

Electromyography. Many different muscles are involved in breathing activity. The diaphragm is the principal muscle for pulmonary ventilation, but the accessory muscles of the chest wall including the intercostal muscles are also involved, Electromyographic activity of the diaphragm and intercostal muscles can be sensed from electrodes on the chest surface. By measuring these signals, one can determine if respiratory efforts are being made, although such measurements cannot be quantitative with regard to the extent of the effort or the volume of gas moved (19). Unfortunately, other muscles in the vicinity of the electrodes that are not involved in breathing also produce electromyographic signals. These signals can severely interfere with those associated with respiration, and this is especially true when the infant is moving. This represents a serious limitation of this method for clinical infant respiration monitoring.

Breath Sounds. Listening to chest sounds through a stethoscope is an important method of physical diagnosis for assessing breathing. The technique can be used for infant monitoring by placing a microphone over the chest or trachea at the base of the neck and processing the electrical signals from this sensor. In addition to the sounds associated with air transport and ventilation, the microphone will pick up other sounds in the body and the environment. Thus, for this type of monitoring to be efficacious, it must be done in a quiet environment. This puts a serious constraint on the practical use of this technique, and it has only been used in limited experimental protocols.

Intraesophageal Pressure. The pressure within the thorax decreases with inspiratory effort and increases with expiratory effort. These changes can be measured by placing a miniature pressure sensor in the thoracic portion of the esophagus or by placing a small balloon at this point and coupling the balloon to an external pressure transducer through a small diameter flexible tube. While this method is invasive, it is not considered a direct method since there is no contact with the flowing air.

An important aspect of intraesophageal pressure measurement is that it represents a standard method that is accepted by physiologists as a measure of respiratory effort. Thus, by combining intraesophageal pressure measurement and the pneumotachograph, one is able to monitor both gas flow and breathing effort. Although both of these methods are generally too complicated for clinical monitoring, they can be used in conjunction with other monitoring methods described in this article as standards against which to assess the other devices.

Transthoracic Electrical Impedance. The electrical impedance across the chest undergoes small variations that are associated with respiratory effort. The measurement of these variations is the basis of the most frequently used infant respiration and apnea monitoring technique. The following section describes the basic principle of operation, the methods of signal processing, and sources of error for this technique.

RESPIRATION MONITORING BY TRANSTHORACIC ELECTRICAL IMPEDANCE

The chest contains many different materials ranging from bone to air. Each of these materials has its own electrical properties and of its own unique location in the thorax. One can roughly represent a cross section of the infant chest as shown in Fig. 3, where the major components consist of chest wall, lungs, heart, and major blood vessels. The various tissues contained in these structures range in electrical conductivity from blood, which is a relatively good conductor, to air, which is an insulator. Both of these materials in the thoracic cavity show a change in volume with time over the cardiac and breathing cycles. Blood varies in volume over the cardiac cycle due to changes in the amount of blood in the heart and the vascular compartments. Air undergoes wide volume changes in the lungs during normal breathing. Thus, the electrical impedance of the lungs and heart will change as the volume of air and blood in each, respectively, changes. If we want to measure the impedance variation due to these volume changes, this can be done by placing electrodes on the surface of each structure. If it were practical to do this, we would see large changes in impedance as the volumes of the respective structures change. Unfortunately, it is not possible to place electrodes on the structures that are to be measured and so these large impedance differences are not seen in practice.

Electrodes must be placed upon the surface of the skin for practical electrical impedance measurements on infants. Most of the current passing between the electrodes will travel through the chest wall and will not pass through the heart and lungs because of the low resistivity of the tissues in the chest wall. Thus, the changes in impedance of the heart and lungs will only represent a small proportion of the impedance measured between the electrodes. Fig. 3 schematically illustrates the relative distribution of the current through the chest when electrodes are placed on the midclavicular lines at the fourth intercostal space. It is seen that most of the current is conducted along the chest wall, so the chest wall impedance will dominate any measurement.

The actual impedance measured by the monitor consists of more than just the impedance between the electrodes on the chest surface. Since an ac electrical signal is needed to measure the impedance, this signal will affect the measurement as well. Generally, a signal in the frequency range from 20–100 kHz is used. At these frequencies, impedances associated with the electrode, the interface between the electrode and the body, and the lead wires contribute to the measured value along with the actual transthoracic impedance. This is illustrated schematically in Fig. 4. The actual impedances for each block are dependent upon the excitation frequency and the actual structures used, but for most clinical applications the net impedance seen by the monitoring circuit is nominally 500 Ω. Of this, the variation associated with respiration is generally no $>$ 2 Ω and frequently even less. The impedance

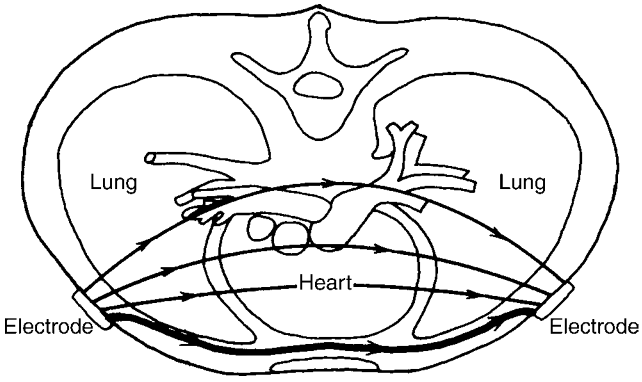


Figure 3. Cross-sectional view of the thorax of an infant showing the current distribution from electrodes placed on the chest wall and excited by a transthoracic impedance type of apnea monitor.

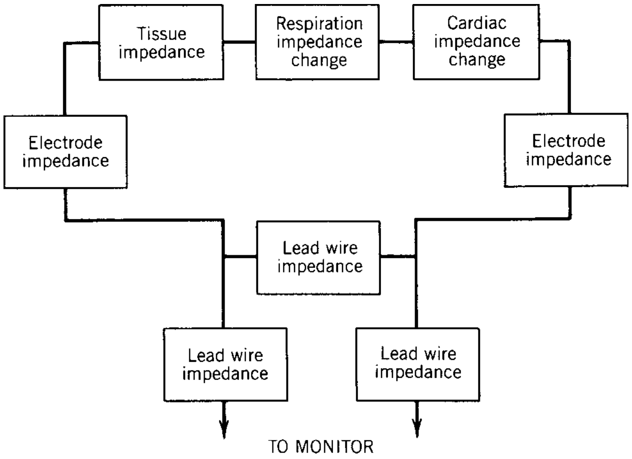


Figure 4. Block diagram of the various impedances seen at the terminals of a transthoracic electrical impedance apnea monitor looking along the lead wires to the patient.

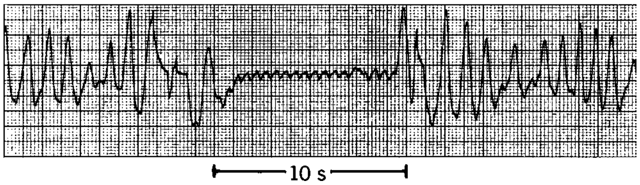


Figure 5. An example of cardiogenic artifact on infant respiration signals from a transthoracic impedance type of monitor illustrating cardiogenic artifact during apnea.

variation associated with the beating heart can be of the same magnitude, although it is generally a little less. Thus, it is seen that a fundamental problem in the indirect measurement of respiration by the transthoracic impedance method is the relatively small changes in impedance associated with the measurement.

To further complicate the situation, each of the nonthoracic impedance components of the circuit illustrated in Fig. 4 can vary in electrical impedance by at least as much if not more than the variation due to respiration. The impedance between the electrode and the infant's skin is strongly dependent on the electrode–skin interface. As electrodes move with respect to the skin, this impedance can vary by amounts much $> 2\ \Omega$. This is also strongly dependent on the type of electrode used and the method that electrically couples it to the skin.

Cardiogenic Artifact

The volume of the heart varies during the cardiac cycle, and so the contribution of the blood to the overall transthoracic impedance will change from systole to diastole. To a lesser extent the vascular component of the chest wall and lungs will also change in blood volume during the cardiac cycle, and this will have some influence on the transthoracic impedance as well, Cardiogenic artifact is illustrated in Fig. 5, which shows a recording of transthoracic impedance from an infant during breathing and during a period of apnea. The cardiogenic artifact is best seen during the apnea, where it appears as a smaller impedance variation occurring at the heart rate. This can be seen by comparing the impedance waveform with a simultaneously recorded electrocardiogram. One notes that the cardiogenic artifact is also present during the breathing activity and appears as a modulation of the respiration waveform.

In the example in Fig. 5 the cardiogenic artifact is relatively small compared to the impedance changes due to breathing, and it is possible to visually differentiate between breathing and apnea by observing this recording. This is not always the case when recording transthoracic impedance as Fig. 6 illustrates. Here one observes periods of breathing and apnea with much stronger cardiogenic artifact. It is difficult to determine what impedance variations are due to breathing and what are due to cardiovascular sources. It is only possible to identify periods of respiration and artifact when the recording is compared with a simultaneous recording of respiration from a recording of abdominal wall movement using a strain gage as shown in Fig. 6. Note that in the case of the impedance signal in this figure, the cardiogenic artifact has two components during each cardiac cycle.

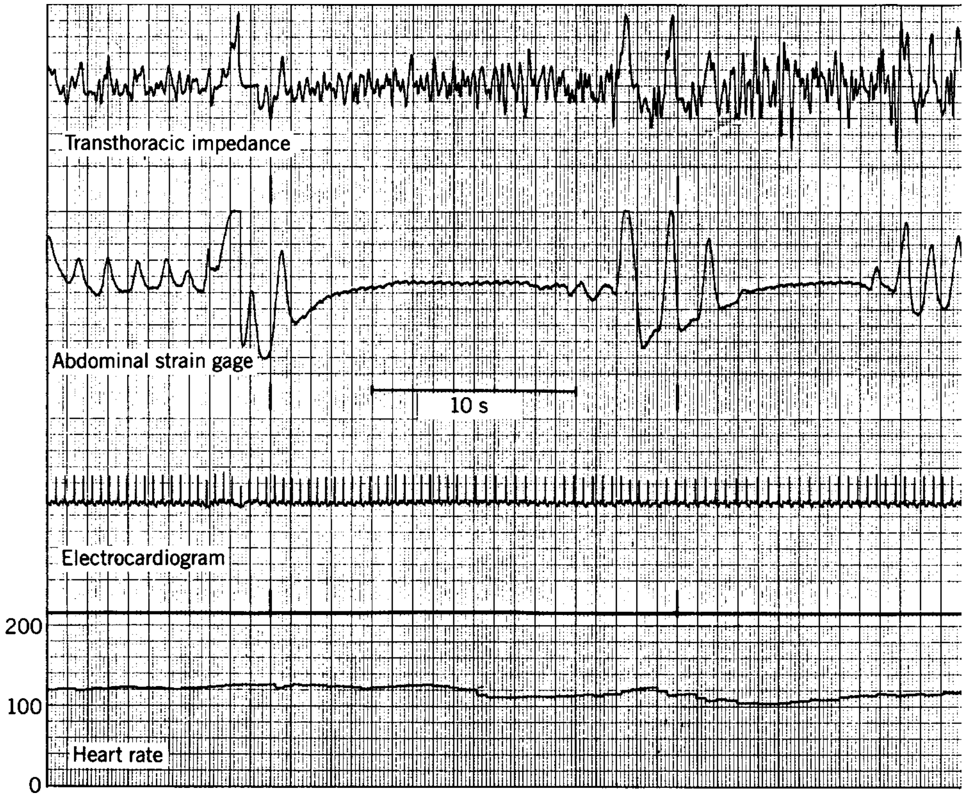


Figure 6. High amplitude cardiogenic artifact is shown on the transthoracic impedance tracing from this recording of multiple signals from a newborn infant. In this case, the transthoracic impedance changes correspond to the electrocardiogram shown on the third trace from the top. Simultaneous recordings from a nasal thermistor and an abdominal strain gage do not show these high frequency variations.

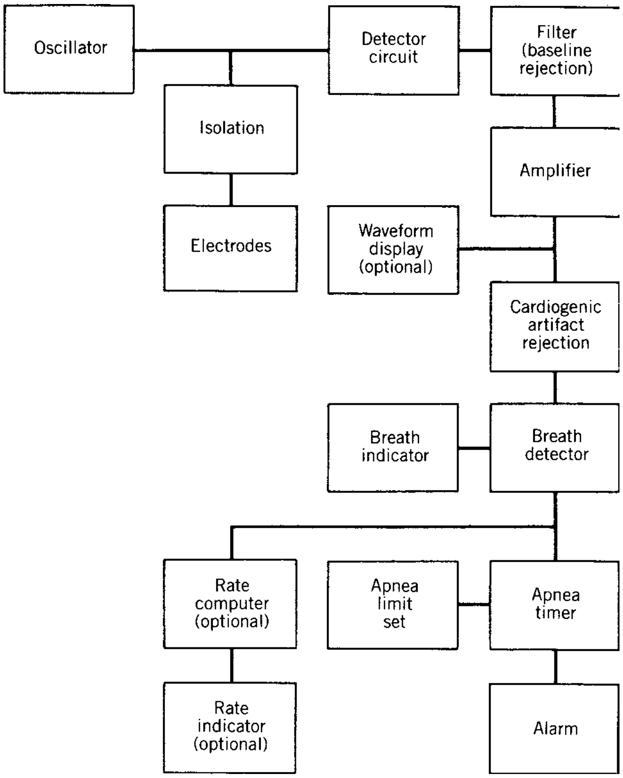


Figure 7. Functional block diagram of a transthoracic impedance infant apnea monitor.

Signal Processing

The electrical signals from the electrical impedance sensor, or any of the other respiration sensors, must be processed to recognize breathing activity and to determine when apnea is present. Different sensors require processors of differing complexity because of different signal characteristics, but the general method of signal processing is the same no matter what sensor is used. The signal processing associated with the electrical impedance method of apnea monitoring will be described in the following paragraphs, since it is one of the most complex as well as most highly developed monitoring systems.

A block diagram of the generalized sections of a transthoracic electrical impedance type of apnea monitor is shown in Fig. 7. The basic functions of the system can be broken down into impedance measurement, breath detection, artifact rejection, apnea identification, and alarm functions. Each of these can be carried out with varying degrees of complexity, and sophisticated signal processing techniques can be used to get the most information out of a less than optimal signal.

A signal generator in the impedance measurement portion of the system produces the excitation signal that is applied to the electrodes. This can either be a sinusoidal or a square wave, and frequently will have a high source impedance so that it behaves as though it was generated by a constant current amplitude source. Passing this current through the lead wire–electrode–body system causes a voltage amplitude proportional to its impedance to appear at the monitor input. Variations in this voltage reflect the variation in impedance. It is therefore important that the current amplitude of the excitation signal remain constant during a measurement. Excitation signal frequency is chosen to be in the range of 20–100 kHz so that electrode–body interface impedances are relatively low, thereby producing less artifact. Detection of individual breaths from a complex breathing signal represents a major task for the respiration monitor. While the design of electronic circuits to carry out such a function on a regular, noise-free, nearly constant amplitude respiration signal such as seen in Fig. 8a presents no problem; very often the respiration waveform is much more complicated and not so easily interpreted, as illustrated in Fig. 8b. Cardiogenic artifact also helps to complicate the signal detection problem since in some cases it can masquerade as a breath. Some of the basic methods of identifying breaths are listed in the following paragraphs. Often individual monitors will use more than one of these in various unique signal processing algorithms.

Fixed Threshold Detection. A breath can be indicated every time the respiration signal crosses a predetermined fixed threshold level. It is important to carefully choose this level so that nearly all breaths cross the threshold, but

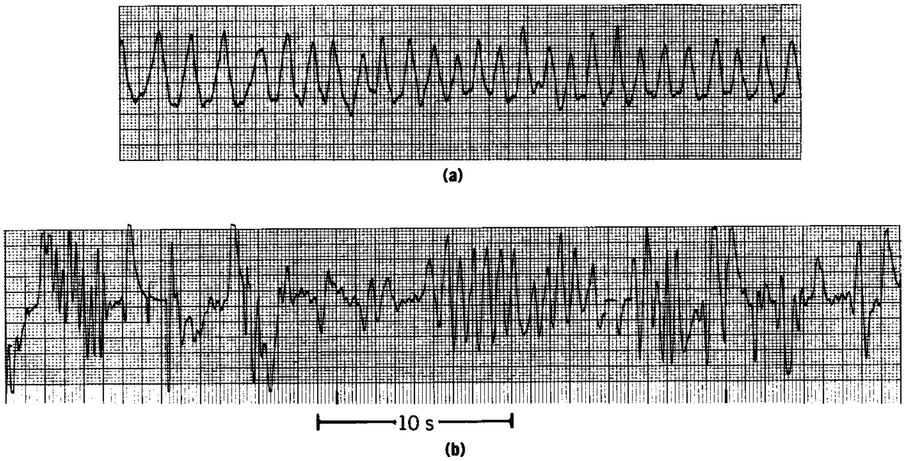


Figure 8. Typical infant respiration signals obtained from infant apnea monitors. (a) The top trace illustrates a relatively quiet signal that can be processed to determine respiration rate and apnea. (b) The bottom trace is a typical example of a noisy signal resulting from infant movement. In this case it would not be easy to determine respiration rate.

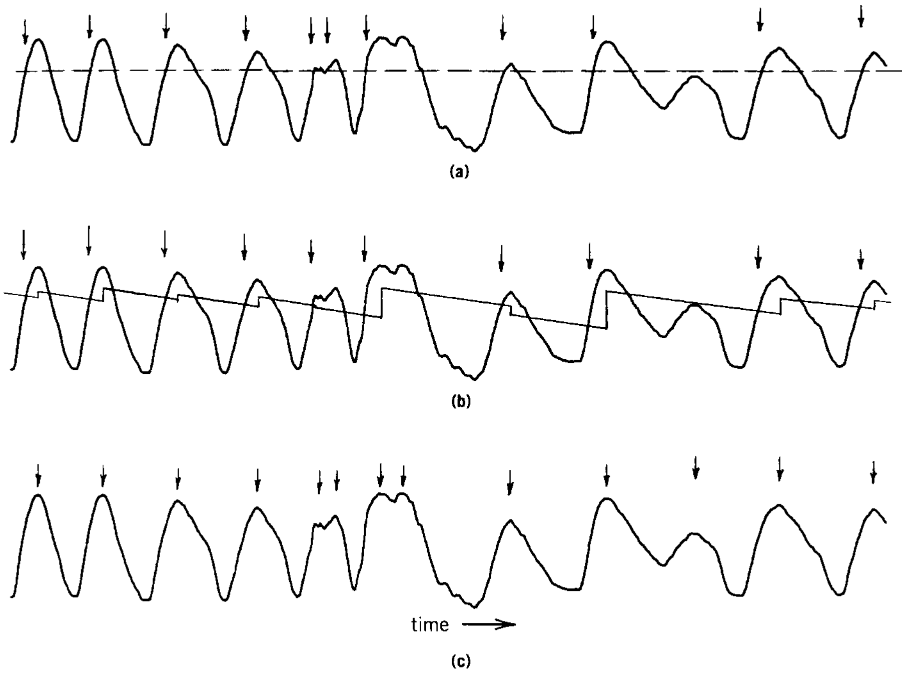


Figure 9. Three identical respiration signals in which different methods of detecting a breath are used. The arrows indicate when the apnea monitor would detect a breath for each method. (a) Fixed threshold detection, note missing breath when signal fails to cross a threshold. (b) Adaptive variable threshold detection, note that breaths can be missed with this method when the amplitude from one breath to the next is significantly different. (c) Peak detector, note that this method can result in double breath detection for signals with multiple peaks.

practically no noise or artifact does. Figure 9a illustrates the basic threshold breath indicator system in which a breath is indicated whenever the signal is greater than the threshold level.

Automatic Gain Control. The fixed threshold method of detection can be improved by preceding the threshold detector with an amplifier that has an automatic gain control. In this way the weaker signals are amplified more than the stronger ones so that all signals appearing at the fixed threshold detector circuit have roughly the same amplitude and will be detected. Although this method makes the fixed threshold detection scheme more reliable, there is also the possibility that noise or cardiogenic artifact will be amplified until it is strong enough to masquerade as a breath during periods of apnea thereby causing the monitor to fail to identify the apnea.

Adaptive Threshold Detection. A variable threshold level can be set by the monitor based upon a preprogrammed algorithm. One common example of this is to have the monitor determine the threshold level based upon the amplitude of the previous breath. This is illustrated in Fig. 9b, where the threshold is set at 80% of the peak amplitude of the previously detected breath. Since this threshold may still be too high if the previous breath had a large amplitude and subsequent breaths were of a relatively low amplitude, this threshold is not fixed, but rather it slowly decreases so that eventually a breath will be detected and the threshold level can be reset. The risk with this type of system is that the threshold will eventually get low enough to detect noise or cardiogenic artifact during an apnea resulting in a breath detected in error. Thus, the algorithm for this adaptive system must have minimum threshold levels that are still well above the noise or cardiogenic artifact level for it to work effectively.

Peak Detector. This circuit recognizes the maximum value of a signal over a short interval of time regardless of the overall amplitude of that signal. The way that a peak detector detects the breaths from a typical respiration waveform is illustrated in Fig. 9c. The basic peak detector can recognize more than one peak in a complex respiration wave. This can give errors if the monitor is used to determine respiration rate. Again, by adding complexity to the signal processing algorithm, this type of error can be greatly reduced.

Filtering. Frequency spectral analysis of infant respiration signals shows that most of the information is contained in the frequency band of 0–6 Hz, and in many cases the band is even narrower (20). Since artifactual signals can exist both within and outside of this frequency range, most apnea monitors filter the respiration signals so that only the frequencies containing information are processed. The type of filtering used depends on the particular monitor design, but any process of filtering can distort the waveforms and may itself introduce artifact. This is especially true when high pass filtering is used to remove the baseline. Thus, filtering can affect the performance of the breath detection method used in the instrument.

Although filtering is an important aspect of the breath detection circuitry, it can in some cases cause motion artifact to begin to look similar to a respiration signal and thus allow the detection circuit to recognize artifact as a breath. Often under the best conditions it is difficult to discriminate between artifact and true breathing signals, and the filtering only further complicates this problem. Nevertheless, without filtering breath detection would be much more difficult.

Pattern Recognition. Computer technology allows algorithms for recognizing various features of the respiration waveform to be applied for breath detection in infant respiration monitors. Features, such as threshold crossing, peaks and valleys, slopes, amplitudes, width, and interval of a respiration wave can be readily detected. More sophisticated algorithms can be trained to recognize breaths that are similar in appearance to preprogrammed waveforms or based on the appearance of previous breaths for a particular patient. Another important aspect of computer recognition of patterns is that the computer can be programmed to ask various questions: Is the measured value physiologically possible? Does the waveform look more like artifact than information? Is the rate too fast? Does the signal correspond too closely to the cardiac cycle so that it might be cardiogenic artifact? Is there more than one peak per breath? All of these techniques of breath detection have advantages and disadvantages for infant monitoring. Each technique, however, imposes constraints on the signal that determine whether it will also detect artifact or miss some true breaths. Even the most sophisticated computer methods suffer from faults such as these and present limitations in breath detection.

Cardiogenic Artifact Rejection. Although cardiogenic artifact represents a major problem when breathing efforts are measured by the transthoracic electrical impedance method, this interference can be seen at times in the output of other indirect sensors of respiration as well. Usually, for these other sensors this artifact is small and does not pose any problem in breath or apnea recognition. Several methods have been used to reduce the problems associated with cardiogenic artifact in the transthoracic electrical impedance type of apnea monitor. Cardiogenic artifact occurs at the heart frequency and its harmonics, which can be different from the periodicity of the respiration signal. In infants the heart rate is usually higher than the respiration rate, although this is not always the case since infants can breath quite rapidly. If the respiration signal containing cardiogenic artifact is passed through a low pass filter having a cutoff frequency that is higher than the expected respiration rates but lower than the heart rates likely to be encountered, much of the cardiogenic artifact can be removed without seriously distorting the respiration signal. The problem with this approach is the selection of a cutoff frequency for the filter. It is generally not possible to find a frequency that is greater than the maximum respiration rate yet less than the minimum heart rate for small infants. Estimated values of such a frequency have to be changed according to the age of the infant, and since bradycardia can be associated with apnea, it is possible that the heart frequency will drop below the filter cutoff frequency during times of apnea, allowing cardiogenic artifact to get into the respiration channel just at the very time when it should be avoided.

The approach of using a filter, however, has merit if the above limitations can be taken into consideration in the design of the filtering system. Although there is no way that a filter can be useful when the heart rate is less than the respiration rate, the filter can help if its cutoff frequency is based upon the apparent respiration and heart rates of the infant. Such adaptive filtering techniques have been successfully used to minimize the effects of cardiogenic artifact.

Since most transthoracic electrical impedance apnea monitors also determine heart rate from the electrocardiogram, this cardiac signal can be used to help identify when a respiration signal consists primarily of cardiogenic artifact. The temporal relationship between the cardiogenic artifact and the electrocardiogram should be constant since both come from the same source. If the respiration signal consists only of cardiogenic artifact, as would be the case during a period of apnea, it is possible to identify the fixed temporal relationship between the signal and the electrocardiogram and therefore reject the signal from being accidentally detected as a breath. The only limitation with this technique is that in rare cases the infant can breath at the same rate as the heart is beating, and the monitor would indicate that an apnea had occurred when in fact it had not.

COMBINATION TRANSTHORACIC IMPEDANCE AND CARDIAC MONITORS

Most commercially available infant apnea monitors take advantage of the fact that the same sensor system, a set of biopotential electrodes, can be used for both transthoracic electrical impedance respiration monitoring and cardiac monitoring. Since the excitation signal for transthoracic impedance monitoring has a frequency of 20 kHz or greater and the highest frequency component of the infant electrocardiogram is < 200 Hz, the excitation signal can be applied to the same electrodes used for obtaining the electrocardiogram. By connecting a low pass filter between the electrodes and the heart rate monitor circuit, this excitation signal can be kept out of the cardiac monitor, and a bandpass filter in the respiration monitor centered at the excitation signal frequency will keep the electrocardiogram and biopotential motion artifact out of the transthoracic impedance monitor circuit.

The combination of respiration and heart rate monitoring in a single instrument helps to identify life-threatening events. If for some reason the respiration monitor fails to recognize prolonged apneas, bradycardia will often be associated with such episodes, and the heart rate monitor will recognize the reduced heart rate and set off an alarm.

MEASUREMENT OF BLOOD GASES

Blood gases refer to the oxygen and carbon dioxide transported by the blood. Acid–base balance is also included in discussions of blood gases since it is closely related to respiratory and metabolic status. Thus, measurements of blood gases are frequently combined with measurements of blood pH. There are invasive and noninvasive methods of measuring blood gases. Both can be used for hospital monitoring of critically ill infants. The principal methods that are used are described in this section.

Invasive Methods

Invasive blood-gas measurement techniques involve direct contact with the circulatory system so that blood samples can be drawn and measured in a laboratory analyzer or a miniature sensor can be placed within the blood stream for continuous measurements. Some of these methods are described in the following paragraphs.

Intraarterial Catheter. The newly born infant has an advantage over other medical patients in that the vessels of the umbilical cord stump can accept a catheter for several hours after birth. Thus, it is possible to introduce a fine-gage, flexible, soft catheter into an umbilical artery of a cardiac or respiratory compromised infant and advance the tip into the aorta so that samples of central arterial blood can be obtained for analysis. Blood samples with a volume of only 50 μL can be analyzed for pH, P_{O_2}, and P_{CO_2} by means of specially designed miniaturized versions of standard analytical chemistry sensors of these variables. Such microblood analyzers are also used for analyzing fetal scalp blood samples. It is important in neonatal applications that only microblood analyzers be used since the total blood volume of very small infants is limited. Since an infant's blood gas status can be labile, it is often necessary to draw many blood samples during the clinical course of care, thus significant blood loss can occur unless very small samples are taken.

Microblood analyzers generally use the inverted glass electrode for pH measurement, a miniaturized Clark electrode for P_{O_2} measurement, and a miniaturized version of the Stowe–Severinghaus sensor for P_{CO_2}. The technology of microblood gas analyzers is well developed, and devices perform reliably in the intensive care situation. Instrumentation is frequently located within the neonatal intensive care unit itself, and respiratory therapists for collecting samples and carrying out the analyses as well as calibrating and maintaining the analyzers serve round the clock.

The major limitation of this sampling technique is that the sample only represents the blood gas status at the time it was taken. Thus, frequent samples must be taken during periods when variations can occur to track these variations, and even with microblood analyzers this can sometimes result in significant blood loss for very small infants. If the method for drawing the blood sample from the infant is stressful, such as a painful vascular puncture or heel stick, the blood gases of the sample will probably not reflect the quiescent status of the patient. As a matter of fact the very act of obtaining the blood sample may be of some risk to the infant since it can temporarily increase hypoxia (21).

The umbilical vessel canulation is not without problems itself. In placing the catheters, one must be careful not to damage the lining of the vessels or perforate a vascular wall resulting in severe bleeding or hemorrhage. Catheters must be made of materials that do not promote thrombosis formation. When catheters are not used for drawing blood, they must be filled with a physiological solution containing an anticoagulant such as Heparin so that blood that diffuses into the tip of the catheter does not clot. Any thrombi formed on the catheter wall or within its lumen can break off and cause embolisms further downstream. For arterial catheters this can be in the blood supply to the lower periphery of the infant, and it is possible to see under perfused feet in infants having an umbilical artery catheter. Catheters in the umbilical vein or peripheral veins can also produce emboli. In this case the clots are returned to the right side of the heart and can go on to produce pulmonary emboli.

Peripheral Blood Samples. Although it is frequently possible to introduce a catheter into an umbilical artery in a newly born infant, this is not always the case, or the need for blood gas monitoring may not arise until the infant is sufficiently old that the umbilical vasculature has permanently closed. In this case it is necessary to canulate a peripheral artery to obtain frequent arterial blood samples. On very small infants this is no minor task since these vessels are very small and difficult to canulate transcutaneously.

An alternative to drawing an arterial blood sample is to take a sample of capillary blood from the skin under conditions where the capillary blood flow has been significantly increased so that the capillary blood appears to be similar to peripheral arterial blood. This can be done, for example, in the heel by first warming an infant's lower leg and foot by wrapping it with warm, wet towels. A blood sample of sufficient size for a microblood analyzer can then be obtained by making a small skin incision with a lancet and collecting the blood sample in a capillary tube in a fashion similar to the technique for obtaining a fetal scalp blood sample (see FETAL MONITORING). Although this technique is not as reliable as sampling from an umbilical artery catheter, it can be used when only a single blood sample is desired and an umbilical catheter would be inappropriate or where it is not possible to place such a catheter. An important limitation of the technique is that the infant's heels can become quite bruised when frequent samples are required and suitable locations for additional samples might no longer be available. When frequent samples are required, it is generally better to attempt canulation of a peripheral artery.

Internal Sensors. Blood gases can be continuously monitored from invasive sensors. Generally, these sensors are incorporated into umbilical artery catheters (22), but tissue measurements have also been demonstrated (23). The most frequently applied technique involves the incorporation of an amperometric oxygen sensor into a catheter system. This can be done either by incorporation of the sensor within the wall of the catheter, by using a double lumen catheter with the sensor in one lumen and the second lumen available for blood samples or infusion, or by using a conventional single lumen catheter with a sensor probe that can be introduced through the lumen so that the sensor projects beyond the distal tip of the catheter.

Oximetry, the measurement of hemoglobin oxygen saturation, can be carried out continuously by means of optical sensors coupled to intravascular catheters or probes. Optical fibers can be incorporated in the wall of a catheter or in an intraluminal probe and used to conduct light to the catheter's distal tip. The light illuminates the blood in the vicinity of the catheter tip, and an adjacent fiber or bundle of fibers collects the backscattered light and conducts it to a photo detector where its intensity is measured. By alternately illuminating the blood with light of two or more different wavelengths, one of which is close to an isosbestic point, and measuring the backscattered light, it is possible to determine the hemoglobin oxygen saturation in the same way as done in laboratory instruments for *in vitro* samples.

Advantages and Disadvantages of Invasive Techniques. The methods described in the previous sections represent direct measurements in that the sensor that is used is in direct contact with the body fluid, usually blood, being measured. This direct contact improves the possibility of accurate measurements. When the sensor is not located in the blood itself but is used to measure samples of blood drawn from the patient, instruments can be frequently calibrated using laboratory standards. Sensors that are used within blood or other tissues have the requirement that they must be small enough to fit in the tissue with minimal damage, either as a part of a catheter or some other probe. The miniaturization process must not compromise accuracy or reproducibility. In cases where microelectronic technology can be used to miniaturize the structures, reproducibility can even be improved in the mass-produced miniature devices as compared to their piece-by-piece-produced larger counterparts. The continuous invasive sensors are also limited in where and when they can be applied. While the umbilical arteries are convenient conduits to the central arterial circulation, they are only patent for a few hours after birth in most newborn infants. Following this time it is very difficult to obtain arterial samples since other vessels must be used. The use of intravascular sensors, and those in tissue as well, also increases the risk of infection and mechanical damage. Care must be taken with intraarterial sensors to avoid serious hemorrhage due to system components becoming disconnected.

Noninvasive Methods

In noninvasive measurement of blood gases, there is no direct contact between the blood or other tissue being measured and the sensor. In this way there is usually less risk to the patient and the technique is easier to apply clinically. The major noninvasive methods used in neonatal monitoring are now described.

Transcutaneous Blood Gas Tension Measurement. One of the major advances in neonatal intensive care monitoring technology was the development of transcutaneous blood gas measurement instrumentation. This allowed the oxygen tension and later the carbon dioxide tension of infants at risk to be continuously monitored without invading the circulatory system (24). These methods make use of a heated sensor placed on the infant's skin that measures the partial pressures of oxygen or carbon dioxide of the blood circulating in the dermal capillary loops under the sensor. The heating of the skin to temperatures of 44 °C arterializes the capillary blood in a manner similar to that used for obtaining capillary blood samples with heel sticks. Although the heating of the blood increases the blood gas tensions in the capillary blood, oxygen consumption by the viable epidermis surrounding the capillaries and diffusional drops through the skin compensate for this increase resulting in good correlations between the transcutaneously measured blood gas tensions and those determined from arterial blood samples in neonates. Sensors can be left in place on neonates for up to four hours, but for longer periods of time it is recommended to move the sensor to a new location to avoid tissue damage due to the elevated temperature. Multiple sensors have been developed in which the heating element is switched between several sensors in the same package periodically so that the overall sensor can be left in place for longer periods of time without producing damage (25).

Although transcutaneous instrumentation can give good correlations between transcutaneous and central arterial blood gas measurements in neonates, it would be misleading to suggest that the transcutaneous instrument is measuring the same thing as is measured from arterial blood samples. Indeed in infants with unimpaired circulatory status, the transcutaneous blood gases and those in the central circulation are similar; however, when there is cardiovascular compromise, heating of the sensor can no longer completely arterialize the capillary blood, and there are significant differences between the transcutaneous and central measurements. Thus, when one makes both transcutaneous and central measurements, differences can be used as a means of identifying shock-related conditions (26).

Transcutaneous Mass Spectrometry. Another noninvasive method for measuring blood gas tensions involves the use of a transcutaneous mass spectrometer (27). A sensor similar to the transcutaneous blood gas sensor in that it contains a heater to arterialize the capillary blood under it is made of a gas-permeable membrane in contact with the skin. This is connected to the mass spectrometer instrument through a fine-bore flexible tube through which an inert carrier gas is circulated to bring the gases that diffuse from the skin into the sensor to the instrument. (For details of this instrument see MASS SPECTROMETERS IN MEDICAL MONITORING). At the present time, mass spectrometry instruments are far more expensive than instruments for electrochemically determining the transcutaneous blood gas tensions. The advantage of the mass spectrometer, however, is that it can simultaneously measure more than a single blood gas component. It can also measure other gases in the blood stream, such as anesthetic agents or special tracers.

Pulse Oximetry. The use of optical techniques to determine the hemoglobin oxygen saturation in blood is well known and is the basis for routine clinical laboratory instrumentation along with the fiber optic catheter oximeter described in the previous section on internal sensors. Oximeters have also been developed for measuring the oxygen saturation transcutaneously. Initial devices measured the continuous steady-state reflection of light of different wavelengths from the surface of the skin. Pigmentation of the skin, unfortunately, limited this technique to qualitative measurements unless the instrument was specifically calibrated to a particular individual at a particular site. Upon examining the backscattered optical signal from the skin, one can notice a small pulsatile component at the heart rate. This is due to the changing blood volume in the capillary beds reflecting the light, and it can be seen for transmitted light as well. By looking at this pulsatile component of the transmitted or reflected light, it is possible to measure only the effect of each fresh bolus of blood entering the capillary bed at systole. This allows the principle of oximetry to be adapted to the transcutaneous measurement of arterial blood hemoglobin oxygen saturation (28). This technique is used for continuously monitoring tissues that can be transilluminated, such as the hand, foot, fingers, toes, ears, and nasal septum. These pulse oximeters have the added advantage that in most applications it is not necessary to arterialize the capillary blood by heating; thus, sensors can be left in place for longer periods of time without risk of tissue injury.

Pulse oximeters have rapidly achieved a major role in neonatal and adult intensive care medicine. It is important to point out that oximetry differs from oxygen tension measurement in that it tells how much oxygen is carried by the hemoglobin. To know total oxygen transport one needs to know the amount of hemoglobin in the blood as well as the profusion of the tissue in question. Thus, oximetry can with some additional data be quite useful in determining whether adequate amounts of oxygen are being supplied to vital tissues. There is one aspect of neonatal monitoring, however, where oximetry is of little assistance. The condition, known as retinopathy of prematurity, is found in premature infants and thought to be related to the newly formed capillaries in the retina, which are exposed to blood of elevated oxygen tension in infants who are receiving oxygen therapy. An important aspect of oxygen monitoring in premature infants is to determine if the arterial blood oxygen tension becomes elevated, so that the amount of oxygen that the infant breathes can be reduced to protect the eyes. If retinopathy of prematurity occurs as a result of elevated oxygen tensions, blindness can result. Thus, to truly protect the patient from this condition, one must measure oxygen tension not hemoglobin oxygen saturation.

Pulse oximeters in routine clinical use are primarily based on the the transmission mode of operation, although backscatter oximeters have also been developed (29,30). The clinical instruments, therefore, are limited in terms of where they can be attached to the subject. Generally, these positions are found on the periphery and are, unfortunately, the first to experience diminished circulation under shock or preshock conditions. Another limitation of currently available pulse oximeters is their great sensitivity to motion artifact. Signal processing algorithms have been developed to reduce the effect of motion on the pulse oximetry signal and to detect motion artifact and prevent it from being indicated as data (31). Nevertheless, since the oxygen saturation values presented represent averages over several heartbeats, movement can result in an apparent decrease in oxygen saturation that in fact has not occurred.

TEMPERATURE MONITORING

An important aspect of treating premature infants is to the maintenance of their thermal environment. A premature infant is not well adapted to extrauterine life, and its temperature control system is not fully developed since it normally would be in a temperature regulated environment in the uterus. Thus, an artificial environment must be provided to help the neonate control its body temperature. This environment is in the form of convective incubators and radiant warmers. Another reason for providing an elevated temperature environment for premature infants is that very often these infants suffer from problems of the respiratory system that limit the amount of oxygen that can be transported to the blood by the lungs. This oxygen is utilized in the metabolic processes of the infant, and among these are the generation of heat to maintain body temperature. By placing the infant in an environment at a temperature greater than normal room temperature, less energy needs to be expended for thermal regulation. A neutral thermal environment can be found where the temperature and relative humidity are such that the infant utilizes a minimum amount of energy to maintain its temperature, and oxygen and nutritional substrates that would normally go into heat generation can be utilized for metabolic processes related to growth and development. Thus, to maintain this environment, it is necessary to monitor both the temperature of the infant and that of the environment.

Temperature monitoring instrumentation in the nursery is relatively straightforward. The sensor is a thermistor that can be in one of two basic forms, an internal probe or a surface probe. The former consists of a semiflexible lead wire with a thermistor mounted at its distal tip. An electrically insulating polymer with good thermal conductivity covers the thermistor and is contiguous with the lead wire insulation. This probe can be placed rectally to give a neonatal core temperature measurement. The surface probe is a disk-shaped thermistor ~ 6 mm in diameter with lead wires coming out in a radial direction. The sensitive surface of the probe is metallic and is in intimate contact with the thermistor, while the other surface of the probe is covered with a thermally insulating polymer so that the thermistor is well coupled to the infant surface it contacts through the metal but poorly coupled to the environmental air. The surface temperature measured is not necessarily the same as core temperature and is strongly dependent on the infant's environment. Often the surface mounted probe is placed over the liver since this organ is highly perfused and is close to the skin surface in small infants. To aid and maintain a good thermal contact between the surface probe and the infant skin, the lead wires, especially near the probe, should be highly flexible so that the wires do not tend to force the thermistor to come loose from the skin as the infant moves. As was mentioned for surface mounted biopotential electrodes, the skin of premature infants is sensitive to many factors, and strong adhesive can produce severe irritation. Weaker adhesives, however, can allow the thermistor to come off, and the use of flexible lead wires greatly reduces this tendency.

The remainder of the instrumentation in temperature monitoring devices is straightforward. An electronic circuit senses the resistance of the thermistor and converts this to a display of its temperature. In some cases alarm circuits are incorporated in the monitors to indicate when the

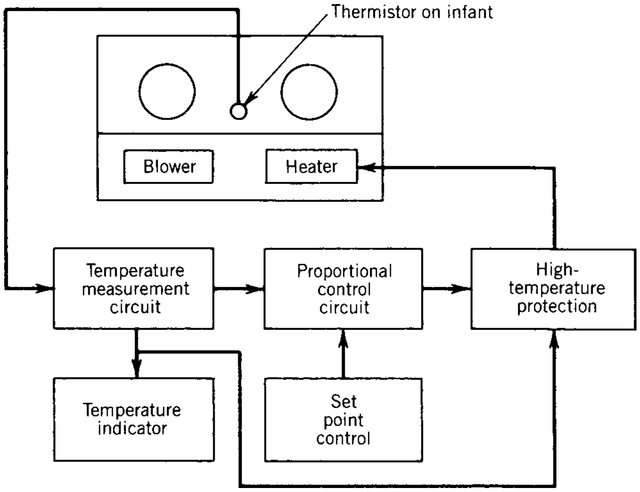

Thermistor on infant
Blower
Heater
Temperature measurement circuit
Proportional control circuit
High-temperature protection
Temperature indicator
Set point control

Figure 10. A servo-temperature control system for incubator temperature directed toward maintaining infants at a preset temperature.

temperature lies outside of a preset range. Temperature instruments are used not only for indicating infant surface and core temperatures, but also for the control of incubator or radiant warmer temperature. Although convective incubators have internal control systems to maintain the air temperature at a preset point, the purpose of the incubator is not as an air temperature controller. Instead the incubator is used to maintain the infant's body temperature at a certain point and to minimize thermal losses from the infant. For this reason some incubators have servo systems that control incubator temperature based on infant temperature rather than air temperature. A block diagram of such a system is illustrated in Fig. 10. Here a thermistor is the sensor and is positioned on the infant's skin or internally. If a radiant warmer is used, it is important that any surface mounted thermistor is not in the radiant field and thus directly heated by the warmer. Frequently thermistors are covered with an insulating disk that has a highly reflective outer surface of aluminum foil so that no direct radiant energy from the heater falls on the thermistor. An electronic circuit determines the thermistor resistance, and hence, its temperature, which is assumed to be equivalent to the infant's temperature. This drives a control circuit that provides proportional control to the heating element of the convective or radiant source. In some cases integrating and differential control is added to the system for optimal response. Additional safety circuits are included in the system to prevent it from overheating or underheating, both of which are undesirable for the infant. The final block of the system is the heater itself. The control system must take into account the time response of this element so as to provide optimal control. An indicator is frequently included in the system to show infant temperature and heater status.

PRESSURE MEASUREMENT

The measurement of the pressure in fluids is important in many aspects of medical care. This is especially true in neonatal monitoring, and instrumentation for the intermittent or continuous measurement of blood pressure is frequently used in the intensive care unit. There are also situations where the monitoring of intracranial pressure is important in the care of infants. Instrumentation for measuring these pressures is similar to other monitoring instrumentation described in this article in that measurements can be made by direct and indirect means. These are described in the following paragraphs.

Blood Pressure

The direct measurement of blood pressure consists of coupling the arterial or venous circulations to a pressure transducer that is connected to an electronic instrument for signal processing, display, and recording. The direct methods used are similar to those used in adult intensive care (see BLOOD PRESSURE MEASUREMENT). Generally, measurements are only made on the arterial circulation, and wherever possible an umbilical artery is used for access to this circulation. An umbilical artery catheter, such as described in the section on direct measurement of blood gases, is filled with a physiologic saline solution containing an anticoagulant. The proximal end of this catheter is connected to an external pressure sensor that is positioned in the incubator near the infant. This is usually a disposable semiconductor pressure sensor that is used only on a single infant and then discarded so that there is no risk of cross-contamination from one patient.

Indirect blood pressure monitoring in the neonate presents special problems not seen in the adult. It is generally not possible to measure an infant's blood pressure using a sphygmomanometer and the auscultation technique because Korotkoff sounds cannot be detected. Thus other, more complicated methods of indirectly measuring blood pressure must be used. If a sphygmomanometer cuff around a limb is still employed, it is important to use the correct size of cuff for the infant being studied. The width of the cuff should be from 45 to 70% of the limb circumference (32). Cuffs of several different sizes are, therefore, available for use with infants. These frequently are inexpensive disposable cuffs designed for use with a single infant.

Systolic and diastolic pressures can be sampled noninvasively using the oscillometric or the kinarteriography methods. Both techniques (see BLOOD PRESSURE MEASUREMENT) are based upon blood volume changes in the section of artery under or distal to the sphygmomanometer cuff. In the case of the oscillometric measurement, the actual volume changes are determined, while the kinarteriography method measures the radial velocity of pulsations in the arterial wall. In the former case pressure variations in the cuff itself are sensed, and signal processing allows mean arterial pressures as well as systolic and diastolic pressures to be determined.

The kinarteriographic technique utilizes an ultrasonic transducer under the cuff. Continuous wave ultrasound is beamed at the brachial artery, when the cuff is on an arm, and some ultrasonic energy is reflected from the arterial wall. This is picked up by an adjacent ultrasonic transducer, and an electronic circuit determines the frequency differences between the transmitted and reflected waves. When the arterial wall is in motion, the reflected ultrasound is shifted in frequency thereby giving a frequency difference between the transmitted and reflected waves. Motion of the arterial wall is greatest when the cuff pressure is between systolic and diastolic pressures; and thus by measuring changes in the frequency shift of the reflected wave, it is possible to determine the systolic and diastolic pressures. Unlike the oscillometric technique, it is not possible to determine the mean arterial pressure with kinarteriography.

Monitoring of Intracranial Pressure

As was the case with blood pressure, intracranial pressure (the pressure of the cerebrospinal fluid and brain within the cranium) can be determined by direct and by indirect methods. The former involves the placement of a tube within the brain such that its distal tip communicates with the intraventricular fluid. The proximal end is connected to a low compliance pressure transducer. Although this technique is highly accurate, a significant risk of infection is associated with its application, and it is only used under extreme circumstances when no other technique is possible.

Noninvasive techniques of monitoring neonatal intracranial pressure are much safer than the direct technique but, unfortunately, are not as accurate. The newborn infant not only has special access available to the central circulation through the umbilical cord, but also has a means of accessing the intracranial contents through the anterior fontanel. This gap in the calvarium means that only soft tissue lies between the scalp surface and the dura mater. Thus, it is possible to assess intracranial pressure through this opening by various techniques.

A skillful clinician can palpate the anterior fontanel and determine to some extent whether the pressure is elevated or not (33). Another clinical method that can be used is to observe the curvature of the scalp over the fontanel as the position of the infant's head with respect to its chest is changed (34). The curvature should flatten when intracranial and atmospheric pressures are equivalent. These techniques are highly subjective and not suitable for neonatal patient monitoring; however, they can be the basis of sensors for more objective measurement.

Various forms of tonometric sensors have been developed for noninvasively measuring and monitoring intracranial pressure (35). These all consist of some sort of probe that is placed over the anterior fontanel in such a way that the curvature is flattened and formed into a plane normal to the surface of the probe itself. guard rings, calibrated springs, and other techniques have been used to achieve this applanation. Ideally once this is achieved the pressure on either side of the membrane consisting of the soft tissue between the dura and the scalp surface should be equal. Thus, by sensing the pressure on the probe side, one can determine the pressure on the other side of the dura. Unfortunately, such a situation only holds in practice when the membrane is thin with respect to the size of its planar portion. In the case of the soft tissue between the dura and the scalp, the membrane thickness can often be close to the size of the planar portion because of the limitations imposed by the opening of the fontanel. Thus, the technique has some definite limitations. The method of attachment of the probe to the infant and the structure of the probe itself are critical to the efficacy of the resulting measurements.

Many investigators have considered different approaches to making an appropriate probe for transfontanel intracranial pressure measurement. These range from strain-gage-based force sensors with guard rings to transducers that attempt to achieve applanation by means of a compliant membrane mounted upon a chamber in which air pressure can be varied. In the case of this latter technique, the position of the membrane is detected and a servo control system is used to adjust the pressure within the chamber so that the membrane presses the tissue of the fontanel into a flat surface. Such a device is shown is schematically is Fig. 11. The position of the membrane is established optically by means of a shutter attached to the membrane such that it varies the amount of light passing from a fiber optic connected to a light source to one connected to a light detector. A servo system controlling the air pressure within the structure adjusts it so that the diaphragm, and hence the tissue of the fontanel, is flat. At this point, the pressure of the air within the sensor should theoretically equal that of the tissue within the fontanel, which in turn should give the intracranial pressure.

Elevated intracranial pressure in infants can be the result of volume occupying lesions, excessive secretion of fluids within the epidural space, the brain tissue itself, or fluid in the ventricles of the brain. A frequent form of lesion is bleeding or hemorrhage within one of these volumes, a condition that is far too often seen in premature infants. Another form of elevated intracranial pressure results from hydrocephalus, a condition in which intraventricular fluid volume and pressure become elevated. By continuous monitoring or serial sampling of intracranial pressure, it

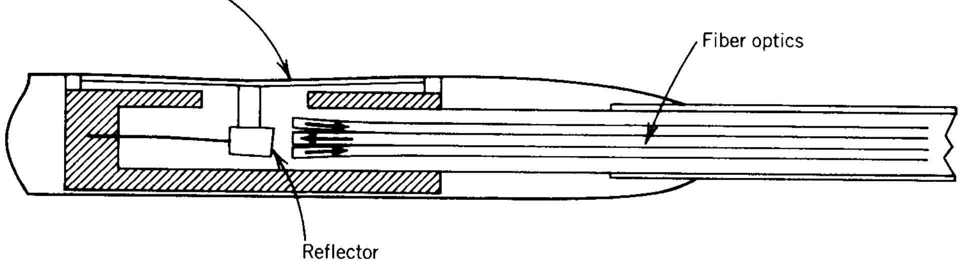


Figure 11. A pressure sensor used far intracranial pressure measurements in the newborn based upon measuring membrane deflection by means of fiber optics.

is possible to provide better control of therapy for these problems.

Monitoring of Intracranial Hemorrhage

As pointed out in the previous section, one cause of elevated intracranial pressure in the newborn is intracranial hemorrhage. It is important to be able to detect when this occurs so that therapeutic measures can be immediately taken to minimize irreversible damage. Although no technique is suitable for continuous monitoring of infants at risk of intracranial hemorrhage, several techniques can be used for surveillance and periodic sampling especially of infants showing possible signs and symptoms of intracranial bleeding.

Devices for the noninvasive measurement of intracranial pressure have been described in the previous section. In addition to these, intracranial hemorrhage can be identified by measurement of transcephalic electrical impedance (36). In this case, it is the baseline value of the impedance that is important. For this reason, a tetrapolar method of measurement must be used to minimize the effects of electrode and lead wire impedances. An alternating current at an excitation frequency of 20–100 kHz is passed between a frontal and occipital electrode placed on the infant's head. A second pair of electrodes are located near the excitation electrodes, but far enough to avoid voltage drops due to the spreading current at the excitation electrodes. The signal is picked up by these electrodes and detected by an electronic circuit to give a voltage proportional to the baseline impedance value. Since the specific impedance of blood is ~ 2.5 times greater than the specific impedance of the cerebral spinal fluid, one should see an elevated transcephalic impedance when the ventricles are filled with blood rather than cerebral spinal fluid. Similarly, if the ventricles have grown in volume because of excess cerebral spinal fluid as in hydrocephalus, the transcephalic impedance should be lower than expected.

In practice, one can only look for changes in transcephalic impedance in infants and not at absolute baseline values because of differences in geometry from one subject to the next. Typically, the technique requires one or more measurements to be taken during the first 24 h of life on each infant and using these measurements to obtain baseline intraventricular, hemorrhage-free data for that particular infant. Subsequent measurements through the infant's hospital course are compared to these initial measurements, and deviations are noted. Significant elevations in impedance have been associated with the occurrence of intraventricular hemorrhage. Studies have been carried out to show that this impedance shift correlates with intraventricular hemorrhage as found using other diagnostic techniques or, in the case of infants who expire, at autopsy (36). The principal advantage of this method is its relative simplicity and ease of measurement on infants in the intensive care unit.

Ultrasonic determination of intraventricular hemorrhage involves making a B scan of the infant's head and locating the ventricular system (37,38). If the ventricles are filled with cerebral spinal fluid alone, the fluid in the ventricles does not reflect ultrasound, and the ventricles appear clear on the image. If there is blood in the ventricular fluid, ultrasonic echoes are produced by the cellular components of the blood. This causes reflections to appear within the ventricle on the image and allows for a definite diagnosis of intracranial hemorrhage.

As with the transcephalic impedance method, the technique of making measurements on infants is straightforward and can be carried out in the neonatal intensive care unit. The equipment necessary for the measurement, however, is more costly than the impedance; but the results are far more specific to identifying intracranial hemorrhage, and thus this is the current method of choice.

An additional method that can be used for detecting bleeding within the ventricles is the use of computerized tomography (CT) scans (39). While this technique is highly efficacious from the standpoint of identifying intracranial bleeding, it is undesirable because it exposes the developing nervous system to significant amounts of X radiation. It also is necessary to transfer infants to the radiology department where the scanning equipment is located to make the measurement. For severely ill, premature infants, this transfer can significantly compromise their care.

MONITORING BILIRUBIN

Bilirubin is a product of the biochemical degradation of the heme moiety that occurs in the hemoglobin molecule as well as other proteins. It is normally found as a result of red blood cell turnover, but it can be elevated in some hemolytic diseases. This form of bilirubin, known as unconjugated bilirubin, enters the circulation and then the skin and other tissues. When it is present in the skin in sufficient concentration, it causes a yellow coloration known as jaundice. It also enters nervous tissue where, if it reaches sufficient concentration, it can cause irreversible damage.

Increased serum bilirubin can occur either as the result of increased production or decreased clearance by the liver. The former situation can occur in normal and premature infants and is referred to as physiologic jaundice of the newborn. It is usually more severe in prematurely born infants and generally peaks about the third day of neonatal life. There are several modes of therapy that can be used to reduce serum bilirubin once elevated values are detected. The simplest way to detect jaundice in the neonate is to observe the infant's skin and sclera for yellow coloration. Quantitative assessment can be carried out by drawing a blood sample and extracting the cellular components from the serum. The absorption of light at a wavelength of 450 nm in such a serum sample is proportional to its bilirubin concentration. Photometric instrumentation for doing this is readily available in the clinical laboratory and some intensive care units (38). The problem with this method is the need for obtaining a blood sample and the time necessary to transport the sample to the laboratory and analyze it in the photometer. A method for the rapid assessment of serum bilirubin in all infants would represent an improvement over these techniques. Fortunately, a relatively simple optical instrument has been developed for assessing serum bilirubin in infants (40). It is illustrated schematically in Fig. 12 and consists of a xenon flash tube

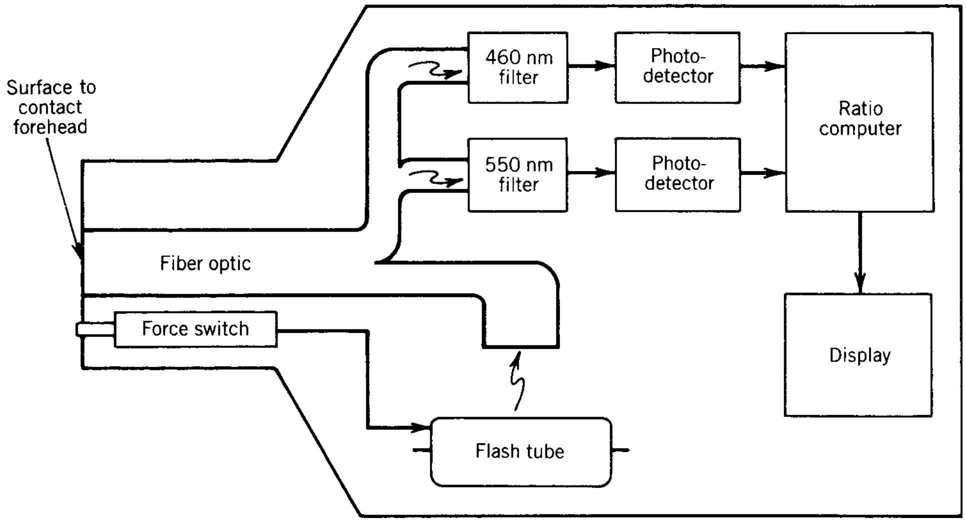


Figure 12. Schematic diagram of a transcutaneous bilirubin instrument.

light source and photometers with filters to measure the reflected light at 460 and 550 nm. A special feature of the instrument is a pressure switch on the portion of the probe that is pressed against the infant's forehead to make the measurement. This probe contains fiber optics that couple the xenon flash tube and the photometers to the skin surface. As the probe is pressed against the skin, the force squeezes the blood from the cutaneous capillaries. Once the force is great enough to sufficiently blanch the skin under the fiber optics so that the blood will not interfere with the measurement, a switch activates the flash tube and readings are taken from the photometers. An internal microprocessor analyzes the reflected light and compensates for any residual hemoglobin. It gives a number proportional to skin bilirubin on a digital display. The proportionality constant, however, differs for different types of neonatal skin; thus, proportionality constants have to be determined for infants of different age, race, and whether they have received phototherapy or not. This instrument involves a sampling technique and thus is only suitable for trend rather than continuous monitoring. Since changes in serum bilirubin are relatively slow, such a technique is entirely appropriate. The principal advantage of this instrument is that it can be readily applied to all infants in the nursery with little effort on the part of the clinical staff. Readings can be made quickly to identify those infants at risk of hyperbilirubinemia.

MONITORING LIFE SUPPORT SYSTEMS

Although one usually associates neonatal monitoring with the measurement of physiologic variables from the newborn or prematurely born infant, an aspect of neonatal monitoring that should not be overlooked is associated with monitoring the patient's environment. Various life support systems are important in neonatal intensive care, and electronic instrumentation for assessing and maintaining the function of these is also important. There should be alarm systems so that when the conditions of life support are inappropriate for neonatal care, care takers are alerted and the problem can be corrected before it causes any harm to the patient. There are many examples of life support system monitoring in the neonatal intensive care unit, and some of the major ones will be described in the following paragraphs.

Maintaining an appropriate thermal environment for the neonate is an important aspect of neonatal intensive care. Incubators and radiant warmers need to have internal temperature instrumentation to ensure that the environment is appropriate for the infant and so that the clinicians providing care can be aware of environmental conditions. Convection incubators frequently have temperature sensors for measuring the environmental air temperature and indicating it on the control panel of the device. These temperature sensors are often a part of the thermal control system in the incubator itself, and as indicated earlier the infant's own temperature can be used as a control signal for some incubators. Built into this incubator temperature monitoring function is an alarm function that can indicate when the incubator temperature becomes too high or too low; potentially life threatening conditions.

It is sometimes necessary to intubate the trachea of a patient and to use a ventilator to control breathing. A gas with elevated oxygen content is often used to help provide additional oxygen to infants who require it. There are many points in a support system such as this where monitoring devices can be useful to assess and in some cases control the function of the device. Where gases of elevated oxygen tension are given, instrumentation to measure the partial pressure of oxygen within the gas to indicate the oxygen fraction of inspired gas is desirable. Various types of oxygen sensors are, therefore, placed in the air circuit to either continuously or intermittently measure and display this quantity. The temperature and humidity of the inspired air are also important, and appropriate sensors and instrumentation for these variables can be included as a part of the respiratory support system. Continuous positive airway pressure is a mode of therapy used in infants requiring ventilatory support. It is necessary to measure and control the positive pressure in such systems to minimize the risk of pneumothorax and to ensure that the desired levels are maintained.

The use of arterial and venous catheters in infants for blood pressure and blood gas monitoring as well as fluid therapy and hyperalimentation represents a safety risk to the infant. Arterial catheters and associated plumbing can become disconnected and cause serious losses of blood that if not quickly checked can result in severe injury or death to the patient. Gas bubbles inadvertently infused along with intravenous fluids can, if sufficiently large, compromise the circulation by producing gas embolisms. Fluid therapy in very small infants must be precisely controlled so that excessive or insufficient amounts of fluid are not provided to the infant. Electronic instrumentation for controlling all of these variables and producing an alarm when dangerous conditions are encountered have been developed (41). Some of these, such as intravenous infusion pumps, are routinely used in neonatal intensive care units. Safety devices for over- or underpressures can be built into the pumps, as can sensors, to indicate when the fluid source has been depleted so that additional fluid can be attached to the pump.

Phototherapy is a technique of illuminating the baby with blue light in the wavelength range of 420–500 nm to oxidize bilirubin to compounds that can be eliminated from the body. Phototherapy units consisting of a group of 20 W fluorescent lamps 30–40 cm above the infant must be used cautiously because there are risks associated with this radiation. Therefore, it is important to determine the amount of radiant energy received by the infant in the 420–500 nm band so that minimal exposure times sufficient to oxidize the bilirubin can be given. Instrumentation consisting of a small probe containing a photosensor that can be held just above the neonate has been developed for this purpose. It is not necessary for this instrumentation to be used to continuously monitor the phototherapy units, but frequent testing of the therapy devices helps not only to determine the appropriate exposure times, but also to indicate when the fluorescent lights become ineffective and need to be changed.

DIAGNOSTIC RECORDINGS

The role of infant monitoring is to determine when events requiring therapeutic intervention occur so that optimal care can be provided. Hard copy recordings from electronic monitoring devices can also be useful in diagnosis of illness and identification of infants who may benefit from electronic monitoring over a longer period of time. Two types of diagnostic recordings are currently used in neonatology: polysomnograms and oxycardiorespirograms although both are still considered experimental and are not routinely used.

Polysomnography. Multiple channel, simultaneous, continuous recordings of biophysical variables related to the pulmonary and cardiovascular systems taken while the newborn or infant sleeps are known as polysomnograms (41–43). These recordings are often made overnight, and 8–12 h is a typical length. The actual variables that are monitored can vary from one study to the next as well as from one institution to the next. However, these usually include the electrocardiogram and/or heart rate. One or more measures of respiratory activity, such as transthoracic electrical impedance or abdominal strain gage; measures of infant activity and movement; measures of infant sleep state, such as eye movements and usually a few leads of the electroencephalogram; and measures of infant blood gas status are also recorded. The number of channels of data in polysomnograms is at least 3 and often is 12 or more. Recordings are made using computer data acquisition systems.

The primary application of polysomnography has been as a tool for research evaluating infant sleep patterns and related physiologic phenomena during sleep. Some investigators feel that polysomnographic recordings are useful in evaluating infants considered to be at risk to sudden infant death syndrome, but at present there is no conclusive evidence that this technique has any value in such screening. There may, however, be specific individual cases where such evaluations may contribute to overall patient assessment.

Oxycardiorespirogram. This technique is used for continuous computer monitoring of infants in the neonatal intensive care unit (44). Four or five physiologic variables are monitored and continuously recorded on a multichannel chart recorder (45). These are the electrocardiogram from which the beat-to-beat or instantaneous heart rate is determined; the respiration waveform or pattern as generally determined by transthoracic impedance monitoring; the respiration rate as determined from the respiration waveform; the oxygen status as determined from a pulse oximeter or transcutaneous blood gas sensors; and at times the relative local skin perfusion as determined by the thermal clearance method from the transcutaneous blood gas sensor.

The importance of the oxycardiorespirogram is that it brings these variables together on a single record, where they can be presented in an organized and systematic fashion. This makes it possible to observe and recognize characteristic patterns between the variables that may be overlooked when all of these quantities are not monitored and recorded together. The oxycardiorespirogram is able to look at variables related to various points along the oxygen transport pathway from the airway to the metabolizing tissue. Thus, it allows a more complete picture of infant cardiopulmonary function than could be obtained by monitoring just one or two of these variables. Typical oxycardiorespirogram patterns have been classified and organized to assist clinicians in providing neonatal intensive care (46).

SUMMARY

Infant monitoring along with adult monitoring under critical care situations involves many individual types of sensors and instruments. Depending on the application, many different types of output data will be produced. In addition, many different conditions for alarms to alert the clinical personnel can occur for each of the different instruments in use. Needless to say that although this provides better assessment of the infant as well as quantitative and in some cases hard copy data, it also requires that this data be integrated to provide manageable information. By combining data from several different pieces of instrumentation, more specific conditions for alarms can be defined and variables can be more easily compared with one another. This can then lead into trend analysis of the data, and the computer certainly represents an important overall controller, analyzer and recorder of these functions. As technology continues to become more complex, it is important not to lose track of the primary goal of this technology, namely, to provide better neonatal and infant care so that critically ill neonates can look forward to a full, healthy, and normal life.

BIBLIOGRAPHY

Cited References

1. Neuman MR. Biopotential Amplifiers. In: Webster JG, editor. Medical Instrumentation: Application and Design. 3rd ed. New York: John Wiley & Sons, Inc.; 1998. pp. 233–286.
2. Neuman MR. Flexible thin film skin electrodes for use with neonates. Proceedings of the 10th International Conference Medical Biological Engineering. Dresden: DDR; 1973.
3. Accidents with apnea monitors. FDA Drug Bull. Aug. 1985; 15:18.
4. Primiano FP. Measurements of the Respiratory System. In: Webster JG, editor. Medical Instrumentation: Application and Design. 3rd ed. New York: John Wiley & Sons, Inc.; 1998. pp. 372–439.
5. Thomas PE. What's the latest on carbon dioxide monitoring? Neonatal Network 2004 July–Aug; 23(4):70–72.
6. Wu CH, et al. Good estimation of arterial carbon dioxide by end-tidal carbon dioxide monitoring in the neonatal intensive care unit. Pediatr Pulmonol 2003;35(4):292–295.
7. Gordon DH, Thompson WL. A new technique for monitoring spontaneous respiration. Med Instrum 1975;9:21.
8. Kulkarni V, et al. AURA: a new respiratory monitor. Biomed Sci Instrum 1990;26:111–120.
9. Neuman MR. A microelectronic biotelemetry system for monitoring neonatal respiration using thermistors. Proceedings of the 21st Annual Meeting Association Advanced Medical Instrumentation. Chicago: 1986.
10. Neuman MR. Multiple Thin-Film Sensor System. US patent No. 5.394,883. 1995 Mar 7.
11. Werthammer J, Krasner J, DiBenedetto J, Stark AR. Apnea monitoring by acoustic detection of air flow. Pediatrics 1983; 71:53.
12. Whitney RJ. The measurement of volume changes in human limbs. J Physiol (London) 1953;121:1.
13. Rolfe P. A magnetometer respiration monitor for use with premature babies. Biomed Eng 1971;6:402.
14. Sackner JD, et al. Noninvasive measurement of ventilation during exercise using a respiratory inductive plethysmograph. Am Rev Respir Dis 1980;122:867.
15. Cohen KP, et al. Design of an inductive plethysmograph for ventilation measurement. Physiol Meas 1994 May; 15(2): 217–229.
16. Brooks LJ, DiFiore JM, Martin RJ. Assessment of tidal volume over time in preterm infants using respiratory inductance plethysmography, The CHIME Study Group. Collaborative Home Infant Monitoring Evaluation. Pediatr Pulmonol 1997 Jun; 23(6):429–433.
17. Fraden J. Piezo/pyroelectric film as a biomedical transducer. Proc Annu Conf Eng Med Biol 1986;28:221.
18. Barrow RE, Colgan FJ. A noninvasive method for measuring newborn respiration. Respir Care 1973;18:412.
19. Prechtl HFR, van Eykern LA, O'Brien MJ. Respiratory muscle EMG in newborns: A non-intrusive method. Early Hum Dev 1977;1:265.
20. Mendenhall RS, Neuman MR. Efficacy of five noninvasive infant respiration sensors. Proceedings IEEE Fronteuro on Engineering Medical Biology. Columbus, (OH): 1983. p 303–307.
21. Peabody JL, et al. Clinical limitations and advantages of transcutaneous oxygen electrodes. Acta Anaesthesiol Scand Suppl 1978;68:76.
22. Eberhart RC. Indwelling blood compatible chemical sensors. Surg Clin North Am 1985;65:1025.
23. Couch NP, et al. Muscle surface pH as an index of peripheral perfusion in man. Ann Surg 1971;173:173.
24. Huch A, Huch R, Lubbers DW. Transcutaneous PO_2. New York: Thieme-Stratton; 1981.
25. Kimmich HP, Spaan JG, Kreuzer F. Transcutaneous measurement of $PaCO_2$ at 37°C with a triple electrode system. Acta Anesthesiol Scand Suppl 1978;68:28.
26. Tremper KK, Waxman K, Shoemaker WC. Effects of hypoxia and shock on transcutaneous PO_2 values in dogs. Crit Care Med 1979;7:526.
27. Reynolds GJ, Goodwin B, Cowen J, Harris F. Simultaneous transcutaneous measurements of O_2, CO_2, and N_2 in neonates with RDS using a mass spectrometer. In: Rolfe P, editor. Fetal and Neonatal Physiological Measurements. London: Pitman; 1980. p 442.
28. Yoshiya I, Shimada Y, Tanaka K. Spectrophotometric monitoring of arterial oxygen saturation in the fingertip. Med Biol Eng Comput 1980;18:27.
29. Mendelson Y, Lewinsky RM, Wasserman Y. Multi-wavelength reflectance pulse oximetry. Anesth Analg (Suppl.) 2002;94(1): S26–30.
30. Reuss JL, Siker D. The pulse in reflectance pulse oximetry: modeling and experimental studies. J Clin Monit Comput 2004; 18(4):289–299.
31. Workie FA, Rais-Bahrami K, Short BL. Clinical use of new-generation pulse oximeters in the neonatal intensive care unit. Am J Perinato 2005 Oct; 22(7):357–360.
32. Darnall RA. Noninvasive blood pressure measurement in the neonate. Clin Perinatal 1985;12(1):31.
33. Wayenberg JL. Non-invasive measurement of intracranial pressure in neonates and infants: experience with the Rotterdam teletransducer. Acta Neurochir (Suppl.) 1998;71: 70–73.
34. Welch K. The intracranial pressure in infants. J Neurosurg 1980; 52:693.
35. Hill A. Intracranial pressure measurements in the newborn. Clin Perinatal 1985;12(1):161.
36. Lingwood BE, Dunster KR, Colditz PB, Ward LC. Noninvasive measurement of cerebral bioimpedance for detection of cerebral edema in the neonatal piglet. Brain Res 2002; 945(1):97–105.
37. Allan WC, Roveto CA, Sawyer LR, Courtney SE. Sector scan ultrasound imaging through the anterior fontanelle. Am J Dis Child 1980;134:1028.
38. Hintz SR, et al. Bedside imaging of intracranial hemorrhage in the neonate using light: comparison with ultrasound, computed tomography, and magnetic resonance imaging. Pediatr Res 1999;45(1):54–59.
39. Goplerud JM, Delivoria-Papadopoulos M. Nuclear magnetic resonance imaging and spectroscopy following asphyxia. Clin Perinatol 1993;20(2):345–367.
40. Strange M, Cassady G. Neonatal transcutaneous bilirubinometry. Clin Perinatol 1985;12(1):51–62.

41. Phillips BA, Anstead MI, Gottlieb DJ. Monitoring sleep and breathing: methodology. Part I: Monitoring breathing. Clin Chest Med 1998;19(1):203–212.
42. Hoppenbrouwers T. Polysomnography in newborns and young infants: sleep architecture. J Clin Neurophysiol 1992;9(1):32–47.
43. Barbosa GA, Keefe MR, Lobo ML, Henkin R. Adaptation of a cardiac monitor for collection of infant sleep data and development of a computer program to categorize infant sleep state. J Nurs Meas 2003;11(3):241–251.
44. Huch R, Huch A, Rooth G. An Atlas of Oxygen-Cardiorespirograms in Newborn Infants. London: Wolfe Medical Publications, Ltd.; 1983.
45. Neuman MR, Huch R, Huch A. The neonatal oxycardiorespirogram. CRC Crit Rev Biomed Eng 1984;11:77.
46. Neuman MR, Flammer CM, O'Connor E. Safety devices for neonatal intensive care. J Clin Eng 1982;7:51.
47. Neuman MR. Therapeutic and prosthetic devices. In: Webster JG, editor. Medical Instrumentation: Application and Design. 3rd ed. New York: John Wiley & Sons, Inc.; 1998. pp. 577–622.

See also BLOOD GAS MEASUREMENTS; INCUBATORS, INFANT; MONITORING, INTRACRANIAL PRESSURE; MONITORING, UMBILICAL ARTERY AND VEIN; TEMPERATURE MONITORING; VENTILATORY MONITORING.

NERVE CONDUCTION STUDIES. See ELECTRONEUROGRAPHY.

NEUROLOGICAL MONITORS

R. R. GHARIEB
Infinite Biomedical Technologies
Baltimore, Maryland

N. V. THAKOR
Johns Hopkins University
Baltimore, Maryland

INTRODUCTION

The electroencephalogram (EEG) is an electrical activity of the brain that is recorded by using electrodes appropriately placed on the scalp, then amplifying and displaying the electrical signal and its clinical relevant feature using a computer, or other suitable monitors. The EEG signal is a wave that varies in time. This wave contains frequency components that can be measured and analyzed. These frequency components have meaning and valuable properties. Table 1 shows the commonly defined waves or rhythms, their frequency, and their properties. Hans Berger, the discoverer of the EEG in humans, observed in 1924 all of the rhythms known today (except the 40 Hz "gamma" band). The described many of their basic properties. Since then, our definitions and understandings of the rhythms have been refined. However, there still remains some uncertainty, and controversy, in how to define and use these bands, for various purposes. Clinicians view the brainwaves for diagnostic purposes and seek to identify patterns that are associated with specific pathologies or conditions. Psychologists also study them in association with mental states, mental processing, and to test concepts of how the brain processes information (1–6).

The EEG is therefore a noninvasive marker for cortical activity. The EEG in humans and animals is used to monitor alertness; coma and brain death; locate area of damage following head injury, stroke, tumor, and so on; monitor cognitive engagement; control depth of anesthesia; investigate and locate seizure origin; test epilepsy drug effects; monitor human and animal brain development; test drugs for convulsive effects; investigate sleep disorder, monitor and track brain ischemia; and so on. Continuous EEG monitoring is a common routine in the intensive care unit (ICU). However, in digital processed EEG, we study the patterns that emerge during various behavioral, as well as introspective, states, and then see what they are defining in terms of a multidimensional representation of some state space. Research that is focused on understanding specific properties, such as attention, alertness, mental acuity, and so on; has uncovered combinations of rhythms, and other EEG properties, that are relevant to these studies. Generally, derived properties are found, that involve computer processing of the EEG, to produce quantification measurements that are useful for research, monitoring, and so on.

Since high speed computers and sophisticated and efficient digital signal processing methodologies have become available. These properties are significant and new features and properties have been extracted from the EEG signal. These features are combined in a system of multivariable representation to formulate various quantitative EEG (qEEG) measures. The features commonly employed are (7–21).

Amplitude
Subband powers
Spectrogram
Entropy and complexity
Coherence
Biocoherence
Power spectrum
Joint-time frequency
Spectral edge frequencies
Coefficient-based EEG modeling
Bispectrum
Etc.

In the following section, the EEG monitors are classified and the main devices of the monitor are described. This section presents two types of monitors. In the common specification of Optimized Monitor section, the general specifications of the optimized EEG monitor are provided.

CLASSIFICATION OF EEG MONITORS

What is an EEG Monitor?

The neurological monitor is simply a display that shows the ongoing neurological activity recorded as the electrical potential by appropriately placing electrodes on the scalp. The conventional monitor goes back to EEG machine, where the electrical activity of the brain could be detected and plotted on scaled paper. Today, the neurological monitors are based on advanced technologies. They are

Table 1. EEG Rhythms their Frequency Bands and Properties

Rhythm Name	Frequency Band, Hz	Properties
Delta	0.1–3	Distribution: generally broad or diffused, may be bilateral, widespread Subjective feeling states: deep, dreamless sleep, non-REM sleep, trance, unconscious Associated tasks and behaviors: lethargic, not moving, not attentive Physiological correlates: not moving, low level of arousal Effects of Training: can induce drowsiness, trance, deeply relaxed states
Beta	4–7	Distribution: usually regional, may involve many lobes, can be lateralized or diffuse; Subjective feeling states: intuitive, creative, recall, fantasy, imagery, creative, dream-like, switching thoughts, drowsy; oneness, knowing Associated tasks & behaviors: creative, intuitive; but may also be distracted, unfocused Physiological correlates: healing, integration of mind/body Effects of Training: if enhanced, can induce drifting, trance-like state if suppressed, can improve concentration, ability to focus attention
Alpha	8–12	Distribution: regional, usually involves entire lobe; strong occipital w/eyes closed Subjective feeling states: relaxed, not agitated, but not drowsy; tranquil, conscious Associated tasks and behaviors: meditation, no action Physiological correlates: relaxed, healing Effects of Training: can produce relaxation Sub band low alpha: 8–10: inner-awareness of self, mind/body integration, balance Sub band high alpha: 10–12: centering, healing, mind/body connection
Low Beta	12–15	Distribution: localized by side and by lobe (frontal, occipital, etc.) Subjective feeling states: relaxed yet focused, integrated Associated tasks & behaviors: low SMR can reflect “ADD”, lack of focused attention Physiological correlates: is inhibited by motion; restraining body may increase SMR Effects of Training: increasing SMR can produce relaxed focus, improved attentive abilities, may remediate Attention Disorders.
Mid-range Beta	15–18	Distribution: localized, over various areas. May be focused on one electrode. Subjective feeling states: thinking, aware of self and surroundings Associated tasks and behaviors: mental activity Physiological correlates: alert, active, but not agitated Effects of Training: can increase mental ability, focus, alertness, IQ
High Beta	15–18	Distribution: localized, may be very focused. Subjective feeling states: alertness, agitation Associated tasks and behaviors: mental activity, for example, math, planning, and so on. Physiological correlates: general activation of mind & body functions. Effects of Training: can induce alertness, but may also produce agitation, etc.
Gamma	40	Distribution: very localized Subjective feeling states: thinking; integrated thought Associated tasks and behaviors: high level information processing,“binding Physiological correlates: associated with information-rich task processing Effects of Training: not known

computer based and display not only the raw EEG, but also various quantitative indexes representing processed EEG. The monitors are EEG processors that have the ability to perform data acquisition, automatic artifact removal, EEG mining and analysis, saving/reading EEG data, and displaying the quantitative EEG (qEEG) measures (indexes) that best describe neurological activity and that are clinically relevant to brain dysfunction.

Neurological Monitor Main Components

As shown in Fig. 1, a typical neurological monitor consists of a few main devices. These devices are connected together through a microcomputer, which supervises and controls the data flow from one device to another. It also receives and executes the user instructions. It implements the EEG methodology routine. The main devices of a typical monitor can be summarized as follows:

ELECTRODES AND ELECTRODE PLACEMENT

Electrodes represent the electrical link between the subject’s brain and the monitor. These electrodes are appropriately placed on the scalp for recording the electrical potential changes. Electrodes should not cause distortion to the electrical potential recorded on the scalp and should be made of materials that do not interact chemically with electrolytes on the scalp. The direct current (dc) resistance of each electrode should measure no more than a few ohms. The impedance of each electrode is measured after an electrode has been applied to the recording site to evaluate the contact between the electrode and the scalp. The impedance of each electrode should be measured routinely before every EEG recording and should be between 100 and 5,000 Ω (2).

The international 10–20 system of electrode placement provides for uniform coverage of the entire scalp. It uses

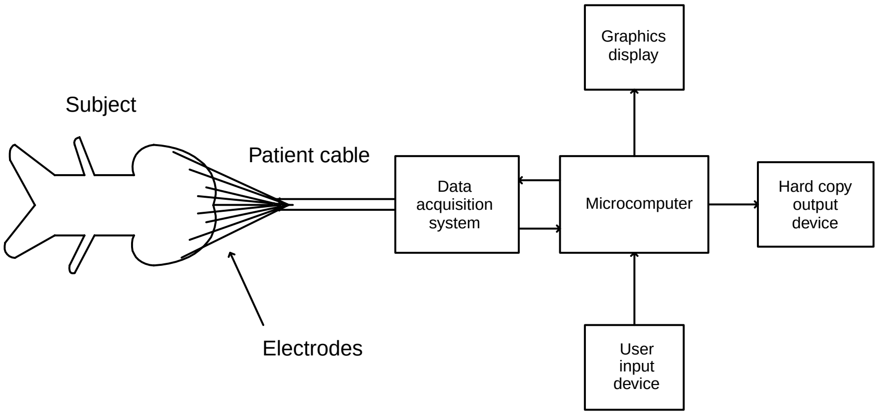


Figure 1. Block diagram of main components of a neurological monitor.

the distances between bony landmarks of the head to generate a system of lines, which run across the head and intersect at intervals of 10 or 20% of their total length. The use of the 10–20 system assures symmetrical, reproducible electrode placement and allows a more accurate comparison of EEG from the same patients, recorded in the same or different laboratories.

Patient Cable

The patient cable assembles the electrode terminals to the recording machine and monitoring instrument. It is preferable that the patient cable be of short length, which assures low impedance and causes no distortion of the electrical potential representing the neurological activity.

Data Acquisition System

It is composed of filters, amplifiers, analog-to-digital converters (ADC), and buffers. Bandpass filters of 0.5–100 Hz band are usually used to enhance the quality of the EEG signal. High gain amplifiers are required since the electrical potentials on the scalp are of microvolt. The input impedance of the amplifiers should be a large value while the output impedance should be a few ohms. The ADC converter digitizes the EEG data by sampling (converts the continuous-time EEG into discrete-time EEG) the data and assign a quantized number for each sample. Figure 2 shows a schematic diagram for the ADC converter while Fig. 3 shows the output–input characteristic of the uniform quantizer. Uniform quantization generates additive white noise to the EEG signal. Portable and wireless units of ADC have been used. The unit is connected to the monitor device through a standard wireless communication routine. This makes the monitor more comfortable and easier to be used.

Microcomputer

The microcomputer represents the master of the EEG monitor. It controls the data flow from one device to another. It reads the EEG data from the ADC buffers. It also hosts the software of the qEEG approaches and the artifact removal programs. Mathematical operations and analysis are carried out in the microcomputer. After processing the EEG data, the microcomputer sends the EEG signal and its qEEG measure (index) to the display. When the microprocessor is instructed to save the EEG session and its qEEG measure, it sends the data to the hard copy device.

Graphics Display

The graphics display displays the contentious EEG signals and online quantitative EEG (qEEG) measure. It helps the neurologists to follow and track in real-time fashion the changes in the brain activity and to monitor the brain development in the intensive care unit (ICU).

Hard Copy Output Device

This device is connected to the microcomputer and stores a version of the EEG data for future use. It could be a hard drive, computer CD, or a printer–plotter for plotting either version of the EEG or the qEEG measure to be investigated by neurologists and to be a part of the patient record.

User Input Device

Through this device, the user can communicate and interact with the monitor. Instructions and various parameters required for the EEG analysis are sent to the microcomputer through this device.

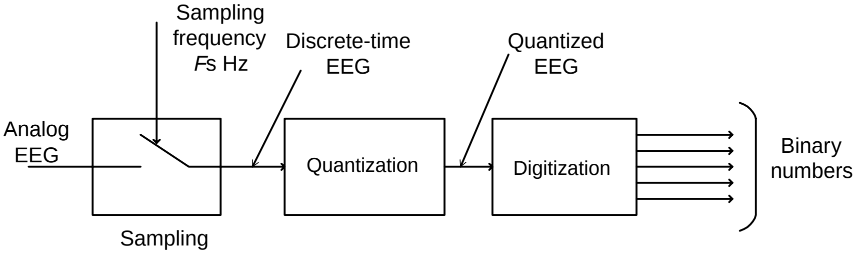


Figure 2. Schematic diagram of the ADC.

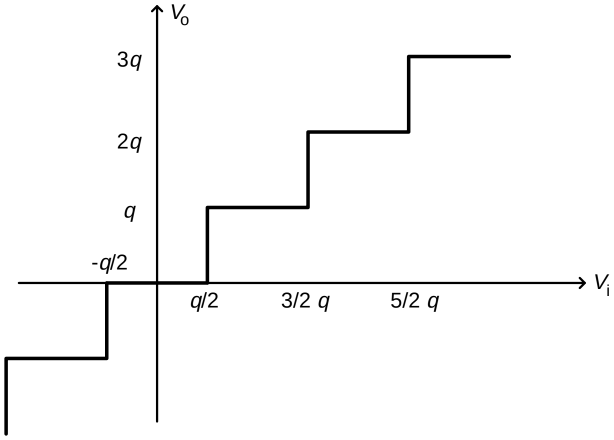


Figure 3. Input–output characteristic of the uniform quantization system.

TYPES OF EEG MONITORS

The EEG monitors can be classified into two main categories based on either their applications or the quantitative EEG index employed for processing and assessment of the brain electrical activity. Accordingly, the most popular monitors can be categorized as follows (7,22):

Application-Based Monitors

Cerebral function monitor (CFM)
Cortical injury monitor (CIM)
Anesthesia monitor
Narcotrend monitor
Patient state analyzer (PSA) with frontal patient state index (PSI)
BrainMaster 2E monitor portable bedside monitor

EEG Index-Based Monitors

Amplitude integrated monitor
Spectral index monitor
Spectral-edge frequency monitor
Bispectral index monitor
Entropy and complexity based monitor

This section presents, a very brief description of both monitor types. This description gives the intuition for the neurological applications of the monitor and the EEG index employed.

Cortical Injury Monitor

The lack of blood and oxygen flow to the brain due to cardiac arrest causes brain ischemia, causing brain cells to die, and consequently affecting (changes) brain activity. It has been demonstrated by many studies that brain ischemia slows the brain electrical activity by suppressing the high frequency and enhances the background activity; the cortical injury monitor has been developed and used for the detection and tracking of brain ischemia. The advantage of the monitor comes from the fact that it provides a quantitative measure extracted from the processed EEG signal for the severity of brain injury after cardiac arrest. It aids neurologists in providing better care for patients with cardiac arrest and provides them with therapeutic intervention, such as hypothermia. The monitor provides assessment of the brain function within the first 4 h after cardiac arrest.

Anesthesia Monitors

Patients receive general anesthesia during surgery. Anesthesia causes reduction of brain activity and concussions. The depth of anesthesia should be evaluated and tracked in real-time fashion to prevent perfect suppression of brain activity. The anesthesia monitor has been developed and used for the assessment of anesthesia and concussions. It provides a quantification measure or index for the depth of anesthesia. The monitor helps patients “rest easy” when they receive general anesthesia for surgery. Of the known anesthesia monitors, the bispectral (BIS) monitor, the narcotrend monitor, and the patient state analyzer (PSA4000) monitor are commonly employed. In the BIS monitor, a qEEG measure based on bispectrum is employed for tracking the depth of anesthesia. The PSA4000 is indicated for use in the operating room (OR), ICU, and clinical research laboratories. The monitor includes the patient state index (PSI), a proprietary computed EEG variable that is related to the effect of anesthetic agent. The narcotrend monitor provides a 6-letter classification from A (awake) to F (general anesthesia with increasing burst suppression). The narcotrend EEG monitor is similar to the BIS monitor positioned on the patient’s forehead. The EEG classification made by the narcotrend monitor are 6 letters: A (awake), B (sedate), C (light anesthesia), D (general anesthesia), E (general anesthesia with deep hypnosis), F (general anesthesia with increasing burst suppression) (23).

Cerebral Function Monitor

The cerebral function monitor (CFM) enables continuous monitoring of the cerebral electrical activity over long periods of time due to slow recording speeds. The cerebral electrical signals picked up by the electrodes attached to the scalp are registered in the form of a curve, which fluctuates to a greater or lesser extent depending on the recording speed. Examination of the height of the curve with respect to zero and its amplitude indicates the voltage of cerebral activity and yields information regarding polymorphism. Thus it is possible to monitor variations in cerebral activity over a prolonged period during anesthesia as well as during the revival phase with the monitor of cerebral function. The CFM is common practice in monitoring the cerebral function in intensive care. To bring the CFM into a polygraphy environment the hardware processing and paper write-out have to be implemented in software. The processor comprises a signal shaping filter, a semilogarithmic rectifier, a peak detector, and low pass filter. After taking the absolute value of the filtered EEG signal, the diode characteristic used to compress the signal into a semilogarithmic value was mimicked by adding a small offset to the absolute value before taking the

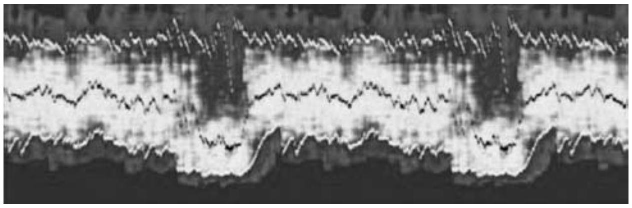


Figure 4. Color CFM of ~ 2.5 h, red is high density, blue low density, and black zero. Vertical scale from 0–5 μV linear, from 5–100 μV logarithmic. The median is given in black and the percentiles in white. As only 1 h of data was available to test the reproducibility of the process, we used a repeated playback mode for this picture. The low median episodes are neonatal State 1 (Quiet Sleep) with the beginning tracé-alternant (high peaks followed by low amplitude EEG) with half way diminishing peak heights, and a neonatal State 2 (REM Sleep) with symmetrical continuous EEG.

logarithm. The envelope of the resulting signal has been made by means of a leaky peak detector and a boxcar averager. Writing the resulting signal on a pixelized computer screen at a speed of 6 cm/h, say 200 pixels per hour gives 18 s per pixel. At a sample rate of 200 Hz, 3600 samples will be written to the same pixel column. Only a line connecting the highest and lowest value of the 18 s period will be seen. All information about local density of the signal between the high and low values will be lost. Therefore there is an amplitude histogram per pixel column and a color plot of this histogram is built. To give even more information, the median and the fifth and ninety-fifth percentile as bottom and peak estimates are shown. The CFM is shown to be useful for seizure detection, neonatal, care in the emergency room, and for the assessment of other brain disorders (18,19,21). The CFM trace may require a specialist for its interpretation. An EEG atlas provides a summary for the interpretation of the EEG based trace. Figure 4 shows an example for neonatal EEG monitoring. Studies have shown that when CFM is used in combination with a standard neurological examination, it enhances the clinician's ability to identify the presence of seizures or to monitor infants EEG and others.

Amplitude-Integrated EEG (aEEG) Monitor

Various brain activities may causes changes in normal EEGs. These changes might be in the amplitude, power, frequency, BIS, entropy or complexity. In fact, since EEG has become available, visual investigation of EEG has been used to asses the neurological function. It is evident that continuous EEG is a sensitive, but nonspecific measure of brain function and its use in cerebrovascular disease is limited. Visual interpretation of EEG is not an easy target and need well-trained expertise, which is not available all the times in the ICU. Besides, information that can be extracted by visual investigation is limited. The EEG amplitude shown in Fig. 5 by the aEEG monitor is the first feature, which has received the attention of neurologists and researchers. It is obvious that there is no clear difference between the aEEG associated with the normal and ischemic injury EEGs. The cerebral function monitor (CFM) uses the aEEG extracted from one channel. The aEEG can show bursts and suppression of the EEG. The CFM is well used for seizure detection and neonatal monitoring (18,19,21). However, the EEG amplitude shows low capability. To clarify this disadvantage, let us ask the question: Does isolectrical EEG mean brain death or even coma? There has been a study seeking the answer of this question (24). In this study, from 15 patients with clinical diagnosis as brain death, EEG was isoelectricity in eight patients while the remaining seven showed persistence of electrical activity. Comatose patients may also show the presence of electrical activity in the alpha band (8–13 Hz). Such diagnosis is referred to as alpha coma. This implies that both investigation and monitoring of EEG amplitude may not be a reliable confirmatory test of brain function and coma. The amplitude assessment of the EEG may then mislead the neurologist's decision.

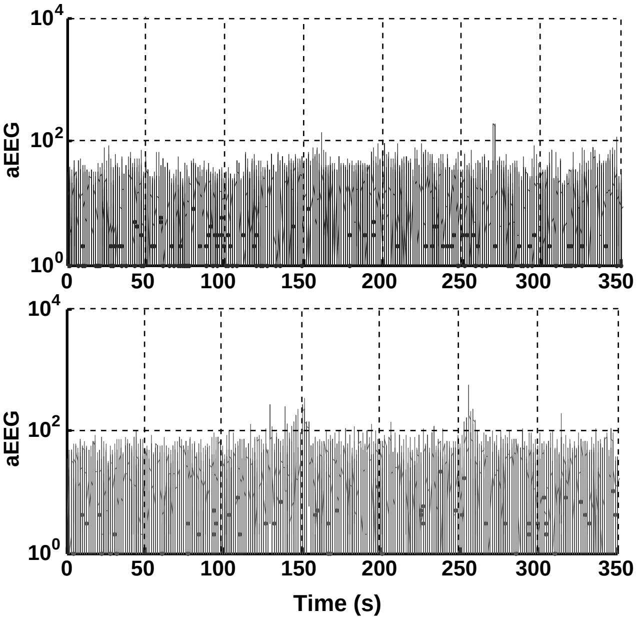


Figure 5. Amplitude-integrated EEG for normal EEG (top) and injury-related EEG of grade CPC 5. It is obvious that no significant differences between the two cases, which implies that amplitude may not be used for injury identification.

Spectrogram-Based Monitor

A number of studies have focussed their attention to the prognostication of frequency contents and the power spectrum of EEG (6,20,25,26). The normal EEG of adults often show three spectral peaks in delta, theta, and alpha, as demonstrated in Fig. 5 (top). The most common observation, in ischemic injury, for example, has been slowing background frequencies by increasing the power of delta rhythm and decreasing the powers of theta and alpha rhythms. Numerous approaches have been employed the frequency contents for developing a diagnostic tool or index. Monitoring the real-time spectrum has also been employed. While this approach gives an indication for ischemic injury, it requires a well-trained specialist. In animal studies, the spectral distance between a baseline (i.e., normal) EEG and the underlying injury-related one was employed as a metric for injury evaluation and

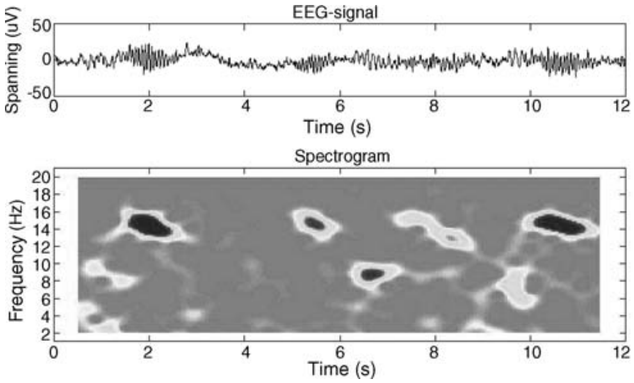


Figure 6. The EEG fragment with isolated 15 Hz spindles, which are clearly visible in the spectrogram. The spectrogram also shows that the spindles are alternated by short periods of 9 Hz activity.

monitoring. However, the spectral distance has the disadvantage of using the whole frequency contents. This is because using the whole frequency contents of the EEG signal increases the likelihood of artifacts-corrupted spectral contents. Time–frequency analysis is a signal analysis technique that provides an image of the frequency contents of a signal as a function of time. Several methods (or time–frequency distributions) can be used, one of which is the spectrogram. The spectrogram is the power spectrum of the investigated signal seen through a time window that slides along the time axis. Figure 6 shows a segment of sleep EEG signal (top) and its spectrogram. It is obvious that the spectrogram shows a sleep spindle at 15 Hz. The spectrogram shows the times where the spindle is activated. The time–frequency analysis can then be a helpful tool to facilitate the EEG interpretation, as is shown in the examples below.

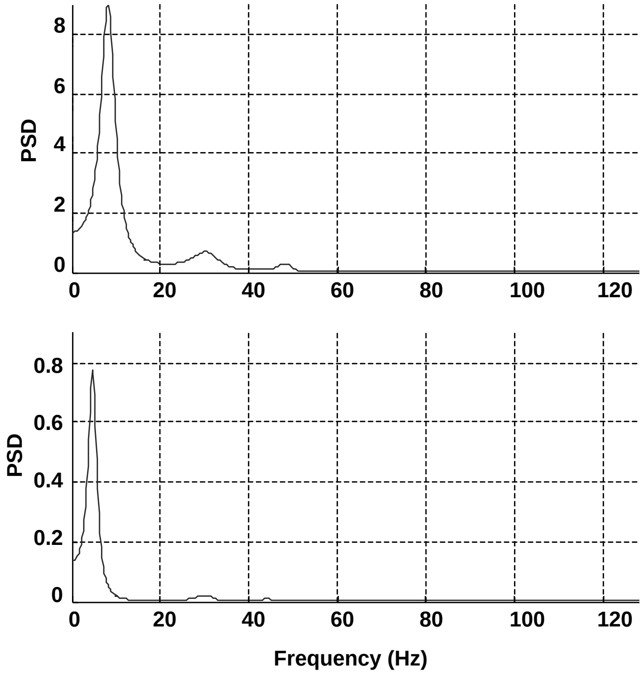


Figure 7. Power spectral density of EEG computed using the AR method applied to 4 s window and averaged >10 windows. (Top) Normal EEG signal and (bottom) abnormal. It is obvious that with abnormal EEG the background frequency gets slower and the high frequencies diminish.

Normalized Separation-Based Monitor

As mentioned, ischemic injury manifests itself in the EEG by slowing the background activity and reducing the high frequency. Such injury-related changes can be used for the separation of normal EEG from injury-related one. Based on this frequency information, a normalized separation was adopted as an qEEG measure. The normalized separation is a spectral-based qEEG measure for assessment of severity of brain injury. It uses the most relevant spectral information related to the normal EEG signal. The normal EEG has a power spectral density showing three fundamental spectral peaks as shown in Fig. 7 (top). It has been demonstrated that employing these three peaks is enough to yield a satisfactory quantitative measure. Moreover, looking selectively at the principal features of the EEG spectrum reduces the sensitivity of the measure to noise and artifacts. This is primarily because a full spectrum-based measure is likely to be susceptible to spectral components related to noise and artifacts. Therefore, the normalized separation employs the principal features of the spectrum and ignores the minor features, which are more sensitive to noise and artifacts. In comparison with amplitude-based measures, such as the aEEG, the aEEG is not a quantitative measure and represents a continuous EEG. This finding implies that well-trained specialist are needed for the interpretation of the aEEG trace. The amplitude is also susceptible to noise and artifacts that mislead the interpretation. In comparison with the higher order spectra-based measures, the normalized separation is enough since most information and features of the EEG are described by the power spectrum. A recent study and clinical investigation supports this claim. The EEG is commonly modeled as a stochastic process and for this reason phase is not important. The phase is the only feature retained in higher order spectra. Figure 8 shows three EEG signals and their corresponding normalized separations. The first EEG is very close to normal and provides a normalized separation of 0.2. In the second EEG spectrum, the third peak is diminished and the normalized

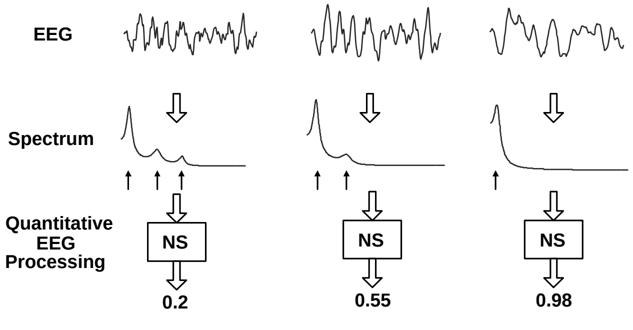


Figure 8. Normalized separation for three EEG cases. The left one is normal EEG where three spectral peaks are shown, the middle one is mildly injury-related, and the right one is a severely injury-related EEG. It is obvious that the spectral-based normalization makes significant separation between these three categories.

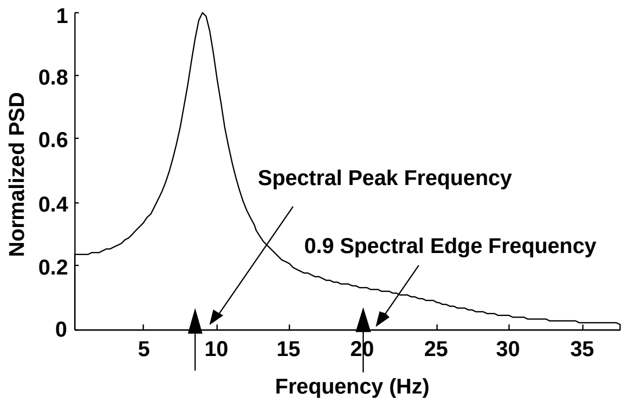


Figure 9. Normalized power spectral density (PSD) showing the spectral peak frequency (SPF) and the 90% spectral edge frequency.

separation is 0.55. In the third EEG spectrum, both second and third spectral peaks are diminished causing the EEG to be separated from the normal EEG by 0.98. The normalized separation ranges from 0 to 1, where the zero value represents and quantifies the normal EEG, while the one value corresponding to the sever abnormal EEG.

Spectral Edge Frequency Monitor

Computers make computation and search for spectral edge frequency of the EEG signal applicable for assessing and monitoring the cortical activity and brain dysfunction. The median frequency and the frequency edges providing 90–95% of the power have been reported to be useful (27,28). The spectral mean and peak frequencies have also been employed (29,30). The success of computing the time-varying spectral edge frequencies depends on the best estimation of the time-varying power spectrum. The fast-Fourier transform (FFT) is a commonly used approach for computing the real-time power spectrum. However, the FFT-based power spectrum provides poor frequency resolution since the resolution is proportional to the reciprocal of the analysis window. The model-based power spectrum estimate, such as the time-varying autoregressive, provides high resolution and low variance estimate of the power spectrum (Fig. 9).

Bispectral Index Monitor

In addition to spectrum, BIS also describes the frequency contents of the EEG. The power spectrum is often used for describing the frequency contents of the EEG modeled as a sum of noncoupled harmonics (17,31–33). In such situations, BIS are identically zero. However, if the focus is on the frequency contents of coupled harmonic (quadratic phase coupling harmonics), BIS is often used. Bispectrum is one of the first successful applications of electroencephalography, which measures the effects of anesthetics on the brain. The BIS index is a number between 0 and 100. It produces a number between 0 and 100 (100 represents the fully awake state, and zero no cortical activity). The BIS correlates with depth of sedation and anesthesia, and can predict the likelihood of response to commands and recall. The BIS values correlate with end-tidal volatile agent concentrations, and with blood and effect-site propofol concentrations. It is not very good at predicting movement in response to painful stimuli. However, there has been a recent study, which shows that BIS information is not necessary and power spectrum is satisfactory to describe an EEG signal. Another fact is that BIS is very sensitive to spike artifacts. The BIS index is a quantitative EEG index developed and employed for measuring the depth of anesthesia. It is based on third-order statistics of the EEG signal, specifically BIS density, and is commercially used for monitoring anesthetic patients. The index quantitatively measures the time-varying BIS changes in the EEG signal acquired from the subject before and during anesthesia. The BIS index will be zero when both the baseline and the underlying signal are either identical or Gaussians. This measure has been demonstrated to be effective for depth of anesthesia measurements. However, in some applications (classification of brain injury due to hypoxic/asphyxic cardiac arrest) the principal information and features of the EEG signal lie in second-order statistics, that is the power spectrum, and only minor information and features are associated with higher order statistics. Therefore, indexes based on higher order statistics may not be best suited to classify brain injury due to hypoxia/asphyxia (33). Besides, that higher order statistics are very sensitive to sparse-like artifacts, which deteriorates the index (34). Therefore, employing higher order statistics-based indexes require an efficient artifact removal approach for preprocessing the EEG signal. Below, a simulated example of BIS is presented. In this example, the BIS density is shown to present information on the coupling harmonics. Let $x(n)$ be a time-series consisting of three sinusoidal components whose frequencies are 64, 128, and 192 Hz. It is obvious that harmonic coupling between the first two sinusoids exists. Fig. 10a

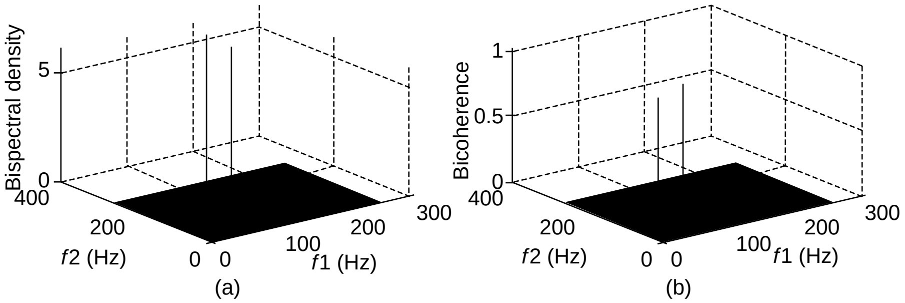


Figure 10. Bispectral density (a) and bicoherence (b) of a simulated sinusoidal signal. Bispectral density shows two lines at the coupling frequencies $(f_1, f_2) = (64, 64)$ and $(f_1, f_2) = (64, 128)$. Bicoherence shows two lines of unity value at the coupling frequencies.

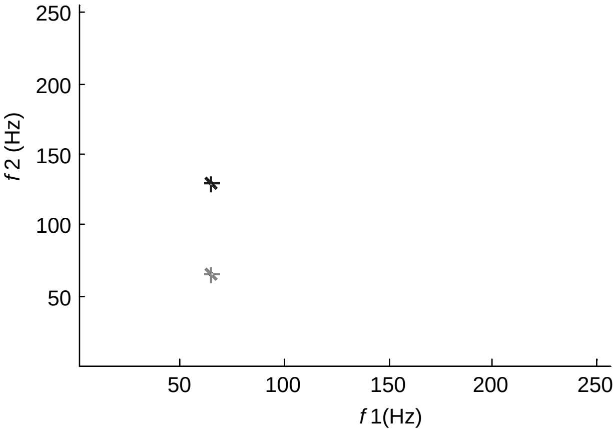


Figure 11. Contour plot of the bicoherence of x(n).

and b shows the mesh plots of the BIS and the two-dimensional bicoherence. Figure 11 shows the contour plot of the bicoherence. It is obvious that the BIS density shows two spectral lines at the frequencies $(f_1, f_2) = (64, 64)$ and $(f_1, f_2) = (64,128)$. The bicoherence is unity (perfect coupling) at these frequencies.

Entropy- and Complexity-Based Monitor

Since the brain processes information, the brain's total electrical activity probably corresponds to information processing in the brain. This assumption was used to study the entropy or self-information in the EEGs of anesthetic patients, postcardiac arrest, in sleep research, and seizure. Entropy as a measure quantifies the disorder of the EEG signal. It represents the complexity and nonlinearity inherent in the EEG signal. It has been shown that normal control subjects provide larger entropy values than those showing ischemic injury postcardiac arrest. The entropy starts to increase with the recovery of brain function. That is, entropy is a relevant indication of the brain order–disorder following cardiac arrest. The subject under anesthesia provides low entropy, while the awake subject shows high entropy since their brain is full of thinking and activity. Numerous approaches for the calculation of entropy have been used, such as Shannon entropy, approximate entropy, Tasllis entropy, and wavelet entropy.

Complexity based on chaotic, state space and correlation dimension has also been employed for assessment and monitoring of brain function (11–13,22,35–41).

COMMON SPECIFICATIONS OF OPTIMIZED MONITOR

The EEG monitor specifications are the hardware and software properties that make the monitor capable of easily and significantly performing assessment and classification of cortical activity. The monitor should satisfy minimum requirements. Common specifications of EEG monitors may include the following: compact design that is rugged and lightweight; automatic classification of EEG; off- and on-line qEEG index; optimized recognition and removal of artifacts; easy operation via friendly touch screen; continues testing of the electrodes to ensure a constant high quality of the EEG signal; variable electrode position; interface to external monitors and documentation systems; wireless communication between various sensors attached to the human; provides a secure way to transmit and store measured data; high-speed data processing; large amount of memory; on-board Ethernet connection.

CONCLUSION

This article presented a descriptive review for commonly known and employed neurological monitors. The typical neurological monitor consists of a few main devices and the software for running these devices. A brief review of the device specifications and their roles have been given. The monitors are classified into two main categories based on their applications and the indexes acquired from the digital EEG signal and employed for monitoring and assessment of cortical dysfunctions. Intuitions of the EEG monitors, with no mathematical details, have been presented. The article concludes by describing the most common specifications for the optimized monitor.

BIBLIOGRAPHY

1. Schneider G. EEG and AEP monitoring during surgery The 9th ESA Annual Meeting, Gothenburg, Swede, April 7–10, 2001.
2. Fisch BJ. EEG Primer- Basic princibles of digital and analog EEG. Fisch & Spehlmann's, Third revised and enlarged edition. New York: Elsevier Science BV; 1999.
3. Collura TF. The Measurement, Interpretation, and Use of EEG Frequency Bands. Report Dec. 7, 1997.
4. Berger H. Uber das elecktroenkephalogram des menchen. Arch Psychiatr Nervenkr 1929;87:527–570.
5. Teplan M. Fundamentals of EEG measurement. Meas Sci Rev 2002;2.
6. Gharieb RR, Cichocki A. Segmentation and tracking of EEG signal using an adaptive recursive bandpass filter. Int Fed Med Biol Eng Comput Jan. 2001;39:237–248.
7. Kong X, et al. Qauntification of injury-related EEG signal-changes using distance measure. IEEE Trans Biomed Eng July 1999;46:899–901.
8. Wendling F, Shamsollahi MB, Badier JM, Bellanger JJ. Time-frequency matching of warped depth-EEG seizure observations. IEEE Trans Biomed Eng May 1999;46:601–605.
9. Mingui Sun, et al. Localizing functional activity in the brain through time-frequency analysis and synthesis of the EEG. Proc IEEE Sept. 1996;84:1302–1311.
10. Ning T, Bronzino JD. Bispectral analysis of the rate EEG during various vigilance states. IEEE Trans Biomed Eng April 1989;36:497–499.
11. Hernero R, et al. Estimating complexity from EEG background activity of epileptic patients-Calculating correlation dimensions of chaotic dynamic attractor to compare EEGs of normal and epileptic subjects. IEEE Eng Med Biol Nov./Dec. 1999; 73–79.
12. Roberts SJ, Penny W, Rezek I. Temporal and spatial complexity measures for electroencephalogram based brain-computer interface. Med Biol Eng Comput 1999;37:93–98.
13. Zhang XS, Roy RJ. Predicting movement during anesthesia by complexity analysis of electroencephalograms. Med Biol Eng Comput 1999;37:327–334.

14. Anderson CW, Stolz EA, Shamsunder S. Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks. IEEE Trans Biomed Eng March 1998;45:277–286.
15. Hazarika N, et al. Classification of EEG signals using wavelet transform. Signal Process 1997;59:61–72.
16. Quiroga RQ, et al. Searching for hidden information with Gabor transform in generalized tonic-clonic seizures. Electroenceph Clin Neurophysiol 1997;103:434–439.
17. Gajraj RJ, et al. Analysis of the EEG bispectrum, auditory potentials and the EEG power spectrum during related transitions from consciousness to unconsciousness. Br J Anesthes 1998;80:46–52.
18. Toer MC, et al. Amplitude integrated EEG 3 and 6 hours after birth in full term neonates with hypoxic-ischemic encephalopathy. Rch Dis Child Fetal Neonatal Ed 1999;81:19–23.
19. Toet MC, et al. Comparison between simultaneously recoded amplitude integrated EEG (Cerebral function monitor) and standard EEG in neonates. Pediatrics 2002;109:772–779.
20. Hassanpour H, et al. Time-frequency based newborn EEG seizure detection using low and high frequency signatures. Physiol Meas 2004;25:935–944.
21. Nageeb N, et al. Assessment of neonatal encephalopathy by amplitude-integrated EEG. Pediatrics June 1999;103:1263–1266.
22. Bezerianos A, Tong S, Thakor N. Time-dependent entropy estimation of EEG rhythm changes following brain ischemia. Ann Biomed Eng 2003;31:1–12.
23. Kreuer S, et al. The narcotrend- a new EEG monitor designed to measure the depth of anesthesia. Anethesit 2001;50:921–925.
24. Paolin A, et al. Reliability in diagnosis of brain death. Intensive Care Med Aug. 1995;21:657–662.
25. Jung TP, et al. Estimating alertness from the EEG power spectrum. IEEE Trans Biomed Eng Jan. 1997;44:60–69.
26. Celka P, Colditz P. A computer-aided detection of EEG seizures in infants: A singular spectrum approach and performance comparison. IEEE Trans Biomed Eng May 2002;49:455–462.
27. McDonald T, et al. Median EEG frequency is more sensitive to increase in sympathetic activity than bispectral index. J Neurosurg Anesthesiol Oct. 1999;11:255–264.
28. Inder TE, et al. Lowered EEG spectral edge frequency predicts presence of cerebral white matter injury in premature infants. Pediatrics Jan. 2005;111:27–33.
29. Rampil IJ, Matteo RS. Changes in EEG spectral edge frequency correlated with the hemodynamic response to laryngoscopy and intubation. Anesthesiol 1987;67:139–142.
30. Rampil IJ, Matteo RS. A primer for EEG signal Processing in anesthesia. Anesthesiology 1998;89:980–1002.
31. Akgul T, et al. Characterization of sleep spindles using higher order statistics and spectra. IEEE Trans Biomed Eng Aug. 2000;47:997–1000.
32. Michael T. EEGs, EEG processing and the bispectral index. Anesthesiology 1998;89:815–817.
33. Miller A, et al. Does bispectral analysis of the electroencephalogram add anything bur complexity? B J Anesthesia 2004;92: 8–13.
34. Myles PS, et al. Artifact in bispectral index in a patient with severe ischemic brain injury. Case Report Int Anesth Res Assoc 2004;98:706–707.
35. Radhakrishnan N, Gangadhar BN. Estimating regularity in epileptic seizure time-series data- A complexity measure approach. IEEE Eng Med Biol May/June 1998; 98–94.
36. Lemple A, Ziv J. On the complexity of finit sequences. IEEE Trans Inf Theory Jan 1976;22:75–81.
37. Tong S, et al. Parameterized entropy analysis of EEG following hypoxic-ischemic brain injury. Phys Lett A 2003;314: 354–361.
38. Lerner DE. Monitoring changing dynamics with correlation integrals: Case study of an epileptic seizure source.
39. Xu-Sheng, et al. EEG complexity as a measure of depth of anesthesia for patients. Yearbook of Medical Informatics. 2003; 491–500.
40. Bhattacharya J. Complexity analysis of spontaneous EEG. Acta Neurobiol Exp 2000;60:495–501.
41. Quiroga RQ, et al. Kullback-Leibler and renormalized entropies application to electroencephalogram of epilepsy patients. Phys Rev E Dec. 2000;62:8380–8386.
42. Mizrahi EM, Kellaway P. Characterization and classification of neonatal seizures. Neurology Dec. 1987;37:1837–1844.

Reading List

Blanco S, et al. Time-frequency analysis of electroencephalogram series. *Phys Rev* 1995;51:2624–2631.

Blanco S, et al. Time-frequency analysis of electroencephalogram series. III wavelet packets and information cost function. *Phys Rev* 1998;57:932–940.

Caton R. The electric currents of the brain. *BMJ* 1875;2–278.

D'attellis CE, et al. Detection of epileptic events in electroencephalograms using wavelet analysis. *Annals of Biomed Eng* 1997;25:286–293.

Franaszczuk PJ, Blinowska KJ, Kowalczyk M. The application of parameteric multichannel spectral estimates in the study of electrical brain activity. *Biol Cybern* 1985;51:239–247.

Gabor AJ, Leach RR, Dowla FU. Automated seizure detection using self-organizing neural network. *Electroenceph Clin Neurophysiol* 1996;99:257–266.

Gath I, et al. On the tracking of rapid dynamic changes in seizure EEG. *IEEE Trans Biomed Eng* Sept. 1992;39:952–958.

Geocadin RG, et al. A novel quantitative EEG injury measure of global cerebral ischemia. *Clin Neurophysiol* 2000;11:1779–1787.

Geocadin RG, et al. Neurological recovery by EEG bursting after resuscitation from cardiac arrest in rates. *Resucitation* 2002;55: 193–200.

Gotman J, et al. Evaluation of an automatic seizure detection method for the newborn EEG Electroenceph Clin. *Neurophysiol* 1997;103:363–369.

Hernandez JL, et al. EEG predictability:adequacy of non-linear forcasting methods. *Int J Bio-Medical Comput* 1995;38:197–206.

Holzmann CA, et al. Expert-system classification of sleep/awake states in infants. *Med Biol Eng Comput* 1999;37:466–476.

Liberati D, et al. Total and Partial coherence analysis of spontaneous and evoked EEG by means of multi-variable autoregressive processing. *Med Biol Eng Comput* 1997;35:124–130.

Pardey J, Roberts S, Tarassenko LT. A review of parametric modeling techniques for EEG analysis. *Med Eng Phys* 1996;18:2–11.

Petrosian A, et al. Recurrent neural network based prediction of epileptic seizures in intra- and extracranial EEG. *Neurocomput* 2000;30:201–218.

Popivanov D, Mineva A, Dushanova J. Tracking EEG dynamics during mental tasks-A combined linear/nonlinear approach, IEEE Eng. *Med Biol* 1998; 89–95.

Quiroga RQ, et al. Performance of different synchronization measures in real data: A case study on electroencephalographic signals. *Phys Rev E* 2002;65:1–14

Sadasivan PK, Dutt DN. SVD based technique for noise reduction in electroencephalogram signals. *Signal Process* 1996;55: 179–189.

Salant Y, Gath I, Hebriksen O. Prediction of epileptic seizures from two-channel EEG, Med Biol. *Eng Comput* 1998;36:549–556.

Schraag S, et al. Clinical utility of EEG parameters to predict loss of consciousness and response to skin incision during total intervention anesthesia. *Anesthesia* April 1998;53:320–325.

Selvan S, Srinivasan R. Removal of ocular artifacts from EEG using and efficient neural network based adaptive filtering technique. *IEEE Signal Process Lett* Dec. 1999;6:330–332.

Vigario RN. Extraction of ocular artifacts from EEG using independent component analysis. *Electroenceph Clin Neurophysiol* 1997;103:395–404.

Zapata A, et al. Detecting the onset of epileptic seizures. *IEEE Eng Med Biol* May/June 1999; 78–83.

Zygierewicz J, et al. High resolution study of sleep spindles. *Clinic Neurphysiol* 1999;110:2136–2147.

Publisher's note. A revised version of this article was submitted after our print production deadline. The revised version will appear in the online edition of this encyclopedia.

See also Electroencephalography; evoked potentials; monitoring in anesthesia; monitoring, intracranial pressure.

NEUROMUSCULAR STIMULATION. See Functional electrical stimulation.

NEUTRON ACTIVATION ANALYSIS

Xiaolin Hou
Risø National Laboratory
Roskilde, Denmark

INTRODUCTION

Neutron activation analysis (NAA) as an elemental analytical method has been used for a long time to determine trace elements in biomedical research and clinical analysis (1,2). The main problems associated with the determination of trace elements in biomedical research and clinical analysis are the very low concentrations of some elements, and the limited amount of sample materials available. It is necessary that the analytical method be sensitive and free of blank contribution.

As a nuclear analytical technique, NAA is based on the excitation of the atomic nucleus of an isotope of the element with neutrons, and the emission of specific radiation, such as gamma rays, from the excited nucleus by decay. Therefore, only trace elements present in the sample during neutron activation will be excited and are able to be measured in this way. The possible contamination of sample during subsequent handling will not influence the result. Thus a comprehensive radiochemical separation of a particular element from interfering elements can be carried out to significantly improve the detection limit, which will not introduce any blank for the measured element. If no pretreatment of sample is completed, almost no blank value will be introduced in the analytical procedure. Therefore, the method can be free of blank contribution. Neutron activation analysis is very sensitive for many trace elements, while many matrix elements such as carbon, hydrogen, oxygen, and nitrogen are less activated by neutrons and produce almost no activity after neutron activation. Therefore, NAA is suitable for the analysis of biomedical samples due to little or no interference from matrix elements. The interference from activated minor elements such as chlorine, sodium, bromine, and potassium may be eliminated by a few days decay of sample due to the short lifetimes of radioactive nuclides of these elements. It is normally used for *in vitro* analysis, but also can be used for *in vivo* analysis of the whole or part of living bodies. This consists of mostly major and minor elements, such as Ca, Cl, K, N, Na, and P, but also some trace elements, such as iodine in thyroid, Si, Cd, Hg in lung and kidney. The contrast with conventional NAA, isotopic and accelerator neutron sources and small reactor are used and prompt gamma rays are measured (3). However, the problem of radiation to living body limits its application. This article will not describe this problem in detail and is an updated version of the previous article published in the first edition of this encyclopedia (4).

THEORY

Neutron activation is a reaction of the nucleus of an element with neutrons to produce a radioactive species, so-called radionuclide. Neutrons used in NAA can be produced by a nuclear reactor, a radioisotope, and an accelerator, in which the nuclear reactor is the most common neutron source used for NAA due to its high neutron flux and suitable neutron energy. Neutrons can exhibit a wide range of energy, which range from thermal neutrons with an average energy of 0.025 eV in a thermal nuclear reactor to fast neutrons of 14 MeV in an accelerator neutron generator. By bombarding an element with neutrons, a neutron is absorbed by the target nucleus to produce a highly energetic state of the resulting nucleus containing an additional neutron. Some excess energy is immediately lost, usually by emission of a gamma ray, a proton, or an alpha particle. In a sample exposed to neutrons, the type of nuclear reactions depends on the energy of the neutrons and on the elements present. The main reaction occurring with thermal neutrons is the (n, γ) reaction. In this case, the highly energetic level of the produced nucleus is de-excited by emission of a gamma ray, while (n, p) and (n,α) reactions are induced by fast neutrons reaction. In conventional NAA, the element is determined by measurement of the radionuclides formed by de-excitation of the produced nucleus. However, the element can be determined also by the measurement of the gamma rays emitted during the de-excitation of the produced nucleus, which is so-called prompt gamma activation analysis due to the very fast de-excitation process (1 ps). In this article, only conventional NAA is discussed. The probability of a particular reaction is expressed by the activation cross-section, σ, with a unit of barn (10^{-28} m^2). An isotope of an element has its specific activation cross-section. The cross-section of a particular nucleus depends on the energy of the neutron. The (n, γ) reaction normally has a higher activation cross-section than (n, p) and (n, α) reactions. The rate of formation of the radionuclide in the neutron activation is expressed as

$$\frac{dN^*}{dt} = \sigma\phi N \tag{1}$$

N is the number of the target nuclei in the atom, so $N = (m/M)N_A\,\theta$. Here, m is the mass of the target element (in g); M is the atomic mass; N_A is Avogadro's number; and θ is the isotope abundance of the target nuclide, N^* is the number of the activated nuclide at time t; ϕ is the neutron flux density (in neutron $m^{-2}{\cdot}s^{-1}$), which is used to express the neutron number pass through a unit area in a unit time, which can be considered also as a product of velocity of a neutron and its concentration.

If the nuclide formed is radioactive, it will decay with time and the decay rate of the radionuclide will be

$$\frac{dN^*}{dt} = -\lambda N^* \tag{2}$$

Here λ is the decay constant of activated nuclide, $\lambda = \ln 2/T_{1/2}$ and $T_{1/2}$ is the half-life of activated nuclide. So, the production rate of an activated nuclide is expressed as:

$$\frac{dN^*}{dt} = (\sigma\phi N) - (\lambda N^*) \tag{3}$$

The activity or disintegration rate (A_0) at the end of irradiation time t_i is then:

$$A_0 = \lambda N^* = \sigma\phi N(1 - e^{-\lambda ti}) \tag{4}$$

The saturation activity of the activated nuclides (Am), that is, the activity when the production of activity is equal to the decay of the activity, can be calculated as:

$$\mathrm{Am} = \sigma\phi N = \frac{m}{M} N_A \theta\sigma\phi \tag{5}$$

If considering the decay of activated radionuclides during the decay (t_d) and counting (t_c), the measured activity of radionuclides is calculated by

$$A = \sigma\phi N(1 - e^{-\lambda ti})e^{-\lambda td}(1 - e^{-\lambda tc})/\lambda \tag{6}$$

The actual number of events recorded by a detector for a particular radionuclide is only a fraction f of the number of decays calculated from Eq. 6, because not every decay can emit a characteristic gamma ray, and once a gamma ray is emitted they may not reach the detector. Considering these findings, the simplest and most accurate way to determine an element by NAA is to irradiate and measure a comparator standard with an exact known content of the element together with the sample. In this case, the ratio of the element content in sample m_s to that in comparator standard m_c is equal to the ratio of their activities, $A_s/A_c = m_s/m_c$. Therefore, the content of target element in the sample can be calculated by

$$m_s = \frac{A_s}{A_c} m_c \tag{7}$$

In addition, from such a single comparator, it is also possible to calculate the sensitivity of other elements by means of the k-factor (5). This factor is an experimentally determined ratio of saturation specific activities expressed in counts:

$$k = \frac{f\theta\sigma}{f^*\theta^*\sigma^*}\,\frac{M^*}{M} \tag{8}$$

Here, the asterisk refers to the element of a single comparator. The factor f is a combination of the emission probability of η and detection efficiency ε. If the relative efficiency function of the detector is known, calibration may be based on k_0 values (5):

$$k = k_0 \frac{\varepsilon}{\varepsilon^*} \tag{9}$$

where

$$k_0 = \frac{\eta\theta\sigma}{\eta^*\theta^*\sigma^*}\,\frac{M^*}{M} \tag{10}$$

These k_0 values are fundamental constants and may be found in tabulation; methods for taking into account the influence on k_0 values of difference in neutron flux spectrum have been developed.

The analytical sensitivity of NAA for various elements can be predicted from Eq. 6, combined with the emission probability of gamma rays of the radionuclide and the counting efficiency of the characteristic energy of the radionuclide. The calculated interference free detection limits of NAA for various elements are listed in Table 1. Note that this data is only applied to radionuclides completely free from all other radionuclides, that is after a complete radiochemical separation. In actual analysis, the activity from other activated elements will interfere with the detection of the target element by increasing the baseline counts under its peaks, so the detection limit will be poorer than that estimated in Table 1. However, in a sample with known composition, practical detection limits may be predicted in advance (6).

Since NAA is based on the excitation of the atomic nucleus of an isotope instead of the surrounding electrons, no information on the chemical state of the element can be obtained. In addition, the de-excitation of the produced nucleus by the emission of high energy gamma rays or other particles gives the formed radionuclide a nuclear recoil energy of several tens of electron volts. This is more than sufficient to break a chemical bond, and the formed radionuclide may no longer be found in its original chemical state. It is therefore not possible to directly use NAA for chemical speciation of an element. However, a NAA

Table 1. Interference Free Detection Limit for Elements by NAA Based on Irradiation for 5 h at a Neutron Flux Density of 10^{13} n/cm$^2{\cdot}$s^{-1} and Typical Counting Conditions[a]

Detection limit, ng	Element
0.001	Dy, Eu
0.001–0.01	Mn, In, Lu
0.01–0.1	Co, Rh, Ir, Br, Sm, Ho, Re, Au
0.1–1	Ar, V, Cu, Ga, As, Pd, Ag, I, Pr, W, Na, Ge, Sr, Nb, Sb, Cs, La, Er, Yb, U
1–10	Al, Cl, K, Sc, Se, Kr, Y, Ru, Gd, Tm, Hg, Si, Ni, Rb, Cd, Te, Ba, Tb, Hf, Ta, Os, Pt, Th
10–100	P, Ti, Zn, Mo, Sn, Xe, Ce, Nd, Mg, Ca, Tl, Bi
100–1000	F, Cr, Zr, Ne
10,000	Fe

[a] See Ref. (6).

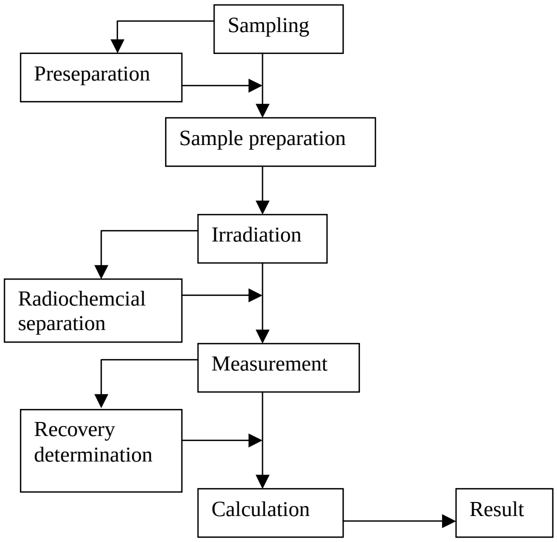


Figure 1. Procedure of neutron activation analysis.

detection coupled with a preseparation of chemical species of elements, so-called molecular activation analysis, can be applied for the chemical speciation of elements for biomedical studies (7). Used in this way, NAA is no longer without a blank problem.

EQUIPMENT AND METHODOLOGY

Optimal utilization of the special qualities of NAA requires an appropriate methodology, which is adapted to the analysis of various biomedical samples for different elements and their species. Figure 1 shows a scheme of NAA, which divides NAA into several steps, some of which are indispensable, while others are supplementary for specific purposes.

Sampling

Sampling is the first step for any analysis, and is the most important step in any meaningful investigation (8,9). There is increasing awareness that the quality of an analytical result may often be influenced more by sample collection than by final measurements. One of the main considerations during sampling should be devoted toward the representativity of the analyzed sample to the object under study. Other considerations should be the avoidance of any contamination during sampling and finding suitable storage of the sample until analysis to prevent the losses of the elements of interest and to avoid contamination. The selection and collection of a representative sample may need to consider two aspects: the type of sample and the size of the selected sample (8). Frequently, valuable data can be obtained from the analysis of readily accessible *in vivo* samples, such as blood and urine. In some cases, the analysis of tissue sample may provide more useful information. A considerable effort has been spent on the analysis of hair and nail (9). They have an advantage of being easily available, however, it may be less representative for most of the elements, and a multitude of factors may affect the observed results. The contamination during the sampling mainly comes from the tools and the container used for samples and the sampling environment. With the exception of excreta, all biomedical materials have to be removed from the organism by means of tools for piercing, cutting, or shearing. The best materials for tools from the point of view of eliminating contamination are polyethylene, Teflon, or other plastic materials, or stainless steel, if Cr, Ni, Co, and Fe are not being considered, but pure titanium is a good material due to its hardness for cutting. For the container, polyethylene, Teflon, and synthetic quartz are the best materials to avoid contamination and absorption of trace elements on the well of the container. The problems of sampling and sample handling have been given more attention in the studies of medical trace elements. A number of sampling protocols were recommended. A detailed discussion of this issue can be found in many articles and books (2,8,9).

Preseparation

Since NAA cannot be directly used for chemical speciation, various chemical species of trace elements have to be separated before irradiation. In recent years, many methods have been developed for the chemical speciation of various trace elements, such as Se, I, Cr, Hg, As, Al, Fe, Zn, Cu, and rare earth elements in biomedical samples (7,10–15). All of these methods are based on preseparation using various chemical and biochemical separation techniques, such as ion exchange, gel chromatography, high-performance liquid chromatography (HPLC), supercritical fluid, electrophoretic, and fractionation techniques. The chemical speciation in biomedical trace elements studies normally includes the following aspects: subcellular distribution of trace elements in various tissues; chemical valence states of trace elements in liquid samples such as urine, blood, and cytosol of tissue homogenate; specific combination of trace elements with various proteins, polypeptides, enzymes, hormones, and small molecular compounds. The separated chemical species of trace elements is then quantitatively determined by NAA. Since the sample is not destroyed and no chemical reagents are used, NAA itself is actually a contamination-free analytical technique. However, in preseparation the utilization of chemical reagents and many types of equipment can cause a high risk of contamination. An enriched stable isotope trace technique was therefore used to overcome this problem (15), which is based on the fact that NAA is actually an analytical technique for isotopes instead of elements. In this case, the contaminations during the separation can be identified and eliminated because of the different isotopic components of the trace elements in the samples with those contaminants from the chemical reagents and other sources.

Due to the very low concentration of most trace elements in biomedical samples, preseparation of elements of interest from main interfering elements and preconcentration is sometimes useful for the improvement of the detection limit of the elements of interest. The simplest preconcentration should be lyophilization and ashing of the samples, which can be directly carried out in an irradiation container, such as a quartz ampoule and aluminum foil to eliminate problems associated with sample transfer. The techniques used for preseparation of the elements of interest are similar as those for radiochemical separation described below. Special attention should be given to avoid contamination and addition of extra interfering elements, such as Cl and Na.

Sample Preparation

Before neutron irradiation, the sample must be transferred to an irradiation container for transport to and from the irradiation position. These containers should not contaminate the sample, nor should they add significantly to the total amount of radioactivity produced during activation. Suitable materials are low density polyethylene for moderate irradiation and high pure aluminum and synthetic quartz for prolonged irradiation. For short-term irradiation, the sample can be wrapped in a thin polyethylene film and inserted in a polyethylene irradiation capsule. After irradiation, the sample with the polyethylene foil is directly measured, since there is no significant contribution of trace elements and radioactivity from the polyethylene film and difficulties on removal of sample from the irradiated film. In the determination of some elements, such as Se, F, B, and Li (16,17), the irradiated samples have to be measured very quickly (<10 s) due to the very short half-lives of the formed radionuclides (0.8 s for ^{8}Li, 17.5 s for ^{77m}Se, 11.0 s for ^{20}F). In this case, the sample with the polyethylene capsule has to be directly transferred to the detector for measurement. For long-term irradiation of sample in a reactor, it is normally required that the sample be dried to avoid any problem of explosion of the container due to the high pressure produced in the container by the radiolysis of water in the samples. In addition, the removal of water from the sample will improve the determination of elements of interest, because an increased amount of sample can be irradiated and the counting geometry can be improved. Lyophilization techniques are normally used for the removal of water, although evaporation by heating can be used. In addition, dry ashing is sometimes used to further improve the determination of elements, but we should watch for loss of elements of interest during this process. The dried sample is normally wrapped in high purity aluminum foil, which is then inserted into an aluminum container for irradiation, since only a short-lived radioisotope of aluminum, ^{28}Al (2.24 min), is formed in the neutron irradiation of Al. This radioisotope quickly disappears by decaying, so it does not interfere with the determination of trace elements that form long-lived radionuclides. For some elements, such as mercury and arsenic, a synthetic quartz ampoule was used because they volatize during irradiation and have very low impurities in quartz materials and are less radioactive from activated SiO_2. To minimize losses and contamination, biomedical samples are often sampled in a high purity quartz ampoule. They are then directly lyophilized and ashed in the quartz ampoule. After irradiation, they can be measured in the same ampoule.

Irradiation

Different neutron sources are available for the activation analysis of samples. Isotopic neutron sources, such as Ra–Be and Am–Be sources that rely on the (α, n) reaction and ^{252}Cf source that is based on spontaneous fission, yield neutron spectra with a limited range of neutron energies. Accelerators can also be used to produce a particular energy of neutrons, such as 14 MeV neutrons produced by (D,T) reaction and 2.5 MeV neutrons by (D,D) reaction. A miniature, sealed-tube neutron generator with a yield of 10^{8}–10^{11} neutron/s has been used widely as a neutron source for NAA (18,19). However, due to low neutron flux produced in these two kinds of neutron sources, they are very seldom used for the NAA of biomedical samples.

Nuclear reactors are much more suitable for NAA of biomedical materials, because they provide much higher neutron flux density (10^{11}–10^{14} n/cm^{2}·s) with correspondingly higher sensitivities for trace elements. In addition, neutrons in the reactor can be well moderated to thermal energies, which is very suitable for NAA because of the high thermal neutron activation cross-sections of trace elements and less interference from fast neutron reactions. For some elements such as iodine, arsenic, and selenium, which are preferably determined by epithermal NAA, an epithermal neutron irradiation channel was set up in some reactors by shielding thermal neutrons with cadmium (20). Almost all research reactors installed pneumatic transfer systems for easy and automatic transfer of samples to and from the irradiation position. Some of them can also quickly transfer the irradiated sample to the counting position above the detector for the determination of elements by counting short-lived radionuclides. The detection limit and analytical precision of NAA can be improved by repeated irradiation and counting of samples, the so-called cyclic NAA. This is particularly useful for the elements determined by short-lived radionuclides (such as Se, F, I, etc.). Cyclic NAA systems were therefore installed in some reactors (17). As a nuclear facility, many nuclear reactors are located in large research centers, normally outside of the city; it makes them less accessible and consequently makes NAA inconvenient for most researchers and medical doctors. In the 1980s, a miniature neutron source reactor (MNSR) was developed especially for NAA in China. Due to its intrinsic safety in design, this reactor can be installed in hospitals and research institutes in a big city. With a combination of its small size and low price, it makes NAA possible as a clinically analytical tool easily and conveniently accessible for many researchers and medical doctors. Ten such reactors were installed in universities and institutes in China, Pakistan, Iran, Syria, Ghana, and Niagara (20,21). A few similar reactors (SLOWPOKE reactor) were developed and installed in Canada.

Measurement

Trace elements are determined by measurement of the activity of the radionuclides formed after neutron irradiation, which can be carried out by counting the number of events taking place in the detector. The choice of the detector used for NAA depends on the type of decay and energies of the radiation emitted by the radionuclides formed. Some radionuclides produced by neutron activation of uranium and thorium decay by emission of neutrons, which are counted by means of a neutron detector filled with BF_3 or ^{3}He gas. Meanwhile, the radionuclides of boron and lithium, which decay by emitting very high energy β particles, may be counted by a Cherenkov detector (16). All these analyses require that the detector be installed at the reactor site because of the short half-lives of these radionuclides, and are consequently not commonly used for biomedical materials.

The most common radionuclides measured by NAA decay by emission of characteristic gamma rays. The most widely used detector is a gamma semiconductor detector, of high purity germanium (HpGe). This detector has high energy resolution; it may separate the gamma rays with energy difference of only 2–3 keV. According to the energies of the radionuclides of interest, different types of germanium detectors can be used. Planar germanium detector and HpGe detector with a thin window are used for the measurement of low energy gamma rays down to 10–20 keV. These detectors have the best energy resolution. In addition, a silicon (lithium) detector can be used for the measurement of low energy gamma rays and X rays. A normal coaxial HpGe detector is used for most radionuclides with energies of gamma rays >60 keV, and well-type HpGe detectors can be used for improvement of the counting efficiency if a very low level of radioactivity needs to be measured. In this kind of detector, the sample sits in the middle of the germanium crystal and almost 4π geometry can be obtained. However, only a small sample can be measured in this detector due to the limited well size of the detector. In connection with appropriate electronic equipment, such as high voltage supply, preamplifier and main amplifier, and a pulse height analyzer with 4,096 or 8,192 channels, it is possible to determine simultaneously many different radionuclides if only their gamma-ray energies differ by 2–3 keV. Additional discrimination is achieved between radionuclides with different half-lives by counting the sample at different decay times.

The measurement of gamma rays is based on the interactions of gamma rays with matter, such as the photoelectric effect, Compton scattering, and pair production. In the photoelectric effect, the gamma ray ejects a shell electron from a germanium atom and produces a vacancy; the ejected electron has a kinetic energy equal to the energy of the gamma ray less the binding energy of the electron. It may interact with other electrons, thus causing secondary ionization, and produce more vacancies in the shell of the germanium atoms. The number of produced photoelectrons and corresponding vacancies is determined by the energy of the gamma ray. In this way, the gamma-ray energy is converted to an electric signal in the detector, which is then amplified by preamplifier and main amplifier, the output from the main amplifier is a peak of nearly Gaussian shape with an amplitude proportional to the gamma-ray energy that enters the detector. This electric signal is finally registered by a multichannel analyzer (MCA) as a count, the number of the channel in the MCA corresponds to the gamma-ray energy and the counts in a channel correspond to the numbers of the gamma rays with the same energy. The peak of the gamma spectrum registered in the MCA is normally also a Gaussian distribution. The width of the peak, or the energy resolution, is an important parameter of the detector. Except for the photoelectric effect, pair production process results in a gamma ray with energy less than 0.511 and 1.02 MeV of the original one. Compton scattering results in a continuum of energy being transferred, which increases the baseline counts of the gamma peaks, and therefore worsens the detection limit of the radionuclide of interest. Most gamma emitting radionuclides also emit beta particles. The interaction of the beta particle with the detector results in a continuum of energy under the gamma spectrum. A beta absorber can be used to reduce the interference of the beta particles. The utilization of anti-Compton techniques will reduce Compton interference; a recently developed multiparameter coincidence spectrum technique significantly improved the detection limit of the trace elements of interest (22).

Radiochemical Separation

NAA for very low level of elements is limited by the presence of other elements in the sample. Some minor elements in the biomedical samples such as Na, Cl, Br, and P contribute to high radioactivity of the irradiated sample, and the signals of the radionuclides produced from many trace elements of interest are overlapped by an increased Compton continuum under the gamma peaks. In order to measure the trace elements of interest and improve the detection limit and analytical accuracy of many trace elements, it is required that these interfering elements be eliminated before counting. Since the irradiated sample is radioactive, this procedure is called radiochemical or post-irradiation separation. Radiochemical separation is normally carried out by the following steps: addition of carrier, decomposition of irradiated sample, chemical separation, and preparation of separated samples for counting. Detailed development and the present progress of RNAA were reviewed by several authors (22–24). Since the activated elements are radioactive, this procedure does not create any blank problem, but may easily result in losses of trace elements to be determined. In order to minimize these losses, a suitable amount of carrier element is always added to the sample before radiochemical separation. Since the amount of carrier element is much higher than the same element from the original sample, it can then be used to monitor the chemical recovery of the elements determined during the radiochemical separation. In many cases, some carriers of interfering elements are also added to improve the removal of these interferences, the so-called holdback carrier. A complete equilibrium between an inactive carrier and radioactive element is necessary to be sure that both undergo the same physical and chemical procedure, and the same chemical recoveries during the radiochemical separation, which is normally carried out by a comprehensive oxidation–reduction cycle.

Both dry ashing and wet digestion are used to decompose the biomedical samples. Dry ashing at 400–700 °C is more suitable for large samples; then the ashed sample can be easily dissolved by diluted acid. However, a considerable loss of several elements may occur during ashing due to volatilization, which can be partly eliminated by ashing in a closed system or by using low temperature ashing at 200–250 °C under vacuum. Iransova and Kucera (25) recently showed alkali fusion at high temperature (800–850 °C) is very effective and rapid for decomposition of the biomedical sample and reduction of the losses of many trace elements. Based on the volatilization of some elements, a combustion method was used to directly separate these elements from the matrix and interfering elements. A particular example is the radiochemical NAA of iodine (26). Acid digestion is a more popular method for decomposition of biomedical samples. This method is normally carried out by boiling a sample in concentrated HNO_3 with $HClO_4$ or H_2SO_4. Sometimes H_2O_2 is added to completely decompose the organic components of the sample. Heating in an open system can also result in losses of some volatile elements such as iodine and mercury; utilization of refluxing can significantly improve the recovery of these elements. Recently, a microwave assisted digestion system was also introduced for the decomposition of sample. The main advantages of this method are rapid digestion and less loss of trace elements due to a closed digestion system. Usually, it is used for small sample analysis.

The main activities of irradiated biomedical material are produced from ^{38}Cl, ^{82}Br, ^{24}Na, and ^{32}P. A complete separation of an element from all other radionuclides is rarely necessary. The removal of main radionuclides and several group separations are very useful for improvement of the determination of most trace elements with adequate accuracy and precision (23,24). Precipitation, ion exchange, and extraction are the most commonly used methods for radiochemical separations. Both ^{38}Cl and ^{82}Br can be removed by anion exchange absorption, precipitation as AgCl and AgBr, and evaporation as HCl and Br_2. A simple and effective separation of ^{24}Na can be carried out by absorption on a hydrated antimony pentoxide (HAP) column (27). Various procedures have also been developed for the group separation (23,24,27). However, the separation of a single element sometimes may be necessary due to their very low level in biomedical materials, such as iodine, vanadium, silicon, uranium, selenium, and mercury (26,28,29). For efficient measurement, the separated sample has to be prepared in small amounts for counting on the detector. This process is normally completed by precipitation to convert the separated elements to a solid form or by evaporation to remove most of the water.

Chemical Recovery

In comprehensive radiochemical separation, the addition of carrier cannot entirely prevent losses of the elements of interest, even in a simple separation procedure. For accurate results, a correction for losses should therefore always be made, which can be carried out by measurement of the chemical recovery of the determined element in the radiochemical separation.

Chemical recovery can be measured by carrier and radiotracer. Before decomposition, carriers are added to the irradiated material in macroamount compared to the element originally present in the sample. When the carriers behave as the probed element in the sample, their chemical recoveries should be the same. Thus the chemical recovery of the carrier element is taken as the recovery of the determined element. The carrier element can be easily measured by classic analytical methods, such as gravimetry, calorimetry, and titration method, especially when a single carrier is added and separated, and the contribution from other elements is negligible. When more than one carrier is added and multielements are separated and determined, a reactivation method could be used for determination of the carrier content in the separated sample. After the measurement, the separated sample containing the activated elements and the carriers is reactivated by irradiation with neutrons, and the amount of carrier is determined by NAA. As the amount of the radionuclide originated from the sample is negligible compared to the added carrier, and the amount of added carrier is known, the chemical recovery can be calculated. Radiotracer is a more direct method for monitoring chemical recovery. In this case, radioisotopes are added to the sample with carrier before the radiochemical separation. These radiotracers are not the same as those produced by neutron activation, and can be measured simultaneously by a gamma detector due to the different energies of gamma rays. Since the radiotracer is another indicator of the same element as the radionuclide produced from the element of interest, the chemical recovery of the radiotracer is the same as the indicator. For example, ^{125}I was used as a radiotracer for monitoring the chemical recovery of iodine in RNAA (29), and ^{57}Co for cobalt.

If the radiochemical separation is carried out in a well-controlled manner, a constant chemical recovery can be assumed, and the recovery correction may be carried out without measurement of chemical recovery for every individual sample. But first the constant chemical recovery has to be determined by processing an unirradiated sample to which is added irradiated carriers or other radiotraces using the same chemical separation procedure. In order to evaluate the precision of the correction, at least 10 determinations should be made.

Analysis of Gamma Spectrum and NAA Calculation

The data acquired by a germanium detector are registered in the MCA, and may be transferred directly or via an analyzer buffer to a computer for processing. At present, MCA can even be made as an interface card, which can be directly inserted to a personal computer. Then the computer can control the gamma detector and the gamma spectra can be acquired and analyzed by computer software. Data acquisition software is chiefly concerned with the handling of the MCA system and its components; the programs will connect the hardware of the spectrometry system with the storage memory for the data. Critical physical parameters such as starting time, duration, and dead-time of the acquisition are recorded together with the spectral data. The acquisition software can also take care of controlling sample changers and automatic sample transfer systems for cyclic NAA. The acquired spectrum can be stored separately or summarized.

Many γ-spectrum analysis programs have been developed to analyze gamma spectra, which can search for γ peaks, calculate their energies, and their peak areas. The special software for NAA can even identify nuclides by the energy of γ peaks and decay time, calculate the radioactivity of a radionuclide, and finally calculate the concentrations of the trace elements in the sample. The energy of a gamma peak can be identified by calibration of the counting system by measurement of some known gamma-ray sources. By comparison with a database of gamma-energy and half-life of radionuclides, the identification of the nuclides can be carried out. The calculation of the peak area is the most critical step in the NAA calculation. Many techniques have been developed to calculate the peak area and to subtract the baseline under the peak, such as Gaussian shape and experimental peak shape fitting. When the peak area and the efficiency of the detector are measured, the activity of the nuclide can be calculated easily by correcting for decay and counting time. If the comparator method is used, the contents of elements of interest can then be calculated using Eq. 7 after the measurement of a comparator element standard. In NAA, overlapping peaks may sometimes occur, when the energy difference between peaks is not large enough, a significant error may result from incorrect separation of peaks and baseline subtraction. However, with proper execution of NAA and correction of results for possible separation losses, NAA is capable of providing unbiased results with known precision for a multitude of trace elements in biomedical materials.

Evaluation and Quality Assurance

NAA has been demonstrated to be reliable and under statistical control due to the absence of unknown sources of variability. This means that it is possible to predict the standard deviation of analytical results so well that the observed and expected variation among replicate data are in agreement. The special qualities of NAA make the method one of the best for the certification of reference materials in the biomedical field. Hardly any certification of trace elements is carried out without considering this method. Its high sensitivity for many elements and the absence of blank values make NAA the preferred method for analysis at an ultratrace level of concentration in many types of biomedical samples. The insensitivity to contamination has proven particularly valuable for the establishment of a normal concentration of a number of elements in human samples. Results by NAA serve as a reference for testing or verifying the reliability of other methods.

The common sources of uncertainty in NAA come from irradiation, counting, spectra acquisition and analysis, blank and interference of nuclear reactions and gamma rays. The uncertainty of irradiation is contributed from the variation of neutron flux during irradiation and inhomogeneous distribution of neutron flux in the irradiation container. The instability of neutron flux may significantly influence the analytical results of elements determined by measuring short-lived radioisotopes, because the sample and standard are not irradiated simultaneously. This problem is normally overcome by on-line monitoring of the neutron flux, and most research reactors, especially a small reactor such as MNSR, installed such a system. In most cases, the uncertainty from this source is low ($<0.5\%$). The neutron flux gradient is sometimes very significant in a big irradiation container, so that a flux monitor foil of Fe, Co, or Au alloy may be used to monitor the different positions to make a neutron flux correction. Uncertainty in counting mainly results from the variation of counting geometry. The uncertainty due to counting geometry can be controlled reliably by well efficiency calibration of the detector for various source shapes; software is available for appropriate calculation. The matrix effects can suppress gamma rays, especially low energy gamma rays, by self-absorption; this effect can be calibrated by measurement of a standard with a similar matrix with sample. Many natural matrix certified reference materials (CRM) have been prepared for this purpose and are commercial available (30).

Uncertainty in the spectra acquisition may come from dead time and pile-up losses of signals. A quick varying and high counting rate often occur in the NAA of biomedical materials due to high concentrations of Cl and Na which may cause a high and varying dead time during the counting. Dead time is normally measured by the gamma spectra system, and live time is used for activity calculation. But when the dead time is quickly changed during counting, the activity may be underestimated by this method. This problem can be solved by using a loss-free counting method, which estimates the number of counts lost during a dead time interval and adds this number to the channel of the just processed pulse, thus presenting a loss-corrected spectrum. The pile-up losses are corrected by electronic or computational mean built in the counting system. Uncertainty in the spectra analysis results mainly from the evaluation of peak area and subtraction of baseline, especially for the analysis of multipeaks. A large attempt has been made to develop effective software to improve the analysis of the gamma spectra.

The blank problem can usually be ignored in NAA. But a blank correction will be necessary, when preseparation and sample preparation are used before irradiation. A given radionuclide can often be produced in more than one way. If the indicator nuclide used in NAA is produced from an element other than the element determined, then nuclear reaction interference occurs, such interference is mainly produced by (n,p) or (n, α) reactions with fast neutrons and elements with an atomic number 1 or 2 above the element to be determined. A typical example of such interference is the formation of ^{28}Al from Al, P, and Si by reactions: ^{27}Al(n, γ)^{28}Al, ^{28}Si(n, p)^{28}Al, and ^{31}P(n, α)^{28}Al. In biomedical materials, the concentrations of Al and Si are very low, but P is high in many kinds of samples. The contribution of P to the activity of ^{28}Al may be very significant, especially when the fraction of fast neutron in the irradiation position is high, it will seriously interfere with the determination of aluminum. This interference can be significantly reduced by using a thermal neutron irradiation channel, where the ratio of thermal/fast neutron flux is very high; in many research reactors, such a thermal neutron channel is available. Since there are no suitable thermal neutron activation products of P and Si, they are determined by using these two fast neutron reactions in NAA, so the interference from Al to P is also very significant due to a much higher cross-section of thermal neutron activation than fast neutrons. The correction can be carried out by the irradiation of the sample with and without thermal neutron shielding. The activity of ^{28}Al in a sample irradiated under thermal neutron shielding (such as in a Cd or B container) mainly contributes from fast neutron reaction, while under conditions without shielding from both thermal and fast neutron reactions.

Double and triple neutron activation reactions can also result in an interference, for example, the interference from ^{197}Au(2n, γ)^{199}Au to the determination of Pt by ^{199}Pt(n,γ)^{199}Pt(β^{-})^{199}Au reaction, and from ^{127}I(3n, γ)^{130}I to the determination of ^{129}I by ^{129}(n, γ)^{130}I reaction (29). But, interference from triple reactions is normally negligible due to their very low probability.

A special type of nuclear reaction interference is caused by the presence of uranium, which yields a large number of radionuclides as a result of fission. The greatest correction is required in the determination of molybdenum and some rare earth elements (REs), but the concentrations of uranium in biomedical materials are usually too low to present any problem.

Gamma-ray spectral interference means that two radionuclides emit gamma rays with the same or nearly the same energy. For example, ^{51}Ti and ^{51}Cr used for NAA of Ti and Cr emitting a single 320.08-keV gamma ray, ^{75}Se and ^{75}Ge used for Se and Ge emitting 264.66 keV gamma ray. For these nuclides, a separation via decay is possible practically due to their different half-lives. However, the interference for the determination of Hg via the 279.19 keV line of ^{203}Hg (46.6 days), which overlapped by the 279.54 keV line of ^{75}Se (119.77 days), cannot be eliminated by decay because their half-lives are not significantly different. This problem can be resolved by measurement of ^{75}Se by other gamma lines and correcting the contribution of ^{75}Se to the peak of 279.54 keV.

As an effective method for analytical quality control, certified reference materials (CRMs) with a matrix similar to the samples, which have appropriate combinations of elements to be determined and with concentration bracketing the range of interest, are normally analyzed to evaluate the analytical accuracy and precision. A large number of CRMs have been prepared by many countries and international organizations (30), which are not only for analytical quality control of NAA, but actually all other analytical methods.

Uncertainty from the sampling procedure and inhomogeneity of the sample should also be considered as a probable contribution to the total uncertainty of the analytical data, although it is not directly related to the NAA method itself. Especially at very low concentration, the trace elements distribution in samples is normally inhomogeneous and a selection of a representative sample is very important for analytical quality. In recent years, the NAA of large samples has been given much attention (31). In this case, a whole sample is analyzed and the uncertainty resulting from the inhomogeneity of elements in the sample can be overcome.

Applications

As an analytical technique for determination of trace elements, the NAA has been widely used in various aspects of biomedical studies and analysis related to trace elements, such as an investigation of the normal trace element level in the human body, trace element related endemic disorder and inheritable diseases, environmental exposure of toxic and nonessential trace elements and their effect on human health, and investigation of trace element nutrition status.

The distribution and normal level of trace elements in blood, urine, and human tissue are basic data for the studies on biomedical trace elements. Some activities have been organized by the International Atomic Energy Agency (IAEA) for establishing the normal level of trace elements in "reference man" with different dietary habits and in different geographic areas. This information was completed by analyzing various normal human tissues collected from different countries in which most of the data was supplied by NAA (32,33). In addition, many NAA were also carried out to analyze a large number of blood, urine, hair, and nail samples for normal trace element level determinations. In combination with the analysis of diets and environmental samples, some trace element related endemic disorders have been investigated, such as Keshan disease found in China, which is related to the deficient intake of selenium, and normally also associated with the deficiency of iodine. Some disorders caused by excessive intake of arsenic from ground water and mercury from contaminated fish and foods have also been investigated by NAA in China and many other countries. Trace element nutrition status by analysis of the trace element level in blood, hair, and diets has also been investigated by NAA in many countries. The results were used as a basis for improved recommendations of safe and adequate, daily average intake of these elements.

Some inheritable diseases have been attributed to the metabolism problem of trace elements, such as Menkes' syndrome, Wilson's disease, and Acrodermatitis enteropathica. Determination of specific trace elements in a small amount of living tissue can be used to diagnose disease and to evaluate the treatment. The ability to analyze very small samples by NAA creates a possibility of following temporal changes in trace element concentrations in a living subject. Menkes' disease was found to be a recessive X-linked disturbance of copper metabolism, resulting in accumulation of copper in several extra-hepatic tissues, including the placenta, by NAA of various human tissues of patients and normal persons (2). The method is being used to diagnose Menkes' disease and identification of female carriers by NAA of the placental tissues (2,34). Wilson's disease was also confirmed to be a copper metabolic defective disease; determination of copper in liver biopsy by NAA was used as a control and verification method for estimation of the adequate treatment of this disease.

A high environmental exposure may cause the increase of some nonessential trace elements in the human body, which may create a potential effect for human health. In recent years, rare earth elements found wide application in industry and agriculture, rare earth trace element fertilizer has widely been used in China for increasing the yield of various agricultural products. It also increases the exposure of humans to these trace elements. Neutron activation analysis was used to investigate the uptake of these elements in the human body via the food chain, their distribution in human tissues, and their combination with different components of tissue, especially brain tissue (7,13).

Since hair and nails are easily obtained compared to blood and tissue samples, they serve as a useful indicator of trace element levels in the body. Neutron activation analysis is very suitable for the analysis of hair and nails, because no decomposition of sample is necessary, therefore they are often used for the analysis of these materials for investigation of the nutrition status of trace elements, the environmental exposure level of toxic elements, and the diagnosis of various diseases related to trace elements (9).

A series international conference entitled Nuclear Analytical Methods in the Life Sciences was held since 1967, and the three most recent conferences in this series were held in 1993, 1998, and 2002. The main achievements in this field can be found in the Proceedings of these conferences (35,36) and other relevant conferences (37,38).

BIBLIOGRAPHY

1. Prasad AS, Oberleas D. Trace elements in human health and disease. New York: Academic Press; 1976.
2. Heydorn K. Neutron activation analysis for clinical trace element research. Boca Raton: CRC Press; 1984.
3. Chettle RD, Fremlin JH. Techniques of in vivo neutron activation analysis. Phys Med Biol 1984;29:1011–1043.
4. Heydorn K. Neutron activation analysis. In: Webster JG, editor. Encyclopedia of Medical Devices and Instrumentation. New York: John Wiley & Sons; 1988.
5. deCorte F, Simnits AS. Recommendation data for use in the $k(0)$ standardization of neutron activation analysis. Atomic data and nuclear data tables 2003;85(1):47–67.
6. Guinn VP, Garzanov E, Cortes E. Further studies in advance prediction of gamma ray spectra and detection limits in instrumental neutron activation analysis. J Radioanal Chem 1978;43(2):599–609.
7. Chai CF, Mao XY, Wang YQ, Sun JX, Qian QF, Hou XL, Zhang PQ, Chen CY, Feng WY, Ding WJ, Li XL, Li CS, Dai XX. Molecular activation analysis for chemical species studies. Fresenius J Anal Chem 1999;363(5–6):477–480.
8. Parr RM. Technical consideration for sampling and sample preparation of biomedical samples for trace element analysis. J Res Nat Bur Stand 1986;91(2):51–57.
9. Chatt A, Katz SA. Hair Analysis, application in the biomedical and environmental sciences. New York: VCH; 1988.
10. Patching SG, Gardiner PHE. Recent developments in selenium metabolism and chemical speciation: A review. J Trace Element Med Biol 1999;13(4):193–214.
11. Shoop DM, Blotcky AJ, Rack EP. Distribution of aluminum species in normal urine as determined by chemical neutron activation analysis. J Radioanal Nucl Chem 1998;236(1–2): 103–106.
12. Hou XL, Chen CY, Ding WJ, Chai CF. Study on chemical species of iodine in human liver. Biological Trace Element Res 1999;69(1):69–76.
13. Chai ZF, Zhang ZY, Feng WY, Chen CY, Xu DD, Hou XL. Study of chemical speciation of trace elements by molecular activation analysis and other nuclear techniques. J Anal Atomic Spectrom 2004;19(1):26–33.
14. Chen CY, Zhang PQ, Hou XL, Chai ZF. Subcellular distribution of selenium and Se-containing proteins in human liver. Biochim Biophy Acta 1999;1427(2):205–215.
15. Feng WY, Li B, Liu J, Chai CF, Zhang PQ, Gao YX, Zhao JJ. Study of chromium-containing proteins in subcellular fractions of rat liver by enriched stable isotopic tracer technique and gel filtration chromatography. Anal Bioanal Chem 2003;375(5):363–368.
16. Heydorn K, Skanborg PZ, Gwozdz R, Schmidt JO, Wacks ME. Determination of lithium by instrumental neutron activation analysis. J Radioanal Chem 1977;37:155–168.
17. Hou XL. Cyclic activation analysis. In: Meyers RA, editor. Encyclopedia of Analytical Chemistry. Applications, theory and instrumentation. Chichester: John Wiley & Sons; 2000.
18. Reijonen J, Leung KN, Firestone RB, English JA, Perry DL, Smith A, Gicquel F, Sun M, Koivunoro H, Lou TP, Bandong B, Garabedian G, Revay Z, Szentmiklosi L, Molnar G. First PGAA and NAA experimental results from a compact high intensity D-D neutron generator. Nucl Instr Meth 2004;A522(3): 598–602.
19. Chichester DL, Simpson JD. Advanced compact accelerator neutron generator technology for FNAA/PGNAA field work, Abstract of 11th International Conference on Modern Trends in Activation Analysis. Guildford, UK; June 2004.
20. Hou XL, Wang K, Chai CF. Epithermal neutron activation analysis and its application in the miniature neutron source reactor. J Radioanal Nucl Chem 1996;210(1):137–148.
21. Khamia I, Al-Somel N, Sarheel A. Neutron flux micro-distribution and self-shielding measurement in the Syrian Miniature Neutron Source Reactor. J Radioanal Nucl Chem 2004;260(2): 413–416.
22. Hatsukawa Y, Toh Y, Oshima M, Hayakawa T, Shinohara N, Kushita K, Ueno T, Toyoda K. New technique for the determination of trace elements using multiparameter coincidence spectrometry. J Radioanal Nucl Chem 2003;255(1):111–113.
23. Heydorn K. Radiochemical neutron activation analysis. In: Meyers RA, editor. Encyclopedia of Analytical Chemistry. Applications, theory and instrumentation. Chichester: John Wiley & Sons; 2000.
24. Alfassi ZB. Determination of trace elements, Balaban Publisher; 1994.
25. Krausov I, Kucera J. Fast decomposition of biological and environmental samples for radiochemical neutron activation analysis, Abstract of 11th International Conference on Modern Trends in Activation Analysis. Guildford, UK; June 2004.
26. Dermelj M, Byrne AR. Simultaneous radiochemical neutron activation analysis of iodine, uranium and mercury in biological and environmental samples. J Radioanal Nucl Chem 1997;216(1):13–18.
27. Girardi F, Sabbioni E. Selective removal of radio-sodium from neutron activated materials by retention on hydrated antimony pentoxide. J Radioanal Chem 1968;1(2):169–178.
28. Heydorn K, Damsgaard E. Simultaneous determination of arsenic, manganese and selenium in biological materials by neutron activation analysis. Talanta 1980;20:1–11.
29. Hou XL, Dahlgaard H, Rietz B, Jacobsen U, Nielsen SP, Aarkrog A. Determination of iodine-129 in seawater and some environmental materials by neutron activation analysis. Analyst 1999;124:1109–1114.
30. Analytical Quality Control Service, International Atomic Energy Agency. www.iaea.org/programmes/aqcs/database/database_search_start.htm, 2004; February 25.
31. Bode P, Overwater RMW, DeGoeij JJM. Large-sample neutron activation analysis: Present status and prospects. J Radioanal Nucl Chem 1997;216(1):5–11.
32. Iyengar G, Kawamura H, Dang HS, Parr RM, Wang JW, Natera ES. Contents of cesium, iodine, strontium, thorium, and uranium in selected human organs of adult Asian population. Health Phys 2004;87(2):151–159.
33. Iyengar GV. Reevaluation of the trace element content in reference man. Rad Phys Chem 1998;51(4–6):545–560.
34. Tonnesen T, Horn N, Sondergaard F, Mikkelsen M, Boue J, Damsgaard E, Heydorn K. Metallothionein expression in placental tissue in Menkes' disease—an immunohistochemical study. APMIS 1995;103(7–8):568–573.

35. Procedings of 7th international conference on nuclear analytical methods in the life science; June 16–22, 2002, Antalya Turkey: J Radioanal Nucl Chem 2004;259:1–539.
36. Proceedings of the International Conference on Nuclear Analytical Methods in the Life Sciences – Sep. 1998, Beijing, China: Bio. Trace Elem; 1999;71(2).
37. Proceedings of the 10th international conference on modern trends in activation analysis, 19–23, April, Maryland: 1999. J Radioanal Nucl Chem 2000;244:1–704 and 245:1–228.
38. Proceedings of the 11th international conference on modern trends in activation analysis; 20–25 June, Guildford, UK: 2004. J Radioanal Nucl Chem 2005; in progress.

Reading List

Afassi ZB. Activation analysis. Boca Raton: CRC; 1990.
Ehmann WD, Vance DE. Radiochemistry and Nuclear Methods of Analysis. New York: John Wiley & Sons; 1991.
Parry SJ. Activation spectrometry in chemical analysis. New York: John Wiley & Sons; 1991.

See also BORON NEUTRON CAPTURE THERAPY; RADIONUCLIDE PRODUCTION AND RADIOACTIVE DECAY; TRACER KINETICS.

NEUTRON BEAM THERAPY

RICHARD L. MAUGHAN
Hospital of the University of
Pennsylvania

INTRODUCTION: THE ORIGINS OF FAST NEUTRON THERAPY

The year 1932 was a remarkable one at the Cavendish Laboratory of the University of Cambridge. It represented a pinnacle of achievement for Lord Ernest Rutherford and his collaborators in the nuclear physics laboratories. The results of three experiments performed in this single year resulted in Nobel Prizes in physics for four of the university faculty. Two of these experiments were of great significance for the production of neutrons. Of course, the most significant was the discovery of the neutron itself by James Chadwick in February, 1932 (1). Chadwick needed a source of neutrons to perform this experiment. At that time, the only sources of energetic heavy particles were naturally occurring isotopes that emitted alpha particles. Chadwick's neutron source was comprised of a polonium source (^{210}Po), which emits a 5.3 MeV alpha particle and a block of beryllium housed in a vacuum chamber. The nuclear reaction

$$\alpha + {}^{9}\mathrm{Be} \longrightarrow {}^{12}\mathrm{C} + \mathrm{n}$$

produces a flux of neutrons with energies of several million electronvolts (MeV). To detect these neutrons, Chadwick used an ionization chamber with a thin entrance window. Pulses could be observed on a oscillograph as the source was brought closer to the ionization chamber. The chamber was air filled and the increasing count rate was interpreted as being due to recoil nitrogen nuclei in the chamber. When a sheet of paraffin wax was placed in between the source and the ionization chamber a further increase in the source count rate was observed and interpreted as resulting from recoil protons from n–p elastic scattering. Further measurements confirmed that the particles producing the recoil protons were neutral particles of the same mass as the proton. With this simple and elegant apparatus Chadwick discovered the neutral nuclear particle, the existence of which Rutherford had postulated in 1919.

Sources of this type give very low neutron fluxes and are not suitable for neutron radiation therapy or radiobiology. It was another of the Nobel Prize winning discoveries of 1932 at Cavendish Laboratory that offered the means for producing intense sources of neutrons. This was the discovery of artificial transmutation of nuclei by John Cockroft and Ernest Walton in April 1932. In this experiment, Cockroft and Walton used their specially designed high-voltage apparatus (now commonly known as a Cockroft–Walton accelerator) to produce protons of energy 700 keV. When these protons interacted with a lithium target, the resulting mass eight compound nuclei disintegrated into two alpha particles with the release of ~17 MeV of energy

$$\mathrm{p} + {}^{7}\mathrm{Li} \longrightarrow {}^{4}\mathrm{He} + {}^{4}\mathrm{He}$$

The alpha particles were detected as scintillations on a zinc sulfide screen and, in order to observe these events, Cockroft or Walton had to sit in a small darkened enclosure at the end of the accelerator tube (2).

Although this first artificial splitting of the atom did not involve the production of neutrons, it was not long before another Rutherford's research team, Mark Oliphant, was to apply the new technology to an experiment to demonstrate the fusion of deuterium nuclei. In this reaction, the products are helium 3 and a neutron

$$^{2}\mathrm{H} + {}^{2}\mathrm{H} \longrightarrow {}^{3}\mathrm{He} + {}^{1}\mathrm{n}$$

The reaction has a positive Q value of 3.27 MeV and a deuteron beam of 400-keV energy produces neutrons with energies up to ~3.5 MeV.

This experiment was performed in 1934 and almost immediately L.H. Gray, yet another of Rutherford's ex-students, working as one of the first British medical physicists at Mount Vernon Hospital in Northwood, England, realized the potential of this reaction as a source of neutrons for radiobiology research. The ability of neutrons to produce ionizing radiation in the form of heavy recoil particles had been realized soon after their discovery and that these particles might be able to produce biological damage similar to that produced by the recoil electrons associated with X ray beams and radium sources had also been recognized. Neutrons are heavy particles and can transfer relatively large amounts of energy to their secondary recoil particles in comparison to the recoil electrons produced when X rays interact with matter. The neutrons are, therefore, capable of high linear energy transfer (LET) to the recoil particles associated with them and for this reason are known as high LET particles. Gray and others observed that the very great difference between the distribution of ions along the track of a recoiling nucleus and a fast electron made it probable that the biological action of neutrons would show interesting differences from that of the other radiations. It was hoped that a study of these differences would throw light on the mode of action of radiations on biological material. There was also the possibility that eventually neutrons might prove a more potent means of treating cancer (3).

Gray obtained funding to build a 400 keV Cockroft–Walton accelerator in which he used the deuteron–deuteron reaction to produce a neutron beam for radiobiology research (3). The capital cost of the unit was $2400 with maintenance costs of $320/annum for 1937 and 1938. The unit was housed in a wooden shed.

Concurrently, progress was being made in accelerator technology in the United States by another renowned physicist, Ernest O. Lawrence. Lawrence's invention of the cyclotron earned him the 1939 Nobel Prize for physics and provided a means for producing ion beams with energies of several tens of million electronvolts, far in excess of the energies obtained by Cockroft and Walton with their accelerator. Ernest Lawrence's brother, John, was a physician and the two brothers soon realized the potential medical benefits of applying neutron beams produced by accelerated protons or deuterons. In 1938, funding was obtained from the National Cancer Institute (NCI) for the construction of a 60 in. medical cyclotron; this was the first NCI research grant. This cyclotron was used to produce a neutron therapy beam using a 16 MeV deuteron beam incident on a thick beryllium target. With the use of this beam, Dr. Robert Stone performed the first clinical trials of fast neutrons.

Hence, by 1938 accelerated beams of particles were being used to produce high intensity neutron beams for radiobiology and/or external beam radiation therapy research in both the United States and the United Kingdom.

RADIOBIOLOGICAL RATIONALE FOR FAST NEUTRON THERAPY

The first clinical trials of neutron therapy were performed at the University of California in Berkeley, and were interrupted by the Second World War, when the Berkeley cyclotron was utilized for nuclear physics research associated with the Manhattan Project. Patients with advanced inoperable head and neck tumors were treated with neutron therapy using the same number of treatment sessions, or fractions, as was normally used for conventional X-ray therapy. Although remarkable tumor regression was seen in a few cases, this was at the cost of excessive damage to the surrounding irradiated normal tissues. In 1947, Stone concluded that " neutron therapy as administered by us has resulted in bad late sequelae in proportion to the few good results that it should not be continued", (4). The trials had been undertaken to test the hypothesis that neutron radiation may be superior to X-ray radiation in curing human cancers. However, at the time little was known about the radiobiological effects of neutron irradiation in comparison to conventional X-ray irradiation. It was not until later, when some basic radiobiological research on the effects of neutron irradiation on mammalian cells had been completed, that the reasons for the failure of this original clinical trial could be understood. Further neutron radiobiology research enabled a firm rationale for neutron therapy to be established.

When mammalian cells are irradiated in vitro by ionizing radiation (X rays or neutrons) cells are killed in proportion to the radiation dose delivered. A plot of the cell kill as a function of the radiation dose, or survival curve, is markedly different in shape depending on whether the cells are irradiated by X rays or fast neutrons. This finding is illustrated in Fig. 1, taken from the work of McNally et al. (5). On a log-linear plot at high doses the response appears linear (i.e., exponential), but at low doses there is a "shoulder" on the survival curve. This shoulder is more pronounced for X ray irradiations than for neutron irradiations. Another point to notice is that for a given radiation dose the cell-kill by fast neutrons is greater than for X rays. The relative biological effectiveness (RBE) of neutrons relative to X rays is defined as the ratio the dose of X rays required to produce a given level of cell-kill compared to the dose of fast neutrons required to give the same cell kill. Because of the different shapes and sizes of the shoulders on the neutron and X ray cell survival curves it can be seen from Fig. 1 that the RBE at a surviving

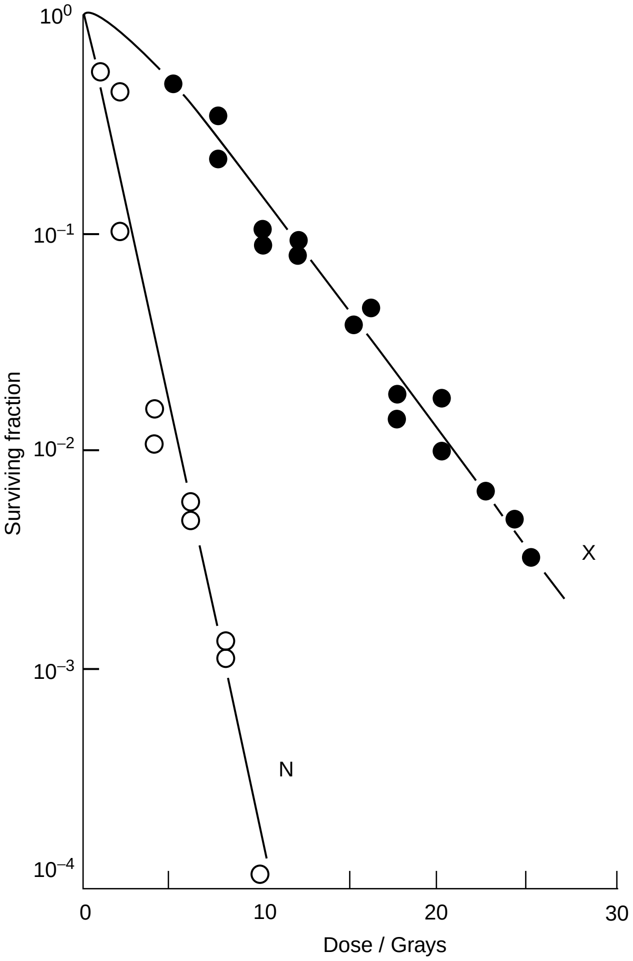


Figure 1. Survival curve for WHFIB mouse tumor cells irradiated with fast neutrons and X rays. Reproduced with permission from McNally, et al. *Rad. Res.* 1982;89:234.

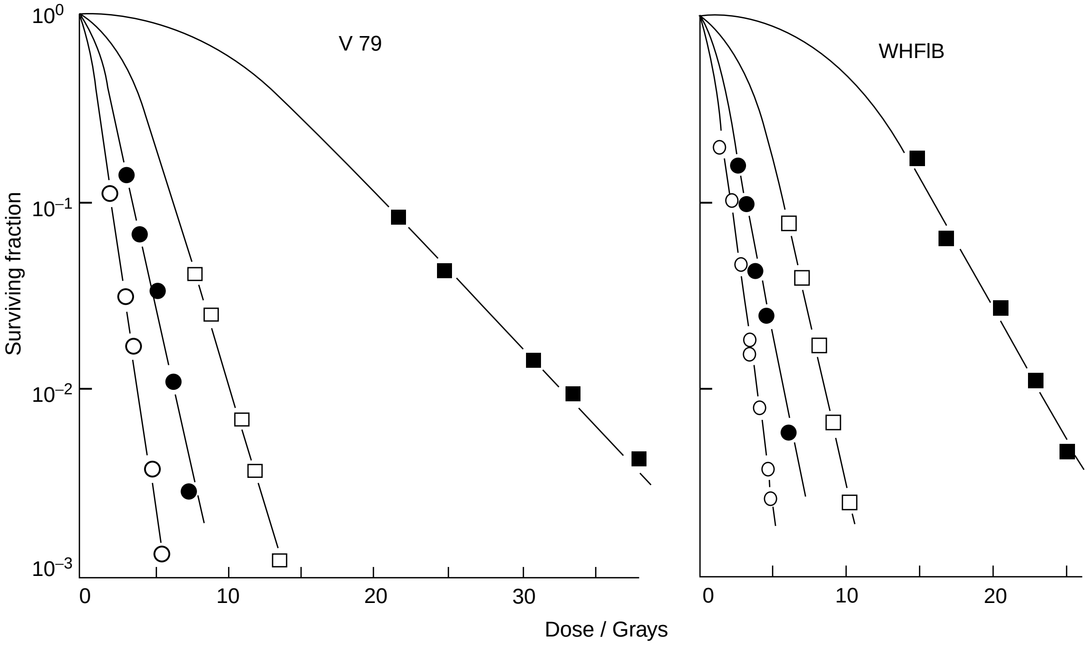


Figure 2. Survival curves for V79 and WHFIB cells, irradiated in air and hypoxia. Squares, X rays; circles neutrons; open symbols, air; solid symbols hypoxia. Reproduced with permission from McNally, et al. *Rad. Res.* 1982;89:232.

fraction of 10^{-2} (i.e., a relatively high dose) is ~3.8, while for a surviving fraction of 3×10^{-1} (i.e., a lower dose) the RBE increases to ~5. John Lawrence measured the RBE in nonhuman biological systems, prior to commencing the Berkeley fast neutron therapy trials with Robert Stone, using large single doses of radiation to produce an observable biological effect. Based on an RBE measured at high single doses, they calculated the required total neutron dose from a knowledge of the total X ray doses delivered to cancer patients at that time. Unfortunately, radiation therapy doses are delivered as a large number of small doses and at these smaller doses the RBE is much larger, hence, the fast neutron dose delivered to patients in the original fast neutron trial was overestimated by a considerable margin. This of course resulted in good tumor control compared to conventional X rays, but also produced an unacceptable level of normal tissue damage. It was not until 1971 that Sheline et al. (6) explained this phenomenon.

In the meantime, L.H. Gray working at Hammersmith Hospital had described a logical rationale for fast neutron therapy. In 1954, in a landmark paper in radiation therapy, Thomlinson and Gray described how, in a poorly vasculated tumor, areas of reduced oxygenation, or hypoxia, can exist as the distance from blood vessels increases (7). If oxygenation drops low enough, the cells become necrotic and die. Gray showed, using diffusion kinetics, that an oxygen concentration gradient can exist within a tumor and that in certain areas of the tumor there may be severely hypoxic, yet viable cells. How this phenomenon can be exploited to advantage in fast neutron therapy is illustrated in Fig. 2. In this figure, cell survival curves for cells irradiated by both X rays and fast neutrons in an oxygen (oxic) and a nitrogen (anoxic) environment are plotted. The anoxic cells are more resistant to radiation than the oxic cells and as tumors are poorly oxygenated and contain anoxic cells this could well result in an inability to deliver sufficient radiation to kill all the tumor cells, leading to tumor recurrence. Normal tissues are well oxygenated and, therefore, are more easily damaged. Hence, it is possible that the doses to the normal tissues surrounding the tumor may reach their acceptable tolerance level before sufficient dose has been delivered to the hypoxic tumor cells to eradicate them. Figure 2 shows that the differential cell killing for oxic and anoxic cells is much greater for X rays than for fast neutrons. Thus for a given level of normal tissue, cell kill or damage neutrons should be more effective at killing hypoxic cells in the tumor than conventional X-ray therapy. It was this hypothesis that lead to the restarting of fast neutron clinical trials at Hammersmith Hospital in London in 1970. The encouraging results obtained at Hammersmith Hospital rekindled interest in fast neutron therapy and by 1980 there were close to 20 centers treating patients.

The clinical results showed that fast neutron therapy appeared to be particularly effective in the treatment of slow growing tumors such as adenocarcinoma of the prostate and bladder. In the meantime, radiobiology research had revealed another important difference between X rays and fast neutrons related to the variation of the radiation sensitivity of mammalian cells at different phases of the mammalian cell cycle. Mammalian cells exhibit well-defined stages in their cycle first described by Howard and Pelc (8). Figure 3 shows a schematic representation of the phases of the cell cycle. Cells spend most of their time in a quiescent phase known as G1 (gap 1), after the G1 phase they move into a phase during which duplicate DNA is synthesized, the S phase. Following DNA synthesis there

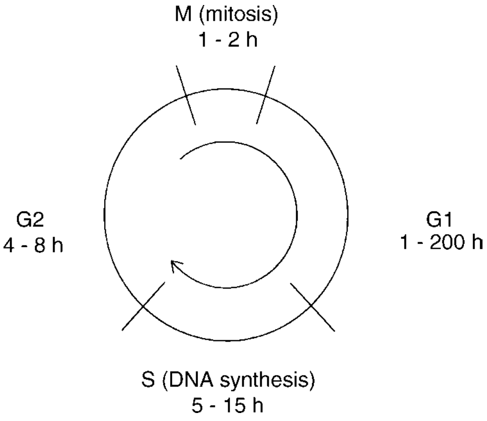


Figure 3. Schematic representation of the phases of the cell cycle.

is another gap phase, G2, which precedes cell division or mitosis (the M phase). The S, G2, and M phase have a similar duration for all mammalians cells, being typically 8, 4, and 1 h, respectively. It is the G1 phase that varies for rapidly and slow growing cells, being as short as 1 h for fast growing cells and as long as 200 h for slow growing cells. The variation in the radiosensitivity of cells irradiated with conventional X rays at different phases of the cell cycle can be considerable (9). The late G1 phase is a relatively radioresistant phase and in cells with a long G1 there is also a period of radioresistance during early G1. As a result, slow growing cells, which spend a larger proportion of the cell cycle in G1 phase than rapidly proliferating cells, would be expected to be more resistant to radiation than the fast growing cells. Withers et al. (10) have shown that the variation of cell sensitivity during the cell cycle is less for fast neutron irradiation than for conventional X-ray radiation. This is a potential advantage for fast neutron therapy, since is means that neutrons are relatively more effective in killing cells in the radioresistant phases of the cell cycle. This would explain the observed efficacy of neutrons in the treatment of slow growing tumors. Another observation, which confirms the efficacy of fast neutrons in treating slow growing tumors, was made by Battermann et al. (11), who showed that the relative biological effectiveness of fast neutrons increases with tumor doubling time.

Thus by the early 1980s it had been established that fast neutron therapy could be justified on radiobiological rationale related to both tumor hypoxia and volume doubling time.

REVIEW OF CLINICAL RESULTS

After the original unsuccessful clinical trials at Berkeley, fast neutron therapy was restarted in 1970 by Dr. Mary Catterall, at Hammersmith Hospital in London. Although Dr. Catterall's studies were not controlled clinical trials they did demonstrate the potential efficacy of fast neutron therapy in the treatment of advanced or recurrent tumors. The success of these studies lead to the establishment of many fast neutron therapy centers around the world and to the instigation of a number of clinical studies of those tumors thought to be the most promising candidates for neutron therapy. These tumors included advanced inoperable tumors of the salivary glands; squamous carcinoma of the head and neck; adenocarcinoma of the prostate, rectum, bladder, and soft-tissue; oesteo- and chondro- sarcomas. The results of these studies demonstrated that fast neutron therapy was particularly promising for the treatment of salivary gland tumors, prostate tumors, and bone tumors. The results for squamous carcinoma of the head and neck, however, were ambiguous. Prompted by these results, the NCI in the United States decided to fund four purpose-built hospital-based neutron therapy facilities to contribute to a definitive series of controlled clinical trials performed at these and other approved centers.

The NCI trials were undertaken in two phases starting with dose searching studies, designated as phase II trials, followed by efficacy studies (phase III trials). The patient accrual into these trials is shown in Table 1. The phase II trials established the general treatment protocol for the phase III studies, that is, a dose of 20.4 Gy would be delivered in 12 fractions over 4 weeks and after 13.6 Gy abdominal and pelvic treatment fields would be reduced in size.

Of the seven phase III trials undertaken, only three were successfully completed; the head and neck, lung, and prostate trials. The salivary gland trial was terminated early; the investigators thought it unethical to continue with randomization because of the excellent results obtained with neutron therapy (12). The remaining three trials were closed because of poor patient accrual. The only other trial to yield a positive result was the prostate trial (13). This trial also highlighted the importance of shaping the beam to conform to the tumor outline. The three centers contributing patients to this trial had facilities with different beam collimation systems of varying degrees of sophistication. The occurrence of normal tissue complications in the bladder and rectum was closely related to the

Table 1. Patient Accrual for NCI–NTCWG Clinical Trials Using the New Generation of Neutron Therapy Facilities in the United States (1984–1991)

Site	No of Patients
Dose searching studies	
Head and neck	59
Thorax	169
Abdomen	78
Pelvis	102
Extremity	92
Subtotal	*500*
Phase III studies	
Salivary gland	9
Head and neck (squamous)	178
Lung	232
Prostate	178
Cervix	28
Rectum	2
Resistant histology	47
Subtotal	*674*
Total accrual	*1174*

sophistication of the beam collimator. The University of Washington, which utilized a sophisticated multileaf collimator for beam shaping, had the least number of treatment related complications, while the M.D. Anderson Hospital in Houston, which used an insert-based collimation with limited shaping capability, had the greatest number of normal tissue complications (13).

An earlier NCI funded Phase III trial for advanced adenocarcinoma of the prostate had shown that a mixed-beam treatment regimen of 60% photons combined with 40% neutrons was superior to conventional photon only therapy (14). With this in mind, Forman et al. (15) undertook a series of in-house neutron therapy trials at Wayne State University using mixed-beam therapy (50% photons and 50% neutrons) for early and some late stage patients. This trial also utilized a sophisticated beam shaping system, a multirod collimator. These trials demonstrated that mixed-beam therapy is as effective as any other state-of-the-art radiation therapy techniques (Intensity Modulated Radiation Therapy or IMRT, proton therapy, brachytherapy) in the treatment of early stage prostate cancer, and that it is probably superior for the treatment of late stage disease (15).

Full reviews of neutron therapy in the treatment of cancer can be found in the work of Wambersie et al. (16) and in a recent International Atomic Energy Agency (IAEA) publication (17). The IAEA report concludes that fast neutrons are superior to photons in the treatment of salivary gland tumors (locally extended, well differentiated), paranasal sinuses (adenocarcinoma, adenoid cystic carcinomas, and possibly other histologies), some tumors of the head and neck (locally extended, metastatic adenopathies), soft tissue sarcomas, osteosarcomas, chondrosarcomas (especially slowly growing/well differentiated), prostatic adenocarcinomas (locally extended), and melanomas (inoperable and recurrent). The IAEA report also identifies tumors for which conflicting or incomplete results have been reported and for which additional studies are necessary, these include inoperable pancreatic tumors, bladder carcinoma, recurrent and inoperable adenocarcinoma of the rectum, tumors of the esophagus, locally advanced tumors of the uterine cervix, and brain tumors for treatment with a neutron boost irradiation before or after X-ray therapy.

NEUTRON SOURCES FOR RADIATION THERAPY

Characteristics of Neutron Sources for Medical Use

In order to be useful for medical applications, fast neutron therapy facilities must meet a set of minimum requirements. As described above, in the early years of the 1970s many neutron therapy centers were set up to carry out clinical trials. Most of these centers made use of existing physics research accelerators (cyclotrons or proton linacs), which were adapted for clinical use. The trials produced results, which were often ambiguous and much of this ambiguity was ascribed to the inadequacies of the equipment. When the NCI in the United States decided to fund a number of therapy facilities, the basic specifications for these accelerators were defined to ensure that the equipment would be adequate to allow meaningful clinical trials to be completed (18); the key requirements are summarized in Table 2. All the facilities funded by the NCI were hospital-based and had rotational isocentric capability, that is, the neutron beam could be rotated around the patient in the treatment position.

Table 2. Summary of Some of the Key Requirements for a Hospital-Based Neutron Therapy Facility as Defined by the NCI[a]

- Neutron beams having build-up and depth-dose characteristics equivalent to 4 MV X rays, penumbra not less sharp than ^{60}Co gamma-ray beams, and dose rates not less than 20 cGy/min.
- Preferably an isocentric beam delivery system and as a minimum one horizontal and/or vertical beam delivery port.
- Access to the neutron beam therapy facility for a minimum of 8 h/day, 4 days/week, 45 weeks/year for patient treatment and 16 h/week additional for physics and biology.
- Methodology for providing a variety of square, rectangular, and irregularly shaped fields ranging in size from 5×5 cm^2 to 20×20 cm^2.
- Capability to shape treatment fields using a variety of wedges and blocks so that any treatment field normally used for conventional X ray therapy can be reproduced on the neutron beam.

[a] See Ref. (18).

Several of these requirements depend critically on the neutron source. In particular, the attenuation (depth-dose) characteristics of the beam, the neutron dose rate (treatment time), and the reliability of the device may be dependent on the neutron source and the means of neutron production.

In order to fully characterize a neutron source it is necessary to have a detailed knowledge of a number of physical characteristics of the neutron beam. Most importantly, the physical data must be sufficient to allow for the accurate calculation of the physical dose delivered both to the tumor site and the surrounding normal tissues in the patient. In addition, physical data may also be necessary to adequately interpret the biological effects that are observed with high linear energy transfer (LET) beams. Table 3 lists the type of data, which are necessary to fully characterize the neutron source. These data are also necessary to fully assess the relative merits, usefulness, and

Table 3. Physical Data Necessary to Characterize Neutron Sources for Radiation Therapy

- Total neutron yield or dose or kerma rate.
- Neutron spectrum (i.e., Neutron yield as a function of neutron energy).
- Neutron yield as a function of angle relative to the forward direction of the beam.
- Neutron dose as a function of depth in a water phantom (i.e, depth-dose data).
- Relative neutron dose as a function of the lateral postion (i.e, dose profiles).
- The microdosimetric properties of the beam (i.e., the LET distribution of the secondary particles).
- Neutron interaction data (cross-sections or kerma) for the interaction for the neutron beam with the constituent nuclei of tissue (C, N, O, H, Ca).

suitability of the various different neutron sources for radiation therapy applications. Much of the data in Table 3 are interconnected. In theory, a detailed knowledge of the neutron spectrum as a function of angle should be sufficient to calculate the other data provided there are comprehensive data on the nuclear cross-sections for the interaction of the neutrons with all the various biological target nuclei involved across the energy range of interest. Although, these data do exist for fast neutron beams they are not comprehensive and performing the necessary calculations with Monte Carlo codes remains a formidable and time consuming task. For this reason, until now it has proved simpler and more efficient to rely on various types of direct measurements to collect the necessary data; this situation may change in the future.

Basic Source Data

Neutron Yield. Neutron yields may be measured using a suitable detector that counts the number of particles arriving at the detector with no regard to the energy of the particle. Such detectors must usually be calibrated in order to determine their absolute efficiency. An excellent account of neutron detectors can be found in the work of Knoll (19). Neutron yields are generally expressed in terms of neutrons per steradian per microcoulomb of incident beam charge ($\mu C^{-1} \cdot sr^{-1}$).

Neutron Spectra. Neutron spectra are measured using a variety of neutron detectors (19,20). For a spectral measurement, the detector must exhibit an energy response in which the magnitude of the detector signal is proportional to the energy of the incident neutrons. Such detectors must also be calibrated so that correction can be made for their counting efficiency.

In a spectrum measurement, the neutron yield is measured as a function of energy ($\mu C^{-1} \cdot MeV^{-1} \cdot sr^{-1}$). The total neutron yield at a particular angle can be calculated from a spectral measurement by integrating over the neutron energy.

Neutron Dose. Neutron dose is most easily and accurately measured using the methods of mixed-field dosimetry (21). A tissue equivalent plastic ionization chamber is used to measure the total neutron plus gamma-ray dose and a Geiger–Muller (GM) tube is used as a neutron insensitive detector. From the two measurements, the neutron and gamma-ray dose can be determined separately. This does not mean that spectral and kerma data are not important. Indeed, the calculation of dose from an ionization chamber reading involves the use of factors, which rely on the exact nature of the neutron spectrum (e.g., kerma ratios and the energy to produce an ion pair). These factors are readily available in the literature in ICRU Report No. 46 (22). In practice, it is sufficient to measure the total dose only, since the percentage of gamma-ray dose is relatively small and its biological effectiveness is ~3 times less than an identical neutron dose.

Microdosimetric Measurements. Microdosimetric data are most commonly measured using a Rossi type A-150 tissue equivalent plastic proportional counter in which the sensitive volume of the counter is filled with a tissue equivalent proportional counter gas at reduced pressure (23). Typically, the internal diameter of these counters is 12.7 mm, and when filled to a pressure of 8.8 kPa with propane-based tissue equivalent gas, the counter simulates a sphere of solid tissue of diameter 2 μm. The counter detects the energy deposited in the gas as recoil particles traverse the gas volume after a neutron interaction has occurred in the plastic wall of the counter. From a knowledge of the counter geometry, it is possible to calculate the energy deposited per unit path length (keV/μm) (24). The microdosimetric spectrum is usually plotted as a single event spectrum in which the event size (y), in units of keV/μm, is plotted on a log scale as the ordinate and the differential dose distribution in event size (y) per unit logarithmic interval, $y \cdot d(y)$, is plotted on a linear scale as the abscissa. A typical microdosimetric spectrum plotted in this form is shown in Fig. 4. In this representation, the area under the curve represents the total dose. The various peaks in the curve can be interpreted as due to the different components of the dose, gamma-rays, recoil protons, alpha particles, and recoil heavy ions. Hence, such plots can be used to distinguish between neutron beams produced by different sources. A detailed account of microdosimetric methods can be found in ICRU Report No. 36 (24).

Beam Characteristics. From a practical radiation therapy physics point of view, the most useful data is that which allows you to calculate the neutron radiation dose distribution within the patient. A detailed knowledge of the neutron fluence and energy is not necessary to make these calculations. In radiation therapy measurements of absorbed dose in a water tank, ~60 × 60 × 60 cm^3 are made using a small volume (typically 0.3 cm^3) ionization chamber; the tank is known as a “water phantom”. Neutron dose at a point in the phantom must be determined at a known depth, field size, and source-to-surface distance

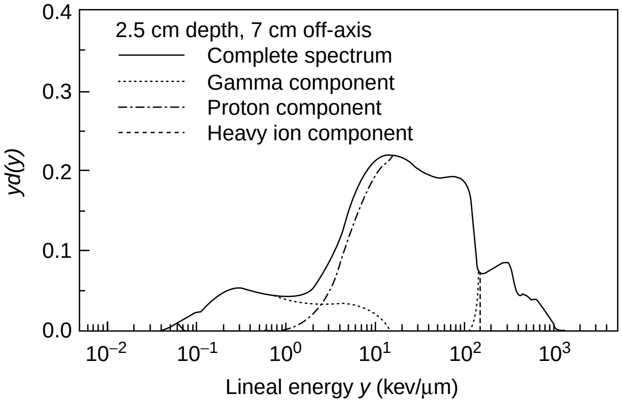


Figure 4. A typical microdosimetric event size spectrum measured with a tissue equivalent plastic Rossi proportional counter. The portions of the spectrum attributable to recoil electrons from gamma-ray interactions and recoil protons, alphas, and heavy ion recoils (C12, N14, and O16) from neutron interactions are identified. Reproduced with permission from Kota and Maughan, Med. Phys. 1996;23:1593.

Table 4. A Comprehensive List of All Neutron Therapy Centers

Facility	Reaction	Type of Machine	Z_{50}, cm	$D_{\max}$, cm	Beam Type	Beam Shaping	First T_x	Status
Berkeley	d(16.0)Be	Cyclotron	8.8	0.2	Fixed	Inserts	1938	Closed
London	d(16.0)Be	Cyclotron	8.8	0.2	Fixed	Inserts/Jaws	1966	Closed
Dresden	d(13.5)Be	Cyclotron	7.9	0.2	Fixed	Inserts	1972	Closed
Houston MDAH	d(50)Be	Cyclotron	13.1	0.8	Fixed	Inserts	1972	Closed
Washington MANTA	d(35)Be	Cyclotron	11.7	0.5	Fixed	Inserts	1973	Closed
Univ. of Washington	d(21)Be	Cyclotron	9.4	0.2–0.3	Fixed	Inserts	1973	Closed
Chiba-Chi	d(30)Be	Cyclotron	10.8	0.5	Fixed	Multileaf	1975	Closed
Fermi Lab	p(66)Be	Proton Linac	16.6	1.6	Fixed	Inserts	1976	Open
Amsterdam	d(0.25)T	D–T	10.3	0.2–0.3	Rotational	Inserts	1976	Closed
Essen	d(14.3)Be	Cyclotron	8.1	0.2	Rotational	Inserts	1976	Open
Glasgow	d(0.25)T	D–T	10.3	0.2–0.3	Rotational	Inserts	1977	Closed
Manchester	d(0.25)T	D–T	10.5	0.2–0.3	Rotational	Inserts	1977	Closed
Heidelburg	d(0.25)T	D–T	10.6	0.3	Rotational	Inserts	1977	Closed
Hamburg	d(0.5)T	D–T	8.8	0.25	Rotational	Inserts	1977	Closed
Cleveland (GLANTA)	p(25)Be	Cyclotron	10.3	0.5	Fixed	Inserts	1977	Closed
Louvain-la-Neuve	p(65)Be	Cyclotron	17.5	1.8	Fixed	Multileaf	1978	Closed
Tokyo	d(14.0)Be	Cyclotron	8.3	0.2	Fixed	Inserts	1978	Closed
Krakow	d(12.5)Be	Cyclotron	7.7	0.2	Fixed	Inserts	1978	Closed
Edinburgh	d(16.0)Be	Cyclotron	8.7	0.2	Rotational	Inserts	1978	Closed
Chicago	d(8.0)D	Cyclotron	9.8	0.15	Fixed	Inserts	1981	Closed
Orleans	p(34)Be	Cyclotron	12.8	0.5	Fixed	Inserts	1981	Open
Cleveland (GLANTA)	p(42)Be	Cyclotron	13.5	2.2	Fixed	Inserts	1982	Closed
Houston (MDA)	p(42)Be	Cyclotron	14	1.2	Rotational	Inserts	1983	Closed
Riyadh	p(26)Be	Cyclotron	10.3	0.5	Rotational	Inserts	1984	Closed
Munster	d(0.25)T	D–T	10.5	0.3	Rotational	Inserts	1984	Closed
Univ. of Washington	p(50)Be	Cyclotron	14.8	1.2	Rotational	Multileaf	1984	Open
Univ. of Pennsylvania	d(0.25)T	D–T	10.3	0.2–0.3	Rotational	Inserts	1985	Closed
Clatterbridge	p(62)Be	Cyclotron	16.2	1.4	Rotational	Jaws	1986	Closed
Seoul	p(50)Be	Cyclotron	14.8	1.2	Rotational	Jaws	1986	Closed
UCLA	p(46)Be	Cyclotron	13.1	1.7	Rotational	Jaws	1986	Closed
Faure S.Africa	p(66)Be	Cyclotron	16.2	1.5	Rotational	Jaws/MLC trim	1988	Open
Detroit	d(48.5)Be	SC Cyclotron	13.6	0.9	Rotational	Multirod	1991	Open
Beijing	p(35)Be	Proton Linac	~13.0	~0.5	Fixed	Inserts	1991	Closed
Nice	p(65)Be	Cyclotron	17.5	1.8	Fixed	Multileaf	1993	Closed

(see the section: Neutron Dose). Measurements are also made along the radiation beam central axis to determine the attenuation of the beam. The parameter Z_{50} is a measure of the beam penetration as defined by the depth in a water phantom at which the neutron dose is reduced to 50% of its maximum value for a 10×10 cm^2 field. The Z_{50} value varies as a function of the energy and type of particle used in the primary beam. As the incident particle energy increases the mean neutron energy increases and the Z_{50} increases; this can be seen from the data in Table 4. This 50% depth-dose point also varies with the dimensions of the irradiation field (normally known as the field size). Some form of collimator is required to set the field size, this may be an attenuating block with a predefined rectangular opening, a pair of attenuating jaws, which provide a variable rectangular opening, or a multileaf collimator, which can define rectangular fields or more complex shapes. Generally, square fields are used for beam data measurements and the field size is defined at the center of the treatment volume (isocenter). The Z_{50} variation with field size arises because the dose at a point depends not only on attenuation of the primary beam in the water phantom, but also on the scattered component of the beam, which is field size dependent. To fully characterize a therapeutic beam it is, therefore, necessary to measure depth dose curves at a variety of field sizes; Fig. 5 illustrates this phenomenon.

Another beam parameter, which varies with the primary beam particle and energy is the beam build-up. When

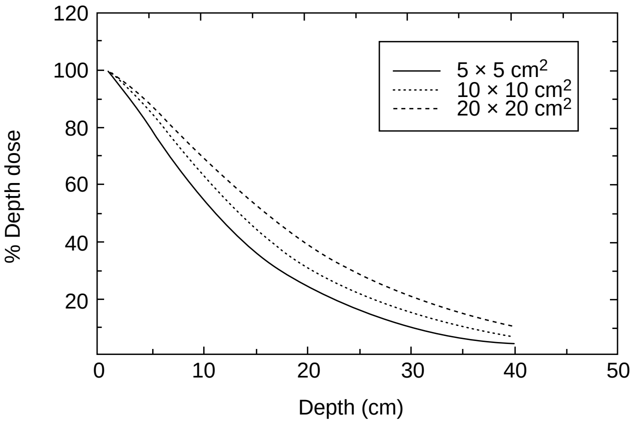


Figure 5. Depth-dose curves for d(48.5)Be neutron beam plotted as a function of field size. Reproduced with permission from Maughan and Yudelev, Med. Phys. 1995;22:1462.

Table 5. The Variation of surface dose and d_{max} as function of the neutron producing nuclear reaction and the energy of the incident particle[a]

Neutron Producing Reaction	Incident Particle Energy, MeV	Surface Dose as a % of the Dose at d_{max}	d_{max}, cm
d + Be	48.5	42	0.9
p + Be	50	38	1.2
p + Be	66	40	1.6

[a] Data taken from Ref. (21) and (25).

an indirectly ionizing radiation beam (neutrons or X rays) passes from air into a solid water medium, the secondary particle fluence in the surface layers is much less than at depth, because in the shallower layers the secondary particles are those that originated in the air and passed into the solid. Since air is much less dense than the solid, there are relatively few secondary particles in the surface layer. As the neutron beam penetrates the solid medium, more and more secondary particles are set in motion in the solid medium, until an equilibrium situation is reached at a depth that is about equal to the average range of the secondary particles in the solid. The energy deposition (dose) reaches a maximum at this depth (known as the depth of maximum dose, d_{max}) and is attenuated beyond this point as shown in Fig. 5. This build-up region of the curve cannot be measured using the instrumentation used to measure the attenuation data in Fig. 5. A specialized thin pill box shaped ionization chamber (known as an extrapolation ionization chamber) is required for these measurements. The build-up region has considerable clinical significance, when treating tumors at depths $>d_{max}$, since the dose in the surface layers of the skin is reduced relative to the tumor dose and, hence, the skin can be spared from excessive radiation damage. Table 5 shows how the surface dose (dose at zero depth), expressed as a percentage of the dose at d_{max}, and d_{max} vary as a function of the neutron producing nuclear reaction and the incident particle energy.

Beam profiles are also measured in the water phantom by scanning the ionization chamber in a direction perpendicular to the radiation beam central axis. A typical beam profile is shown in Fig. 6. As can be seen from this figure, the exact shape of the profiles depends on the depth in the phantom and the radiation field size. The most important feature of the profiles is the sharpness of the beam edges; This parameter degrades with both increase in field size and depth due to increased scattering of the neutrons. The exact sharpness of this penumbra region depends on many factors including, the source size, scattering from the collimator system and beam monitoring components in the beam path, the collimator geometry (i.e., whether the edges of the collimator jaws or leafs are divergent), and finally on neutron scattering in the patient (or phantom). Generally, phantom scatter is the predominating factor.

Depth-dose and profile date are inputted to the computer programs used for calculating the dose distribution in patients. These programs use a variety of different mathematical algorithms to calculate the dose distributions in the patient.

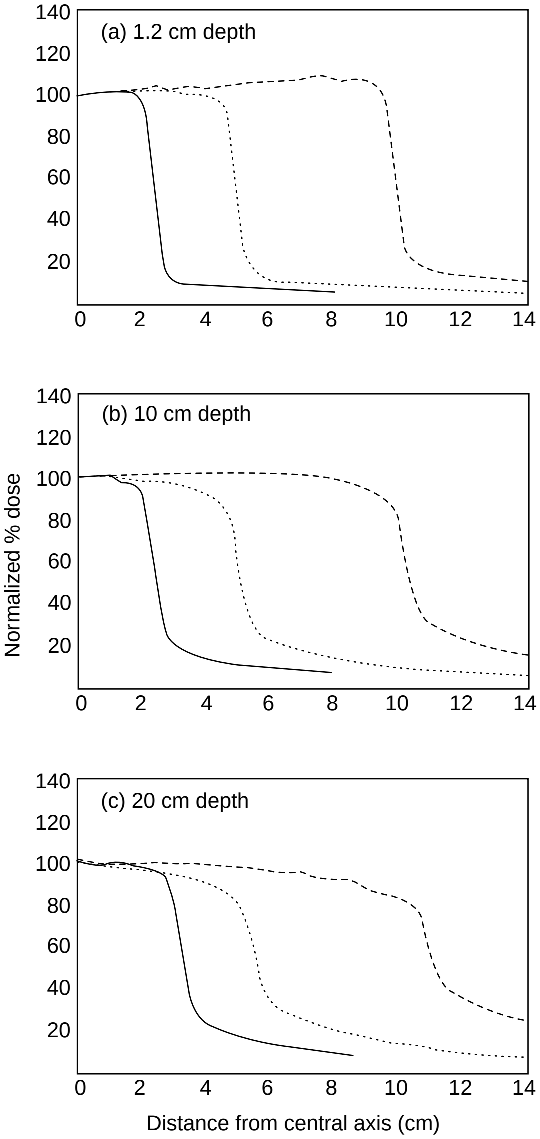


Figure 6. Beam profiles for field sizes of 5 × 5 cm^2 (solid line), 10 × 10 cm^2 (dotted line), and 25 × 25 cm^2 (dashed line) at depths of (*a*) 1.2, (*b*) 10, and (*c*) 20 cm in a water phantom. Reproduced with permission from Maughan and Yudelev, Med. Phys. 1995; 22: 1461.

NEUTRON PRODUCTION

Fast Neutron Therapy Beams

Over the past 65 years, there have been at least 34 centers that have been involved in fast neutron radiation therapy. Table 4 lists these centers and indicates which nuclear reaction has been used to produce the neutron beam. In column 2, the lower case letter indicates the accelerated particle, the number in parenthesis is the energy in megaelectronvolts of this projectile and the final letter(s) represents the target nucleus. Of the facilities, 12 have used the deuteron stripping reaction with a beryllium target, 14 the proton inelastic scattering reaction on a beryllium target, 7 the deuteron–tritium fusion reaction and only 1, the deuteron–deuterium fusion reaction. The relative merits of these various modes of neutron production will be discussed.

The d-Be Reaction

In practice, the deuteron stripping reaction on beryllium is the most prolific neutron producing reaction since a solid beryllium target capable of stopping the full energy of the beam can easily be constructed. There are basic physical reasons for the deuteron stripping reaction on beryllium being more prolific in producing neutrons than the inelastic scattering of protons from a beryllium target. The deuteron is a loosely bound structure of a neutron and a proton in which the two particles spend most of their time at greater distances from each other than the range of the forces between them. Hence, when an energetic deuteron approaches a beryllium target nucleus it is possible for the proton to be absorbed into the target nucleus, breaking free from the neutron that carries on following its original path at its original velocity (i.e., with half the kinetic energy of the original deuteron). The excited ^{10}B nucleus formed may also decay by neutron emission. When a proton beam interacts with a beryllium target the proton is absorbed into the target nucleus to form a compound nucleus, ^{10}B, in an excited state, which may decay by emitting a neutron. Hence, the stripping reaction is a much more prolific source of neutrons, since many neutrons originate from the breakup of the deuteron. In addition, the stripping reaction is very forward-peaked in the laboratory, while the (p,n) reaction on beryllium produces a more isotopic distribution, since it involves the formation of a compound nucleus. The theoretical aspects of the production of intense neutron beams using the deuteron stripping reaction with beryllium targets has been discussed by August et al. (26). Experimental neutron yield data and spectral data on the characteristics of neutrons from beryllium targets bombarded with protons and deuterons with energies of 16, 33, and 50 MeV are available in the work of Meulders et al. (27). The above experiments measure the total neutron fluence at 0° or the differential fluence as a function of neutron energy at various angles relative to the forward direction. In order to estimate the usefulness of a given reaction as a neutron source for radiation therapy, it is necessary to know the neutron dose rate produced in practice by a given attainable beam current. Such information can be calculated from neutron spectrum data using neutron kerma factors for water or body tissue (22).

Another vitally important parameter is the penetration of the neutron beam in tissue. Although in principle it is possible to calculate this information from the spectral and kerma data, in practice there are insufficient data and the calculations are difficult. Therefore, ionization chamber measurements are often more convenient for measuring both the dose rate and penetration of neutron beams. There is extensive data on neutron dose rates and depth dose characteristics of neutron beams across a wide range of energies for the deuteron stripping reaction on beryllium. Smathers et al. (28) reviewed the available dose rate data in the incident deuteron energy range of 11–50 MeV and concluded that the tissue kerma measured free-in-air at a target-to-detector distance of 1.25 m and for a 5×5 cm^2 field size could be fitted by an equation of the form

$$1\mathrm{n} K = 1\mathrm{n}\, a + b \ln E \tag{1}$$

In this equation, E is the energy of the incident beam in million electronvolts, K is the tissue kerma in units of cGy min^{-1}·μA^{-1} and a and b are constants with numerical values of 1.356×10^{-4} and 2.97, respectively.

Later, Wootton (29) reviewing ionization chamber dose rate data for the d-Be reaction quoted the following expression for the dose rate at 1.25 m

$$D \cdot Q^{-1} = 2.49 \times 10^{-2}\, E_{\mathrm{d}}^{2.95} \tag{2}$$

where $D \cdot Q^{-1}$ (in Gy/C) is the absorbed dose to tissue free-in-air at a 1.25 m target to detector distance per unit charge of deuteron beam, and E_{d} is the incident deuteron energy in million electronvolts. If Eq. 1 is rewritten in this form and normalized to the same units it becomes

$$D \cdot Q^{-1} = 2.26 \times 10^{-2}\, E_{\mathrm{d}}^{2.95} \tag{3}$$

At $E_{\mathrm{d}} = 50$ and 10 MeV, these equations agree to within 2 and 5%, respectively. Of course, such equations can only be used to give a rough estimate of the neutron dose output of a neutron therapy device, since the exact output depends on the details of the target, flattening filter, collimator, and dose monitor design.

The spectral data of Lone et al. (30) gives information on the fluence averaged energy of the neutron beam (E_{n}), they derived the following expression:

$$E_{\mathrm{n}} = 0.4\, E_{\mathrm{d}} - 0.3 \tag{4}$$

relating the mean neutron energy to the incident deuteron energy (E_{d}) in million electronvolts.

Data on the penetration of neutron beams, produced by the d-Be reaction, have been published by Shaw and Kacperek (31). In Fig. 7, the values of Z_{50} from Table 4 for the d-Be reaction is plotted as a function of incident deuteron energy. A power law fit to the curve yields the

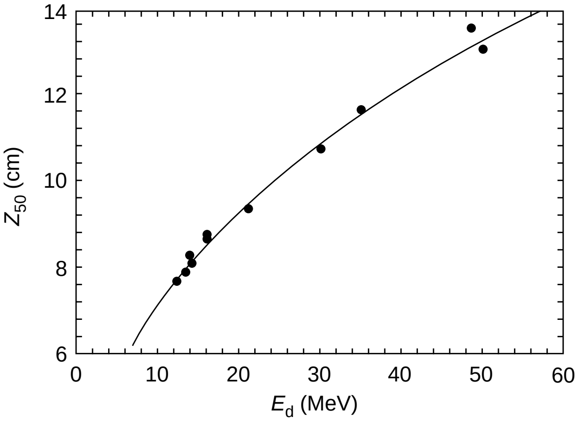


Figure 7. The neutron beam 50% depth-dose value (Z_{50}) for neutrons produced by the d-Be reaction plotted as a function of the incident deuteron energy (E_{d}). The solid line is a power law fit to the data (Eq. 5). The data are for a 10×10 cm^2 field at a range of source-surface distances (SSD) between 91 and 183 cm.

following equation

$$Z_{50} = 2.90\ E_{\mathrm{d}}^{0.39} \quad (5)$$

The exact value of Z_{50} depends on the target structure and the beam filtering effects of the flattening filter and the dose monitor devices.

The clinical relevance of microdosimetric data in neutron therapy planning has been discussed by Pihet et al. (32). The microdosimetric dose distribution shown in Fig. 4 can be analyzed in several ways, the most useful single parameter, which can be used to describe the distribution is the dose mean lineal energy corrected for saturation

$$y_1^* = \int y_{\mathrm{sat}} \cdot d_1(y) \cdot dy \quad (6)$$

This parameter was defined in the dual radiation action theory of Kellerer and Rossi (33). The function y_{sat} is a response function that accounts for the saturation effect that is observed in mammalian cell systems (34); as the LET of the beam is increased the observed RBE decreases due to the overkill effect (35). In Fig. 8, the $y_1{}^*$ values for a variety of different therapy beams are plotted as a function of the mean neutron energy, as defined in Eqs. (4) and (8). The closed circles represent the data for the d-Be reaction.

The p-Be Reaction

From Table 4, it can be seen that most of the early neutron therapy centers (i.e., those operating at lower energies) utilized the deuteron stripping reaction or the fusion reaction as the source of neutrons (see Section: The d-T Reaction). Interest in the p-Be reaction increased when the importance of constructing neutron sources with rotational isocentric capability (i.e., capable of rotating around the patient with the tumor center on the axis of rotation) and with penetration equivalent to 4 MV photon beams was realized (Table 2). Good penetrability requires deuteron or proton beams with energies of 40–50 MeV or greater, and isocentricity requires bending magnet systems capable of bending these beams 180° (a 45° bend followed by a 135° bend). Conventional cyclotrons capable of producing 50 MeV deuterons were too large and expensive as were the magnet systems for bending these beams. Proton cyclotrons of 50 or 60 MeV offered a much less expensive alternative in the late 1970s, when the decision to install a new generation of hospital-based neutron therapy facilities was being made by the NCI in the United States. The problems of switching from the deuteron stripping to the p-Be reaction were soon recognized: the energy spectrum from the reaction of protons on a beryllium target has a significant low-energy tail, which reduces the average neutron beam energy and spoils the penetration. Also the neutron output is much less, therefore, higher beam currents are required with an increase in the problems associated with target cooling and target activation. The penetration problem can be overcome by using nonstopping targets (i.e., beryllium targets in which the incident proton beam does not lose all its energy) in conjunction with polyethylene filters, which filter out the low energy component of the beam. These techniques have been discussed in detail for proton beams with energies between 30 and 60 MeV by Bewley et al. (36) and for a 41 MeV proton beam by Smathers et al. (37). The absorbed dose rate ($D \cdot Q^{-1}$) to tissue at 1.25 m from the target is given by Wootton (29) as

$$D \cdot Q^{-1} = 2.44 \times 10^{-2}\ E_{\mathrm{p}}^{2.37} \quad (7)$$

where $D \cdot Q^{-1}$ is in units of Gy/C and E_{p} is the incident proton energy in million electronvolts.

For the p + Be reaction with a stopping target, the average neutron energy for neutrons with energies >2 MeV (E_{n}) measured at 0° to the incident beam is given by

$$E_{\mathrm{n}} = 0.47\ E_{\mathrm{p}} - 2.2 \quad (8)$$

where E_{p} is the incident proton energy (29).

In Fig. 9, the value of Z_{50} from Table 4 for the p-Be reaction is plotted as a function of incident proton energy, the solid curve is a power law fit to the data that gives the

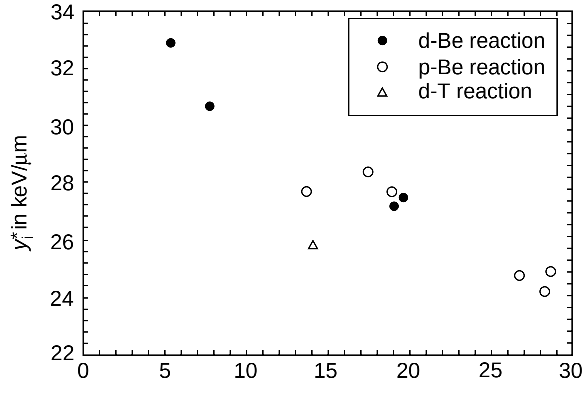


Figure 8. The microdosimetric parameter mean lineal energy corrected for saturation ($y_1{}^*$) plotted as a function of the mean energy of the neutron beam for various neutron producing reactions: d-Be, open circles (E_{n} from Eq. 4); p-Be, open circles (E_{n} form Eq. 8), and d-T reaction, open triangle (E_{n} = 14.1 MeV).

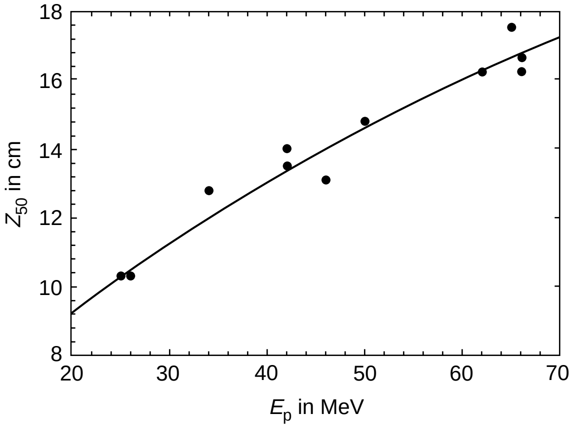


Figure 9. The neutron beam 50% depth-dose value (Z_{50}) for neutrons produced by the p-Be reaction plotted as a function of the incident proton energy (E_{p}). The solid line is a power law fit to the data (Eq. 9). The data are for a 10 × 10 cm^2 field for range of SSD between 125 and 190 cm.

following equation:

$$Z_{50} = 2.06\ E_{\mathrm{p}}^{0.50} \tag{9}$$

The greater spread in the data points, when compared with the similar data plotted in Fig. 7, is a result of the greater variety in the target design (i.e., target thickness and filtration conditions) used at the different facilities.

The microdosimetric data for the p-Be reaction is represented by the open circular data points in Fig. 8. The dose mean lineal energy corrected for saturation correlates with the mean neutron energy for both the p-Be and the d-Be produced neutron beams.

The d-D Reaction

This reaction was used in the neutron radiation therapy facility at the University of Chicago, where a deuteron beam of energy 8.3 MeV was incident on a thick cryogenic deuterium gas target designed by Kuchnir et al. (38). Two reactions predominate when a deuterium target is bombarded with deuterons:

$$\mathrm{d} + \mathrm{D} = {}^{3}\mathrm{He} + \mathrm{n} \qquad Q = +3.27\ \mathrm{MeV}$$

and,

$$\mathrm{d} + \mathrm{D} = \mathrm{d} + \mathrm{n} + \mathrm{p} \qquad Q = -2.22\ \mathrm{MeV}$$

Hence, there are two distinct groups of neutrons produced, the higher energy group resulting from the first of these two reactions. The neutron energy spectrum for bombardment of a thick stopping target exhibits two maxima corresponding to the two groups. The relative magnitude of the two peaks depends on the incident deuteron energy. At an incident deuteron energy of 6.8 MeV the higher energy peak due to the $\mathrm{D(d,n)^3He}$ reaction predominates, but for an incident energy of 11.1 MeV, the two peaks are comparable (39). Waterman et al. (39) calculated the neutron spectra at 6.8, 8.9, and 11.1 MeV from a knowledge of the mass stopping power of deuterons in deuterium and from the cross-sections of the two reactions as given by Schraube et al. (40).

The dosimetric properties of the d–D neutron beam are summarized in the work of Kuchnir et al. (38). Figure 10 shows the variation in absorbed tissue dose rate (Gy/μC) as a function of the incident deuteron energy for a thick deuterium gas target. The measurements were made at a SSD of 126 cm with a 11.1 × 11.1 cm^2 field size. The data can be fitted by a power law expression.

$$D \cdot Q^{-1} = 2.41 \times 10^{-2}\ E_{\mathrm{d}}^{3.28} \tag{10}$$

where $D \cdot Q^{-1}$ (Gy/C) is the absorbed dose to tissue measured free-in-air per coulomb (C) of incident beam current, and E_{d} is the incident deuteron beam energy. Measurements have been made by Weaver et al. (41) at an incident deuteron energy of 21 MeV, but with a transmission gas target. For a target filled to a pressure of 3.33 MPa (33 atm), equivalent to an energy loss of ~3.5 MeV, the measured dose was 2.25×10^{-4} Gy/μC for a 10 × 10-cm field at 1.25 m SSD.

In practice, the University of Chicago neutron therapy facility produced a maximum dose rate of 0.12 Gy/min at an SSD of 91 cm, for a 10 × 10 cm field size. The penetration of the University of Chicago d–D beams in terms Z_{50} is 9.8 cm (Table 4). An interesting feature of the d–D reaction is that as the incident deuteron energy is increased the mean neutron energy produced remains practically constant. This finding is evident in the depth-dose data of Kuchnir et al. (38), where changing the incident deuteron energy has no significant effect on the value of Z_{50}. Even at 21 MeV with a transmission target the Z_{50} remains constant at ~10 cm. Thus, the inherent poor penetration of neutron beams produced by the d–D reaction, combined with the difficulties of producing a cryogenic deuterium gas target discouraged the use of this reaction as a neutron source for radiation therapy.

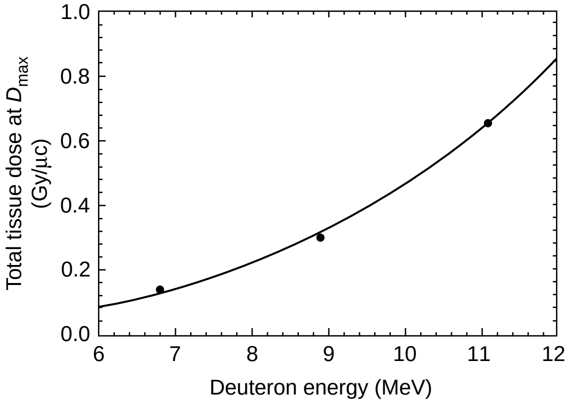


Figure 10. Absorbed dose at the depth of maximum dose as a function of incident deuteron energy for the d–D reaction. The data are from Kuchnir et al.(38). The solid curve is a power law fit to the data (Eq. 10). Measurements were made in a 11.1 × 11.1 cm^2 field at an SSD of 126 cm.

The d–T Reaction

For many years, this reaction was seen as the ideal reaction for producing a relatively inexpensive source of neutrons for radiation therapy. The large positive Q-value for the reaction

$$\mathrm{d} + {}^{3}\mathrm{H} \longrightarrow {}^{4}\mathrm{He} + \mathrm{n} \qquad Q + 16.6\ \mathrm{MeV}$$

results in monoenergetic neutrons of energy ~14 MeV. In principle, a relatively modest deuteron energy of 250–500 keV should be sufficient to produce an intense source of 14-MeV neutrons if sufficient beam current can be obtained. The original intention was to produce the source and target assembly in the form of a sealed tube, which could be easily replaced in the treatment head and would have a lifetime of 1000 h or more. Such a unit would have been similar in this respect to the 250-kVp X-ray units that were in widespread use before the advent of ^{60}Co units and high-energy electron linacs in conventional photon radiation therapy. Initially, the main problem with these devices was that associated with producing a target in the sealed tube configuration that would provide sufficient neutron dose rate. However, many different systems were used in attempts to produce a practical d–T generator and these have been reviewed in detail in ICRU Report No. 45 (21).

Of the five types of commercially available d–T generators, which were used in clinical trials to treat significant numbers of patients, four were of a type that employed some form of sealed tube in which a mixed deuterium–tritium beam was accelerated to an energy of 200–250 keV and used to bombard a tritiated rare earth target (titanium, erbium, or scandium). The characteristics of these four machines are given in Table 4. The Haefely device produced the highest dose rate with the longest average tube life of $\sim$300 h and was installed in Heidelberg and Münster. The operation of the Philips and Elliot tubes are described by Broerse et al. (42). A Philips machine was installed in Amsterdam and the Elliot devices were used in Glasgow and Manchester. An account of the construction of the Haefely machine is given by Schmidt and Rheinhold (43) while a detailed appraisal of its clinical operation can be found in the work of Höver et al. (44). The University of Pennsylvania D–T generator was built by the Cyclotron Corporation (Berkeley, CA).

The fifth commercial unit, installed in Hamburg, was produced as a collaboration between AEG in Germany and Radiation Dynamics Inc. (RCI) in the United States. The machine used a pure deuterium beam accelerated to 500-keV incident on a replaceable rotating tritiated titanium target (45). The source and target design were improved by incorporating an analyzed deuterium beam (to remove molecular D_2^+ beam components) and a larger target (46). With these improvements a dose rate of 0.12–0.13 Gy/min was achieved.

PRACTICAL FAST NEUTRON THERAPY FACILITIES

In fast neutron radiation therapy the need for state-of-the-art neutron facilities, which allow neutron treatments to be delivered with precision and sophistication equivalent to that used in modern conventional X-ray therapy, is well recognized. Modern trends in X-ray therapy are toward conformal therapy with multiple static fields, multileaf collimators, three-dimensional (3D) treatment planning and most recently (IMRT). All these tools must be available for neutron radiation therapy if effective randomized phase III clinical trials are to be completed to compare the two modalities.

An important aspect of this problem is beam penetration. The problem with neutron beams is that it is not possible to increase the mean energy of the neutrons to a point at which the neutron beams have percentage depth-dose characteristics that are equivalent to modern high energy (15–25 MV) photon beams, since as the neutron beam energy increases, the average LET of the beam decreases. If the average LET is decreased too far, the radiobiological advantage of the neutron beam will be significantly diluted (e.g., RBE tends to decrease and neutron beam advantages associated with hypoxia decrease, the radiosensitivity variation within the cell cycle tends to that of low LET radiations). Hence, there is a trade-off between beam penetration and LET effect. This trade-off can be seen in Fig. 8, which illustrates how the effective LET (y_1^*) of the neutron beam decreases as the mean neutron energy (E_n) increases.

The requirement that neutron beams should be at least equivalent to 4 MV photon beams (Table 2) arises in part from this trade-off. Of the 34 facilities listed in this Table 4, only 10 satisfied this penetration requirement. Of the 6 operational facilities 4 satisfy the requirement and the most penetrating beams at the Ithemba Laboratory in South Africa and Fermi Laboratory in the United States, produced by the p(66)Be reaction, are equivalent to an 8-MV photon beam. If all the requirements of Table 1 are considered and in addition a multileaf or multirod collimator for producing irregularly shaped fields is made mandatory, then only three of the operational facilities meet all the requirements. These are at the University of Washington in Seattle, the Ithemba Laboratorty in South Africa and at Harper Hospital, Wayne State University in Detroit. The fact that the neutron beams are less penetrating than the 15–25 MV photon beams that are commonly used for treating deep-seated tumors may not be a problem. In a treatment planning comparison of 3D conformal neutron and photon radiotherapy for locally advanced adenocarcinoma of the prostate, Forman et al. (47) showed that the dose–volume histograms for gross tumor, rectal, and bladder volumes treated with neutrons and photon beams are not significantly different. Wootton (29) suggested that neutron beams with a Z_{50} of >15 cm are required, and that for the d–Be reaction to be useful in this case, an incident deuteron energy of 61 MeV would be required. Forman's data, however, indicate that a Z_{50} of 13.6 cm is adequate for producing acceptable dose distributions for the treatment of pelvic tumors.

In the late 1970s, economic considerations led to the choice of the p–Be reaction as the neutron source for a new generation of hospital-based high-energy proton cyclotrons for clinical trials in the United States, because deuteron producing conventional cyclotrons and the associated bending magnet system required to produce rotational beams were too costly. These machines were installed at the MD Anderson (MDA) Hospital in Houston, at the University of California Los Angeles, and at the University of Washington in Seattle (Table 4). Since this time the development of a compact superconducting deuteron cyclotron for neutron radiation therapy by Henry Blosser and his associates at the National Superconducting Cyclotron Laboratory at Michigan State University has had a significant impact on the technology of neutron therapy. This superconducting facility (25,48) has many innovative features. The accelerator weighs $\sim$25 Mg (25 tons), $\sim$10 times less than a conventional 50 MeV deuteron cyclotron. The unit has an internal beryllium target and is mounted between two large rings (4.3 m outer diameter) in order to provide for 360° rotation around the treatment couch. A 25 Mg counterweight mounted on the rings acts as a primary beam stop, which reduces the required thickness of the shielding walls. The total rotating mass is $\sim$60 Mg (60 tons). Figure 11 is a schematic of the cyclotron and gantry. Figure 12 shows a section through the median plane of the cyclotron indicating its' main components. The unit does not require a separate bending magnet system to produce an isocentric beam and it can be installed in a single shielded room. With no beam extraction or elaborate bending magnet system, the operation is

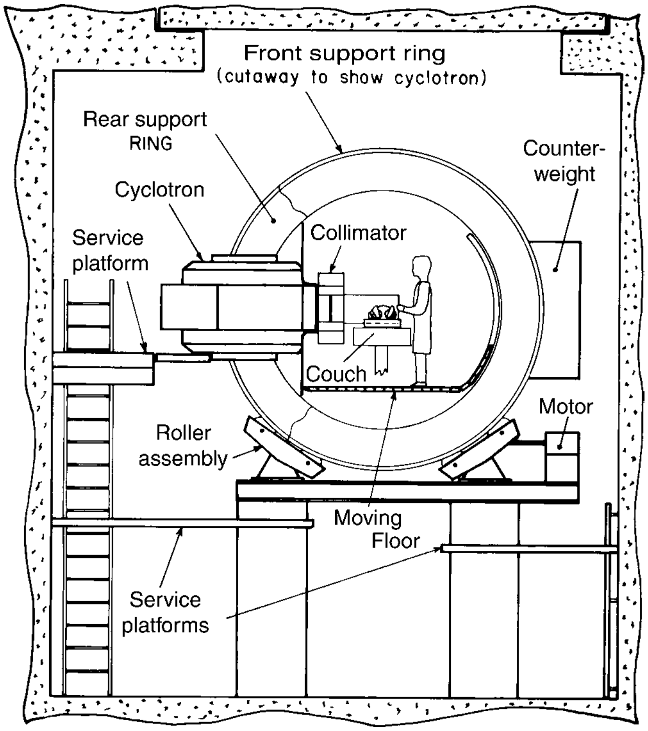


Figure 11. A schematic of the superconducting cyclotron mounted on the rotating gantry at the Wayne State University Facility. Reproduced with permission from Maughan et al., Med. Phys. 1994;21:781.

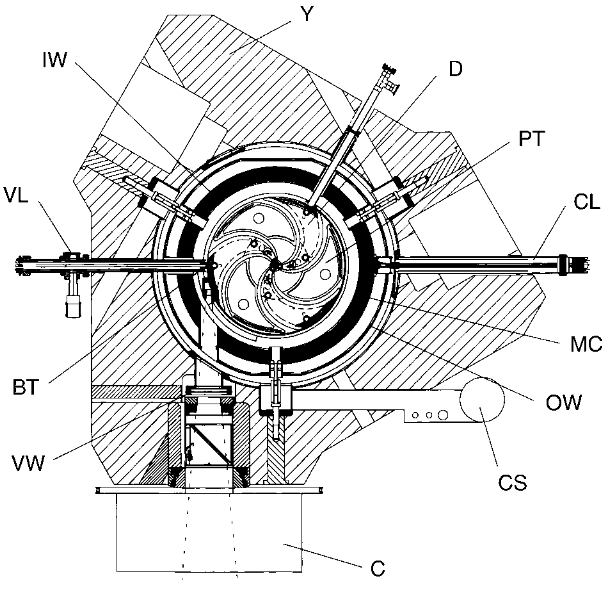


Figure 12. Schematic section through the median plane of the superconducting cyclotron, showing the following features Y = magnet yoke, MC = magnet superconducting coil, PT = magnet hill pole tip, IW = cryostat inner wall, OW = cryostat outer wall, CL = magnet coil electrical leads, CS = cryogen supply and gas return lines, D = radio frequency system dees, BT = internal beryllium target, VL = target vacuum lock, VW = beam chamber vacuum window, and C = neutron beam collimator. Reproduced with permission from Maughan et al., Med. Phys. 1994;21:780.

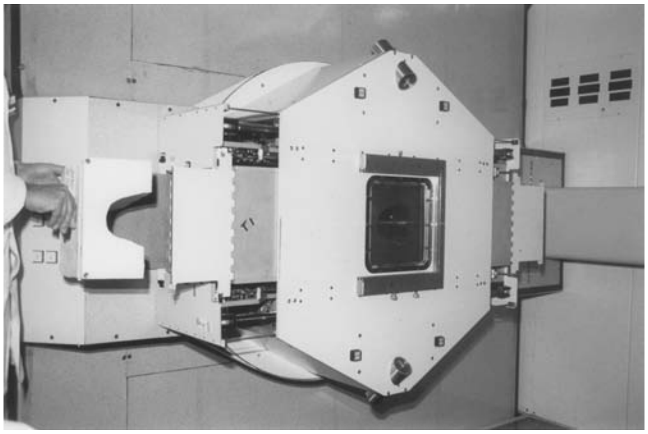


Figure 13. A photograph of the multirod collimator. One-half of the polystyrene foam form used to push the rod array into the desired shape is visible on the left. Reproduced with permission from Maughan et al., Med. Phys. 1994;21:781.

considerably simplified. The unit also incorporates a unique multirod collimator for producing irregularly shaped fields (49), which conform accurately to the tumor volume (Fig. 13). This facility has been in routine clinical use since March of 1992, and up to the end of 2002, ~1800 patients have been treated. Close to 10,000 individual treatment fields have been routinely treated in a single year making this the busiest and most efficient neutron therapy facility in the world.

DISCUSSION AND CONCLUSIONS

Neutron therapy has been demonstrated to be superior to conventional therapy in the treatment of salivary gland tumors, some tumors of the paranasal sinuses and other head and neck sites, soft tissue sarcomas, chondrosarcomas, osteosarcomas, advanced adenocarcinoma of the prostate, and inoperable and recurrent melanoma (17). For a range of other sites, further investigation is necessary to establish the efficacy of neutron therapy; these sites include pancreas, bladder, rectum, esophagus, uterine cervix, and brain.

However, in spite of these successes, neutron therapy appears to be in decline with only six centers actively treating patients (three in the United States and one each in Germany, France, and South Africa). The emphasis on precision radiation therapy has resulted in the development of intensity modulated radiation therapy techniques in conventional X ray therapy. These techniques allow for highly conformal dose delivery, maximizing the dose to the tumor volume and minimizing the dose to the surrounding normal tissues. There is also a considerable increase in the number of proton beam therapy centers, using the unique energy deposition patterns associated with proton beams to achieve even greater conformality than is achievable with IMRT.

In Europe and Asia, there is interest in developing ^{12}C ion beams for radiation therapy. These developments are spurred by the superior results achieved with neutron therapy in the cases outlined above. Heavy ion beams, such as ^{12}C beams, are high LET beams, which combine the biological advantages of neutrons with the dose distribution advantages of protons. Such beams are extremely expensive to produce. The application of intensity modulated radiation therapy techniques in neutron therapy (IMNT) could improve the conformality of neutron therapy. Compact superconducting cyclotrons with computer controlled multileaf collimators, which allow IMNT to be delivered, could be an attractive and less expensive alternative to ^{12}C therapy.

The superconducting technology could be applied to designing a compact 60–70 MeV gantry mounted proton cyclotron to provide a beam with better depth dose characteristics than the existing Wayne State University cyclotron. The possibility of building a compact conventional 50 MeV proton cyclotron in a similar configuration to the superconducting deuteron cyclotron has been suggested (Jongen, unpublished data). A computer controlled MLC is at present under construction at Wayne State University with the intention of using it to implement IMNT (50). Such advances in neutron therapy technology are important if it is to achieve its full potential and remain competitive with the other radiation therapy modalities (i.e., conventional X rays and electrons, protons, and heavy ions).

In the >80 years since its discovery in Cambridge by James Chadwick, the neutron has found an important place in radiation therapy research, and much has been done to improve the means of neutron production and delivery.

BIBLIOGRAPHY

1. Chadwick J. The existence of a neutron. Proc R Soc London, Ser A 1932;136:629–708.
2. Cockroft JD, Walton ETS. Experiments with high velocity positive ions. I.-The disintegration of elements by high velocity protons. Proc R Soc London, Ser A 1932;137:229–242.
3. Gray LH, Read J, Wyatt JG. A neutron generator for biological research. Br J Radiol 1940;147:82–94.
4. Stone RS. Neutron therapy and specific ionizations. Am J Roentgenol 1948;59:771–779.
5. McNally NJ, Maughan RL, de Ronde J, Roper MJ. Measurements of some radiobiological and physical properties of neutrons produced by a 4-MV Van de Graaff accelerator. Rad Res 1982;89:227–237.
6. Sheline GE, Phillips TL, Field SB, Brennan JT, Raventos A. Effects of fast neutrons on human skin. Am J Roentgenol 1971;111:31–41.
7. Thomlinson RH, Gray LH. The histological structure of some human lung cancers and possible implications for radiotherapy. Br J Cancer 1955: 9:539–549.
8. Howard A, Pelc SR. Synthesis of deoxyribonucleic acid in normal and irradiated cells and its relation to chromosome breakage. Heredity 1953;6 (suppl): 261–273.
9. Sinclair WK, Morton RA. X-ray sensitivity during the cell generation cycle of cultured Chinese hamster cells. Rad Res 1966;29:450–474.
10. Withers HR, Mason K, Reid BO. Response of mouse intestine to neutrons and gamma rays on relation to dose fractionation and division cycle. Cancer 1974;34:39–47.
11. Battermann JJ, Breur K, Hart GAM, VanPeperzeal HA. Observations on pulmonary metastases in patients after single doses and multiple fractions of fast neutrons and cobalt-60 gamma rays. Eur J Cancer 1981;17:539–548.
12. Griffin TW, Pajak TF, Laramore GE, Duncan W, Richter MP, Hendrickson FR. Neutron vs photon irradiation of inoperable salivary gland tumors: results of an RTOG-MRC cooperative study. Int J Rad Oncol Biol Phys 1988;15:1085–1090.
13. Russell KJ, Caplan RJ, Laramore GE, Burnison CM, Maor MH, Taylor ME, Zink S, Davis LW, Griffin TW. Photon versus fast neutron external beam radiotherapy in the treatment of locally advanced prostate cancer: results of a randomized trial. Int J Radiat Oncol Biol Phys 1994;28:47–54.
14. Laramore GE, Krall JM, Thomas FJ, Russell KJ, Maor MH, Hendrickson FR, Martz KL, Griffin TW. Fast neutron radiotherapy for locally advanced prostate cancer: final report of an RTOG randomized clinical trial. Am J Clin Oncol 1993;16: 164–167.
15. Forman JD, Yudelev M, Bolton S, Tekyi-Mensch S, Maughan R. Fast neutron irradiation for prostate cancer. Cancer Metastasis Rev 2002;12:131–135.
16. Wambersie A, Auberger T, Gahbauer RA, Jones DTL, Potter R. A challenge for high-precision radiation therapy: the case for hadrons. Strahlenther Onkolog 1999;175 (Suppl II): 122–128.
17. IAEA Report TECDOC-992, Nuclear data for neutron therapy: Status and future needs, Vienna: International Atomic Energy Agency, 1997.
18. National Cancer Institute Request for Proposal NCI-CM-97282. Clinical Neutron Therapy Program 1979.
19. Knoll GF. Radiation Detection and Measurement. 2nd ed. New York: Wiley; 1989.
20. Cross WG, Ing H. Neutron spectroscopy. In: Kase KR, Bjärngard BE, Attix FH, editors. The Dosimetry of Ionizing Radiation. Volume 2, Orlando: Academic Press; 1987. pp. 91–167.
21. ICRU Report No. 45, Clinical neutron dosimetry part I: determination of absorbed dose in a patient treated with external beams of fast neutrons, Bethesda, International Commission on Radiation Units and Measurements, 1989.
22. ICRU Report No. 46, Photon, electron, proton and neutron interaction data for body tissues, Bethesda, International Commission on Radiation Units and Measurements, 1992.
23. Rossi HH, Rosensweig W. A device for the measurement of dose as a function of specific ionization. Radiology 1955;64: 404–441.
24. ICRU Report No. 36, Microdosimetry. Bethesda, International Commission on Radiation Units and Measurements, 1983.
25. Maughan RL, Yudelev M. Physical characteristics of a clinical d(48.5) + Be neutron therapy beam produced by a superconducting cyclotron. Med Phys 1995;22:1459–1465.
26. August LS, Theus RB, Shapiro R. Stripping theory analysis of thick target neutron production for D + Be. Phys Med Biol 1976;21:931–940.
27. Meulders J-P, Leleux P, Macq PC, Pirart C. Fast neutron yields and spectra from targets of varying atomic numbers bombarded with deuterons from 16 to 50 MeV. Phys Med Biol 1975;20:235–243.
28. Smathers JB, Otte VA, Smith AR, Almond PR. Fast neutron dose rate vs. energy for the d-Be reaction—a reanalysis. Med Phys 1976;3:45–47.
29. Wootton P. Neutron therapy facilities and their specification. Radiat Protection Dosim 1988;23:349–355.
30. Lone MA, Ferguson AJ, Robertson BC. Characteristics of neutrons from Be targets bombarded with protons, deuterons and alpha particles. Nucl Instrum Methods 1981;189:515–525.
31. Shaw JE, Kacperek A. Fast neutron beams, in Central Axis Depth Dose Data for Use in Radiotherapy: 1996. Br J Radiol 1996; (Suppl 17): 97–104.
32. Pihet P, Gueulette J, Menzel HG, Grillmaier RE, Wambersie A. Use of microdosimetric data of clinical relevance in neutron therapy planning. Radiat Protection Dosim 1988;23:471–474.
33. Kellerer AM, Rossi HH. The theory of dual radiation action. Curr Top Radiat Res Q 1972;8:85–158.

34. Barendsen GW. Responses of cultured cells, tumors and normal tissues to radiations of different linear energy transfer. Curr Top Radiat Res Q 1968;4:332–356.
35. Hall EJ. Radiobiology for the Radiologist. 3rd ed. Philadelphia: J.B. Lippincott; 1988. p 170.
36. Bewley DK, Meulders JP, Page BC. New neutron sources for radiotherapy. Phys Med Biol 1984;29:341–349.
37. Smathers JB, Graves RG, Earls L, Otte VA, Almond PR. Modification of the 50% maximum dose depth for 41 MeV (p^+, Be) neutrons by use of filtration and/or transmission targets. Med Phys 1982;9:856–859.
38. Kuchnir FT, Waterman FM, Skaggs LS. A cryogenic deuterium gas target for production of a neutron therapy beam with a small cyclotron. In: Burger G, Ebert EG, editors. Proceedings of the Third Symposium on Neutron Dosimetry. Luxembourg: Commission of the European Communities; 1978. pp. 369–378.
39. Waterman FM, Kuchnir FT, Skaggs LS, Bewley DK. Page BC, Attixs FH. The use of B10 to enhance the tumor dose in fast-neutron therapy. Phys Med Biol 1978;23:592–602.
40. Schraube H, Morhart A, Grünauer F. Neutron and gamma radiation field of a deuterium gas target at a compact cyclotron. In: Burger G, Ebert HG, editors. Proceedings of the Second Symposium of Neutron Dosimetry in Biology and Medicine, EUR 5273. Luxembourg: Commission of the European Communities; 1975. pp. 979–1003.
41. Weaver KA, Eenmaa J, Bichsel H, Wootton P. Dosimetric properties of neutrons from 21 MeV deuteron bombardment of a deuterium gas target. Med Phys 1979;6:193–196.
42. Broerse JJ, Greene D, Lawson RC, Mijnheer BJ. Operational characteristics of two types of sealed-tube fast neutrons radiotherapy installments. Int J Radiat Oncol Biol Phys 1977;3: 361–365.
43. Schmidt KA, Rheinhold G. The Haefely–GFK fast neutron generator. Int J Radiat Oncol Biol Phys 1977;3:373–376.
44. Höver KH, Lorenz WJ, Maier-Borst W. Experience with the fast neutron therapy facility KARIN under clinical conditions. In: Burger G, Ebert HG, editors. Proceedings of the Fourth Symposium on Neutron Dosimetry. EUR 7448EN. Volume II, Luxembourg: Commission of the European Communities; 1981. pp. 31–37.
45. Offerman BP, Cleland MR. AEG/RDI neutron therapy unit. Int J Radiat Oncol Biol Phys 1977;3:377–382.
46. Hess A, Schmidt R, Franke HD. Technical modifications at the DT neutron generator for tumor therapy at Hamburg-Eppendorf. In: Schraube H, Burger G, editors. Proceedings of the Fifth Symposium on Neutron Dosimetry EUR 9762EN. Volume II, Luxembourg: Commission of the European Communities; 1985. pp. 1019–1026.
47. Forman JD, Warmelink C, Sharma R, Yudelev M, Porter AT, Maughan RL. Description and evaluation of 3-dimensional conformal neutron and proton radiotherapy for locally advanced adenocarcinoma of the prostate. Am J Clin Oncol 1995;18:231–238.
48. Maughan RL, Blosser HG, Powers WE. A superconducting cyclotron for neutron radiation therapy. Med Phys 1994;21: 779–785.
49. Maughan RL, Blosser GF, Blosser EJ, Yudelev M, Forman JD, Blosser HG, Powers WE. A multi-rod collimator for neutron therapy. Int J Radiat Oncol Biol Phys 1996;34: 411–420.
50. Farr JB, Maughan RL, Yudelev M, Forman JD, Blosser EJ, Horste T. A multileaf collimator for neutron radiation therapy. In: Marti F, editor. Cyclotrons and Their Applications 2001. Melville, NY: American Institute of Physics; 2001. pp. 154–156.

See also BRACHYTHERAPY, HIGH DOSAGE RATE; IONIZING RADIATION, BIOLOGICAL EFFECTS OF; NUCLEAR MEDICINE INSTRUMENTATION; RADIOTHERAPY, HEAVY ION.

NEUROSTIMULATION. See SPINAL CORD STIMULATION.

NMR. See NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY.

NONIONIZING RADIATION, BIOLOGICAL EFFECTS OF

ANDREW WOOD
Swinburne University of Technology
Melbourne, Australia

INTRODUCTION

Non-ionizing radiation (NIR) refers to that portion of the electromagnetic (EM) spectrum in which the characteristic wavelength is greater than around 180 nm. Radiation of shorter wavelength than this has sufficient quantum energy (given by hc/λ, with h = Planck's constant, c = wavespeed in vacuo, and λ = wavelength) to remove outer electrons from neutral atoms to cause the atom to become ionized, hence, the term "ionizing radiation." NIR consequently does not have the same intrinsic potential for atomic and molecular alteration or the health effects consequent to this. For this reason, damage to DNA and other biomolecules due specifically to the removal of electrons is difficult to envisage. The main groupings of NIR, with increasing wavelength (and decreasing frequency) are ultraviolet (UVR), visible, infrared (IR), radio frequency (RF), and extremely low frequency (ELF). The RF spectrum can be further divided as shown in Table 1 to include microwaves (MW), millimeter waves (MMW), terahertz radiation (THzR), as well as the conventional divisions for broadcast communications. Although not part of the EM spectrum, UVR is normally considered to be part of NIR, as are static (0 Hz) electric, and magnetic fields. The application of the term "radiation" to the ELF portion is also of little consequence, because the wavelength is several thousand kilometers at 50/60-Hz power frequencies.

NON-IONIZING RADIATION PROTECTION

Guidelines on NIR radiation protection are developed by the International Commission on NIR Protection (ICNIRP). In North America, other bodies have developed standards, such as the IEEE International Committee on Electromagnetic Safety and the American National Standards Institute (ANSI), or guidelines, such as the American Conference of Government Industrial Hygienists (ACGIH). Some jurisdictions have chosen to incorporate these (or related) guidelines into legislation.

The mechanism of interaction of NIR with living tissue varies with the groupings just mentioned. These are summarized below, along with effective protection measures against overexposure.

UVR

UVR exposure from the sun outweighs that from all other sources except for a small group of persons in exceptional circumstances. Solar UVR over-exposure is a worldwide problem, leading to increased skin cancer, and by World Health Organization estimates, up to 3 million people are made blind through cataracts. Burning of the skin is a direct indicator of overexposure, at least in the short term. Solar radiation and other UVR sources can initiate photochemical reactions, such as the breakdown of atmospheric oxygen to form oxygen-free radicals and ozone. UVR also has a role in vitamin D control and production. Of greater relevance to adverse health effects, biomolecules (such as DNA components and proteins) can undergo resonant UVR absorption to give rise to dimers (where two similar molecules join to form a single unit). For example, adjacent thymine bases in DNA can fuse to cause an abnormal form. The cell repair mechanisms can sometimes fail to detect this, leading to mutations. The initial response of the skin to UVR within hours of exposure is reddening (erythema or sunburn) due to increased blood flow and edematous changes. The role of photochemical reactions in erythema is unclear. In addition, the immune response can also be suppressed by UVR, increasing risk of infection. On the other hand, the socially attractive tanning of the skin is caused by UVR-induced increase in melanin pigmentation. Chronic exposure leads to skin aging and increased risk of skin cancer. Non-melanoma skin cancers (NMSC) include basal cell carcinoma (BCC: 80%) and squamous cell carcinoma (SCC). The risk of NMSC varies with annual solar UVR dose to the power of between 2 and 3. Melanoma, which has a poor prognosis due it its ability to metastasize, is related to the amount of sun exposure or sunburn during childhood. Chronic eye exposure leads to increased cataract risk. Certain pharmaceutical and other agents lead to photosensitization, in which absorption of longer wavelength UVR can lead to resonant absorption usually associated with shorter wavelengths. The UVR range is usually divided into UV A, B, and C, as indicated in Table 1. The rationale for this is that (1) biological photoreactions are less important above 315–320 nm, and (2) there is virtually no terrestrial solar radiation below 280–290 nm. The boundaries between the ranges are somewhat imprecise. UVA has less capability to cause erythema (by a factor of around 1000) than UV B, but because UVA radiation is the predominant form of solar radiation, it contributes around one sixth of erythemal dose. A minimum erythemal dose (MED) is the UVR exposure (in joule per centimeter squared), which gives rise to just noticeable reddening in the skin of previously unexposed persons. Overexposure is defined as that which leads to erythema within 3 hours or less in a normal population. MEDs have been determined experimentally for narrow bandwidths in the range 180–400 nm, giving a minimum of $30\ \mathrm{J{\cdot}m^{-2}}$ at 270 nm. A set of values $S\lambda$, which denote the relative effectiveness of UVR to cause erythema at a specific wavelength λ, are then derived. For example, because at 180 nm, $2500\ \mathrm{J{\cdot}m^{-2}}$ is required for the occurrence of erythema compared with $30\ \mathrm{J{\cdot}m^{-2}}$ at 270 nm, S_{180} is 30/2500 or 0.012. As exposures are usually to a range of wavelengths (and mainly in the UVA range), a weighted sum for each wavelength component according to its capacity to cause erythema can be obtained. The standard erythemal dose (SED) is then defined such that 1 SED is $100\ \mathrm{J{\cdot}m^{-2}}$. This measure is independent of skin type, because MED measurements relate to fair-skinned subjects. Most commonly, overexposure is a result of being outdoors without skin protection, but it can also result from artificial sun-tanning

Table 1. The Non-ionizing Radiation Spectrum

Name of Range	Frequency Range	Wavelength Range	Common Sources
Ultraviolet	UVC 1.07–3 PHz[a]	100–280 nm	Germicidal lamps, Arc welding
	UVB 0.95–1.07 PHz	280–315 nm	Solar radiation, Arc welding
	UVA 750–950 THz[b]	315–400 nm	Solar radiation, Solarium
Visible	430–750 THz	400–770 nm	Solar radiation, indoor and outdoor illumination
Infrared[c]	Near IR (IR A) 214–430 THz	0.7–1.4 μm	Furnaces
	Mid IR (IR B) 100–214 THz	1.4–3 μm	Night photography
	Far IR (IR C) 0.3–100 THz	3 μm–1 mm	Infrared spectroscopy
Terahertz			
Microwave (including millimeter wave)	Extremely High Freq 30–300 GHz	1 mm–1 cm	Satellite, radar, and remote sensing
	Super High Freq 3–30 GHz	1–10 cm	Speed radar guns, Communications
	Ultra High Freq 1–3 GHz	10–30 cm	Mobile telephony
Radio frequency	Ultra High Freq 0.3–1 GHz	30 cm–1 m	Mobile telephony
	Very High Freq 30–300 MHz	1–10 m	TV, FM Radio Broadcasting
	High Freq 3–30 MHz	10–100 m	Electro-welding equipment
	Medium Freq 0.3–3 MHz	100 m–1 km	AM Radio
	Low Freq 30–300 kHz	1–10 km	Long-wave radio
	Very Low Freq 3–30 kHz	10–100 km	Navigation and time signals
Extremely low frequency	< 3 kHz	> 100 km	Electrical power, Electrotherapy
Static	0 Hz		Geomagnetic field, Magnetic Resonance Imaging systems

[a] PHz = peta-Herz, or 10^{15} Hz

[b] THz = tera-Herz, or 10^{12} Hz

[c] The boundaries between near-, mid-, and far-IR are imprecise, as is the terahertz range indicated.

(in a solarium), proximity to tungsten halogen lamps (without filtering glass covers), proximity to UVR lightboxes in scientific and industrial applications, and certain forms of flame welding, with the main possibility of eye damage in these latter sources. Cases of erythema from fluorescent tubes have been reported in extreme cases of photosensitization. The main forms of protection are wearing appropriate clothing, sunblocks (such as zinc oxide cream), sunscreens (based on photo-absorbers, such as para-amino-benzoic acid and cinnamates), and effective sunglasses. Staying out of the sun where this can be avoided is a good behavioral approach for exposure minimization. The "sun protection factor" (SPF) is effectively the ratio of time of exposure before erythema occurs in protected skin to the corresponding time in unprotected skin. A ratio of at least 30 is recommended for effective protection in recreational and occupational exposure to solar radiation. It is important to ensure that sunglasses have sufficient UVR absorption to protect against cataract. Various forms of clothing protect against UVR exposure to differing degrees, ranging from wet open-weave cotton, which offers an ultraviolet protection factor or UPF (which is analogous to SPF) of only around 3–6, to elastane (Lycra) with UPF values of around 100 (99% absorption). It should be noted that these protection factors are computed as the ratio of effective dose (ED) with and without protection (ED/ED_m). The ED is the sum of solar spectral radiance components weighted according to erythemal effectiveness. Here $\mathrm{ED} = \sum E_\lambda S_\lambda \Delta\lambda$, where E_λ is the solar spectral irradiance in watt per centimeter squared per nanometer, S_λ is the relative effectiveness of UVR at wavelength λ causing erythema (as mentioned), and $\Delta\lambda$ is a small bandwidth in nanometers. The units of ED are watt per centimeter squared. ED_m is similar, but it contains a factor T_λ to denote the fractional transmission of the test sunscreen (cream, fabric) at a particular wavelength (i.e., $\mathrm{ED}_m = \sum E_\lambda S_\lambda T_\lambda \Delta\lambda$). The Global UV Index (UVI) is a dimensionless quantity in which the ED is summed over the range 250—400 nm and multiplied by 40 $\mathrm{m}^2\mathrm{W}^{-1}$. In Darwin, Australia, this ranges from 0 to 3 in the early morning and evening to 14 or more at noon on a clear day. At this UVI, erythema will result in fair skin after 6 minutes. See http://www.icnirp.de/documents/solaruvi.pdf for further information.

It is estimated that significant reductions in the incidence of both malignant and benign forms of skin cancer could be achieved by the enforcement of protective measures, particularly in occupational settings involving fair-skinned people in outdoor work in tropical or subtropical regions. Occupational exposures in Australia have recently been measured (1), and UVR safety has been reviewed in several publications (see Reference (2), for example). Indicative exposure limits are given in Table 2. It should be emphasized that for brevity many details are omitted from this table. For full details of limits pertaining to a particular geographical region, local radiation protection authorities should be consulted. The ICNIRP guidelines are readily accessible via downloads from http://www.icnirp.de. These represent reviewed publications originally appearing in *Health Physics*.

Visible Radiation

This is the region of NIR to which the retinal pigments of the eye are sensitive, so understandably, eye injury is the main concern in overexposure. There are two forms of hazard: photochemical and thermal. In addition, if the eye lens has been surgically removed (aphakia), there is an enhanced risk of damage. Photochemical damage becomes more likely with shorter wavelengths and is sometimes referred to as the "blue light hazard." The type of photochemical reaction is bleaching of the visual pigments, leading to temporary loss of vision. Thermal injury can result in permanent impairment of vision, especially if the foveal region, used for fine focus, is involved. Thresholds for these forms of injury have been determined in the wavelength range 400–1400 nm (thus including near infrared, see below) and an assessment of whether these are exceeded, for a particular source takes into consideration the spectral characteristics of the source. For exposures shorter than a few hours, the total radiance should be below $10^6\ \mathrm{W{\cdot}m^{-2}{\cdot}sr^{-1}}$, where sr refers to a unit solid angle tended by the source. Lasers represent the sources most likely to cause injury, and because these emit a small number of discrete wavelengths, this assessment can be straightforward. Eye injury is minimized by the blink reflex, but laser wavelengths outside the visible range are less easy to control, because their paths are difficult to track, especially from incidental reflections. Lasers are classified according to the luminous power, their visibility, and their effective aperture, as described further in a separate entry on **LASERS**. High-power lasers are used in machining, welding, and engraving of a variety of materials, including plastics, metals, and fabrics. They also provide the source of beams in communications and photonics research laboratories. During normal operation, a combination of administrative and engineering controls provide adequate protection for workers. On the other hand, high-power lasers used in "light show" entertainment have sometimes given rise to unintentional beams directed at members of the public. The unrestricted distribution of laser pointers, with a capacity of causing eye damage, has also been a concern in several jurisdictions. Apart from laser sources, welding flames represent the next most common form of visible light hazard ("welder's flash"). Hazard can be minimized by the use of appropriate goggles. Recently, high-powered light-emitting diode (LED) sources have been evaluated by the ICNIRP for their potential for visible light hazard, particularly those emitting blue light. Although injury is unlikely, the power density of these devices continues to increase as technology develops.

IR

The major sources of IR radiation that are of concern are furnaces and some high-powered non-visible laser devices (femtosecond lasers). Here there is an increased possibility of local thermal injury, but because there is poor penetration of the lens of the eye, the possibility of retinal damage is reduced compared with the visible range. The IR range is divided into three ranges as shown in Table 1. Above 1–2 μm, water is a strong absorber of IR. Whereas guidelines for optical radiation extend up to 1.4 μm (near

Table 2. Approximate Exposure Limits for NIR: Exact Limits Vary Between Countries and In Some Cases Between Different Contexts of Exposure

Name of Range	Indication of Level Above Which Intervention Is Recommended	Biohazard Forming Basis of Protection	References to *Health Physics* Publications
Ultraviolet	U-shaped over wavelength range: 180 nm–2.5 kJ·m^{-2}; 270 nm–30 J·m^{-2} (minimum); 400 nm–1 MJ·m^{-2}	Skin reddening due to burn (erythema), also prevention of cataract	Vol 71, p 978 (1996) Vol 84, pp 119–127 (2004)
Visible	Depends on viewing position and spectral content of source	Retinal thermal or photochemical damage	Vol 73, pp 539–554 (1997)
Lasers (includes above and below)	Depends on wavelength, exposure duration, and size of aperture. For long exposures (> 100 s), limits are of the order of 1 W·m^{-2}	Retinal (esp. foveal) damage: photochemical or thermal Also skin.	Vol 71, pp 804–819 (1996) Vol. 79, pp 431–440 (2000)
Infrared	100 W·m^{-2} for long exposure* Not well defined	Thermal injury to lens and cornea	Vol 73, pp 539–554 (1997)
Terahertz			
Microwave (including millimeter wave)	6–300 GHz: 50 W·m^{-2} (time averaged) 50 kW·m^{-2} peak[a]	Rise in tissue temperature sufficient to cause protein denaturation	Vol. 74, pp 494–522 (1998)
	10 mJ·kg^{-1} within 50 μs interval[a]	Microwave hearing	As above
Radio frequency	0.1–6,000 MHz: 0.4 W·kg^{-1} for whole-body exposure; 10 W·kg^{-1} for 10 g mass (head and torso)[a].	Rise in tissue temperature sufficient to cause protein denaturation	As above
	3–10,000 kHz: f/100 (f in Hertz) mA·m^{-2} in head and torso[a]	Shocks or burns due to induced current or contact current	As above
Extremely low frequency	Tissue induced field: 18 mV·m^{-1} for f, 20 Hz; 18(f/20) mV·m^{-1} for f between 20 & 800 Hz (IEEE)[b], 10 mA·m^{-2} for range 4–1,000 Hz (ICNIRP)	Magnetophosphenes, micro-shock	As above
Static	0.2 T time weighted average[a], 2 T ceiling, 5 T limbs	Magnetophosphenes associated with movement	Vol 66, pp 100–106 (1994)

[a] These basic restrictions are for occupational exposures: Divide by 5 to get general public limits.

[b] These basic restrictions are for “controlled environment” (i.e., occupational) exposures: Divide by 3 to get general public limits.

infrared), there is some disagreement on the appropriate levels beyond that. Levels of incident radiation above 100 W·m^{-2} are considered as posing an unacceptable thermal hazard. Those at risk of overexposure include foundry workers and welders. Recently, advances have extended telecommunications frequencies into the “terahertz gap,” the region between 0.3 and 3 THz, which has been unexploited by technological applications. The health effects are currently unknown, but they are expected to be similar to those of the contiguous frequency ranges. However, there is a current discontinuity between IR and RF standards or guidelines for a 1 mm wavelength (0.3 THz).

RF

Common sources of high-power RF emissions include welding equipment and induction heaters used in industrial drying processes. Radio, TV, and telecommunications transmitters can involve high broadcast powers (400 kW or more for commercial TV stations). There are two types of potential hazard: thermal injury in the range 100 kHz–300 GHz and neural stimulation due to induced currents or contact with metallic surfaces at frequencies below 10 MHz. At 300 GHz, the effective wavelength in tissue is less than 1 mm, so very little will penetrate below the skin. On the other hand, at 80 MHz, the wavelength is comparable with the long axis of the human body, so absorption is enhanced. Protective measures in terms of incident RF power density (W/cm^{-2}) are thus strictest in the range 10–400 MHz. The basic restriction above 100 kHz is on the rate of energy absorption by tissue (specific absorption rate, or SAR, in W/g of tissue). SAR is related to the RF electric field induced in tissue (E_i V·m^{-1}) such that

$$\mathrm{SAR} = \sigma E_i / \rho$$

where σ is local conductivity in S/m and ρ is tissue density in kg/cm^3. In unperfused insulated tissue, SAR is related to the rate of rise of temperature dT/dt via

$$\mathrm{SAR} = k \cdot dT/dt$$

where k is the specific heat of tissue, 3480 J·kg^{-1}·K^{-1} approximately.

This basic restriction is limited to values for whole-body or localized exposures such that normal thermoregulation would not be compromised, with a 10-fold safety margin. Although there is some variation between standards in place throughout the world, many countries employ a distinction between occupationally exposed persons ("aware users") and the general public, for whom an extra five-fold level of protection is provided. The ICNIRP value for whole-body SAR for the general public is 0.08 $\mathrm{W \cdot kg^{-1}}$, with higher values of 2 $\mathrm{W \cdot kg^{-1}}$ in the head and trunk and 4 $\mathrm{W \cdot kg^{-1}}$ in the limbs, averaged over 10 g of tissue. The power density of incident plane-wave radiation (in watt per centimeter squared), which would give rise to these levels of SAR (for far-field exposures), has been computed by mathematical modeling and animal studies in a conservative manner, such that if these reference levels are complied with, the basic restrictions will be met. As, for free space, the power density S is related to the electric and magnetic field values (E and H, respectively) by $S = E^2/377 = H^2 0.377$, compliance testing can be accomplished by measuring E-field values alone. Reference levels at particular frequencies can be found by reference to the ICNIRP guideline as indicated in Table 2.

Induced current density restrictions are imposed at 10 MHz and below. Above this frequency, it is considered that the fields vary too quickly to produce neural stimulation. Again, there is a safety factor of 10 between occupational levels and the level at which mild stimulatory effects can be noted in 1% of the population. This ranges from 100 $\mathrm{A \cdot m^{-2}}$ at 10 MHz to 10 $\mathrm{mA \cdot m^{-2}}$ at 4 Hz–1 kHz, in the ICNIRP guidelines. This will be discussed further in the ELF section.

At frequencies between 0.2 and 6 GHz, a phenomenon of "microwave hearing," due to thermoelastic expansion of brain tissue in response to pulsed radiation, occurs. Additional restrictions are in place in the ICNIRP guidelines to prevent this from occurring.

Overexposure to RF radiation, leading to serious burns, is usually due to the failure of control measures, such as guards on RF seam welding apparatus or work on RF antennas mistakenly thought to be nonoperational.

The safety of communications equipment, including mobile telephony handsets and base stations, is a major community concern. There is little substantive evidence of harm from long-term exposure at so-called "non-thermal" levels, but because there are many young users of handsets, many countries have endorsed a precautionary approach, encouraging use only for necessity. The scientific evidence for the possibility of "non-thermal" effects has been reviewed in the United Kingdom by the Independent Expert Group on Mobile Phones (IEGMP) (3) and by other bodies. The IEGMP concluded that although "the balance of evidence to date suggests that (low levels of RF radiation) do not cause adverse health effects" that "gaps in knowledge are sufficient to justify a precautionary approach." Some national standards (for example, Australia and New Zealand) incorporate a "precautionary" clause; that is, exposures incidental to service delivery should be minimized (but taking other relevant factors into consideration). The limiting of mobile phone use by children was recommended by the IEGMP (3), but the Health Council of the Netherlands sees no convincing scientific argument to support this (4).

ELF and Static

The range of frequencies (0–3 kHz) includes power transmission and distribution systems (50/60 Hz) as well as transportation systems (0, 16.7, 50, and 60 Hz), surveillance systems, and screen-based visual display units. Here the main potential hazard from exposure to fields (rather than direct contact with conductors) seems to be from inappropriate neural stimulation due to induced current (as in the case of RF, above). Consequently, treating ELF as a special case may seem out of place, but because the ELF range is precisely that of biogenic currents due the operation of nerves and muscles, its separate treatment is justified. The susceptibility of cells to the influence of exogenous currents is related to the time constants for the operation of cell membrane channels, which are typically of the order of milliseconds. At lower frequencies, cell membranes tend to adapt to imposed electrical changes, so restrictions need to be strictest in the range 10–1000 Hz. In humans, the retina of the eye represents a complex network of interacting nerve-cells, giving rise to sensations of pinpoints of light when stimulated by external electric and magnetic fields (EMFs). As this gives a guide to the levels at which stimulatory effects could become an annoyance, or could possibly be interpreted as a stressor, a basic restriction for occupational exposure of 10 $\mathrm{mA \cdot m^{-2}}$ (which corresponds to an induced field of around 100 $\mathrm{mV \cdot m^{-1}}$) has been adopted by the ICNIRP for the range 4–1000 Hz. This restriction rises above and below this range. In particular, at 0 Hz (static fields), levels are restricted to 40 $\mathrm{mA \cdot m^{-2}}$. Levels for the general public are less by a factor of 5. Reference levels for magnetic fields are derived from these basic restrictions by considering the body to be simple geometric objects, but more advanced modeling yields similar results. For sinusoidally varying fields, the reference magnetic fields can be derived from basic restrictions via the formulas

$$B = E/(\pi f r) \quad \text{or} \quad B = J/(\sigma \pi f r)$$

where E refers to the basic restriction in terms of induced tissue electric field (in volt per meter), J is the basic restriction in induced current density ($\mathrm{A/cm^2}$), f is the frequency in Hertz, σ is the tissue conductivity (S/m), and r is the radial distance from the center of symmetry (in the same direction as the external magnetic field B).

Electric field reference levels are derived more from considerations of avoiding "microshocks," which may occur, for example, if an arm with finger extended is raised in an intense electric field. Details of these reference levels can be found (for the ICNIRP limits) at http://www.icnirp.de. As it is possible to exceed the electric field reference levels in electrical switchyard work, special precautions need to be taken. Exceeding magnetic field reference levels is rare. Some government and other organizations have advocated a much more prudent approach to limiting exposure, particularly to the general public. This comes from some dozen or so well-conducted epidemiological studies linking exposure of children to a time-weighted average magnetic field of 0.4 μT or more, to an approximate doubling of leukemia incidence. The possibility of low-level health effects of ELF has been the topic of research for nearly three decades. As there is no agreed mechanism for how elevated leukemia rates could be brought about, nor is there adequate evidence from long-term animal studies, there is doubt that magnetic fields are the causative agent. Nevertheless, time-varying ELF magnetic fields (but not electric fields, nor static fields) have been categorized by the International Agency for Research in Cancer (IARC) as a "possible carcinogen" (category 2B) (5). Essentially, the U.S.-government funded EMF-RAPID (Electric and Magnetic Field Research and Public Information Dissemination) program, whose Working Group reported in 1998 (6), came to a similar conclusion. The final report of the NIEHS Director (7), on the other hand, concluded that "the scientific evidence suggesting that ELF-EMF pose any health risk is weak" but also acknowledged that "exposure cannot be recognized as entirely safe because of weak scientific evidence that exposure may pose a leukemia hazard." The report also advocated "educating both the public and regulated community on means aimed at reducing exposures." There is intense debate on how a policy of prudence should actually be interpreted, because approximately 1% of homes would be in the "over 0.4 μT" category (8,9) (this percentage varies widely between and even within countries). Several moderate cost engineering measures can be employed to reduce field levels from transmission lines, and electric power companies often employ these in new installations.

PERCEIVED ELECTRO-SENSITIVITY

Several persons claim debilitating symptoms associated with proximity to electrical installations or appliances or in association with the use of mobile (cell) phones. Despite several well-conducted, independent, "provocation studies," in which sufferers have been subjected to energized and not energized sources in random order, no association between exposure status and occurrence of symptoms has been established. A recent Dutch study of psychological sequelae of mobile phone use implied that the overall baseline responses in a group of "electro-sensitives" differed from a similarly sized group of "normals," but that the changes associated with mobile phone use were similar in both groups.

ULTRASOUND

Few processes and devices outside of clinical medicine involve the possibility of human exposure to ultrasound if normal protective guarding measures are in place. Airborne ultrasound is used in surveying instruments and in a variety of drilling, mixing, and emulsification industrial processes. Ultrasonic descalers are used in dentistry and to clean jewelry. Reports of injury are rare. For industrial applications, the frequency range of 20–100 kHz is covered by ICNIRP limits and is based on the pressure amplitude of the ultrasound in air (these are of the order of 110 dB, referenced to 2×10^{-5} Pa). In clinical applications, ultrasonic energy is usually delivered across the skin via coupling gel and is in the frequency range 1–25 MHz. Diagnostic ultrasound is designed to prevent tissue temperature rising above 41 °C for sustained periods (10,11). Effectively, beam intensities are capped at 1000 $\mathrm{W{\cdot}m^{-2}}$ (spatial peak, temporal average), except for short periods of insonation. Higher intensities are possible if the energy density is below 500 $\mathrm{kJ{\cdot}m^{-2}}$. This gives a large margin below established hazardous effects. Therapeutic ultrasound exposure is usually limited by patients reporting excessive heat, but use on patients with limited sensation is of concern. Intensities of 10 $\mathrm{kW{\cdot}m^{-2}}$ are common in therapeutic applications. Tissue damage occurs above 10 $\mathrm{MW{\cdot}m^{-2}}$.

SERIOUS INJURY FROM NIR

From above, it would appear that NIR is fairly innocuous. It should be stressed, however, that high-power devices, if inappropriately used or modified, can cause serious injury. UVC is routinely used as in germicidal devices, and the micro-cavitation produced by intense ultrasound beams is used to disrupt tissue. Laser skin burns occasionally occur in research laboratories. Severe injury and fatalities have resulted from surgical uses of lasers in which gas embolisms have become ignited within body cavities. Early unshielded microwave ovens were associated with severe kidney damage. Cases of severe burns are still too common in small businesses using RF heat sealers, often due to the removal of guards. Serious burns result from an accidental or ill-advised approach to broadcast antennas and other communications equipment (12). In addition to burns, severe chronic neurological deficits can also result from overexposure to RF currents (13).

ACHIEVING ADEQUATE PROTECTION AGAINST NIR

Opinion is divided about the need to control NIR exposure by legislation. Communications equipment manufacturers have to comply with rigid requirements related to health guidelines and standards, and many countries have the power to prosecute in instances where equipment is tampered with or altered such that the guidelines would be exceeded. Codes of practice often have provisions for marking "no go" areas where levels could be exceeded, with appropriate signage. In terms of the potential for preventing debilitating illness or early death, the link between solar UVR and skin cancer and cataract represents the area where intervention is most warranted. It is estimated that adequate sun protection could perhaps save tens of lives per million of population per annum with over $5M pa per million in savings in health costs. The costs of ensuring employers of outdoor workers and the workers themselves complying with measures of UVR exposure reduction are hard to estimate, but they are likely to be high. Whereas compliance with a limit of 30 $\mathrm{J{\cdot}m^{-2}}$ equivalent (or MED) is achievable in relation to artificial sources, this level can be exceeded in less than an hour's exposure to intense solar radiation around noon in low latitudes. Employers can be required to educate their workforce to use appropriate measures to reduce the risk of becoming sunburnt, but it is virtually impossible to eliminate this from actually occurring. It would seem unreasonable to require employers to be responsible for an overexposure to a familiar and essential source of energy to which we have all been exposed since the dawn of time.

As several forms of NIR carry with them an uncertainty of possible harm in the long term, several national radiation protection authorities have espoused the "Precautionary Principle." This entails taking measures to reduce exposure, even where exposures are well within levels set by scientific evaluation of the available research. It is recognized that reducing exposure might itself introduce new hazards or increase other hazards (such as being unable to use a cell-phone in an emergency because of extra power restrictions), so an evaluation of the need to be "Precautionary" with respect to NIR should be in the wider context of overall risk management. In general, the introduction of arbitrary extra margins of safety, in order to appease public outcry, is not warranted.

USES OF NIR IN MEDICAL DIAGNOSIS AND THERAPY

UVR

The UVR-induced photochemical reactions form the basis of an effective treatment of the disease psoriasis, which is marked by widespread red itchy scales on the skin. This is caused by an accelerated cell cycle and DNA synthesis in skin cells. The drug psoralen is preferentially taken up by these dividing cells, which on subsequent exposure to UVA radiation, leads to binding with DNA and subsequent inhibition of synthesis and cell division. A normal course of treatment consists of 25 monthly visits to a clinic, with 8-methoxypsoralen taken orally, followed 2 h later by a UVR exposure of 10–100 $\text{kJ}\cdot\text{m}^{-2}$ per visit. This is usually delivered via a bank of 48 or so high-intensity fluorescent tubes.

A second use of UVR in biological and clinical analysis and research is in the identification of biomarkers through fluorescence. One technique involves placing electrophoretic gels over a UVR lightbox to localize the fluorescent regions. As mentioned, the possibility of overexposure in those who perform multiple observations is a matter of concern.

Lasers

The high intensity of laser radiation, particularly if it is pulsed, provides a means of tissue ablation, carbonization, coagulation, and desiccation. High-intensity short pulses produce photomechanical disruptions of tissue. At longer pulse lengths (~ 1 s), thermal and photochemical processes become more important. Excimer (= excited dimer) laser radiation has proved to be useful in the surgical treatment of defects in vision. This technique, radial keratotomy or keratectomy, reshapes the corneal surface to alter the effective focal length of the eye and thus do away with the need for spectacles or contact lenses. Laser ablation is also useful in the treatment of ocular melanoma, Barratt's esophagus, removal of "port wine" stains on the skin, and (using an optical fiber delivery system in a cardiac catheter) the removal of atheromatous plaque in coronary arteries. A second property of intense laser light, that of photo-activation, is exploited in a range of treatments known as photodynamic therapy (PDT). In this, several compounds are known to be preferentially taken up by tumor tissue but also have the property of resonant absorption of light to produce free radicals, such as singlet oxygen and oxygen radical, which ultimately lead to endothelial cell membrane damage, blood supply shutdown, and hence necrosis of tumor tissue. These photosensitizing compounds are injected, or in some cases taken by mouth. Intense laser light (of 600–770 nm wavelength) is then directed at the tumor to produce this photo-activation. Energy thresholds are of the order of 1 $\text{MJ}\cdot\text{m}^{-2}$. Although used mainly on superficial tumors (depth less than 6 mm), optical fiber delivery into deeper tissue (such as the breast) has also been trialled. As the tumor tissue becomes fluorescent on uptake of these compounds, diagnostic techniques (photodynamic diagnosis or PDD) are based on a similar principle. Suitable compounds are related to hemoglobin (hematoporphyrin derivative or HpD), rhodamine, amino levulinic acid, bacteriochlorins, and phthalocyanines. The herb St John's Wort also yields hypericins that have similar properties. The ability to use scanning optics in association with optical fibers has provided ways of making microscopic endoscopy possible.

Incoherent sources of blue light are used in the treatment of neonatal jaundice (hyperbilirubinemia). Bilirubin is decomposed during the exposure of the neonate to fluorescent tubes (filtered to remove wavelengths shorter than 380 nm).

IR

Infrared reflectivity from the skin and from layers immediately below the skin varies with skin temperature. Thermography has been used to identify regions of enhanced or reduced peripheral blood flow, occurring, for example, in mammary tumors. The high false-positive rate has inhibited its use in mass screening for this disease. On the other hand, breast imaging using time-of-flight IR transmission methods shows promise. Blood oxygen saturation is easily measured noninvasively via the ratio of reflectances at two wavelengths, 650 and 805 nm (the wavelengths showing greatest and least sensitivity to the degree of saturation, respectively). This forms the basis of the pulse oximeter, which clips on the finger and gives an indication of pulse rate in addition to oxygen saturation. Laser Doppler blood flow meters give an indication of capillary blood flow via the autocorrelation of reflected light signals. Wavelengths of 780 nm are selected because of the good depth of penetration of skin.

IR spectroscopy has a wide range of industrial and research applications, because of specific molecular stretching, bending, and rotational modes of energy absorption.

Terahertz

Several medical applications have been proposed for terahertz radiation, arising out of differential reflection from cancerous/normal skin and from its relatively good transmission through bones and teeth. Its use in biosensing is also being investigated.

RF

The tissue heating and consequent protein denaturation has been used in catheter-tip devices for ablating accessory conduction pathways in the atria of the heart, giving rise to arrhythmias. The use of focused RF in cancer hyperthermia treatment has been used in conjunction with conventional radiotherapy to improve the hit rate of the latter, most likely due to the increased available oxygen via thermally induced blood flow increase. Increased blood perfusion is also thought to underlie the use of RF diathermy in physiotherapy, although this has now been almost entirely replaced by therapeutic ultrasonic diathermy (see below). RF exposures are part of magnetic resonance imaging (MRI), where some care has to be taken to avoid "hot spots" during investigation. SARs can exceed 2 $W{\cdot}kg^{-1}$ at frequencies in the region of 100 MHz. If we can extend the term "radiation" to include the direct application of RF currents, then electrical impedance tomography (EIT) should be included. In this technique, current of approximately 50 kHz is applied via a ring of electrodes to the torso or head, essentially to identify differential conductivity values in different organs and thus track shifts in fluid content, post-trauma, for example.

ELF

In clinical diagnosis, nerve conduction and muscular function studies are performed by examining responses to electrical stimulation (by single pulses or trains of pulses of the order of a few milliseconds in duration) of particular groups of nerve fibers. Electrical stimulation of specific regions of the body are also reported to give rise to beneficial effects. For example, or transcutaneous electrical nerve stimulation (TENS) is of some efficacy in controlling pain by raising the threshold for pain perception. Interferential therapy, which consists of a combined exposure of regions of the skin to low currents at two narrowly separated frequencies (for example, 4 kHz and 3.7 kHz) are claimed to be useful for a range of muscular and joint pain conditions and for circulatory disorders, but the mode of interaction is unclear. The currents are of the order of 50 mA, and the tissue is reportedly performing a demodulation of the 4 kHz carrier to produce a TENS-like deep current of a few hundred Hertz. Similarly, pulsed magnetic fields (PEMFs) are claimed to be effective in speeding healing in bone fractures, despite the small magnitude of induced currents. On the other hand, electroconvulsive therapy (ECT), in which pulses of current of several milliamperes are passed through the head, cause general nerve activation. This therapy is of proven value in cases of severe depression, but the origin of this benefit is an enigma. Transcranial magnetic stimulation (TMS) can be used both in diagnosis by eliciting specific responses and in therapeutic mode, in a manner analogous to ECT. However, the therapeutic efficacy of TMS still awaits clarification.

Static

The application of permanent magnets to painful joints is claimed to have beneficial effects, but the evidence for efficacy is equivocal. It has been suggested that the Lorentz-type forces on flowing electrolytes (such as blood) produce electric fields and currents. However, at typical blood flow velocities of 0.1 $m{\cdot}s^{-1}$, a 1 mT magnet will only induce 0.1 $mV{\cdot}m^{-1}$, which is well below levels shown in Table 2.

Ultrasound

The use of ultrasound in the range 1–25 MHz in diagnosis originates from the wavelength (and, hence, resolution) being of the order of a few millimeters. Acoustic mismatch between tissue layers gives radar-type echoes that form the basis of 2D and 3D imaging. The Doppler shift due to flowing fluid forms the basis of its use in blood flow measurements. Differential absorption provides a means for tissue characterization. In therapeutic ultrasound, the warmth is produced by adiabatic expansion and contraction within the tissue, to a depth of several centimeters. At higher intensities, cavitation and mechanical movement of organelles can occur.

BIBLIOGRAPHY

1. Vishvakarman D, Wong JC, Boreham BW. Annual occupational exposure to ultraviolet radiation in central Queensland. Health Phys 2001;81:536–544.
2. Diffey BL. Solar ultraviolet radiation effects on biological systems. Phys Med Biol 1991;36:299–328.
3. Stewart W. Mobile phones and health, Independent Expert Group on Mobile Phones, National Radiation Protection Board, Chilton, Didcot, U.K., 2000.
4. van Rongen E, Roubos EW, van Aernsbergen LM, Brussaard G, Havenaar J, Koops FB, van Leeuwen FE, Leonhard HK, van Rhoon GC, Swaen GM, van de Weerdt RH, Zwamborn AP. Mobile phones and children: is precaution warranted? Bioelectromagnetics 2004;25:142–144.
5. IARC, Non-ionizing radiation, Part 1: static and extremely low-frequency (ELF) electric and magnetic fields. International Agency for Research on Cancer, Lyon, France Monographs, Vol. 80, 2002.
6. Portier CJ, Wolfe MS. Assessment of health effects from exposure to power-line frequency electric and magnetic fields. National Institute of Environmental Health Sciences, Research Triangle Park, NC, NIH Publication 98-3981, 1998.
7. Olden K. NIEHS Report on health effects from exposure to power-line frequency electric and magnetic fields. National Institute of Environmental Health Sciences, Research Triangle Park, NC, NIH Publication No. 99-4493, 1999.
8. Greenland S, Sheppard AR, Kaune WT, Poole C, Kelsh MA. A pooled analysis of magnetic fields, wire codes, and childhood leukemia. Childhood Leukemia-EMF Study Group. Epidemiology 2000;11:624–634.
9. Ahlbom A, Day N, Feychting M, Roman E, Skinner J, Dockerty J, Linet M, McBride M, Michaelis J, Olsen JH, Tynes T, Verkasalo PK. A pooled analysis of magnetic fields and childhood leukaemia. Br J Cancer 2000;83:692–698.
10. Abramowicz JS, Kossoff G, Marsal K, Ter Haar G. Safety Statement, 2000 (reconfirmed 2003). International Society

of Ultrasound in Obstetrics and Gynecology (ISUOG). Ultrasound Obstet Gynecol 2003;21:100.
11. Safety statement, 2000. International Society of Ultrasound in Obstetrics and Gynecology (ISUOG). Ultrasound Obstet Gynecol 2000;16:594–596.
12. Hocking B, Joyner KJ, Newman HH, Aldred RJ. Radiofrequency electric shock and burn. Med J Aust 1994;161:683–685.
13. Schilling CJ. Effects of exposure to very high frequency radiofrequency radiation on six antenna engineers in two separate incidents. Occup Med (Lond) 2000;50:49–56.

Further Reading

Australian Radiation Protection and Nuclear Safety Agency: http://www.arpansa.gov.au
Dennis JA, Stather J, editors. Non-ionizing radiation. Radiation Protection Dosimetry. 1997. 72:161–336.
ICNIRP references from *Health Physics*: http://www.icnirp.de.
National Radiological Protection Board (NRPB) U.K.: http://www.nrpb.co.uk.

See also BIOMATERIALS, SURFACE PROPERTIES OF; IONIZING RADIATION, BIOLOGICAL EFFECTS OF; RADIATION THERAPY SIMULATOR.

NUCLEAR MAGNETIC RESONANCE IMAGING. See MAGNETIC RESONANCE IMAGING.

NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY

WLAD T. SOBOL
University of Alabama
Birmingham, Alabama

INTRODUCTION

When two independent groups of physicists (Bloch, Hansen, and Packard at Stanford and Purcell, Torrey, and Pound at MIT) discovered the phenomenon of nuclear magnetic resonance (NMR) in bulk matter in late 1945, they already knew what they were looking for. Earlier experiments by Rabi on molecular beams, and the attempts of Gorter to detect resonant absorption in solid LiF, seeded the idea of NMR in bulk matter. Fascinating stories describing the trials and tribulations of the early developments of NMR concepts have been told by several authors, but Becker et al. (1) deserve a special citation for the completeness of coverage.

For his achievements, Rabi was awarded a Nobel Prize in 1944, while Bloch and Purcell jointly received theirs in 1952. What was the importance of the Bloch and Purcell discoveries to warrant a Nobel Prize despite an abundance of prior work offering numerous clues? It was not the issue of special properties of elementary particles, such as a spin or magnetic moment: This was first demonstrated by the Stern–Gerlach experiment. It was not the issue of the particle interactions with magnetic field: This was first illustrated by the Zeeman effect. It was not even the magnetic resonance phenomenon itself: This was first demonstrated by Rabi. It was the discovery of a tool that offered a robust, nondestructive way to study the dynamics of interactions in bulk matter at the atomic and molecular level that forms the core of Bloch and Purcell's monumental achievements. However, despite the initial excitement at the time of their discovery, no one could have predicted just how extensive and fruitful the applications of NMR would turn out to be.

What is the NMR? The answer depends on who you ask. For Bloch's group at Stanford, the essence of magnetic resonance was a flip in the orientation of magnetic moments. Bloch conceptual view of the behavior of the nuclear magnetic moments associated with nuclear spins was, in essence, a semiclassical one. When a sample substance containing nuclear spins was kept outside a magnetic field, the magnetic moments of individual spins were randomly oriented in space, undergoing thermal fluctuations (Brownian motion). The moment the sample was placed in a strong, static magnetic field, quantum rules governing the behavior of the spins imposed new order in space: the magnetic moments started precessing around the axis of the main magnetic field. For spin ½ particles (e.g., protons), only two orientations w.r.t. static magnetic field were allowed; thus some spins precessed while oriented somewhat along the direction of the external field, while other spun around while orienting themselves somewhat opposite to the direction of that field. To Bloch, a resonance occurred when externally applied radio frequency (RF) field whose frequency matched the precessional frequency of the magnetic moments, forced a reorientation of precessing spins from parallel to antiparallel (or vice versa). They called this effect a nuclear induction.

As far as the Purcell's group was concerned, NMR was a purely quantum mechanical phenomenon. When a diamagnetic solid containing nuclei of spin I is placed in a static magnetic field, the interactions of nuclear magnetic moments with the external magnetic field cause the energy levels of the spin to split (the anomalous Zeeman effect). When an external RF field is applied, producing quanta of energy that match the energy difference between the Zeeman levels, the spin system would absorb the energy and force spin transitions between lower and upper energy states. Thus, they described the phenomenon as resonance absorption.

It can be proven that these two concepts of NMR phenomenon are scientifically equivalent. However, the two views are psychologically very different, and have been creating a considerable chasm in the accumulated body of knowledge. Some aspects of NMR applications are intuitively easier to understand using Bloch's semiclassical vector model, while other naturally yield themselves to the quantum picture of spin transitions among energy states. The details of this dichotomy and its impact on the field of NMR applications are fascinating by themselves and have been extensively discussed by Ridgen (2).

At the time of the NMR discovery, nobody had any inkling that this phenomenon might have any applications in medicine. To understand how NMR made such a big impact in the medical field, one has to examine how the NMR and its applications evolved in time. Nuclear magnetic resonance was discovered by physicists. Thus it is not surprising that the initial focus of the studies that followed was on purely physical problems, such as the structure of materials and dynamics of molecular motions in bulk matter. During a period of frenzied activities that followed the original reports of the discovery, it was very quickly understood that interactions among nuclear spins, as well as the modification of their behavior by the molecular environment, manifest themselves in two different ways. On the one hand, the Zeeman energy levels could shift due to variations in the values of local magnetic field at different sites of nuclear spins within the sample. This causes the resonant absorption curve to acquire a fine structure. Such studies of NMR lineshapes provide valuable insights into the structure and dynamics of molecular interactions, especially in crystals. This branch of NMR research is customarily referred to as radiospectroscopy.

On the other hand, when a sample is placed in the external magnetic field, the polarization of spin orientations causes the sample to become magnetized. When the sample is left alone for some time, an equilibrium magnetization develops. This equilibrium magnetization, $\boldsymbol{M}_0$, is proportional to the strength and aligned in the direction of the external static magnetic field, $\boldsymbol{B}_0$. An application of RF field disturbs the equilibrium and generally produces a magnetization vector, $\boldsymbol{M}$, that is no longer aligned with $\boldsymbol{B}_0$. When the RF field is switched off, the magnetization returns over time to its equilibrium state; this process is called a relaxation. The process of restoring the longitudinal component of the equilibrium magnetization requires that the spins exchange energy with their environment; thus, it is commonly referred to as spin–lattice or longitudinal relaxation. The characteristic time that quantifies the rate of recovery of the longitudinal component of magnetization toward its equilibrium value, $\boldsymbol{M}_0$, is called the spin–lattice relaxation time and denoted T_1 or, in medical applications, *T1*. At equilibrium, the transverse magnetization component is zero. Thus, any nonzero transverse component of nonequilibrium magnetization must decay back to zero over time. This process tends to be dominated by interactions among spins and is thus called a spin–spin or transverse relaxation. The characteristic time that quantifies the rate of decay of the transverse component of magnetization is called the spin–spin relaxation time and denoted T_2 or, in medical applications, *T2*. Both T_1 and T_2 strongly depend on the nature of the molecular environment within which the spins are immersed, thus offering a robust probe of molecular dynamics and structure in a variety of materials (solid, liquid, and gaseous) over a range of conditions (temperature, phase transitions, chemical reactions, translational and rotational diffusion, etc.). Studies of relaxation times are referred to simply as NMR relaxation studies, and sometimes as relaxometry.

In solids, dipole–dipole interactions among spins are dominant, which for proton NMR (^{1}H NMR) studies results in fairly wide lineshapes (with a width of several kHz) with very little fine structure. In most liquids, however, the substantially faster molecular reorientations average the dipole–dipole interactions, effectively suppressing them to produce a vanishing net effect on the NMR absorption curves that become much narrower (typically of the order of Hz). This feature has led to a discovery of chemical shift phenomenon.

The most dramatic demonstration of the chemical shift was the observation made in 1951 by Arnold et al. (3) who showed separate spectral NMR lines from nonequivalent protons in a sample containing a simple organic substance, ethanol. This gave birth to high-resolution NMR spectroscopy or HR NMR, a powerful tool that assists chemists in nondestructive analysis of organic compounds. This *in vitro* technique underwent massive developments over the years and almost overshadowed the NMR applications in physics. An exhaustive overview of HR NMR applications has been published by Shoolery (4). Today, HR NMR spectroscopy plays a major role in studies of biological materials *in vitro* and in drug development research. This research, although not directly used in clinical care, nevertheless is having a major impact on the development of medical arts. A comprehensive review of biological applications of NMR spectroscopy has been provided by Cohen et al. (5).

Standard NMR studies are performed *in vitro*: The sample is placed in the bore of a laboratory magnet, and the signal is collected from the entire volume of the sample. Samples are relatively small: The typical NMR tube vial is $\sim$ 5 mm outside diameter (OD) and holds $\sim$ 0.5 mL of sample material. Nuclear magnetic resonance magnets have relatively small active volumes [typical bore size of modern NMR cryomagnets is $\sim$ 70 mm inside diameter (ID)], but very high magnetic field homogeneity, routinely $> 10^{-9}\,\boldsymbol{B}_0$.

In the early 1970s, a revolutionary concept emerged from the pioneering work of Lauterbur in the United States and Mansfield, Andrew, and Hinshaw in the United Kingdom. They discovered that by using judiciously designed magnetic field gradients it was possible to retrieve an NMR signal from a small localized volume (called a voxel) within a much larger sample (e.g., a human body). This started a new field of NMR applications, called magnetic resonance imaging (MRI) that greatly enhanced the practice of diagnostic medicine (see the section **Magnetic Resonance Imaging**).

One of the frustrating limitations of MRI applications was the ambiguity of lesion characterization. The development of MRI focused on the noninvasive visualization of soft tissues within the living human body; as a result, the technical and engineering trade-offs made in the process of improving the quality of images have essentially rendered the method nonquantitative. In essence, MRI proved to be extremely sensitive in the detection of various lesions within the patient's body, but not very robust in providing information needed to fully identify the characteristics of the tissue within the lesion. Thus, in addition to basic NMR tissue characteristics (proton density, *T1*, and *T2*), the interpreters of medical MR images came to rely on morphology of lesions (size, shape, and location) to draw conclusions about lesion pathology. In this context, the concept of localized NMR spectroscopy experiments, where MRI techniques are used to locate the lesion and localize the volume of interest, while NMR spectroscopic techniques are used to acquire NMR spectra of the tissue within a lesion, becomes intuitively evident. However, while the concept may appear naturally obvious, implementations have proven to be extremely difficult. Despite first successful experiments in acquiring localized phosphorus ^{31}P NMR *in vivo* spectra from a human forearm, performed in 1980 by a group led by Chance, the true clinical applications of localized NMR spectroscopy have only recently begun to appear. While first attempts focused on ^{31}P NMR spectroscopy using surface coils to localize the signals within the human body, current clinical applications almost exclusively utilize ^{1}H NMR spectra to gain additional, clinically relevant information about the lesion of interest. The methodology used in this approach is referred to as magnetic resonance spectroscopy (MRS), magnetic resonance spectroscopic imaging (MRSI), or chemical shift imaging (CSI). Techniques of this particular application of the NMR phenomena in medical practice will be the subject of further discussion here. While the interest in exploring clinical applications of MRS of nuclei other than protons (e.g., ^{31}P, ^{13}C, ^{19}F, and ^{23}Na) still remains, a vast majority of current clinical applications uses ^{1}H MRS and thus only this particular nucleus will be considered in further discourse. Readers interested in other nuclei are encouraged to explore literature listed in the ***Reading List*** section.

THEORY

In this section, the quantum mechanical approach of formalism is used, since this formalism is most naturally suited to explain the various features of NMR spectra. To begin with, consider an ensemble of noninteracting protons, free in space where a strong, perfectly uniform magnetic field $\boldsymbol{B}_0$ is present. Because all spins are considered identical, the Hamiltonian of the system includes a single spin and all quantum mechanical expectation values are calculated over the entire assembly of spins. Under those conditions, the Hamiltonian ($\mathcal{H}$) describes the Zeeman interaction of the nuclear magnetic moment μ with the external magnetic field and has a form

$$\mathcal{H} = -\boldsymbol{\mu} \times B_0 = -g\hbar B_0 I_z \quad (1)$$

where γ is the gyromagnetic ratio, $\hbar$ is the Planck's constant, and I_z is the z component of the nuclear spin operator $\boldsymbol{I}$, which for protons has an eigenvalue value of ½. Because I_z has only two eigenvalues, ± ½, the system's ground energy level is split into two sublevels, with the energy gap proportional to $\boldsymbol{B}_0$, as shown in Fig. 1. Now assume that spins are allowed to weakly interact with their molecular environment, which are collectively described as the *lattice* (regardless of the actual state of the sample; e.g., in liquids the lattice refers to thermal diffusion, both rotational and translational, of the atoms or molecules that host the spins). When the system is left undisturbed over time, it will reach a thermal equilibrium, where the spin populations at the higher and lower energy levels are described by a Boltzmann distribution, as shown in Fig. 2. The resultant sample equilibrium magnetization is equal to

$$M_z^{\mathrm{eq}} = M_0 = \frac{g^2\hbar^2 N B_0}{4kT} \quad (2)$$

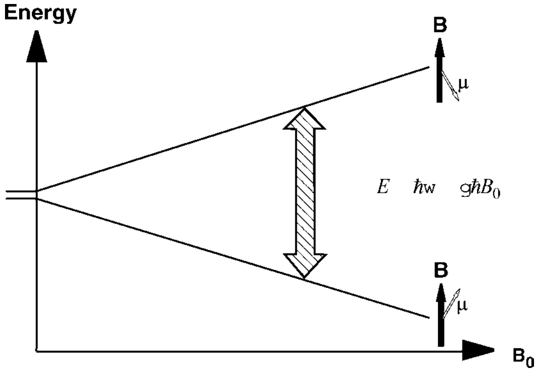


Figure 1. The Zeeman splitting of the ground-state energy levels for the spin ½ system as a function of the external magnetic field strength, B_0.

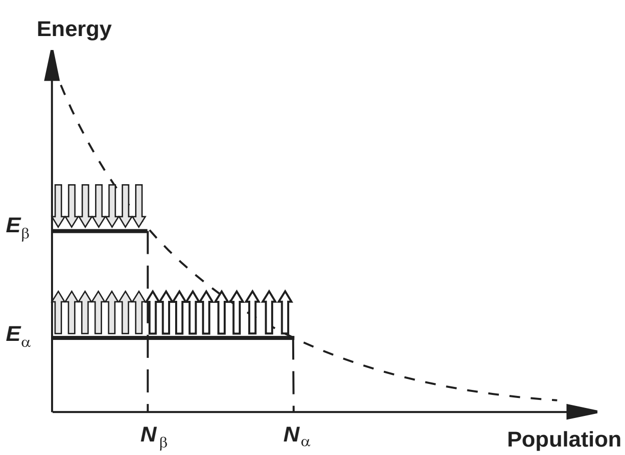


Figure 2. An illustration of Boltzmann distribution of spin populations for an ensemble of identical spins ½, weakly interacting with the lattice, at thermal equilibrium.

where T is the absolute temperature of the sample, N is a number of spins in the sample, and k is the Boltzmann constant. At normal conditions, this equilibrium magnetization is too small to be detectable, but when a resonance phenomenon is exploited by applying a short burst of RF energy at resonance frequency ω_0 (called an RF pulse), the magnetization can be flipped onto a transverse plane, perpendicular to the direction of $\boldsymbol{B}_0$. This transverse magnetization will precess at the resonant frequency of the spins and thus will generate an oscillating magnetic field flux in the receiver coils of the NMR apparatus, which will be detected as a time-varying voltage at the coils terminals. This signal is called a free induction decay (FID) and its time evolution contains information about the values of resonant frequency of the spins, ω_0, the spin-spin relaxation time, *T2*, and the distribution of local static magnetic fields at the locations of the spins, $T2^*$. The local static magnetic fields, experienced by spins at different locations in the sample, may vary from spin site to spin site, chiefly due to the inhomogeneity of the main magnetic field B_0. In addition, in heterogeneous samples, commonly encountered in *in vivo* experiments, local susceptibility variations may contribute to $T2^*$ effects. For a variety of reasons, the chief one being the ease of

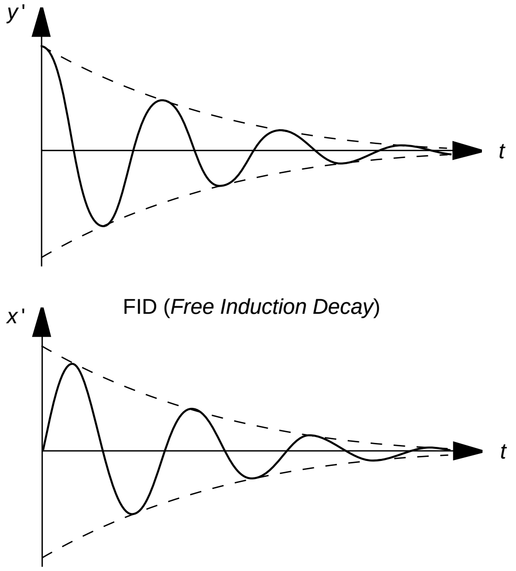


Figure 3. Real and imaginary components of an FID signal.

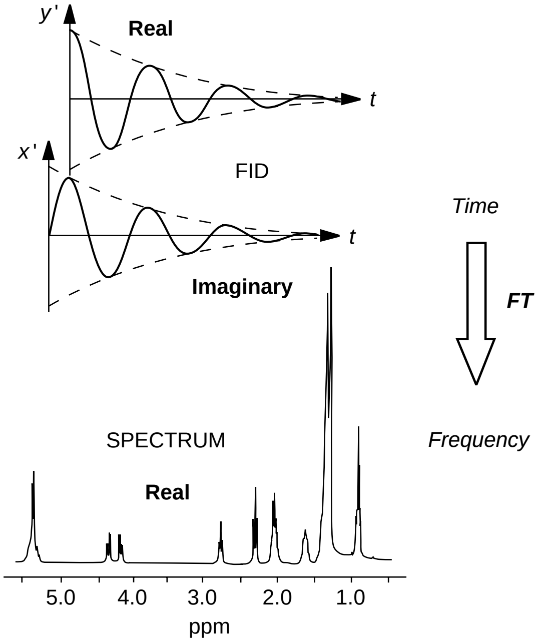


Figure 4. A method of generating an NMR spectrum from the FID signals.

interpretation, the FID signal is always recorded in the reference frame that rotates at the frequency ω close to ω_0, (called "the rotating frame"). From the engineering point of view, recording the signal in the reference frame, rotating at the frequency ω, is equivalent to frequency demodulation of the signal by frequency ω.

The recorded FID has two components: one, called real, or in-phase, is proportional to the value of transverse spin magnetization component aligned along the y' axis of the rotating frame (the [*x*', *y*', *z*] notation is used to denote rotating frame, as opposed to the stationary, laboratory frame $[x, y, z]$). The other FID component is proportional to the value of transverse spin magnetization projected along the *x*' axis and is referred to as imaginary, or out-of-phase signal (see Fig. 3). To a human being, the FID signals can be difficult to interpret; thus an additional postprocessing step is routinely employed to facilitate data analysis. A Fourier transform (FT) is applied to the FID data and the signal components having different frequencies are retrieved, producing an NMR spectrum (see Fig. 4).

The chemical shift, mentioned earlier, is responsible for a plethora of spectral lines (peaks) seen in a typical NMR spectrum. Consider a simple organic molecule that contains hydrogen localized at three nonequivalent molecular sites. Chemists call molecular sites equivalent if the structure of chemical bonds around the site creates an identical distribution of electron density at all proton locations. When the sites are nonequivalent, different distributions of electron cloud around the protons will have a different shielding effect on the value of the local magnetic field, experienced by individual protons. The strength of electron shielding effect is proportional to the value of B_0 and is accounted for in the Hamiltonian by using a shielding constant, σ.

$$\mathcal{H} = -\gamma\hbar \sum_{i=1}^{3} (1 - \sigma_i) B_0 I_{zi} \tag{3}$$

In this example, a molecule with only three nonequivalent proton sites has been considered; in general, the summation index in Eq. (3) must cover all nonequivalent sites, however, many may be present. It is evident from Eq. (3) that individual protons located at different sites will have slightly different resonant frequencies, which will give rise to separate resonant peaks located at different frequencies in the observed NMR spectrum, as shown in Fig. 4. This accounts for a fine structure of NMR spectra that consists of multiple lines at different frequencies, identifying all nonequivalent molecular sites in the sample studied.

The interaction of the nuclear spin with the electron cloud surrounding it has a feedback effect, resulting in a slight distortion of the cloud; the degree of this alteration is different depending on whether the spin is up or down w.r.t. the magnetic field B_0. This distortion has a ripple effect on surrounding nonequivalent spins, and consequently they become coupled together via their interactions with the electron cloud; this phenomenon is called a spin–spin coupling or *J* coupling, and is accounted for by adding another term to the spin Hamiltonian:

$$\mathcal{H} = -\sum_{i=1}^{n} \left[\gamma\hbar (1 - \sigma_i) B_0 I_{zi} + \sum_{j<i} J_{ij} I_i \times I_j \right] \tag{4}$$

where J_{ij}, known as a spin–spin coupling constant, describes the strength of this effect for each pair of nonequivalent protons. The presence of spin–spin coupling leads to a hyperfine structure of the NMR spectra, splitting peaks into multiplet structures, as shown in Fig. 5 that contains two fragments of an NMR spectrum of lactic acid. The structure of each multiplet can be explained using simple arrow diagrams, visible next to NMR lines. The signal at 1.31 ppm is generated by 3 equiv protons located in the CH_3 group that are linked to the proton spin in the CH group via *J* coupling. The spin of the proton in the CH group can have only two orientations: up or down, as indicated by arrows

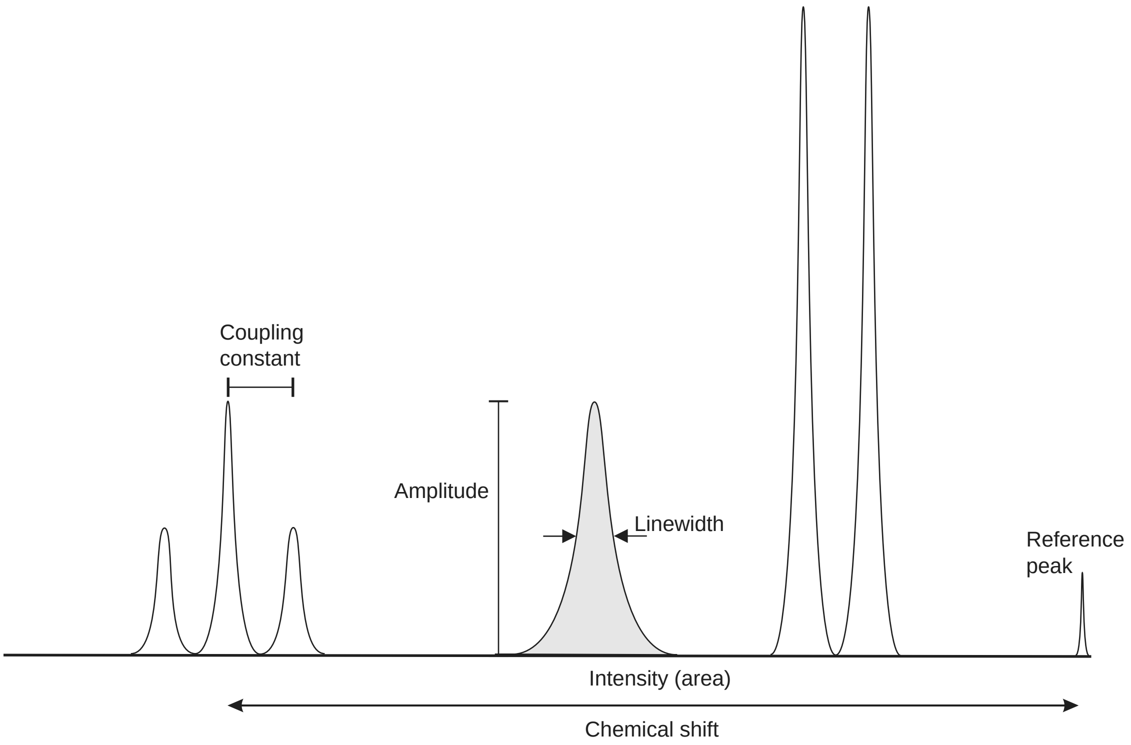


Figure 5. A fragment of experimental spectrum (lactic acid at 500 MHz) showing resonances from a CH_3 group (at 1.31 parts per million, ppm) and a CH group (at 4.10 ppm), respectively. The splitting of CH_3 resonance into a doublet and the CH resonance into a quadruplet is caused by the J coupling. The structure of each multiplet can be derived using simple rule of grouping spins according to their orientations, as shown by groups of arrows.

that follow the bracket next to the CH_3 label. Thus, the signal from CH_3 protons is expected to split into a doublet with relative signal intensities 1:1, as indeed is seen in the recorded spectrum. Similarly, the signal at 4.10 ppm is generated by a single proton in the CH group that is linked via J coupling to 3 equiv protons in the CH_3 group. The spins within this group of three protons can assume eight different configurations, depending on their orientation w.r.t. to the magnetic field $\boldsymbol{B}_0$. Some of those orientations are equivalent (i.e., they have the same energy) and thus can be lumped together, as shown by groups of arrows that follow the bracket next to the CH label. Simple counting leads to a prediction that the signal from the CH proton should split into a quadruplet with relative signal intensities 1:3:3:1. Again, this is clearly visible in the recorded spectrum.

As illustrated in Fig. 6, these simple considerations show that each peak in the spectrum can be fully characterized by specifying its position w.r.t. an established reference peak (chemical shift), amplitude, intensity (intensity, or the area under the peak, is proportional to the concentration of spins contributing to the given peak), linewidth (provides information $\sim T2$ and magnetic field homogeneity), and multiplet structure (singlet, doublet, triplet, etc., carries information about J coupling). Thus, the NMR spectra, like the one in Fig. 4 showing an experimental spectrum of vegetable (maize) oil, contain a wealth of information about the structure and conformation of molecules found within the sample.

Strictly speaking, the linewidths of the peaks in the NMR spectrum are determined by the values of $T2^*$ and thus are sensitive to the homogeneity of the main magnetic field and other factors contributing to the distribution of local static magnetic fields seen by the spins. Wider distributions of local fields, lead to shorter $T2^*$ values and broader corresponding peaks in the NMR spectrum. Broad peaks make spectra harder to interpret due to overlap between peaks located close to each other. This feature puts a premium on shimming skills of the NMR spectrometer operator (shimming includes methods to improve the homogeneity of the static magnetic field). For *in vivo* studies, the shimming tasks are made even more difficult by tissue heterogeneity that causes local variations in the magnetic field (referred to as susceptibility effects within the MRS community).

There is a peculiar feature in the way the NMR spectra are recorded that routinely confounds the NMR newbies. For historical reasons, the NMR spectra are plotted as if the spectrometer frequency was kept constant and the magnetic field B_0 was varied (swept) over a range of values to record all the characteristic peaks (see Fig. 7a). In this approach, the horizontal axis of the plot represents the values of the external magnetic field necessary to reach a resonant condition for a given group of spins. However, all modern NMR spectrometers use an acquisition method in which the magnetic field B_0 is kept constant and differences in resonant frequencies of various nuclei (due to their

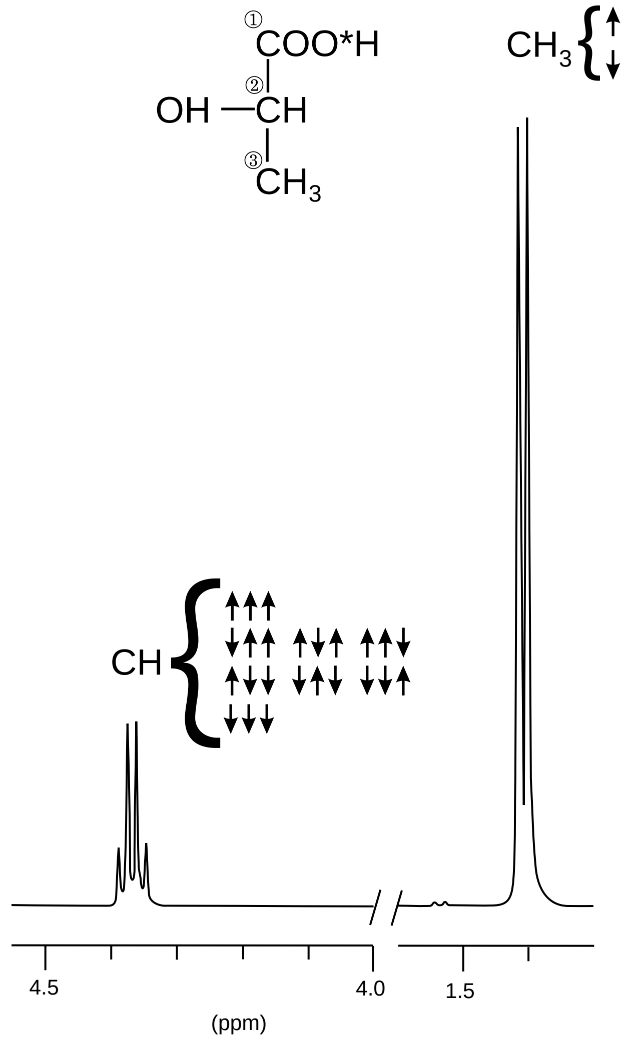


Figure 6. A schematic representation illustrating various characteristic parameters used to describe elements of an NMR spectrum.

diverse chemical shifts) are recorded instead. It is fundamental to understand that if the horizontal axis of an NMR spectrum was meant to represent the resonant frequencies of nuclei with different chemical shift, all residing in the same external magnetic field B_0, then the lines to the right would represent signals with *lower* frequencies, which at best is counterintuitive. This paradox is resolved when one looks at the relationship between magnetic field and NMR resonance frequency for nuclei located at sites with varying electron shielding. This relationship is derived from Eq. (3), leading to the well–known Larmor equation:

$$w = g(1 - s)B_0 \qquad (5)$$

This equation is plotted in Fig. 7b for three sites with different values of the chemical shielding constant σ. As σ increases, the slope of the line decreases. Thus, if the RF frequency is kept constant and the magnetic field is swept to reach subsequent resonant conditions, the weakly shielded nuclei will resonate at lower field, and as the strength of the shielding effect increases, the external magnetic field must be increased to compensate for the increased shielding. It must be reiterated that the term magnetic field used in this context refers to the *external* magnetic field B_0, produced by the spectrometer's magnet, and not to the value of the local magnetic field that the spin is actually experiencing. Therefore, the signals from heavily shielded nuclei will appear at the high end of the spectrum, as illustrated by the horizontal line in Fig. 7b. On the other hand, if the external magnetic field B_0 is kept constant and the frequency content of the FID signal is looked at, it will be noticed that nuclei at heavily shielded sites resonate at lower frequency, which reflects the decrease of the *local* magnetic field due to the shielding effects. Therefore, the signals from heavily shielded nuclei will appear at the low end of the spectrum, as illustrated by the vertical line in Fig. 7b. Historically, magnetic field sweeping was used in the early days of NMR spectroscopy and a sizeable volume of reference spectra were all plotted using the fixed-frequency convention. This standard was retained after pulsed NMR technology replaced the earlier continuous wave NMR spectroscopy, despite the fact that the fixed-field approach would have been far more natural.

The size of the chemical shift varies linearly with the strength of the magnetic field B_0, which makes comparison of spectra acquired with NMR spectrometers working at different field strengths a chore. To simplify matters, chemists introduced a concept of relative chemical shift, which is defined as follows: It is realized that the term ω in Eq. 5 represents a resonant frequency of a group of equivalent spins in their local magnetic field, B_L. Thus, Eq. (5) can be used to define a variable τ as

$$\tau = s \times 10^6 = \frac{B_L - B_0}{B_0} \times 10^6 \qquad (6)$$

The value of τ is dimensionless and expresses the value of the shielding constant σ in ppm. It is interpreted as a change in the local magnetic field *relative* to the strength of the main magnetic field produced by the spectrometer's magnet. As seen from Eq. (6), the τ scale is directly proportional to σ, that is, heavily shielded nuclei will have a large value of τ (see Fig. 7a). This also makes the τ scale collinear with the B_0 axis, which is inconvenient in modern NMR spectroscopy, which puts a heavy preference on spin resonant frequencies. To address this awkward feature, chemists use a more practical chemical shift scale, called δ, that is defined as

$$d = 10 - \tau \qquad (7)$$

This scale is a measure of the change in local resonant frequency relative to the base frequency of the NMR spectrometer, ω_0. The factor 10 in the above relationship arises from the fact that a vast majority of observed proton chemical shifts lie in the range of 10 ppm; by convention, tetrametylsilane (TMS), which exhibits one of the strongest shielding effects, has been assigned a value of $\delta = 0$. This was done after careful practical consideration: all 12 protons in TMS occupy equivalent positions, and thus an NMR spectrum of TMS consists of a strong, single line. The referencing procedure has evoked a considerable amount of

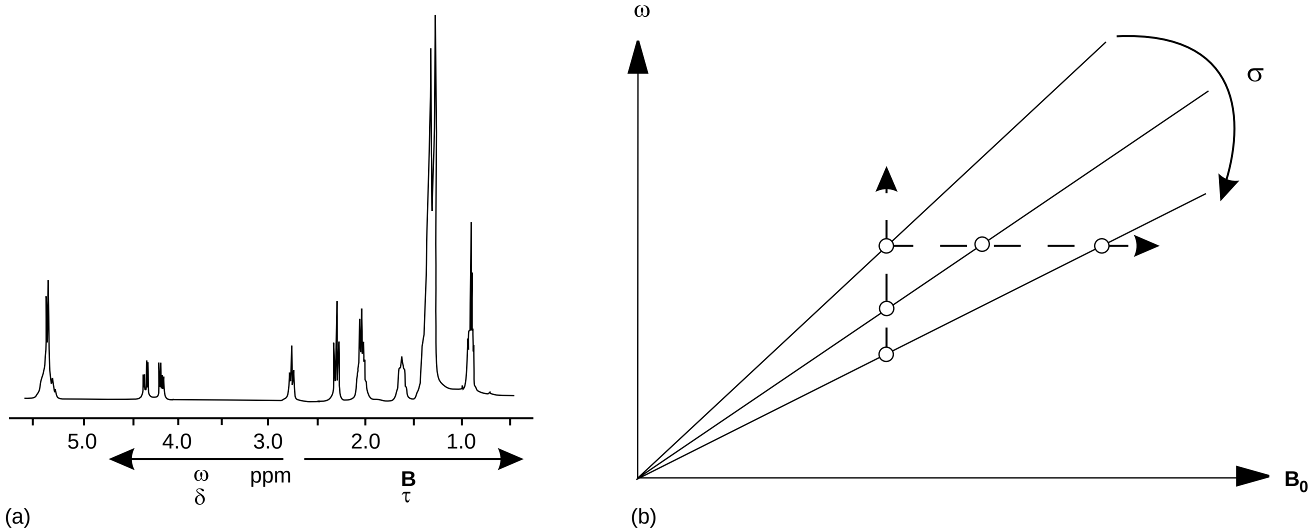


Figure 7. An apparent paradox in the interpretation of the abscissa calibration for an NMR spectrum: (a) the same spectrum can have a value of magnetic field B assigned to the abscissa so that the values of B increase when moving to the right, or have the value of frequency ω assigned to the same abscissa, but increasing when moving to the left. (b) The diagram on the right provides an explanation of this effect (see the main text for details).

debate over the years and currently the International Union of Pure and Applied Chemistry (IUPAC) specifies that shields are to be reported on a scale increasing to high frequencies, using the equation

$$d = \frac{w_x - w_{\text{ref}}}{w_{\text{ref}}} \times 10^6 \tag{8}$$

where ω_x and ω_{ref} are the frequencies of the reported and reference signals, respectively.

Since TMS easily dissolves in most solvents used in NMR spectroscopy, is very inert, and is quite volatile, it is a very convenient reference compound. In practice, a small amount of TMS can be added to the sample, which will produce a well-defined reference peak in the measured spectrum (see Fig. 5). After the experiment is completed, TMS is simply allowed to evaporate, thus reconstituting the original composition of the sample.

In MRS applications, the δ scale is used exclusively to identify the positions of the peaks within the spectra. For example, the water peak is known to have $\delta = 4.75$ ppm; therefore, if an MRS spectrum is acquired on a machine with the main magnetic field of 1.5 T and the base RF frequency of 63.86 MHz, the water line will be shifted by $4.75^{*}63.86 = 303$ Hz toward the higher frequency from the reference TMS peak. It also means that the protons in water experience weaker shielding effects than protons in the TMS. Finally, if the spectrum is plotted according to the accepted conventions, the water line appears to the left of the reference peak. Unfortunately, TMS cannot be used as an internal reference in MRS applications (it cannot be administered to humans). Thus in practice, the signal from NAA is used as a reference and has an assigned value of $\delta = 2.01$ ppm, which is the chemical shift of the acetyl group within the NAA NMR spectrum, acquired *in vitro*.

EQUIPMENT AND EXPERIMENTS

In medical applications, standard MRI equipment is used to perform MRS acquisitions. Thus, in contrast to standard *in vitro* NMR experiments, *in vivo* MRS studies are performed at lower magnetic field strength, using larger RF coils, and with limited shimming effectiveness due to magnetic susceptibility variations in tissue. As a result, the MRS spectra inherently have lower signal-to-noise characteristics than routine *in vitro* spectra; this is further aggravated by the fact that in MRS signal averages are accumulated using fewer scans due to examination time constraints. This creates a new set of challenges related to the fact that when the MRI equipment is used to perform MRS, it is utilized outside its design specifications that focus on the imaging applications of MR technology. Fortunately, many hardware performance characteristics that are absolutely crucial to the successful acquisition of spectroscopy data, such as magnetic field uniformity and stability, or coherence and stability of the collected NMR signals, are appreciated in MRI as well. Thus, steady improvements in MRI technology are contributing to the emergence of clinical applications of MRS. Since this article focuses on MR spectroscopy, the following considerations will describe features of data acquisition schemes that are unique to MRS applications, and disregard those aspects of hardware and pulse sequences design that form the core of the MRI technology (see the section on **Magnetic Resonance Imaging**).

The first challenge of MRS is volume localization. Obviously, an NMR spectrum from the entire patient's body, while rich in features, would be of very little clinical utility. Over the years, many localization techniques have been proposed, but in current clinical practice only two methods are routinely used. The first one collects NMR spectra from a single localized volume and is thus referred to as single voxel MRS, or simply MRS. The other allows collection of spectra from multiple voxels arranged within a single acquisition slab. It is often referred to as multi voxel MRS, MRS imaging (MRSI), or chemical shift imaging (CSI). With this method, more advanced visualization techniques can be used, such as generation of specific metabolite concentration maps over larger regions of interest (ROI).

Single voxel (SV) MRS is simpler to implement. The volume of interest (VOI, or voxel) is selected using one of the two alternative data acquisition pulse sequences. The first one uses a phenomenon known as a stimulated echo to produce a signal used to generate the spectroscopic FID that is then transformed into the NMR spectrum. To produce a stimulated echo, three RF pulses are used, each rotating (flipping) the magnetization by 90°. The theory of this process is too complex to be discussed in detail here; an interested reader is referred to the original paper by Hahn (6) or to more specific texts on NMR theory, such as those listed in the ***Reading List*** section. The application of the stimulated echo method for *in vivo* MRS was first proposed by Frahm et al. (7) who coined an acronym STEAM (**St**imulated **E**cho **A**cquisition **M**ode) to describe it. A simplified diagram of the stimulated echo MRS acquisition pulse sequence is shown in Fig. 8. The three RF pulses are shown on the first line of the diagram, the echo signal from the localized voxel can be seen on the bottom line of the diagram. How does this sequence allow the selection of a specific VOI as a source of collected signal? The key to successful VOI localization is to use a slice-selective RF excitation. This technology is taken straight from mainstream MRI, and thus the reader is referred to MRI-specific information for further details (see the section **Magnetic Resonance Imaging** or MRI monographs listed in the ***Reading List***). Briefly, the slice selection technique relies on the use of band-limited RF pulses (notice the unusual modulation envelope of RF pulses shown in Fig. 8) applied in the presence of tightly controlled magnetic field gradients (MFG), shown as pulses labeled G_x, G_y, and G_z in Fig. 8. When a band-limited RF pulse is played in the presence of an MFG, only the spins in a narrow range of positions, located within a thin layer of the studied object (a slice) will achieve the resonance conditions and respond accordingly; all the spins outside the slice will be out of resonance and remain unaffected. Thus, the spin magnetization within the selected slice will be flipped by the RF, and magnetization outside the slice will remain unaffected. An analysis of this process shows that the slice orientation is perpendicular to the direction of the MFG, slice thickness is controlled by the amplitude of the MFG pulse, and location (offset from the magnet's isocenter) is determined by the shift in carrier frequency of the RF pulse. Thus, the first pulse in Fig. 8 will excite spins in a slice that is perpendicular to the z axis of the magnet (G_z was used), the two remaining pulses will excite spins in slices perpendicular to the x and y directions, respectively. Since the condition required to produce a stimulated echo is that the spins be subject to all three pulses, only the matter located at the intersection of the three perpendicular slices fulfills the criterion, and thus only the spins located within this volume will generate the stimulated echo signals. The dimensions of the VOI selected in such a way are determined by the thickness of individual slices, and the location of the VOI is determined by the position of individual slices, as illustrated in Fig. 9.

The other single voxel MRS protocol uses a spin echo sequence to achieve volume selection. To produce the localized signal, three RF pulses are used, as before. However, while the first RF pulse still rotates the magnetization by 90°, the remaining two RF pulses flip the magnetization by

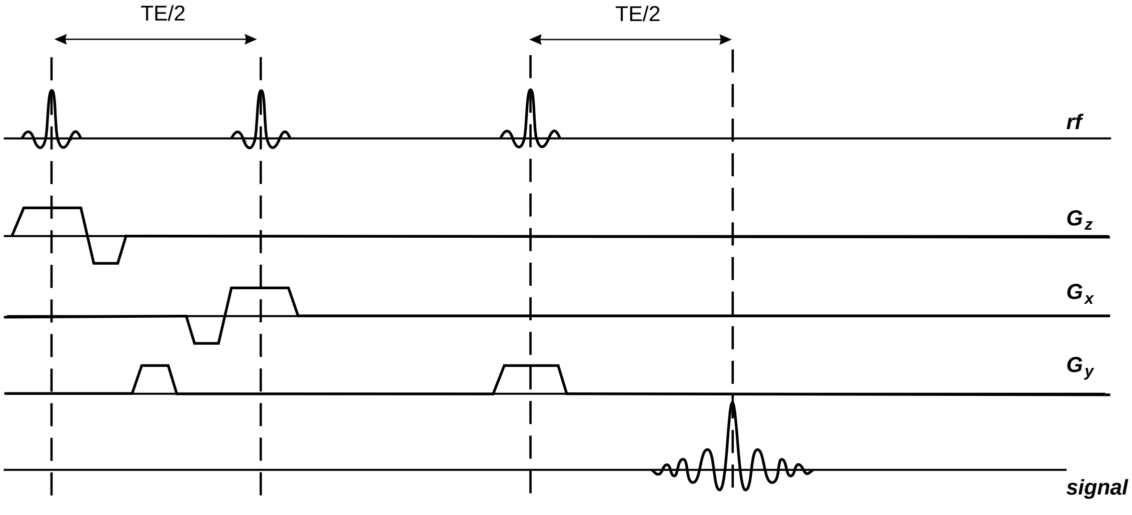


Figure 8. A simplified diagram of a basic STEAM MRS sequence. Time increases to the right, the time interval covered by this diagram is typically ~ 100 ms. On the first line the RF pulses are shown; the next three lines show the timing of the pulses generated by the three orthogonal magnetic field gradient assemblies (one per Cartesian axis in space); the last line, labeled signal, shows the timing of the resulting stimulated echo NMR signal.

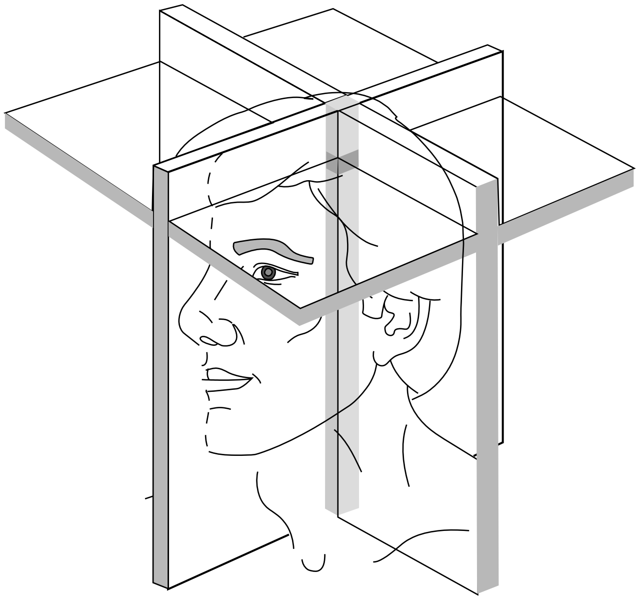


Figure 9. The principle of VOI selection in STEAM protocol (see text for details).

180° each. As a result, two spin echoes are produced, as shown on the bottom line of Fig. 10. The first echo contains a signal from a single column of tissue that lies at the intersection of the slices planes generated by the first and second RF pulses, while the second echo produces a signal from a single localized VOI within this column. As with the stimulated echo, the theory of this process is too complex to be discussed here; an interested reader is referred to the original paper by Hahn (6) or to more specific texts on NMR theory, such as those listed in the ***Reading List*** section.

The application of dual echo spin echo method to *in vivo* MRS has been proposed by Bottomley (8) who coined an acronym **P**oint **R**esolved **S**pectroscopy (PRESS) to describe it. A simplified diagram of the PRESS acquisition pulse sequence is shown in Fig. 10. The three RF pulses are shown on the first line of the diagram; but only the second echo signal seen on the bottom line of the diagram comes from the localized voxel region.

What are the advantages and weaknesses of each volume localization method? First, in both methods an echo is used to carry the spectroscopy data. Since echoes are produced by transverse magnetization, the echo amplitudes are affected by *T2* relaxation and the time delay, *TE*, that separates the center of the echo from the center of the first RF pulse that was used to create the transverse components of magnetization. The longer the *TE*, the stronger the echo amplitude attenuation due to *T2* effects will be. Since the amplitude of the echo signal determines the amplitude of the spectral line associated with it, the SV MRS spectra will have lines whose amplitudes will depend on the selected *TE* in the acquisition sequence. For the STEAM protocol, it is possible to use relatively short TE values; mostly because the magnetization contributing to the stimulated echo signal is oriented along the z axis during the period of time between the second and third RF pulses, and thus it is not subject to *T2* decay (in fact, it will grow a little due to *T1* relaxation recovery). Therefore, the time interval between the second and the third RF pulse is not counted toward TE. Furthermore, 90° RF pulse have shorter duration than 180° ones, offering an additional opportunity to reduce TE. Consequently, in routine clinical applications the STEAM protocols use much shorter TE values (~ 30 ms) than PRESS protocols (routine TE values used are ~ 140 ms). Therefore, when using the

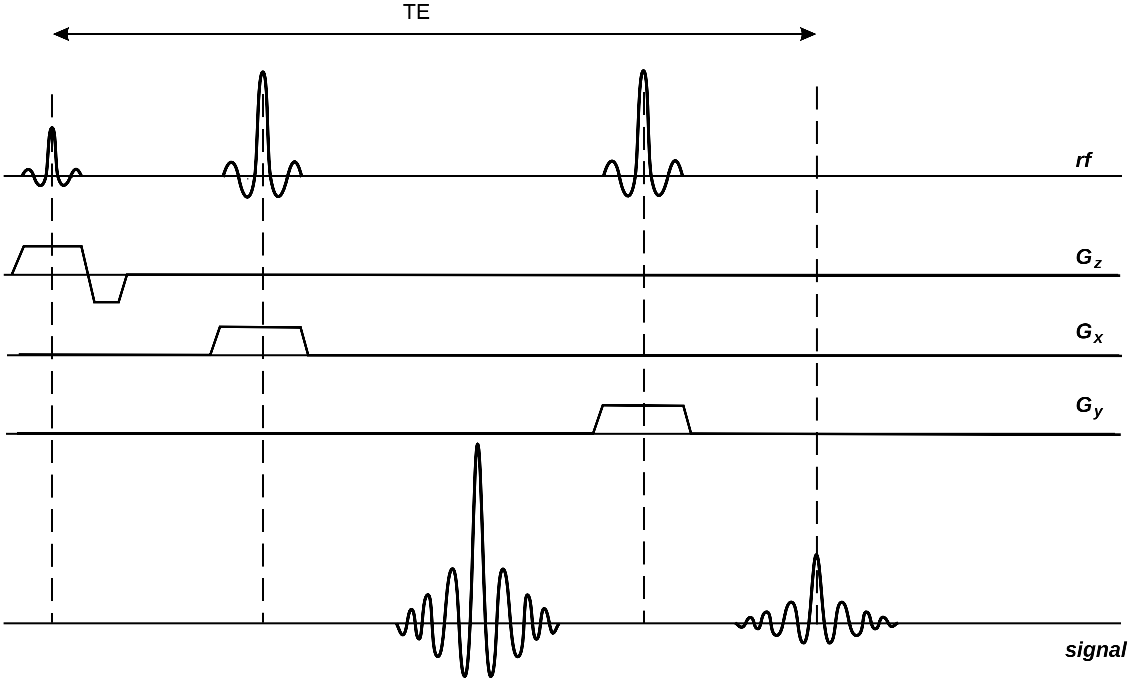


Figure 10. A simplified diagram of a basic PRESS MRS sequence. Notice similarities in VOI selection algorithm within both STEAM and PRESS protocols.

STEAM protocol one is rewarded with a spectrum whose line intensities are closer to the true metabolite concentrations in the studied tissue. However, theoretical calculations show that the signal intensity of a stimulated echo is expected to be 50% less than that of a spin echo generated under identical timing conditions (i.e., *TE*). This is an inherent drawback of the STEAM protocol, since the spectra tend to be noticeably noisier than those produced using PRESS.

PRESS, by design, uses two spin echoes to generate a signal from localized VOI. This limits the minimum *TE* possible for the second echo to $\sim$ 80 ms or so, depending on the MR scanner hardware. Since the magnetization never leaves the transverse plane (except during RF pulses), the *T2* effects are quite strong. As a result, metabolites with shorter *T2* will decay down to the noise levels and vanish from the final spectrum, which has an ambivalent impact on clinical interpretations, simplifying the spectrum on one hand while removing potentially valuable information on the other. This effect can be further amplified by signal intensity oscillations with varying TE, caused by *J* coupling. For further details on this topic the reader is referred to the specific texts on MRS theory and applications, listed in the ***Reading List*** section.

One word of caution is called for now. The numbers quoted here reflect capabilities of MR hardware that represent a snapshot in time. As hardware improves, these parameters rapidly become obsolete.

There are other, finer arguments supporting possible preferences toward either method (STEAM or PRESS). These include such issues as suppression of parasitic signals arising from outer-volume excitation (residual signals coming from outside the VOI), sensitivity to baseline distortion of the spectra due to the use of solvent suppression, accuracy of VOI borders defined by each method, and so on. Thus, in current clinical practice there are strong proponents of both STEAM and PRESS techniques, although lately PRESS seems to be gaining popularity because of simpler technical implementation issues, such as magnetic field homogeneity correction (shimming), or compensation of effects caused by eddy currents induced in the magnet cryostat by magnetic field gradient and RF pulses.

While SV MRS is relatively straightforward, and thus preferred by novices in the practice of clinical MRS, most routine applications today demand spectroscopic data collected from multiple locations within the organ studied. Therefore, a solution that would allow collection of MR spectra from multiple locations at the same time has a natural appeal to physicians. To achieve such a task is no small matter, and many schemes have been proposed before a method that today is considered most practical in daily use has been found. The method was first proposed by Brown et al. (9). Their method is both conceptually simple and revolutionary. It utilizes a method that encodes both space–(localization) and time–dependent (NMR spectra) information using mechanisms that manipulate the phases of signals emitted by individual spins at different locations. The acquisition sequence is schematically shown in Fig. 11, which illustrates the two-dimensional (2D) CSI principle using a PRESS pulse sequence. Of course, this approach is equally applicable to STEAM method as well. An astute reader will immediately recognize that the spatial encoding part of the protocol is virtually identical to dual phase encoding techniques used in 3D MRI acquisitions. At this point, readers less familiar with advanced MRI techniques probably would want to read more about this method elsewhere (see **Magnetic Resonance Imaging** or MRI monographs listed in the ***Reading List***). The enlightenment occurs when one realizes that with this scheme the acquired signal is not frequency encoded at all. Therefore, after 2D FT processing in the two orthogonal spatial directions, one ends up with a collection

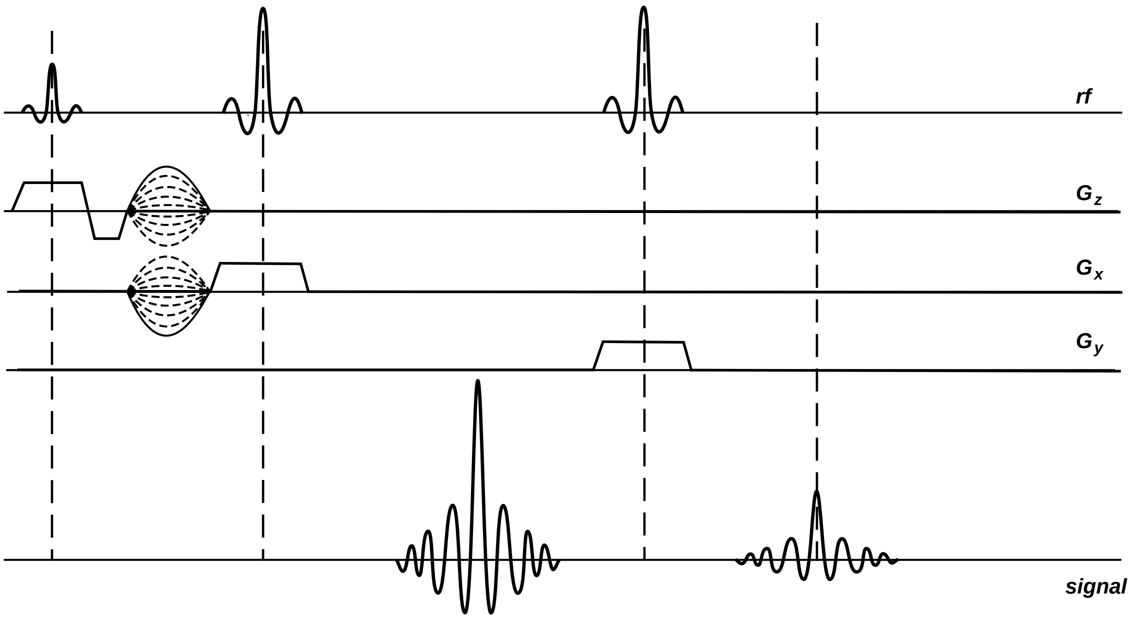


Figure 11. A simplified diagram of a basic 2D CSI MRS sequence. The dotted gradient lobes represent phase encoding gradients responsible for multivoxel localization.

of FIDs that are free from any residual phase errors due to spatial encoding, but nevertheless represent NMR signals from localized VOIs, anchored contiguously on a planar grid. The size and orientation of the grid is determined by the spatial encoding part of the protocol; thus, localization is conveniently decoupled from NMR frequency beats caused by the chemical shifts of studied metabolites.

If this method is so simple, why do people still want to acquire SV spectra? First, the method is quite challenging to implement successfully in practice, despite the seeming simplicity of the conceptual diagram shown in Fig. 11. Both spatial and temporal components affect the phase of collected NMR signals, and keeping them separated requires a great deal of MR sequence design wizardry and the use of advanced MR hardware. Due to peculiarities of the FT algorithm, even slight phase errors (of the order of 1°) are capable of producing noticeable artifacts in the final spectra. Second, the VOI localization scheme is ill suited to a natural way of evaluating lesions spectroscopically; most clinicians like to see a spectrum from the lesion compared to a reference spectrum from a site that is morphologically equivalent, but otherwise appears normal. In the human brain, where most spectroscopic procedures are performed today, this means that the reference spectrum is acquired contralaterally to the lesion location. Such an approach represents a drawback in CSI acquisitions where VOIs are localized contiguously, and typically a few wasted voxels must be sacrificed to ensure the desired anatomic coverage of the exam. Finally, the dual-phase encoding scheme requires that each pulsed view (a single execution of the pulse sequence code with set values of both phase encoding gradients) is repeated many times to collect enough data to localize voxels correctly. As many views must be acquired as there are voxels in the grid, which causes the required number of views to grow very fast. For example, even for a modest number of locations, say 8^*8, 64 views must be generated. This leads to acquisition times that appear long by today's imaging standards (typical MRSI acquisition requires 3–8 min).

It is difficult to fully exploit the richness and diversity of technical aspects of localization techniques used in *in vivo* MRS; extended reviews, such as papers by Kwock (10), den Hollander et al. (11), or Decorps and Bourgeois (12), can be used as springboards for further studies.

The second major challenge of *in vivo* ^{1}H MRS arises from the fact that the majority of tissue matter is simply water, which for humans can vary from 28% in the bones to 99% in CSF, with an average of $\sim$ 75–80% in soft tissues (e.g., brain matter, muscle, lung, or liver). Thus, if a proton spectrum from just about any soft tissue (except adipose) is recorded, the result would look like the one presented in Fig. 12a, where an experimental spectrum from a muscle tissue of a young (and lean) rat, collected *ex vivo* on a 300 MHz NMR spectrometer, is shown. At a first glance, the result is boring: Only a single, strong peak from water is visible. Closer inspection, however, uncovers some additional details: First of all, the background of the spectrum is not totally flat, but composed of a broad peak, much wider than the normal range of chemical shifts expected in HR ^{1}H NMR spectra. This is illustrated in Fig. 12b, where the background of the spectrum in Fig. 12a has been enhanced by amplifying the background and cutting off most of the strong, narrow water peak. The broad spectrum is produced by protons in macromolecular components of the tissue: the proteins, DNA, RNA, and thousands of other compounds responsible for function and control of cellular activities. The spectrum broadening is caused by the residual dipolar interactions that were not fully averaged out because large molecules move more slowly than the small ones. *Note:* This component of the NMR signal is normally not visible in MRI and MRS applications; since the line is so broad, the relaxation time $T2$ of this component is quite short (on the order of hundreds of μs), thus by the time the MR signals are collected at TE that are in the range of milliseconds, this signal has decayed out. One can see some small blips on the surface of the broad line in the Fig. 12b: Those are signals from tissue biochemical compounds whose molecules are either small enough, or have specific chemical groups that are free to move relatively fast due to conformational arrangements. These clusters of protons have $T2s$ long enough to produce narrow lines, and their chemical environment is varied enough to produce a range of chemical shifts. In short, those little blimps form an NMR spectral signature of the tissue studied. As such, they are the target of MRS. To enhance their appearance, various solvent suppression techniques are used. In solvent suppression, the goal is to suppress the strong (but usually uninteresting signal) from solvent (in case of tissue, water), thus reserving most of the dynamic range of signal recorder for small peaks whose amplitudes are close to the background of a tissue spectrum. This point is illustrated in Fig. 12c, which shows a spectrum from the same sample as the other spectra in this figure, but acquired using a water suppression technique. Now, the metabolite peaks stand out nicely against the background, in addition to some residual signal from water peak (it is practically impossible to achieve 100% peak suppression).

The most common implementation of water suppression in MRS *in vivo* studies uses a method known as CHESS: **Ch**emical **S**hift **S**elective suppression, first proposed by Haase and colleagues at the annual meeting of the Society of Magnetic Resonance in Medicine in 1984. It consists of a selective 90° pulse followed by a dephasing gradient (homogeneity spoiling gradient, or homospoil). The bandwidth of the RF pulse is quite narrow, close to the width of the water line, and the carrier RF frequency offset is set to the water signal center frequency. When such a pulse is applied, only the water protons will be at resonance and they will flip by 90°, leaving magnetizations of all other protons unchanged. The resultant FID from the water signal is quickly dispersed by using the homospoil gradient. When the CHESS segment is followed by a regular spectroscopy acquisition sequence (STEAM or PRESS), the first RF pulse of those spectroscopic sequences will tip all the magnetizations from metabolites, but will not create any transverse magnetization from water. The reason is that at the time the spectroscopic acquisition routine starts, the longitudinal magnetization for water protons is zero, as prepared by the CHESS routine. The description of technical details of various solvent suppression methods can be found in the paper by van Zijl and Moonen (13). The side effect of solvent suppression is a baseline distortion of the resulting spectrum;

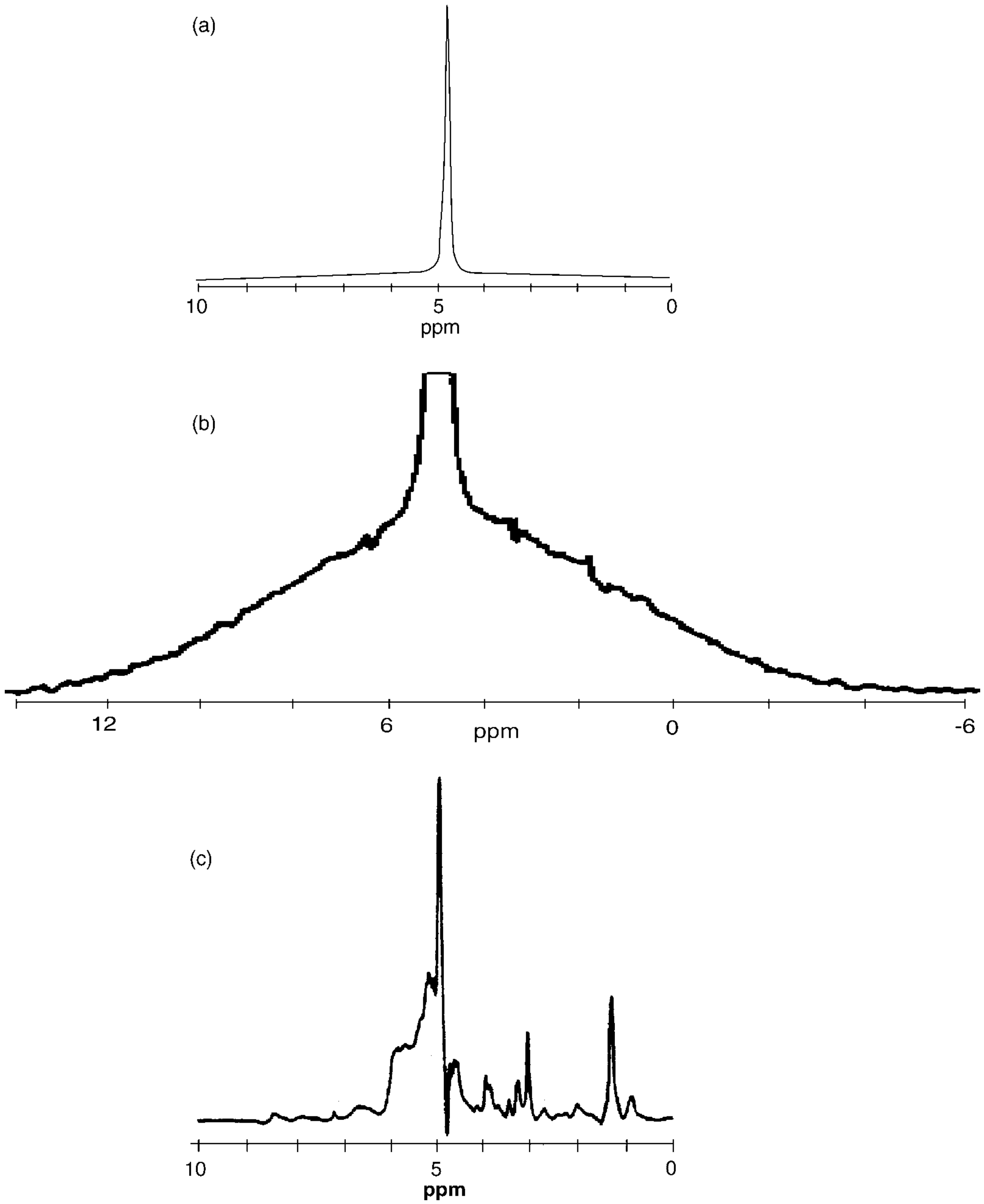


Figure 12. An example of NMR spectra obtained from the same sample (a lean muscle from a young rat's leg) collected *in vitro*: (a) standard spectrum obtained using a single RF pulse and performing an FT on the resulting FID; (b) the same spectrum scaled to reveal a wide, broad line generated by protons within macromolecules, notice small humps on top of this broad line: these are signals from small mobile macromolecules; (c) the high-resolution spectrum of the same sample, obtained using a spin–echo method and applying water suppression, both water and macromolecular peaks are suppressed, revealing small narrow lines that are subject to clinical MRS evaluations.

this distortion can be particularly severe in the vicinity of the original water peak, (i.e., at 4.75 ppm). To avoid difficulties associated with baseline correction, the MRS spectra in routine clinical applications are limited to the chemical shift range from 0 to $\sim$ 4.5 ppm.

In this short description of clinical MRS, the discussion had to be limited to its main features. There are many interesting details that may be of interest to a practitioner in the field, but had to be omitted here because of space constrains; the reader is encouraged to consult comprehensive reviews, such as the work of Kreis (14), to further explore these topics.

APPLICATIONS

Now that there is a tool, it is necessary to find out what can be done with it. The clinical applications of MRS consist of three steps: first, an MR spectrum is acquired from the VOI; second, specific metabolite peaks are identified within the spectrum; and third, clinical information is derived from the data, mostly by assessing relative intensities of various peaks. The first step has been covered, so let us now look at step two. The easiest way to implement the spectral analysis is to create a database of reference spectra and perform either spectrum fitting or assign individual peaks by comparing the experimental data to the reference spectra. This is easier said than done; the reliable peak assignment in NMR spectra acquired *in vivo* has been one of the major roadblocks in the acceptance of MR spectroscopy in clinical practice.

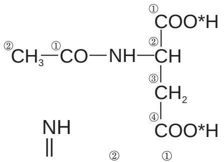
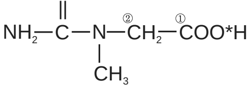
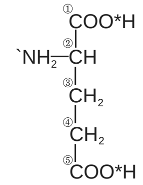
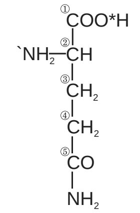
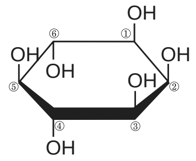
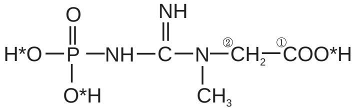
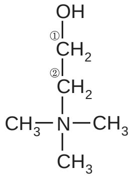
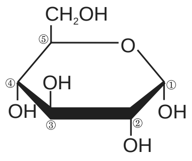
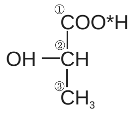
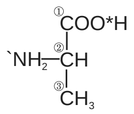
Currently, the majority of clinical MRS studies are performed in the human brain. Over the past several years, a database of brain metabolites detectable by proton MRS *in vivo* has been built and verified. This process is far from finished, as metabolites with lower and lower concentrations in the brain tissue are identified by MRS and their clinical efficacy is explored. A list of components routinely investigated in current clinical practice is shown in Table 1. Even a short glance at this list immediately reveals the first major canon of the MRS method: more than a cursory knowledge of organic and biochemistry is required to fully comprehend the information available. An appreciation of the true extent of this statement can be quickly gained by taking a closer look at the first molecule listed in that table: *N*-acetylaspartate, or NAA. This compound belongs to a class of *N*-alkanoamines; the italic letter *N* represents the so-called locant, or a location of the secondary group attached to the primary molecule. In this case, N is a locant for a group that is attached to a nitrogen atom located within the primary molecule. The primary molecule in this case is an L-aspartic acid, which is a dicarboxylic amino acid. Dicarboxylic means that the molecule contains two carboxylic (COOH) groups. As an amino acid, L-aspartic acid belongs to the group of the so-called nonessential amino acids, which means that under normal physiological conditions sufficient amounts of it are synthesized in the body to meet the demand and no dietary ingestion is needed to maintain the normal function of the body. The *N*-acetyl prefix identifies a molecule as a secondary amine; in such compounds the largest chain of carbon compounds takes the root name (aspartic acid), and the other chain (the acetyl group, CH_3CO–, formed by removal of the OH group from the acetic acid CH_3COOH) becomes a substituent, whose location in the chain (the *N*-locant) identifies it as attached to the nitrogen atom. But what about the L prefix in the L-aspartic acid mentioned above? It has to do with a spatial symmetry of the molecule's configuration. The second carbon in the aspartic chain has four bonds (two links to other carbon atoms, one link to the nitrogen atom, and the final link to a proton). These four bonds are arranged in space to form a tetrahedron with the four atoms just mentioned located at its apexes. Such a configuration is called a chiral center, to indicate a location where symmetry of atom arrangement needs to be documented. There are two ways to distribute four different atoms among four corners of a tetrahedron, one is called levorotatory (and abbreviated by a prefix L-), and the other is called dextrorotatory (and abbreviated by a prefix D-). It turns out that the chirality of the molecular configuration has a major significance in biological applications: Since successful mating of different molecules requires that their bonding sites match, only one chiral moiety is biologically active. In our case, L-aspartic acid is biologically active; the D-aspartic acid is not. Last, but not least, the reader has probably noticed that the name of the molecule is listed as *N*-acetylaspartate, while we keep talking about aspartic acid... well, when an acid is dissolved in water, it undergoes dissociation into anions and cations; the molecule of aspartic acid in the water solution looses two protons from the carboxylic groups COOH (the locations of the cleavages are indicated by asterisks in the structural formulas shown in Table 1), and becomes a negatively charged anion. To reflect this effect, and suffix -ate is used. Thus, the name *N*-acetylaspartate describes an anion form of a secondary amine, whose primary chain is a levorotatory chiral form of aspartic group, with a secondary acetyl group attached to the nitrogen atom. The NAA is so esoteric a molecule that most standard biochemistry books do not mention it at all; its chief claim to prominence comes from the fact that it is detectable by MRS. A recent review of NAA metabolism has been recently published by Barlow (15).

The example discussed above emphasizes that much can be learned just from a name of a biological compound. To gain more literacy in the art of decoding the chemical nomenclature of biologically active compounds, the reader is encouraged to consult the appropriate resources, of which the Introduction to Subject index of CAS (16) is one of the best.

As mentioned earlier, routine clinical MRS studies focus on proton spectra spanning the range from 0 to ~ 4.5 ppm; the NMR properties of compounds listed in Table 1 are presented in Table 2, which identifies each molecular group according to carbon labeling used in structural formulas shown in Table 1. For each molecular group, the chemical shift of the main NMR peak is listed, along with the spectral multiplet structure characterizing this line. A simulated theoretical spectrum shows all lines in the range from 0 to 5 ppm, to give the reader an idea where the molecular signature peaks are located in the spectra acquired *in vivo*. Finally, information is provided whether, for a particular line, the signal acquired in standard MRS *in vivo* studies is strong enough to emerge above the noise levels and become identifiable. This information is further supplemented with comments indicating whether a particular line is expected to be visible on spectra acquired with short or long TE values. It is evident that the information provided in Table 2 is absolutely critical to successful interpretation of clinical MRS results; unfortunately, space limitations prevent us from further dwelling into details of spectral characteristics of clinically important metabolites. This information can also be difficult to locate in the literature since most of the data still reside in original research papers; fortunately, a recent review by Govindaraju et al. (17) offers an excellent starting point.

An examination of data shown in Table 2 quickly leads to a realization that only a limited number of metabolite signature lines can be successfully used in the

Table 1. Basic Properties of Metabolites Most Commonly Detected in MRS Spectra of the Human Brain

Metabolite	Full Name	Acronym	Formula	CAS[a] Number	Structure	Molecular Weight	Normal Concentration Range in brain, mmol
N-Acetylaspartate	*N*-Acetyl-L-aspartic acid; amino acid	NAA	$C_6H_9NO_5$	[*997-55-7*]		175.14	8–17
Creatine	(1-Methylguanidino) acetic acid; non-protein amino acid	Cr	$C_4H_9N_3O_2$	[*57-00-1*]		131.14	5–11
Glutamate	L-Glutamic acid; amino acid	Glu	$C_5H_9NO_4$	[*56-86-0*]		147.13	6–13
Glutamine	L-Glutamic acid-5-amide; amino acid	Gln	$C_5H_{10}N_2O_3$	[*56-85-9*]		146.15	3–6
myo-Inositol	1,2,3,5/4,6-Hexahydroxy-cyclohexane	m-Ins	$C_6H_{12}O_6$	[*87-89-8*]		180.2	4–8
Phosphocreatine	Creatine phosphate	PCr	$C_4H_{10}N_3O_5P$	[*67-07-2*]		211.11	3–6
Choline	Choline hydroxide, Choline base, 2-Hydroxy-*N*,*N*,*N*-trimethyl-athanaminium	Cho	$C_5H_{14}NO$	[*62-49-7*]		104.20	0.9–2.5
Glucose	D-Glucose, dextrose anhydrous, corn sugar, grape sugar	Glc	$C_6H_{12}O_6$	[*50-99-7*]		180.15	1.0
Lactate	L-Lactic acid, 2-hydroxypropanoic acid	Lac	C_3H_6O	[*79-33-4*]		90.07	0.4
Alanine	L-Alanine, 2-amino-propanoic acid	Ala	$C_3H_7NO_2$	[*65-41-1*]		89.09	0.2–1.4

[a] CAS = Chemical Abstracts Service Registry Number of the neat, nondissociated compound. In structural formulas, an asterisk * indicates a site where, upon dissociation, a proton is released; apostrophe ` indicates a site where, upon dissociation, a proton is attached. Metabolite names refer to dissociated (ionic) forms of the substances since this is the form they are present in the *in vivo* environment.

Table 2. The NMR Properties of Metabolites Most Commonly Detected in MRS Spectra of the Human Brain

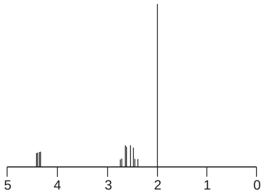
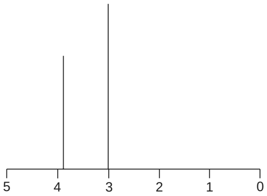
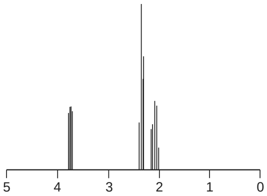
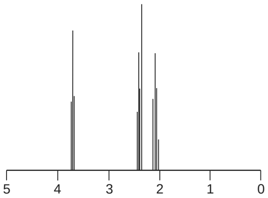
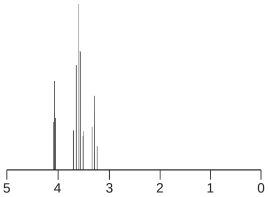
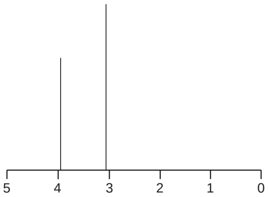
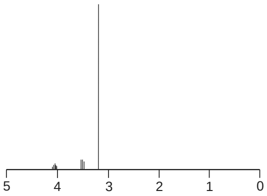
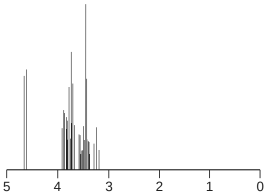
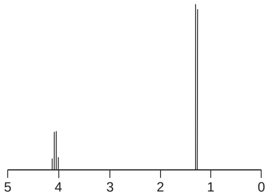
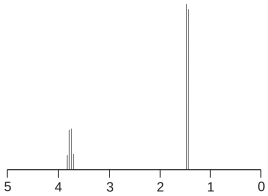
Metabolite	Acronym	Molecular Group	Chemical Shift δ ppm[a]	Multiplet Structure[b]	Visibility *in vivo*[c]	Theoretical Spectrum –Range (0,5) ppm
N-Acetylaspartate	NAA	• CH_3 Acetyl	**2.00**	s	Yes	
		• CH_2 Aspartate	2.49	dd	No	
		• CH_2 Aspartate	2.67	dd	No	
		• CH Aspartate	4.38	dd	No	
Creatine	Cr	N-CH_3	**3.03**	s	Yes	
		• CH_2	**3.91**	s	Yes	
Glutamate	Glu	• CH_2	**2.07**	m	Yes, on short TE	
		• CH_2	**2.34**	m	Yes, on short TE	
		• CH	3.74	dd	No	
Glutamine	Gln	• CH_2	**2.12**	m	Yes	
		• CH_2	**2.44**	m	Yes	
		• CH	**3.75**	t	No	
Myo-inositol	m-Ins	• CH	3.27	t	No	
		• CH, • CH	**3.52**	dd	Yes, on short TE	
		• CH, • CH	**3.61**	t	Yes, on short TE	
		• CH	4.05	t	No	
Phosphocreatine	PCr	N-CH_3	**3.03**	s	Yes	
		• CH_2	**3.93**	s	Yes	
Choline	Cho	N-$(CH_3)_3$	**3.18**	s	Yes	
		• CH_2	3.50	m	No	
		• CH_2	4.05	m	No	
Glucose	Glc	β–• CH	4.63	d	Not visible in normal brain due to low concetration	
		All other CH	**3.23–3.88**	m		

Table 2. (*Continued*)

Metabolite	Acronym	Molecular Group	Chemical Shift δ ppm	Multiplet Structure	Visibility *in vivo*	Theoretical Spectrum –Range (0,5) ppm
Lactate	Lac	• CH_3 • CH	**1.31** 4.10	d q	Not visible in normal brain due to low concetration	
Alanine	Ala	• CH_3 • CH	**1.47** 3.77	d q	Not visible in normal brain due to low concetration	

[a] The bold type indicates a dominant line in the spectrum.

[b] s = singlet, d = doublet, dd = doublet of doublets, t = triplet, q = quadruplet, m = multiplet.

[c] Reference to short TE indicates that those signals have short *T2*s and thus will be suppressed in acquisitions with long TE.

interpretation of clinical proton MRS spectra. In normal volunteers, five major markers can routinely be detected and evaluated:

The *N*-acetylaspartate peak at 2.0 ppm, commonly referred to as NAA.

The combination of creatine (Cr) and phosphocreatine (PCr); reported together as two lines positioned ~ 3.0 and 3.9 ppm, respectively. Mostly referred to as Cr, but some people use a label tCr (for total creatine). The peak at 3.0 ppm is often identified as Cr, and the peak at 3.9 ppm as Cr2,

The combination of glutamine (Gln) and glutamate (Glu) at 2.2–2.4 ppm; since the peaks from those two compounds strongly overlap and are typically unresolvable, the are routinely reported together and referred to as Glx.

The choline peak at 3.25 ppm, referred to as Cho; The primary contributions to this peak are from bound forms of choline: phosphorylcholine (PC) and glycerophosphorylcholine (GPC), with only minor signal from free choline.

The *myo*-inositol group at 3.6 ppm is visible only in spectra acquired with short *TE* (due to *J* coupling effects that suppress the intensity of lines forming this signal at long *TE*s). *Myo*-inositol name poses evidently a challenge to acronym creators, since it can be found labeled as m-Ins, MI, mI, and In.

In addition, the following markers are routinely evaluated in a variety of diseases:

Lactate (Lac), often visible as a doublet at 1.3 ppm.

Mobile lipds (Lip) whose methyl (CH_3) groups are visible at 0.9 ppm and methylene (CH_2) groups provide signal at 1.3 ppm.

In special cases other metabolites may become visible, and we list here two examples:

Alanine (Ala) with a signature peak at 1.5 ppm.

Glucose (Glc), mostly showing as a broad peak ~ 3.5 ppm; note that the theoretical spectrum shown in Table 2 is a superposition of α- and β-anomers that occur in 1:2 ratio, respectively, in equilibrium *in vivo* conditions.

Examples of typical MRS spectra obtained from healthy subjects are shown in Fig. 13. The first thing to notice is that the signal noise ratio in those spectra is poor, so indeed, only the strongest lines from metabolites listed in Table 2 have a chance to become visible.

In the context of this discussion, the following questions have been discussed: What is MRS, how to perform it, who is interested in those studies, and why? If an actual clinical MRS examination were being performed, at this stage of MRS study we would have localized the lesion, collected an *in vivo* MR spectrum from the tissue within this lesion, identified the signature metabolite peaks, and analyzed their relative intensities. Now would have come the time to ponder what the results mean and what is their clinical significance. It is not an easy task, since the last analytic step listed above produces results that must be compared to "normal" baseline references. These reference data are obtained by performing statistical analysis of multiple results obtained from healthy people: a challenging task given biological diversity of normal subjects. Thus, MRS results are reported using such imprecise terms as unchanged, elevated, or suppressed. Sometimes, when clinicians want to be particularly precise, they will report ratios, such as the NAA/Cr ratio, which essentially renormalizes all observed signal intensities by assigning an arbitrary intensity of one unit to the Cr peak. In other words, this approach uses the Cr peak as an internal reference. As mentioned in Table 3, the rationale for such

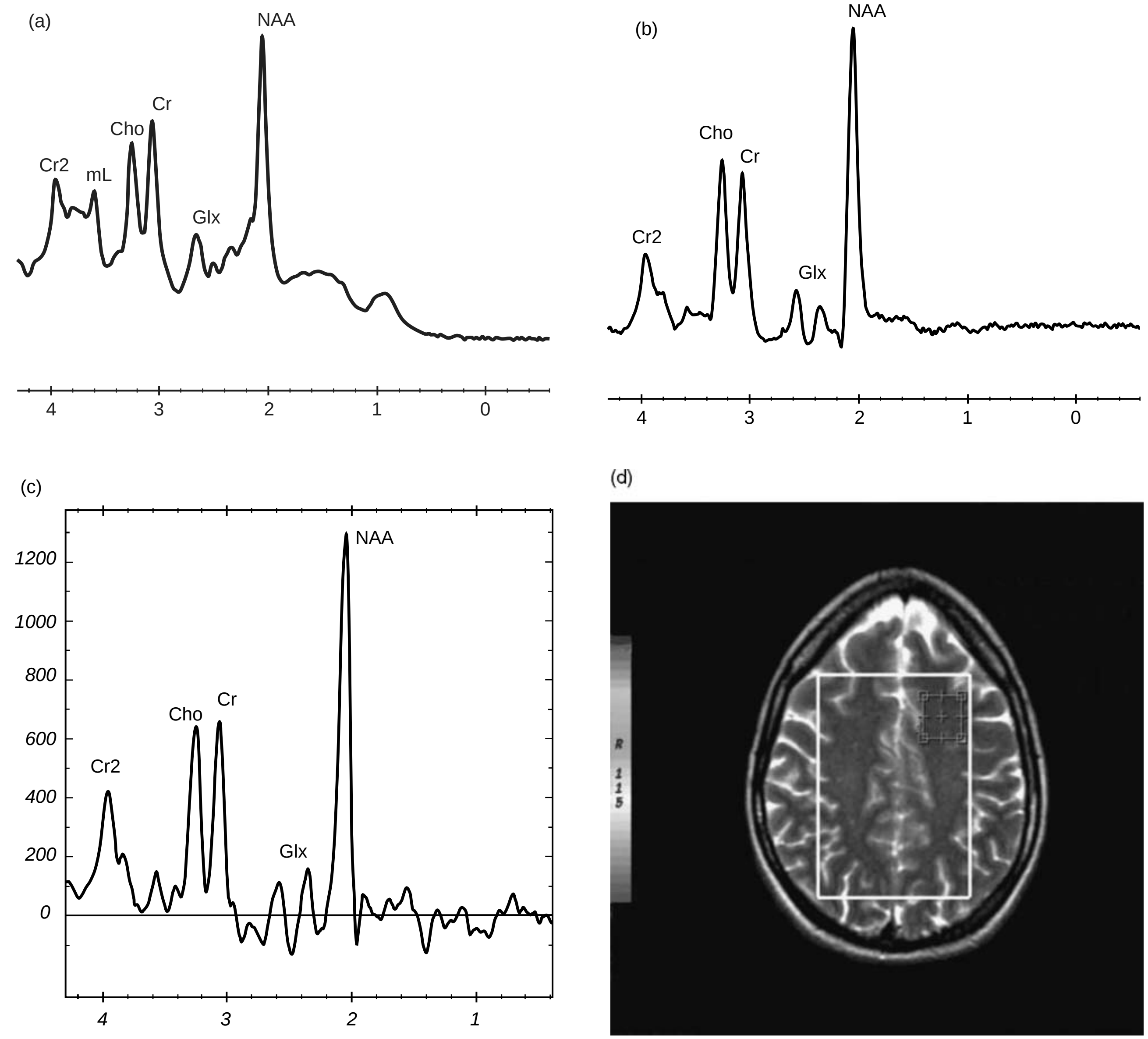


Figure 13. Examples of typical MRS spectra obtained *in vivo* for a normal volunteer: (a) STEAM SV protocol with $TE = 30$ ms and TR of 2000 ms, magnitude mode; (b) PRESS SV protocol with $TE = 144$ ms and $TR = 1500$ ms, magnitude mode; (c) PRESS 2D CSI with $TE = 135$ ms, $TR = 1000$ ms, $16^{*}16$ voxel locations, real mode; (d) Image reference showing a location of selected VOI associated with the spectrum shown in (c) The VOI is generated by combining voxel locations. All data acquired within a single study, lasting $\sim$ 15 min; spectra (a) and (b) were acquired at the same location using the same VOI size, no signal averaging was used (NEX = 8).

an approach is that the levels of Cr tend to be relatively stable under normal physiologic conditions. Such findings are purely phenomenological and their value is established over time by practicing evidence-based medicine, that is by performing statistical analysis of a large number of findings and looking for correlations between MRS results and the patient's clinical status. The drawback of such an approach is that in the early stages of new methodology, there is no established consensus regarding data interpretation, thus one is forced to read a large number of published clinical reports and develop one's own approach to the inclusion of MRS findings into the clinical decision making process. A short summary of the current understanding of clinical findings related to MRS results is provided in Table 3. However, since these findings are still subject to frequent updates, it would have been pointless to try to provide in-depth coverage of these issues here, knowing full well that the data are likely to be obsolete by the time this publication appears in print. Instead, the reader is strongly encouraged to survey the current literature for reviews of clinical MRS topics; some excellent, previously published papers can serve as a

Table 3. Clinical Properties of Metabolites Most Commonly Detected in MRS Spectra of the Human Brain

Metabolite	Full Name	Acronym	Function	Pathological Variation
N-Acetylaspartate	*N*-Acetyl-L-aspartic acid; amino acid	*N*AA	Plays osmoregulatory function in the intercompartmental system, thought to be responsible for the removal of intracellular water, against the water gradient, from myelinated neurons	Decrease in NAA levels is indicative of reversible axonal injury or neuronal loss. In adults it has been associated with neoplasia, multiple sclerosis, hypoxia, ischemia, stroke, trauma, epilepsy, encephalitis, and neurodegenerative syndromes
Creatine	(1-Methylguanidino)acetic acid	Cr	Synthesized in the liver, it is exported to muscle and brain where it is phosphorylated into phosphocreatine and used as an energy store	Levels of creatine and phosphocreatine are tightly controlled under physiologic conditions, and thus this peak has been suggested as an internal reference for metabolite amplitude or area ratios. Increased levels of Cr and PCr have been observed, however, in hyperosmolar states, as well as in trauma. They also increase with aging
Glutamate	L-Glutamic acid; amino acid	Glu	Most abundant amino acid found in the human brain, acts as an excitatory neurotransmitter	Plays a role in detoxification of ammonia in the hyper ammonemic states
Glutamine	L-Glutamic acid-5-amide; amino acid	Gln	Plays a role in glutamate regulation; astrocytes metabolize glutamate to glutamine, thus preventing excitotoxicity	Elevated levels of glutamate and glutamine have been observed in hepatic encephalopathy, Reyes syndrome, meningiomas, and rare inherited enzyme deficiency. Reduced levels have been associated with Alzheimer's disease
myo-Inositol	1,2,3,5/4,6-Hexahydroxycyclo hexane	m-Ins	A component of membrane phospholipids, functions as a cerebral osmolyte. It is also thought to play an essential role in cell growth	Concentration fluctuates more than that of any other major metabolite in the brain. Increased levels of myo-inositol have been observed in neonates, Alzheimer's disease, diabetes, and hypersomolar states. *myo*-Inositol levels are decreased in hepatic encephalopathy, hypoxia, stroke, and some neoplasms
Phosphocreatine	Creatine phosphate	PCr	A major energy storage in the body	See notes for creatine.
Choline	Choline hydroxide, Choline base, 2-Hydroxy-*N*,*N*,*N*-trimethylathanaminium	Cho	Choline is important for normal cellular membrane composition and repair, normal brain function and normal cardiovascular function	Increased levels of choline have been detected in normal infants, and aging adults. They were also associated with neoplasia, gliosis, demyelinating disease, inflammation or infection, trauma, diabetes, chronic hypoxia, and AIDS. Decreased levels of choline have been found in hepatic encephalopathy, stroke, and dementias, including Alzheimer's disease
Glucose	D-Glucose	GLc	A major energy carrier throughout the body.	Increased levels notes in diabetes mellitus, parental feeding, hypoxic encephalopathy
Lactate	L-Lactic acid, 2-hydroxypropanoicacid	Lac	Lactate is a signature byproduct of carbohydrate catabolism and thus when normal cellular oxidative respiration mechanisms are active, its levels in the brain tissues are very low	Visible in a variety of diseases. Increased levels of lactate were observed in some tumors, during the first 24 h after infarction, in hypoxia, anoxia, near-drowning, and hypoventilation
Alanine	L-Alanine, 2-Aminopropanoic acid	ALa	Alanine is a nonessential amino acid of uncertain function in the brain	The alanine peak is difficult to detect since it is easily overshadowed by lactate. Alanine levels might be elevated in meningiomas
Lipids	Fatty acids, glycerides, glycolipids, lipoproteins	Lip	Normally the lipid signals are not visible in the MRS spectra of the brain, but might appear due to fat contamination (voxel bleed)	Mobile protons from lipids (0.9 ppm for CH_3 and 1.3 ppm for CH_2) are not normally visible in brain spectra, but can appear in diseased conditions. Lipid signals are suppressed at long *TE*s. Elevated lipid levels are observed in cellular necrosis, high-grade astrocytoma, and lymphoma

starting point (18–21). The last paper on this list, by Smith and Stewart (21), also offers excellent overview of spectroscopy studies in organs other than the brain.

It is impossible, in one short article, to offer a fully comprehensive coverage of such an advanced and rapidly evolving discipline as clinical magnetic resonance spectroscopy. The author can only hope that this brief overview offers sufficient information and enough reference pointers to let the readers start exploring this new and exciting field on their own.

BIBLIOGRAPHY

Cited References

1. Becker E, Fisk CL, Khetrapal CL. The development of NMR. Vol 1: 2-158. In: Grant DM, Harris RK, editors. Encyclopedia of NMR. Chichester (UK): John Wiley & Sons, Inc.; 1996.
2. Ridgen JS. Quantum states and precession: The two discoveries of NMR. Rev Mod Phys 1986;58:433–488.
3. Arnold JT, Dharmati SS, Packard ME. Chemical effects of nuclear induction signals from organic compounds. J Chem Phys 1951;19:507.
4. Shoolery JN. The development of experimental and analytical high resolution NMR. Progr NMR Spectrosc 1995;28:37–52.
5. Cohen JS, et al. A history of biological applications of NMR spectroscopy. Progr NMR Spectrosc 1995;28:53–85.
6. Hahn EL. Spin echoes. Phys Rev 1950;80:580–594.
7. Frahm J, Merboldt KD, Hänicke W. Localized proton spectroscopy using stimulated echoes. J Magn Reson 1987;72:502–508.
8. Bottomley PA. Spatial localization in NMR spectroscopy *in vivo*. Ann NY Acad Sci 1987;508:333–348.
9. Brown TR, Kincaid BM, Ugurbil K. NMR chemical shift imaging in three dimensions. Proc Natl Acad Sci USA 1982; 79:3523–3526.
10. Kwock L. Localized MR Spectroscopy — Basic Principles. Neuroimaging Clin N Am 1998;8:713–731.
11. den Hollander JA, Luyten PR, Mariën AJH. ^{1}H NMR Spectroscopy and spectroscopic imaging of the human brain. In: Diehl P, et al., editors. NMR — Basic Principles and Progress. Vol. 27. Berlin: Springer-Verlag; 1991.
12. Decorps M, Bourgeois D. Localized spectroscopy using static magnetic field gradients: comparison of techniques. In: Diehl P, et al., editors. NMR — Basic Principles and Progress. Vol. 27. Berlin: Springer-Verlag; 1991.
13. van Zijl PCM Moonen CTW. Solvent suppression strategies for *in vivo* magnetic resonance spectroscopy. In: Diehl P, et al., editors. NMR — Basic Principles and Progress. Vol. 26. Berlin: Springer-Verlag; 1991.
14. Kreis R. Quantitative localized ^{1}H MR spectroscopy for clinical use. Progr NMR Spectr 1997;31:155–195.
15. American Chemical Society, Naming and Indexing of Chemical Substances for Chemical Abstracts, Appendix IV. Chemical Abstracts Service, *Chemical Abstracts Index Guide*, Columbus: American Chemical Society; 2002.
16. Baslow MH. *N*-acetylasparate in the vertebrate brain: metabolism and function. Neurochemical Res 2003;28:941–953.
17. Govindaraju V, Young K, Maudsley AA. Proton NMR chemical shifts and coupling constants for brain metabolites. NMR in Biomed 2000;13:129–153.
18. Blüml S, Ross B. Magnetic resonance spectroscopy of the human brain. In: Windhorst U, Johansson H, editors. Modern Techniques in Neuroscience Research. Berlin: Springer-Verlag; 1999.
19. Ross B, Bluml S. Magnetic resonance spectroscopy of the human brain. Anat Rec (New Anat) 2001;265:54–84.
20. Smith JK, Castillo M, Kwock L. MR spectroscopy of brain tumors. Magn Reson Imaging Clin N Am 2003;11:415–429.
21. Smith ICP, Stewart LC. Magnetic resonance spectroscopy in medicine: clinical impact. Progr NMR Spectrosc 2002;40:1–34.

Reading List

Bernstein MA, King KF, Zhou XJ. Handbook of MRI Pulse Sequences. Burlington (MA): Elsevier Academic Press; 2004. This is a “geek’s delight”. A detailed, step-by-step presentation and discussion of problems and solutions encountered in routine MRI practice today. A must-read for anyone seriously interested in learning MRI beyond the popular level.

deGraaf R. In Vivo NMR Spectroscopy: Principles and Techniques. Chichester (UK): John Wiley & Sons Inc.; 1999. This book covers both theoretical and practical aspects of MRS and is widely considered to be one of the best textbooks available on the subject. The book is particularly well suited for people involved in MR research outside a clinical medical environment, since it focuses on physics and engineering aspects of the methodology. Students, beware: it is very expensive ($350 for 530 pages), so it is best to seek it out at the library.

Ernst RR, Bodenhausen G, Wokaun A. Principles of NMR in One and Two Dimensions. Oxford (UK): Clarendon Press; 1987. Fundamental monograph on modern theory of NMR spectroscopy. Very comprehensive coverage, but definitely not for beginners.

Fukushima E, Roeder SBW. Experimental Pulse NMR: A Nuts and Bolts Approach. Reading (MA): Addison-Wesley; 1982. The best introduction to practical NMR. The book is out of print, but libraries still carry it and it is relatively easy to purchase second-hand, since it has been hugely popular among graduate students starting up in the NMR field at a graduate level.

Goldman M. Quantum Description of High Resolution NMR in Liquids. Oxford (UK): Clarendon Press; 1988. Very methodical and thorough coverage of HR NMR Spectroscopy in liquids, but sometimes unconventional formalism requires an extra effort on the part of the reader to really understand all aspects of discussed subject matter.

Grant DM, Harris RK, editors. Encyclopedia of NMR. Chichester (UK): John Wiley & Sons Inc.; 1996. This is a monumental piece of work (eight volumes and one update volume thus far) that rightly deserves the title of the most comprehensive review of the field to date.

Martin ML, Martin GJ, Delpuech J-J. Practical NMR Spectroscopy. London (UK): Heyden & Son, Ltd.; 1980. This book is out of print, but copies are available at libraries. It is one of the best “hands-on”, practical texts on HR NMR spectroscopy. Covers practical hints on hardware, experiment setup, sample preparation, various techniques of spectral editing, and so on.

Slichter CP. Principles of Magnetic Resonance. Berlin: Springer-Verlag; 1990. Considered by many as the “the Bible” of MR theory. It is an advanced textbook that is meant to provide the beginner a necessary background to get started in the field of MR.

Young IR, editor. Methods in Biomedical MRI and Spectroscopy. Chichester (UK): John Wiley & Sons Inc.; 2000. This two-volume set contains most entries that have been originally included in the Encyclopedia of NMR, but they have been expanded and updated by the original contributors.

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

NUCLEAR MEDICINE INSTRUMENTATION

Lawrence E. Williams
City of Hope
Duarte, California

INTRODUCTION

Nuclear medicine exists as a clinical specialty due to two basic reasons involving signal detection. Of primary importance is the high sensitivity of tissue measurements. In principle, a single labeled molecule or nanostructure may be detected upon the decay of its attached radiolabel. A second reason is the possibility of using radiolabeled materials of interest to study the physiology of animals and eventually patients. While imaging is the primary application of nuclear techniques, targeting implies an associated therapeutic strategy. All three traditional forms of radioactive emission, alpha (α), beta (β^- and β^+), and gamma radiation (γ) are available to the investigator. Negative betas are identical to the electrons found external to the atomic nucleus and are the antiparticle to β^+ (positron). Penetration distances in soft tissue for α and β rays range from μm and up to several millimeters, respectively, and so limit imaging use to organ samples or perhaps very small intact animals. Both of these particles are, however, employed in radiation therapy.

It is the photon emitter that is most valuable as an imaging label since it can be used *In vivo* on relatively large animals and patients. One exception to this general rule is the application of positron emitters (β^+) in imaging. Notice that a β^+ annihilates with a local atomic electron to form two or three photons of high energy. Thus, the positron emitter is effectively giving off quanta of a detectable type although up to several millimeters away from the site of the original decay. Because of momentum conservation, emission of two annihilation photons is essentially back-to-back; that is, at 180° separation, so as to define a line in space. This fact allows positron emitters to be an almost ideal label for 3D imaging.

Labeling Strategies

Radioactive labels may be used, in principle, to locate and quantitatively measure pharmaceuticals within excised samples, intact animals, and patients. Several strategies of labeling are possible. The radioactive tag may be used directly in the atomic form, such as ^{123}I as a test species replacing the stable isotope ^{127}I for evaluation of the patient's thyroid physiology. A secondary method is to replace a stable atom in a biological molecule by a radioactive isotopic form as ^{14}C in lieu of stable ^{12}C in a sugar. Finally, as is most common, the label is simply attached by chemical means to a molecule or engineered structure of interest. One can tag an antibody with radioactive ^{131}I or use ^{111}In inside a 50 nm phospholipid vesicle to track their respective movements inside the body of a patient. Because of protein engineering and nanotechnology, such radiolabeled manmade structures are of growing importance. Table 1 gives an outline of the three types of labeling and examples of associated clinical studies.

Applications of nuclear tagging can literally go far beyond clinical assays. When the 1976 Viking landers came down on the surface of Mars, a test for living organisms was performed using various ^{14}C labeled nutrients. An assay was then performed on a scoop of Martian soil mixed with the radiotracers using a radiation detector sampling emitted gases. It was thought that ^{14}C-methane would prove metabolism (i.e., life). While a weak positive signal was detected in the reaction chamber, these results have yet to be verified by other test procedures. Methane has, however, been found as an atmospheric gas by more recent exploratory spacecraft.

Limitations of Radioactive Labels

In the last two types (II and III) of labeling, radionuclides can become separated from the molecule or structure of interest. This disassociation may occur during preparation and/or delivery of the pharmaceutical or later *In vivo*. Responsible processes include reversible binding of the radionuclide, enzymatic action, or even competition with stable isotopes of the same element. Nuclear medicine specialists must recognize such limitations in any resultant analyses: a subtlety often overlooked in a report or document.

A second important logical issue associated with nuclear imaging is tissue identification and anatomic localization. Nuclear imaging physicians are very analogous to astronomers in that entities may be observable, but indeterminate as to type or location. Relatively strong (hot) sources appearing against a weak background in a nuclear image may be coming from a number of tissues. The physician may not, in fact, be able to identify what structure or organ is being observed. Hybrid imaging devices combining nuclear and anatomic imagers such as computed tomography, (CT) are being implemented to correct for this ambiguity and are discussed below.

Lack of specific radiopharmaceuticals has been the greatest limitation to the growth of nuclear medicine. Many tracer agents owe their discovery to accidental events or the presence of a traditional metabolic marker for a given tissue type. Yet, these historical entities may target to several different organs *In vivo* and thus lead to ambiguous images. More recently, molecular engineering, computer modeling and the generation of specific antibodies to tissue and tumor antigens have improved production of novel and highly specific agents. The most specific of these entities is the monoclonal antibody binding to a particular sequence of amino acids in the target antigen's structure.

Therapy Applications

Detection and imaging via tracers are not the only clinical tasks performed in nuclear medicine. Of increasing importance is the provision of radiation therapy when there is preexisting imaging evidence of radiopharmaceutical

Table 1. Methods and Examples of the Three Types of Nuclear Medicine Labeling

Method	Label Example	Clinical Study	Detector Device
I. Substitution of radioactive atom for common stable atom	^{123}I for ^{127}I (stable)	Thyroid uptake	Single probe or gamma camera
II. Insertion of radioactive atom in a molecule	^{14}C for ^{12}C in glucose	Glucose metabolism	Liquid Scintillator (LS) detecting exhalation of $^{14}CO_2$
III. Attachment of radioisotope to a structure	^{111}In attached to an liposome	Planar or SPECT image of cancer patient	Gamma camera

targeting to the lesion(s) in question. The oldest such treatment is the use of ^{131}I as a therapy agent for thyroid cancers including both follicular and papillary types. Here, the radionuclide emits imaging photons and moderate energy beta radiation so that localization can be demonstrated simultaneously with the treatment phase of the study. In some applications, the therapy ligand is intentionally a pure beta emitter so as to limit radiation exposure to the medical staff and patient's family. In this case, no gamma photons are available to the imaging devices. The therapist must use the coadministration of a surrogate tracer to track the position of the pure beta therapy agent. An example is the use of ^{111}In-antibodies to cancer antigens to track the eventual location of the same antibody labeled with the pure beta emitter ^{90}Y.

RADIONUCLIDE PRODUCTION

Reactor Production of Radionuclides

Production of radionuclides that are useful in nuclear medicine relies on several different methodologies. The most common nuclear medicine radiolabel, ^{99m}Tc, is produced as a decay product of its parent ^{99}Mo. Production of ^{99}Mo is generally done via nuclear fission occurring inside a nuclear reactor. Radioactive ^{99}Mo is taken into the radiopharmacy where it is attached to an alumina (Al_2O_3) column. By washing physiological saline through this generator device, the user may elute the technetium that is chemically dissimilar from the ^{99}Mo, and so comes free of the column. Possible breakthrough or leakage of Mo is measured upon each so-called "milking" procedure to assure the pharmacist that the eluted material is indeed technetium. While other generator systems are available, obtaining specific radionuclides generally requires provision of the appropriate reaction using a suitable accelerator.

Cyclotron Production of Radionuclides

A more general way to produce radioactive species of a given type is via a designated nuclear reaction. For example, while many isotopes of iodine can be found in fission reactor residues, their chemical identity makes separation a difficult problem. For that reason, ^{123}I has been obtained with the nuclear transmutation:

$$\text{Proton} + {}^{124}\text{Te} \rightarrow {}^{123}\text{I} + 2\,\text{neutrons}$$

More than one reaction can occur given the same initial conditions. In the above case, production of ^{124}I is possible when only one neutron is generated by the bombarding proton in an isotopically pure ^{124}Te target. This contamination is intrinsically present in any ^{123}I product resulting from the bombardment. Since ^{124}I has a 100 h half-life that is much longer than that of ^{123}I (13 h), the relative amount of this impurity increases with time and may become difficult to correct for in resultant gamma camera images.

While a variety of particle accelerators may be used, the most common device to produce a given radionuclide by a specific reaction is the industrial or clinical cyclotron. This is a circular accelerator invented by Lawrence and Livingston in which large electromagnets hold the proton (or other charged particle) beam in a circular orbit of increasing radius as its energy is enhanced twice per cycle with radio frequency (rf) radiation. Circulation of the beam is permitted over extended acceleration times as the volume between the magnetic poles is kept in a relative high vacuum condition.

Straight-line machines, such as tandem Van de Graaff units and linear accelerators (linacs), in which the beam moves in a geometric line from low energy ion source to the reaction site, have some disadvantages compared to a cyclotron design. In linear devices, length is generally proportional to the desired energy so as to make the machine difficult to house: particularly in a clinical setting. The clinical cyclotron is small enough to fit within a medium-sized room as shown in Fig. 1. Second, the high voltage needed to accelerate the proton or other ion may be difficult to maintain over the length of the straight-line device. Electric breakdowns not only interrupt accelerator operation, they may also damage the internal electrodes.

In order that the appropriate nuclear reaction is possible, the proton beam must strike an isotopically purified target. This may occur within the cyclotron or in a separate chamber external to the accelerator. The latter method is preferred as it permits easier access to the resultant product and rapid switching of one target with another as the reactions are varied. External target locations also reduce the radiation level inside the accelerator. In the ^{123}I example shown above, the target is a foil of highly purified Te metal; this is an isotope that is $\sim 5\%$ abundant in natural tellurium.

Unlike linear machines, beam extraction into the target chamber can be problematic for a cyclotron since the ion being accelerated is moving in a stable circular orbit. Traditionally, extraction was done using an electrode. A more effective way to extract protons from the vacuum chamber is to initially attach two electrons to each proton to form an H^- ion. This molecular species is accelerated until it reaches the correct reaction energy and a corresponding outer orbit. At this point, the circulating negative hydrogen ion is allowed to hit a so-called stripper foil that removes both electrons and converts the ion back to an ordinary proton (H^+). The proton is not geometrically stable at that radius and field and is magnetically led out of the cyclotron's vacuum chamber and into the target chamber for the desired reaction.

In addition to longer lived radionuclides, such as ^{123}I, ^{67}Ga, and ^{201}Tl, cyclotrons are conventionally used to manufacture short-lived radionuclides for positron emission tomography (PET) imaging. The latter include ^{11}C (20 min half-life), ^{13}N (10 min), and ^{15}O (2 min). Commercially, the most common product is ^{18}F (110 min) for use in fluorodeoxyglucose (FDG) as described below. Because of the several minute half-lives of the first three of these labels, it is necessary that the cyclotron is available on-site within the nuclear pharmacy. With ^{18}F production, the accelerator may be more remote; perhaps as far as an hour's drive from the clinical site so that fluorine decay does not appreciably reduce the delivered activity.

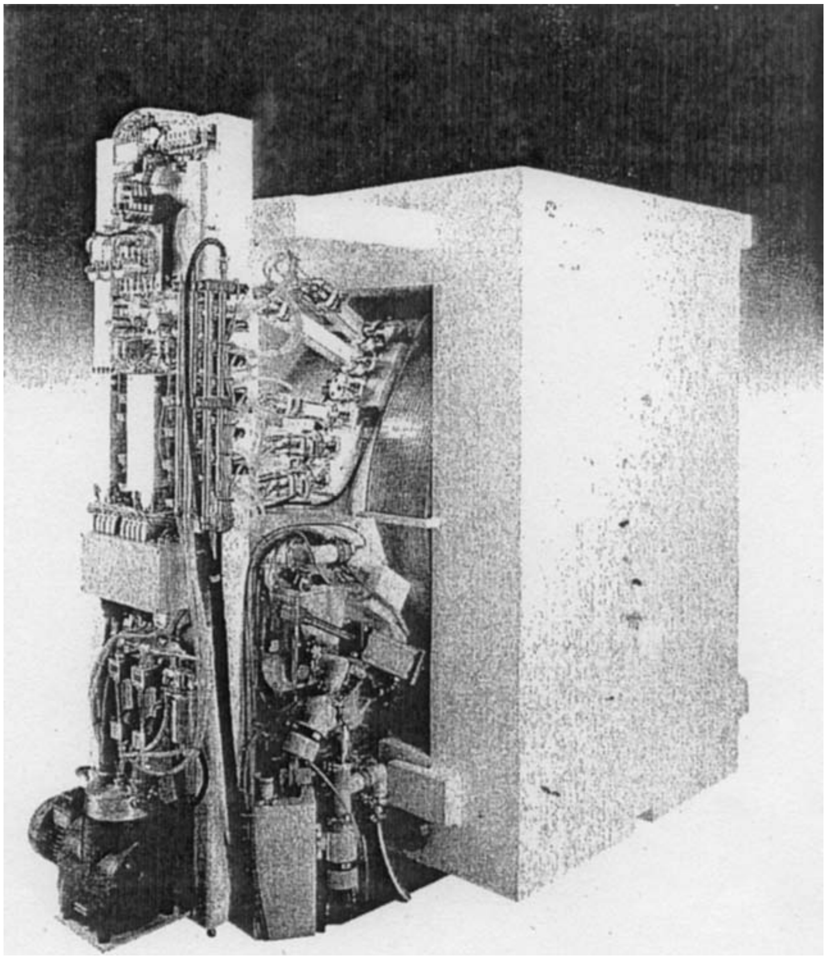


Figure 1. Medical cyclotron.

SYSTEMATICS OF RADIATION DETECTION

Detection Methods for Ionizing Radiation

Ionizing radiation is detected using electrons liberated within a sensitive volume of a detector material. All three classical states of matter, gas, liquid, and solid have been used as an ionization medium. Table 2 lists examples of each state and the devices associated with it. Most materials have ionization energies on the order of 30 eV per electron–ion pair. In solid-state semiconductors, such as Si or Ge, electron–hole pairs can be formed using ~3 eV. This lower value means that semiconductors can provide many more (~10×) ionization events for a given photon or electron energy. Such an increased number of events in turn yields improved statistical certainty that the particle has activated the counter. High thermal noise levels and elevated costs of large arrays of semiconductors have limited their use clinically.

Spectrometry

Signals of various sizes can arise in the detection process. Radionuclide counting depends on selection of the appropriate signal in a milieu of background radiation and other sample decays. For example, the technologist may have to count several beta emitters simultaneously or to detect a given gamma ray energy among many other emissions. Figure 2 shows a gamma spectrum from ^{137}Cs; both

Table 2. Detector Materials used to Measure Ionizing Radiation

State of Matter	Material	Energy per Ionization, eV	Device	Application
Gas	Argon	32	Dose calibrator	Photon activity assay
	Air	32	Ion chamber	Exposure level measurement
Liquid	Scintillation fluid (toluene)	30	Liquid counter	Beta assay in biological samples
Solid	NaI (Tl)	30	Gamma camera or probe	Photon counting
	Si (Li)	3	Solid-state probe	Photon and beta counting
	LiF (Tl)	30	TLD (Thermoluminescent dosimeter)	Radiation safety

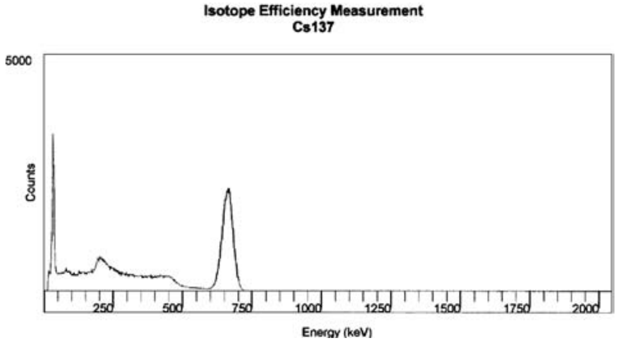


Figure 2. Energy spectrum of ^{137}Cs as measured by a NaI(Tl) probe.

Compton scattering and photoelectric effect (PE) are observed in this probe made of NaI (Tl).

In the PE, all of the photon energy is given over to an electron–ion pair in the absorbing material. Compton scatter may go on inside the patient prior to the photon coming into any detection system. In such cases, the direction and energy of the quantum may be changed so that an unwanted source may contribute to the counting process. Photon energy analysis is used to guard against such events in imaging; if the energy is seen to be reduced from that of the expected value, an electronic discriminator circuit rejects the ionization event. This pulse height analysis (PHA) is common to all nuclear detector systems and is described for imaging devices below.

ONE-DIMENSIONAL NUCLEAR MEDICINE DETECTORS

Well Counters

The most primitive instrument for photon detection is the counter or probe. In this case, a NaI(Tl) crystal is generally used to form a single large scintillation detector. In the scintillation process, the ionization event within the crystal is converted to visible light with a decay time on the order of 2 μs or less. Note that NaI is hygroscopic so that isolation of the crystal from the atmosphere is required. A reflective cap of Al is generally used as part of this hermetic seal. Resultant scintillation light is amplified by photomultiplier (PM) tubes to yield an electric signal proportional to the total amount of visible light. Well counters have the crystal in a hollow (cup) shape with the sample within the cup to maximize geometric sensitivity. Shielding is provided by an external layer of lead so as to reduce background counts. This is particularly important in a laboratory or clinical context. A mobile combination of well counter and probe system is shown in Fig. 3. Applications include sample assay using a standard source to give absolute values to the amount of detected activity. Counting experiments may involve patient tissue specimens obtained from the surgeon or animal organs obtained during measurement of biodistributions. Radiation protection is an additional application, whereby surface swab samples are counted to see if contamination is removable and possibly being spread around a lab or clinical area.

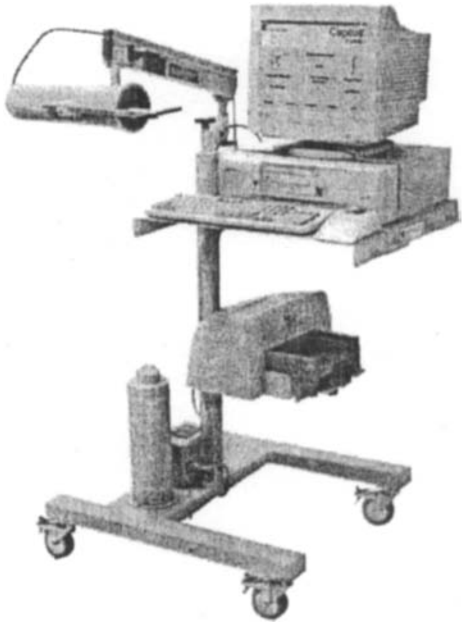


Figure 3. Well counter and probe mounted on a mobile chassis.

A second type of well counter, using high pressure Argon gas as the detector, is the dose calibrator. This device is used in all nuclear pharmacies and clinics to measure the amount of radioactivity (mCi or MBq) in the syringe prior to administration via injection into a patient's vein. A curie is defined as 3.7×10^{10} decays per second and a bequerel is one decay per second. Standards are used to calibrate the device at the relevant energy of the radiopharmaceutical. Since the walls of the counter stop alpha and beta radiation, a dose calibrator generally may be used only for the photon component of the decay radiation. One exception is the assay of very high energy beta emitters, such as ^{90}Y or ^{32}P. In these cases, the betas give off a continuous spectrum of X rays of appreciable energy while they are decelerated before coming to rest. Such brake radiation (bremsstrahlung in German) may be detected quantitatively to calculate the amount of high energy beta emitter present in the syringe. Lower energy beta emitters, however, present difficulty in quantitative assays and generally require a different strategy for detection.

Liquid scintillation (LS) counters are a third form of well counter. Here, the beta emitter is dissolved into a liquid hydrocarbon that has been doped so as to produce scintillations suitable for PM detection. These devices have wide application in the quantitative assay of low energy beta emitters used for *In vitro* biological research. Radionuclides of interest include ^{3}H (E_{beta} max $= 18$ keV), ^{14}C (155 keV) and ^{35}S (167 keV). Energies cited refer to the kinetic energy of the betas. These labels are generally used in type II labeling as shown in Table 1. Multiple samples are sequentially measured for a fixed counting interval by lowering the tube containing mixed scintillator and radioactive material into a darkened space viewed by one and probably two PM tubes. The sample is dissolved in a liquid (usually toluene), which is activated with small amounts of fluors, such as PPO (2,5-diphenyloxazole) and POPOP (4-bis-2,5-phenyloxazolyl) benzene as solutes so as to provide visible light upon being struck by the electrons released during decay. Standards are included in the experimental

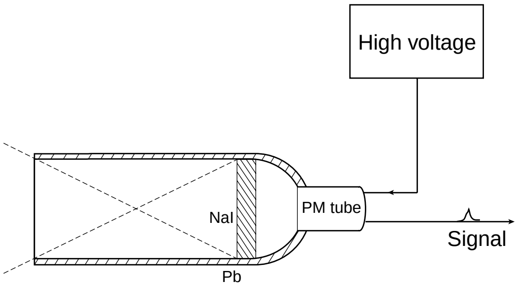


Figure 4. Operating principle of the clinical probe. Note that the observed field of view increases with distance from the opening of the collimator.

run to give absolute values for the activity. Efficiencies may approach 90% or more for moderately energetic betas. Reduction (quenching) of the light output due to solvent impurities and biological molecules within the sample can significantly affect the results and are accounted for by using standards.

The addition of fluors may not be needed to count very high energy beta particles. If the beta speed exceeds that of light in the solvent (E_β max > 0.26 MeV in water), a photon shock wave is produced in the medium. Termed Cerenkov radiation, the emitted light is analogous to that of the acoustic wave or sonic boom produced by an aircraft exceeding the speed of sound. An observer may use Cerenkov light, which includes a continuum of visible and ultraviolet (uv) photons, to directly quantitate beta activity in the sample.

Probes

Clinical probes contain a planar crystal, usually a right circular cylinder, placed at the end of a long, shielded tube called a collimator. The central field of view is typically on the order of a circle 10 cm in diameter. The collimator is another right circular cylinder so that the total field observed increases with distance from the opening. Since the patient has a relative small thickness, on the order of 30 cm or less, this expanding view is not detrimental to the resultant clinical counting experiments. Such static NaI (Tl) devices are routinely used in measurements of thyroid uptake of radioactive ^{123}I as described above. Figure 4 contains a cross-section through a typical probe. With a single probe giving a result of activity for a relatively small, fixed field of view, it is necessary that sets of several probes be employed for measurement in an extended or spatially variable organ.

Conformal arrays have been used to yield information on regional cerebral blood flow (rCBF) in patients. Sets of 10 or more detectors have been arranged around the patient's skull so as to measure regional accumulation of perfusion tracers, such as ^{133}Xe in the brain. Because of the low gamma energy (81 keV) of ^{133}Xe, a given probe essentially views only physically adjacent tissues in rCBF counting. Such arrays led to the discovery that regional brain–blood flow varied with the mental task that the patient was performing during the time of observation. In this application, it is necessary that the intervening scalp blood flow be subtracted from the time–activity curves for each region. Tomographic methods such as PET do not suffer from this limitation. The PET flow measurements have confirmed the probe rCBF results and generalized them to other aspects of brain blood flow and metabolism during conscious and subconscious thought processes.

A more recent probe application is the detection and uptake measurement of so-called sentinel lymph nodes in melanoma, breast, and other cancer patients. These sites are defined as the first draining node associated with the lesion. They are located following a near-primary injection of a ^{99m}Tc-labeled cluster of sulfur colloid particles. Particle sizes up to 1×10^3 nm may be used. Of necessity for spatial resolution, the hand-held probe has a greatly reduced field of view, on the order of a few millimeters, and may be driven by battery power for convenience in the operating room (OR). Because of size limitations at incision sites, such probes may be of the solid-state type, whereby the ionizing event is converted to an electronic signal directly without the necessity of PM tube signal amplification. At present, CdTe and CsI(Tl) detectors have been incorporated into clinical probe systems. In the latter case, a photodiode is used in lieu of a PM tube to provide miniaturization of the device. An example of a surgical probe is shown in Fig. 5. For use in the OR, the device is usually gas sterilized, and then placed into a plastic sleeve before being put into an operating field.

Similar probe applications can involve radiolabeled antibody proteins used to locate small metastatic lesions in cancer patient after removal of their primary tumor. This has been termed radioimmune-guided surgery (RIGS). By measuring the gamma activity per gram of excised tissue, the radiation oncologist may estimate the radiation dose achievable with that patient's disease if radioimmunotherapy (RIT) were eventually utilized. In the case of RIT, a beta label is attached to the antibody in lieu of the gamma label used in localization if the radionuclide label does not emit both types of ionizing radiation. Probe-guided biopsy allows direct treatment planning for the RIT procedure that may follow.

Probes are also available for positron detection in the OR. This measurement assures the surgeon that the resection has taken out all of the suspect tissue that has been previously located using a FDG imaging study and a PET scanner. Because of the presence of both annihilation 511 keV photons and positrons, some correction mechanism is necessary for these instruments. A dedicated microprocessor attached to the detector system will provide this information if the probe has separate sensitive elements for positrons plus photons and for photons alone so that a subtraction may be done in real time.

TWO-DIMENSIONAL DETECTORS

Rectilinear Scanners

Because of the limited field of view of single probes, it was once considered clinically relevant for such devices to be mounted on a motor-driver chassis so as to pass in raster fashion over an entire organ. The trajectory of the probe in this context is the same as a gardener mowing the lawn. A simple thyroid probe in this application would prove problematic since it is focused at infinity; that is, observes all

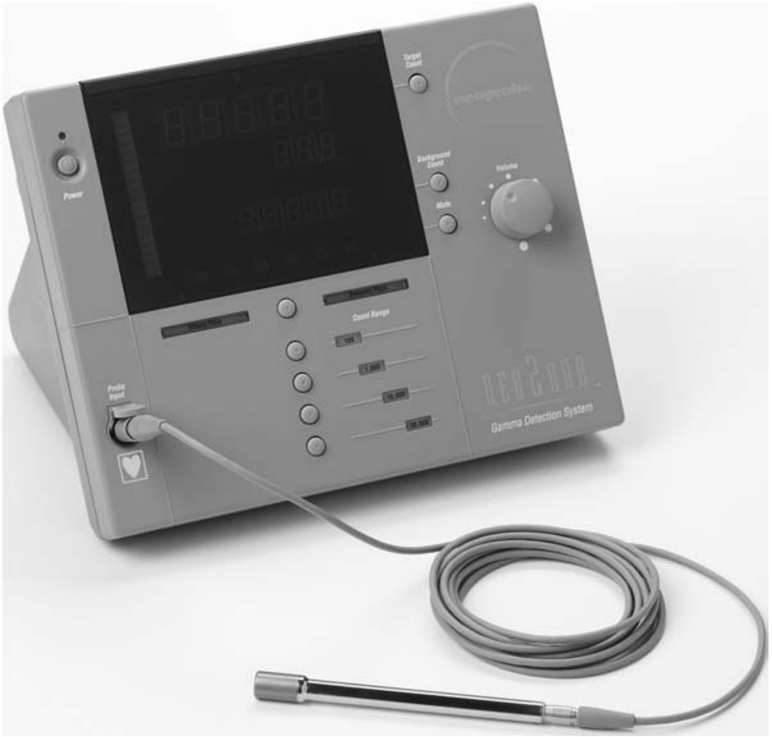


Figure 5. Operating room probe. Miniaturization is dictated by the need to minimize the incision site at the sentinel lymph node. With robotic developments, even smaller designs will be necessary.

tissues from one side of the patient through to the opposite side. It can be used on the thyroid since no other organ taking up radioiodine usually lies within the neck region. In order to generally restrict the depth of the field of view, focused (converging) collimation was developed for raster-driven rectilinear scanner systems so that only emitters at a fixed distance were detected with relatively high efficiency. Dynamic studies, whereby activity was imaged during its physiological motion within the body, were difficult with this device unless the kinetics were significantly slower than the total raster scan time. Today the rectilinear scanner is a historical artifact that is no longer used in the clinic because of the development of the gamma camera. A camera allows both static and dynamic imaging over a reasonably large field (50 cm) without requiring movement of the detector assembly.

Gamma Cameras

H. Anger, in the late 1950s, avoided most of the scanner problems by inventing a gamma camera. As in the probe example, a right circular cylinder of NaI(Tl) was used to detect the photon. However, instead of a single PM tube, a hexagonal array of such tubes was employed to determine (triangulate) location of a given scintillation within the detector's lateral (x, y) dimensions. This fundamental principle is illustrated in Fig. 6. In order to spread the light somewhat more uniformly over the PM cathode, a light pipe (diffuser) is generally interposed between scintillation crystal and photomultiplier array. Localization was originally done with an analogue computer measuring the relative signal strength from each of a set of PM tubes. A second type of processing occurs with the sum of the PM signals. An energy window is set so that only photons having energy within a prespecified range are recorded as true events. The window is sufficiently wide, for example, ±10%, that most signals arising from PE absorption of a monoenergetic gamma are recorded, but other photons, such as those scattered in the patient, are rejected. If the radionuclide emits several different photons, separate energy windows are set to count each energy level. The sum of all counts within all windows is then taken as the clinical result.

The original camera had cylindrical geometry arising from the single-crystal shape. Modern cameras generally have rectangular NaI(Tl) detectors made by combining annealed crystals of relatively large size allowing the

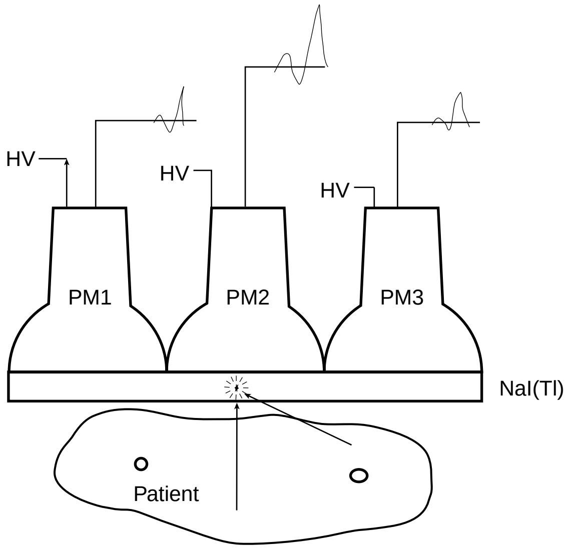


Figure 6. Principle of the Anger gamma camera. If no collimation is included (as shown) there is ambiguity of decay position.

entire width of the patient to appear in one field of view. The ensemble of crystal, multiple PM tubes, and associated computer electronics is referred to as the camera head. It is usually in the form of a rectangular solid and is mounted on a gantry allowing rotation and translation with respect to the patient bed. In the latter case, the motion is one dimensional (ID).

In the absence of directional information, a photon coming from anywhere within the entire hemisphere above the detector may impact the same position on the camera face. To remove the ambiguity, it is necessary that a collimator be provided between the detector crystal and the radioactive object(s). A collimator projects the activity distribution onto the crystal face. Essentially, this is a shadow or projection of the radioactivity distribution. Four standard types of collimators are shown in Fig. 7. The most common of these in clinical use is the parallel-hole type that is focused at infinity; that is, only passes parallel photons (rays) coming from the tissue of interest. Notice that the image and object size are equal in this case (magnification, M, $= 1$). This is essentially the same geometry used in the thyroid probe. Divergent collimators minify ($M < 1$) and convergent collimators magnify ($M > 1$) radioactive objects being imaged. The terms divergent and convergent refer to the point of view of the camera crystal. Convergent collimation is focused at a point in space; this is the same type of system used in the rectilinear scanner described above. However, in the camera case, the focal point is on the other side of the patient where this is no activity. Pinhole collimation may lead to either magnification or minification depending on the location of the object relative to the pinhole aperture.

Efficiencies of all collimators are relatively poor with pinholes becoming the worst at extended distances from the camera face. Typical values are on the order of 1×10^{-4} for commonly used parallel-hole types. Thus, if an experimenter deals with a very flat (essentially 2D) source, such as a thin radioactive tissue sample, it is better to simply remove all collimation and use the intrinsic localizing capability of the bare crystal and attached PM system. A transparent plastic sheet should be placed between source and camera fact to minimize possible contamination.

Every collimator is designed to be effective at a given photon energy. Lead septae in the device are effectively four to five half-value layers for the quantum of interest. A half-value layer is that thickness of material that reduces the intensity of gamma radiation by a factor of 2. Thus, using a collimator designed for high energy photons in the case of a relatively low energy emitter will lead to both lower efficiency as well as poorer image quality. For radionuclides emitting several different photons, the collimation must be appropriate for the highest gamma ray energy being measured. If this is not done, a hazy background of events due to these photons passing through the collimator walls will obscure the image.

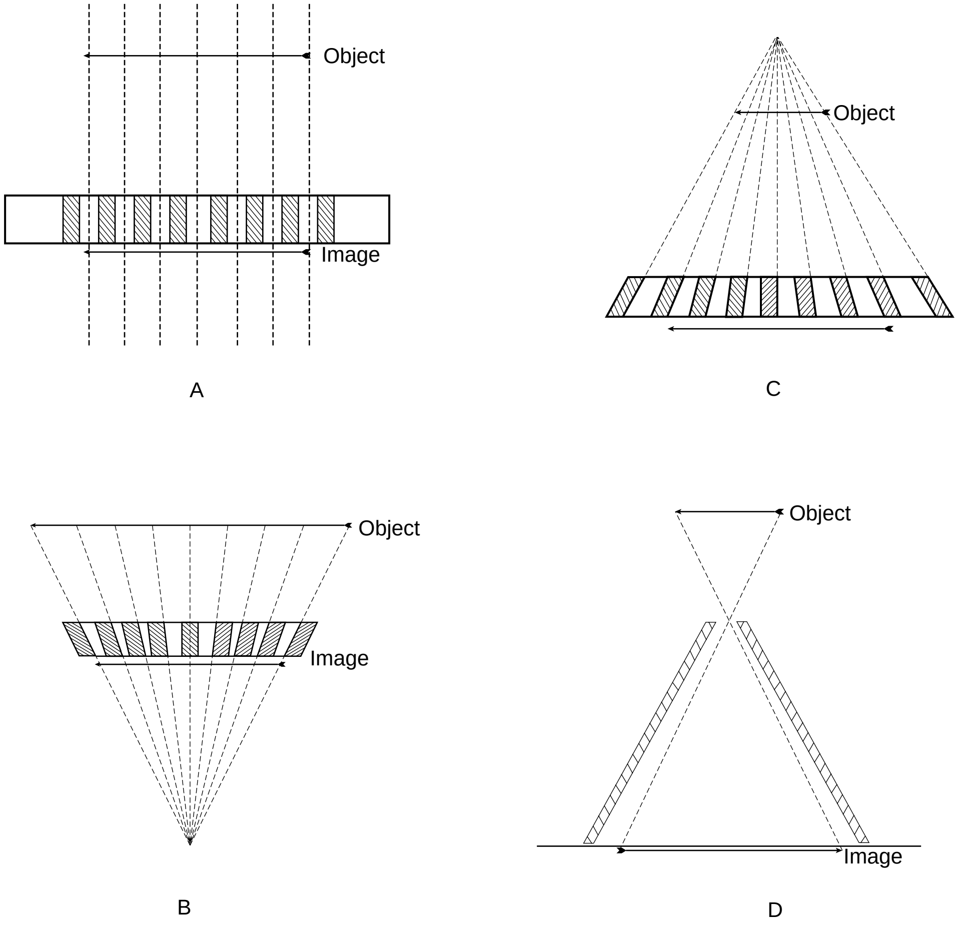


Figure 7. Four standard types of collimators used on the gamma camera.

Spatial resolution of gamma camera systems is on the order of 1 cm near the collimator surface, but generally becomes worse with increasing source depth inside the patient. When the count rate becomes extremely elevated, however, the localizing algorithms of such devices can be confused by multiple simultaneous scintillations with resulting imaging artifacts and reduced resolution. Most clinical protocols recognize this limitation by keeping the count rate at or below 5×10^4 counts per second (cps).

Because absolute measurement of resolution as well as object (organ) size is important, it is useful to image point sources of radiation for testing each camera-collimator system. This test object may be a set of small (1 mm) radiation sources of the imaged radionuclide having a known spacing. Resolution and object size in any resultant film or digital image can be defined directly using such devices. Variation with depth (patient thickness) and distance from the collimator may also be evaluated.

Digital Processor Applications Within the Camera Head

Anger's patented design originally relied on an analog computer to position the scintillation flash within the lateral dimensions of the NaI(Tl) crystal. Each scintillation event was weighted by location and signal amplitude of the several recording PM tubes. One of the original problems of the design has been the non-uniformity of response due to intrinsic and temporal variation in PM tube and other analog circuit components. In modern camera heads, digital processors are used to position the scintillation flash as well as perform spectroscopic analyses in real time on the detected events. Such dedicated processors inside the camera head observing individual PM tubes can greatly improve the uniformity so that the central field of view (CFOV) can have uniformities approaching 2%. Uniformity is particularly important for 3D imaging involving rotation of the camera head as described below. Values for each head are measured regularly with a flat source of radioactivity of an appropriate energy for most of the clinical imaging. Cobalt-57 is the radionuclide of choice for this procedure since it is close in emission energy (120 keV) to the common radiolabel ^{99m}Tc (140 keV) and has an extended half life of 270 d.

Note that communication formats are now available for information transfer between cameras and external computers. The digital imaging and communications in medicine (DICOM) standard is the international format for this transfer of information. This information may be used to produce comparisons of nuclear and other images to improve the diagnostic process.

Types of Acquisition from Gamma Cameras

One very important choice made by an operator prior to any camera study is the method of photon event recording in any external computer or work station memory. It is most common to acquire each scintillation as an event or count at coordinates (x, y). With total time of acquisition fixed at some realistic (patient-derived) limit, these events are added at their spatial positions to form a single digital image. This method of data recording is called frame mode. It is, by far, the most common type of camera data acquisition.

It may be that the timing of the tracer movement is either very rapid or uncertain for the patient–study. In that case, one may *a priori* choose list mode acquisition whereby each event is recorded as a triplet: (x, y, t) with computer clock time (t) included. After all events are list mode recorded, the operator or clinician may reconstruct the study in any sequence of time frames that is desired. For example, the first minute may be assigned to image 1, the second minute to image 2, and minutes 3–10 to image 3. Each of these images would appear to the reader as if they were taken in frame mode over that interval. Such an allotment may be revised subsequently as clinical questions arise. Large memory sizes are clearly useful if list mode imaging is to be pursued. Modern cameras often do not offer the possibility of list mode acquisitions, but instead rely on use of high speed frame-mode data recording.

A special type of frame mode acquisition is the gated study. Here, data are acquired in synchrony with a repeated physiological signal, usually the patient ECG. The R-wave-to-R-wave interval is predivided into a number (n) of equally spaced segments. Data obtained during time segment 1 of the cardiac cycle are placed into image 1, from time segment 2 into image 2, and so on. The result is a closed loop of n images that shows the beating heart when the gating signal is derived from the electrocardiogram (ECG).

External computer processing of camera data has been used to generate an additional type of output referred to as a functional image. For example, the clinician may wish to measure the rate of physiological clearance of a radiotracer from individual pixels within a time sequence of organ images. Using the external computer to calculate regional rate constants and to store this array, the resultant functional image displays the relative magnitudes of the computed kinetic values. Using an arbitrary scale, faster clearing regions are shown as brighter pixels. By looking at the functional image, regions of slower clearance can be readily identified and followed post subsequent therapies such as microsurgery for stroke patients.

Gamma Camera Types

Mobile Cameras. Battery-powered Anger cameras may be mounted on motor-driven chassis for use at the bedside or other remote areas. In such cases, the head is generally smaller than a static camera, on the order of 25 cm in diameter, and the energy range limited to 140 keV (^{99m}Tc) due to shielding weight concerns. Movement up ramps and using elevators would be restricted otherwise. Mobile units are most often utilized in planar heart work and have been involved in the testing of patients under escalating stress such as on a treadmill in cardiology. Patient evaluations in the OR or ICU are other applications of the device. Aside from breast imaging using ^{99}mTc -sestamibi, use of mobile gamma cameras has been limited, however, because of two specific reasons listed below.

Tomographic imaging is generally not possible with the mobile camera due to the difficulty of rotating the device in a rigorous orbit about the patient. In addition, use of high energy gamma labels is not possible for the minimally shielded detector head. Because of the importance of 3D imaging of the heart (see below), clinical usage has dictated that the more optimal study results if the patient is brought to the nuclear medicine clinic in order that optimal tomographic images be obtained.

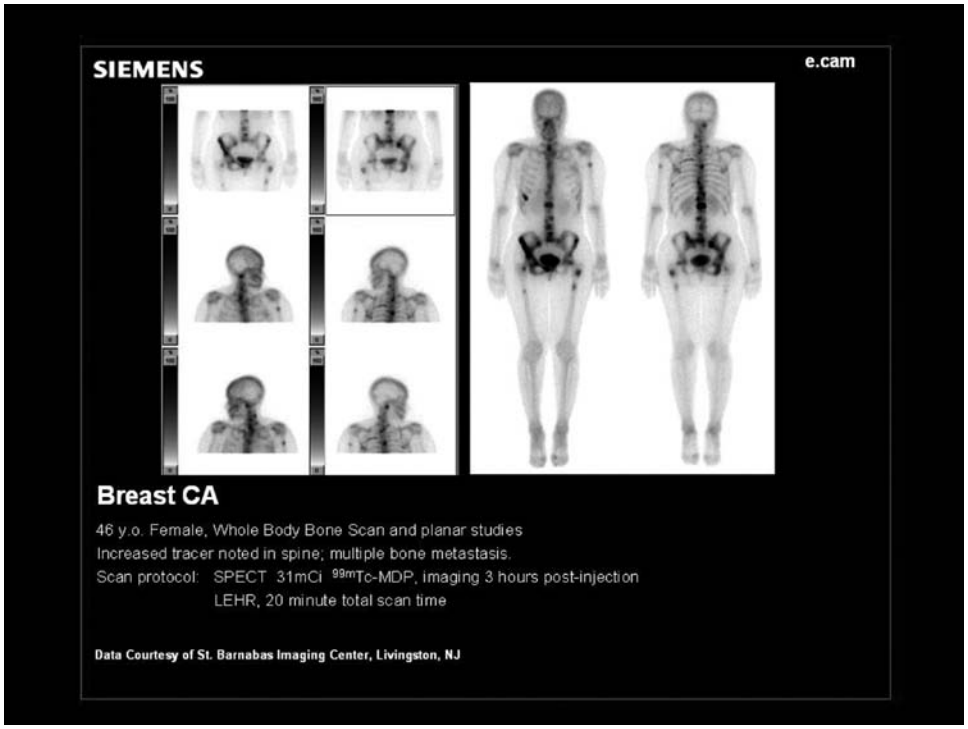
Static Single-Head Cameras. The most common camera type, the static single head, is usually a large rectangular device with a NaI(Tl) crystal having a thickness of ~6–9 mm. Larger thicknesses up to 25 mm may be useful for higher energy gammas, but loss of spatial resolution occurs as the PM location of the scintillation becomes more indeterminate. Lateral crystal dimensions are approximately 35 × 50 cm, although actual external size of the head would be significantly larger due to the necessity of having lead shielding surrounding the detector. This shielding must go both around the detector crystal as well as behind it to prevent radiation coming into the sensitive NaI(Tl) from the direction opposite the patient. Such protection is of importance in a busy clinical situation where more than one study is being conducted simultaneously in a relatively small space. Large rectangular camera heads permit simultaneous imaging over the entire width of a typical patient and allow whole body imaging with a single pass of the detector from the head to the feet. This is essentially an updating of the rectilinear scanner concept although here it is a 1D motion (z).

Images from Single-Head Cameras. Two standard imaging formats are employed with the gamma camera. Regional images, or vignettes, are taken of the organs of interest in the clinical study. A patient complaining of pain in the knee will be placed adjacent to the camera to permit various views of that joint following administration of a bone-seeking radiotracer, such as ^{99m}Tc-MDP. In addition, a whole-body image may be acquired to check for overall symmetry of tracer uptake. An example of the latter is given in Fig. 8. Here, the camera head is driven from the head to the foot of the patient and a series of frame images acquired over a span of 20–30 min. A computer attached to the camera allows these separate images to be seamlessly united to form the whole-body format.

Anger's camera concept has had one of its greatest impacts in cardiac dynamic imaging, whereby the sequential heart images are stored in a repetitive sequence that is correlated to the ECG signal obtained from the patient as described above. Figure 9 includes a continuous loop of 16 images of the left ventricle during a cardiac cycle using a labeled red cell tracer based on ^{99m}Tc. By setting a computer-generated region of interest (ROI) over the ventricle, one can measure the relative amount ejected; that is, the left ventricular ejection fraction (LVEF). Note that absolute amount of the tracer is not needed in the study since it is only a fractional ejection fraction that is of interest to the cardiologist. Irregular heart beats and/or patient motion during the 10–20 min of data taking can make such studies difficult to process.

Other dynamic studies are popular and clinically important. These include the renogram whereby the uptake and clearance of a filtered agent, such as ^{99m}Tc-DTPA is measured over a 1 h period. Both kidneys are followed and characteristic times of tracer accumulation and excretion are estimated by the radiologist: often using external computer software. A partial listing of typical studies involving gamma camera image data is included in Table 3.

Multiple-Head Cameras. It is becoming common to use more than one gamma detector head within a single



Breast CA

46 y.o. Female, Whole Body Bone Scan and planar studies
Increased tracer noted in spine; multiple bone metastasis.
Scan protocol: SPECT 31mCi ^{99m}Tc-MDP, imaging 3 hours post-injection
LEHR, 20 minute total scan time

Data Courtesy of St. Barnabas Imaging Center, Livingston, NJ

Figure 8. Whole-body image of a bone scan patient using translation of the gamma camera from head to foot. A sample of 20 mCi of ^{99m}Tc-MDP was used as the radiotracer for this image taken at 4 h postinjection.

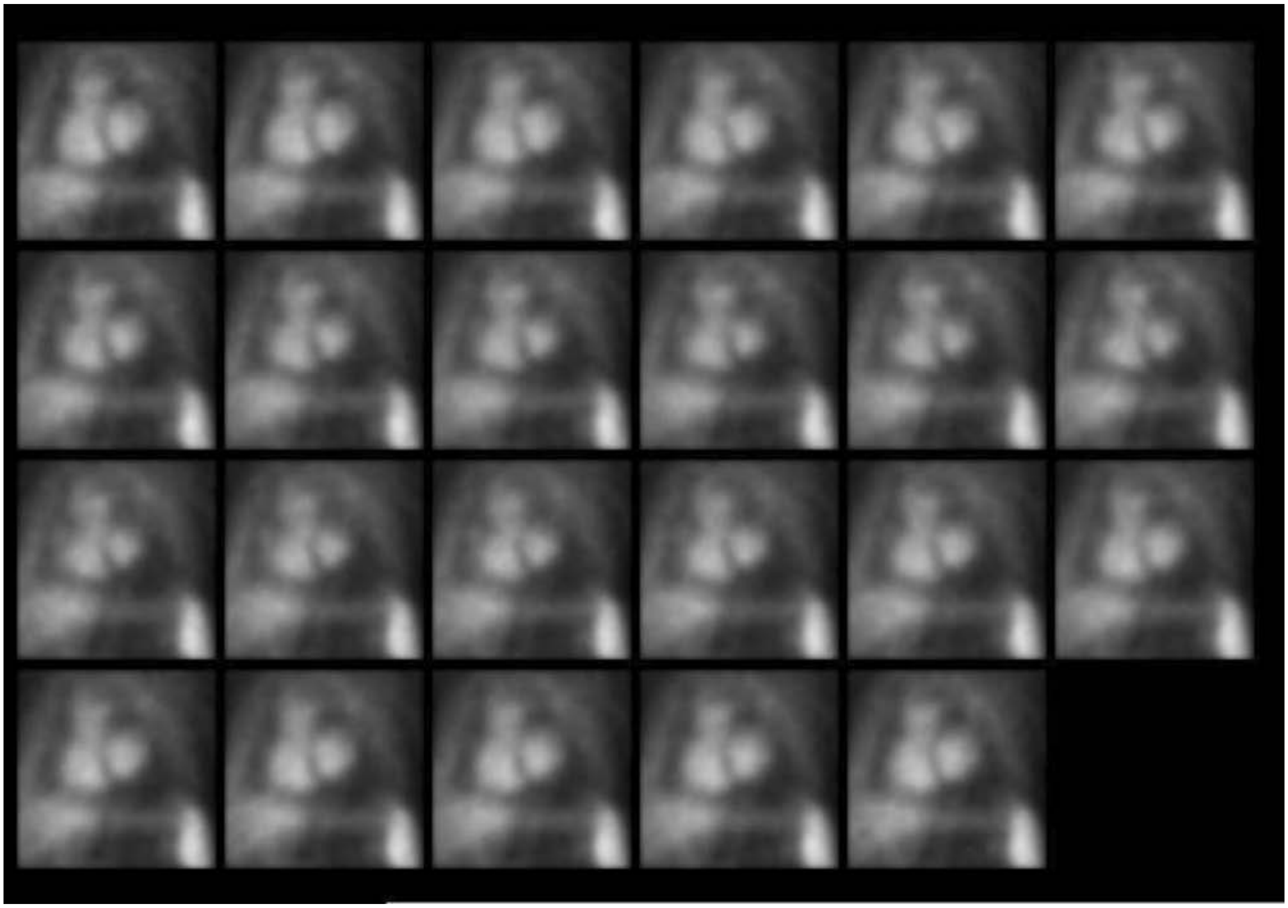


Figure 9. A multiple gated (MUGA) study of the left ventricle. The 16 images acquired over a heartbeat are uniformly assigned in time to the R wave-to-R wave cardiac interval. An ejection fraction of 69% was calculated.

supporting gantry (cf. Fig. 10). Speed of data acquisition, in either 2D or 3D mode, is the most important reason for this augmentation. By using two heads in a 2D study, the patient may be imaged from opposing sides simultaneously. Thus, if the organ of interest or tumor site were closer to the back of the patient, one could obtain information from the posterior head that would be useful even if the anterior head showed no discernible uptake sites. Alternatively, anterior and lateral views of an organ system may be obtained simultaneously and serially in a dynamic study of gastric emptying, for example. A second, and very important, application of multiple head camera systems is in more efficient 3D imaging.

THREE-DIMENSIONAL DETECTORS

There are two quite distinct methods to provide 3D imaging in nuclear medicine. If one uses ordinary (nonpositron) gamma-emitters, the strategy is referred to as single-photon emission computer tomography or SPECT.

SPECT Imaging

Here, the detector head or, more likely the set of two or three heads, is rotated around the patient over an extended arc. This orbit may be a full 360° arc or may be less due to body habitus or tissue location. One uses the rectangular Anger

Table 3. Representative Gamma Camera Imaging Studies Done in Nuclear Medicine

Study	Agent	Label	Device	Results
Renogram	DTPA and MAG3	^{99m}Tc	Camera	Kinetic values
MUGA	Red cells	^{99m}Tc	Camera with EKG gating	Ejection fraction of LV
Myocardium	Sestamibi	^{99m}Tc	Camera	Bulls eye image of LV
Bone scan	MDP	^{99m}Tc	Camera	Fracture location. Tumor location
Lung scan	Aggregated albumin	^{99m}Tc	Camera	Regions of reduced perfusion
Lung scan	Aerosolized albumin	^{99m}Tc	Camera	Regions of reduced ventilation
Lung scan	Xenon gas	^{133}Xe	Camera	Regions of reduced ventilation
Thyroid imaging	Iodine	^{123}I	Camera	Uniformity of uptake in gland

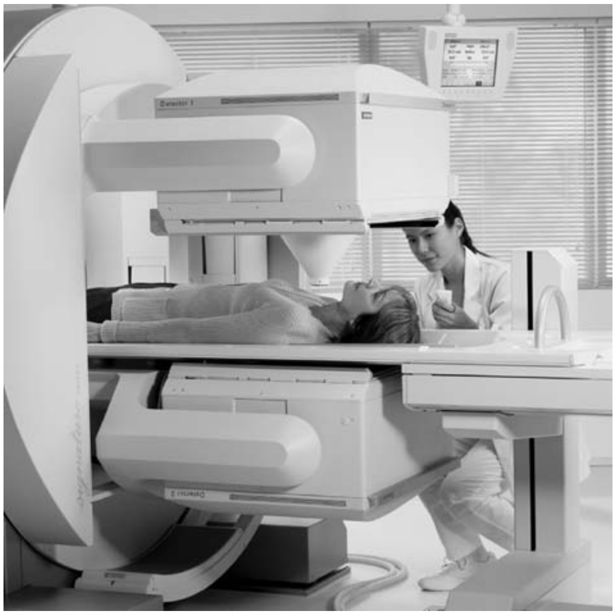


Figure 10. Dual-headed gamma camera. Both detector heads are mounted on the same gantry to allow translation (for whole body) and rotation (for SPECT) of the system. An open geometry permits use of gurneys with this system.

head as described above with the parallel-hole collimation in place. With injected activities on the order of 100–300 MBq, data acquisitions require on the order of 20 min. Patient immobility is necessary. Data may be taken in a shoot-and-step mode at fixed angular intervals or they may be acquired continuously during the rotation. Storage of such vast amounts of information requires a dedicated computer system recording the counts at each spatial position on the head (x,y) and at each angle (θ) during the rotation.

Several reconstruction algorithms are available to the technologist to generate the requisite tomographic images of the patient. Corrections for attenuation and Compton scatter must also be applied for the generation of these images. While pseudo-3D images may appear on the computer monitor as an output of the reconstruction, the radiologist will review and file to the picture archival and communication system (PACS) system the transaxial, sagittal, and coronal projections of the activity. It is important to realize that numerical values usually shown in these various projection images are not absolute, but only relative quantities. Quantitative SPECT, in which the numerical pixel value is equal (or at least proportional) to the activity in Bq, requires, in addition to the above corrections, that a set of standard sources of the same radionuclide be imaged along with the patient. Such calibrations can be done simultaneously with the clinical study, but are usually performed as a separate procedure. Figure 11 shows the three projection sets (axial, sagittal and coronal) in the case of a patient having a ^{99m}Tc sestamibi myocardial scan of the left ventricle.

PET Systems

Back-to-back photon emission (511 keV each) characteristic of positron decay of a labeling radionuclide has led to the development of PET. While paired Anger camera heads have been used as the detectors, it is much more efficient to use a ring of solid-state scintillation detectors arrayed around the patient. Bismuth germinate (BGO) has been the standard material, but LSO (lutetium orthosilicate) is becoming more popular due to its higher light output and shorter pulse length at 511 keV. In the standard situation, each detector block is broken into separate light emitting substructures that act as individual scintillation detectors. By having a few phototubes observing a separate block of such elements, the number of PMs may be reduced using Anger's gamma camera principle. Whole body PET scanners may have $> 10^4$ individual scintillators arrayed in an open circle or set of rings around the patient bed. Multiple rings are conventional so that several axial sections may be acquired simultaneously over a distances of 10–15 cm. Note that no detector rotation is inherently required since the solid-state system completely encircles the patient. If needed, the bed will be driven along the axis of the detector rings in order to perform extended imaging of the subject. The most common study utilizes FDG with ^{18}F as the radiolabel and covers the patient from head to groin. Sites within the body that metabolize glucose are imaged thereby. Brain and possible tumor areas are important applications of PET glucose imaging. Ambiguity with infection sites is a limitation to this protocol; this is particularly the case in the immune-compromised patient.

Because the two emitted photons are coincident in time and define a line in space, the positron detection process does not, in principle, require collimation (Fig. 12). Using contiguous rings of detectors is the most common system design; if the rings act alone or together as a single detector system defines the two types of imaging that are performed on a PET system. Internal (patient) photon attenuation is taken into account in the reconstruction of the PET image set. This is done using a transmission source of positron emitter, usually ^{68}Ge, to evaluate the patient thickness for the various ray directions at each bed position. Typically, the attenuation correction occurs during the scanning procedure with a short time interval given over to use of the source at each bed location.

Two-Dimensional PET Imaging

A clinical PET scanner is shown in Fig. 13. In 2D PET, every ring of detectors is isolated by tungsten collimation from all but single adjacent rings. Thus, each circle of solid-state scintillators is used in isolation to generate a single axial slice through the patient. This approach yields the highest resolution available in positron tomography with systems having spatial resolutions on the order of 5 mm. Reduction in the amount of scatter radiation is also obtained in 2D images. A FDG image is given in Fig. 14. While described as 2D, the result is actually tomographic and gives the usual projections in the three planes intersecting the patient's body. In these planes, the precise estimate of resolution depends on the positron's kinetic energy. One must combine, in quadrature, the positron range in soft tissue with inherent ring resolution to predict the overall spatial distance ambiguity. Higher energy positron emitters will have correspondingly poorer spatial

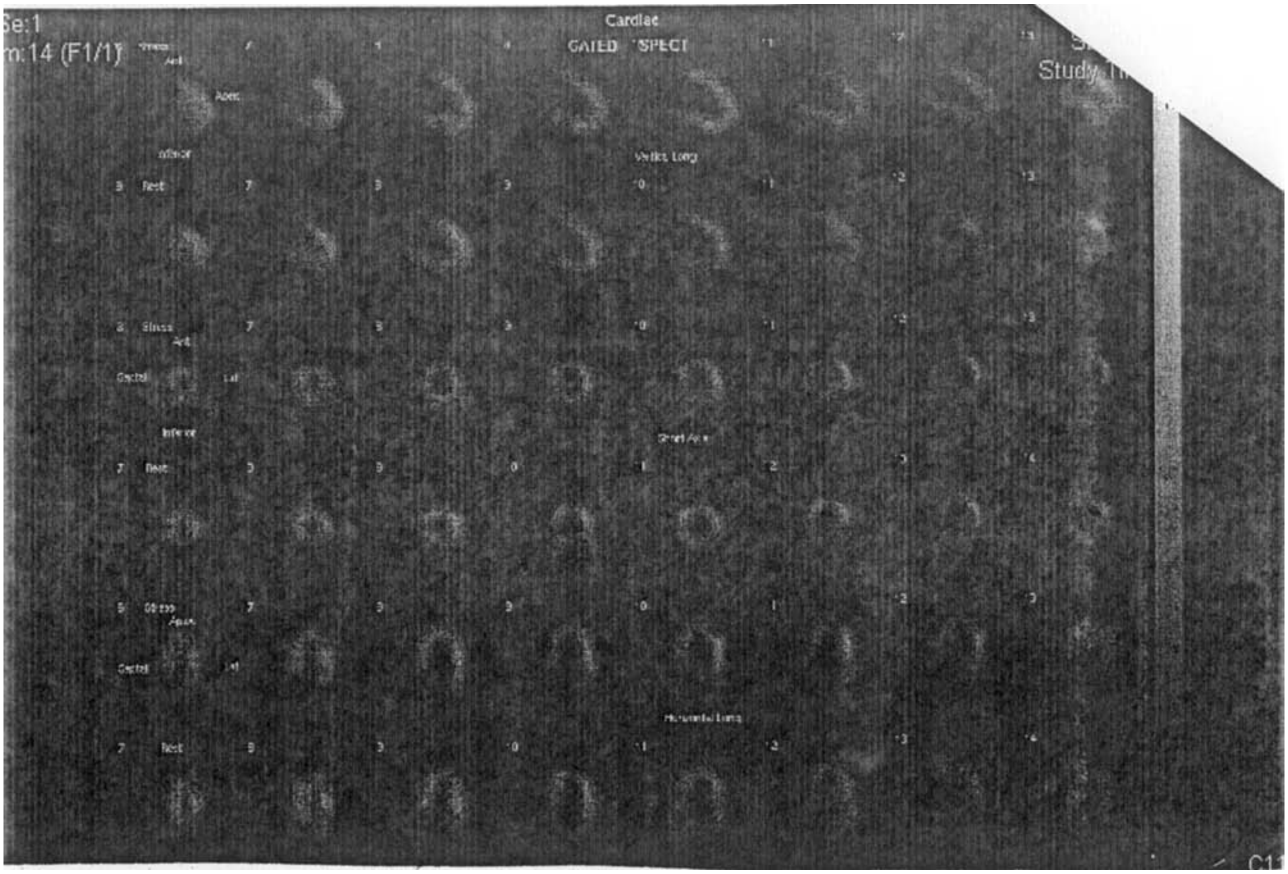


Figure 11. The SPECT image set for a gated myocardial study. In each pair of rows, the upper set of images gives the stress result, the lower set the resting result. The patient received 30 mCi of ^{99m}Tc-sestamibi for the study.

resolution due to the greater range of the positron prior to decay.

Other criteria for the selection of a positron label may be applied; for example, the half-life of the radionuclide. If that lifetime is very short, manufacturing and targeting may take so many physical half-lives that imaging is not possible. Additionally, one should consider the relative probability of β^+ emission in the decay scheme. This likelihood may be reduced because of competition with electron capture from the K shell of the radionuclide. Additionally, there is the possibility that other photons may be emitted along with the positrons so as to cause a background effect in the PET scanner. For example, ^{124}I, along with annihilation radiation at 511 keV, also emits ordinary gamma rays with energy in excess of 2 MeV. Such high energy photons readily penetrate collimators to reduce contrast in the images and make quantitation of the absolute radioiodine activity difficult.

Three-Dimensional PET Imaging

When the collimation between PET scanner rings is removed, each circle of detectors can have coincidences with itself as well as with all other detector rings. This mode of operation is referred to as 3D imaging. Spatial resolution is somewhat worse than that of the collimated (2D) case and may be 1 cm or more. However, the added sensitivity may be very important: particularly if whole body images are to be obtained in a patient with possible multiple sites of interest such as a referral from medical oncology. Sequential PET images of the whole body may be used to evaluate chemotherapy or other interventions. A quantitative method is available for such comparisons.

One feature of PET imaging merits emphasis. In the quality assurance of the positron scanner, the operator will routinely obtain transmission images through a phantom of known size using 511 keV photons from an external source. With this information and calibration using a known activity source, the user may reconstruct radioactivity distributions in the patient with absolute units. Thus, the concentration of positron emitter at a given image voxel can be estimated. Called the specific uptake value (SUV), this parameter is essentially $\%\mathrm{ID}\cdot\mathrm{g}^{-1}$, where ID refers to the injected activity or dose (MBq). The resultant SUV value is a function of time. Two direct consequences result. First, the clinician can make comparisons between organ sites both now and with regard to earlier studies on that patient or relative to normal individuals. Results of therapy may be directly evaluated thereby. The SUV values may even be used to make diagnostic assessments, such as the likelihood of malignancy at the voxel level. In addition, the radiation dose to the entire organ and even to local volumes within the tissue may be directly made with the SUV parameter. This

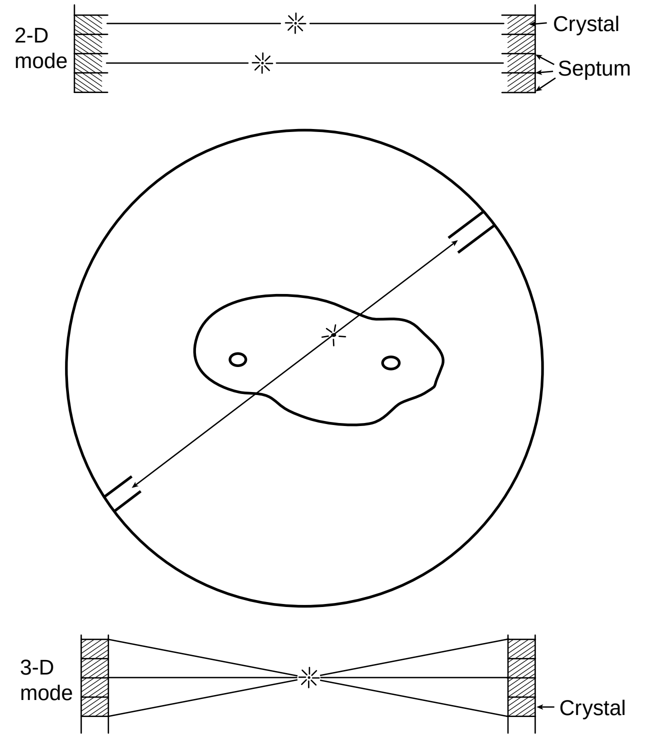
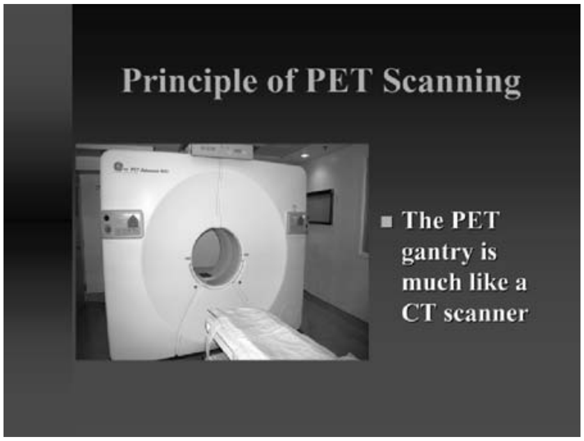


Figure 12. Principle of a PET scanner. Note that the direction of the annihilation radiation defines a line in space. 2D and 3D configurations are accomplished with and without collimation, respectively.

is in contrast to gamma camera planar data whereby the results may be quantified only with associated calculations that depend upon acquiring a set of images from at least two sides of the patient.

HYBRID IMAGING INSTRUMENTS

Nuclear image information, of either gamma camera or PET type, is limited in that regions of elevated (or reduced) activity are not necessarily identifiable as to anatomic location or even organ type. A patient may exhibit a hot spot in a planar gamma camera view that could correspond to uptake in a lobe of a normal organ, such as the liver or perhaps to an adjacent metastatic site. Similar arguments may be made with SPECT or PET images. Clinical decisions and surgical options are difficult to determine in this ambiguous context. Radiologists viewing nuclear medicine images are forced to cloak their patient assessments in correspondingly vague spatial terms.



Principle of PET Scanning

- The PET gantry is much like a CT scanner

Figure 13. A clinical PET scanner.

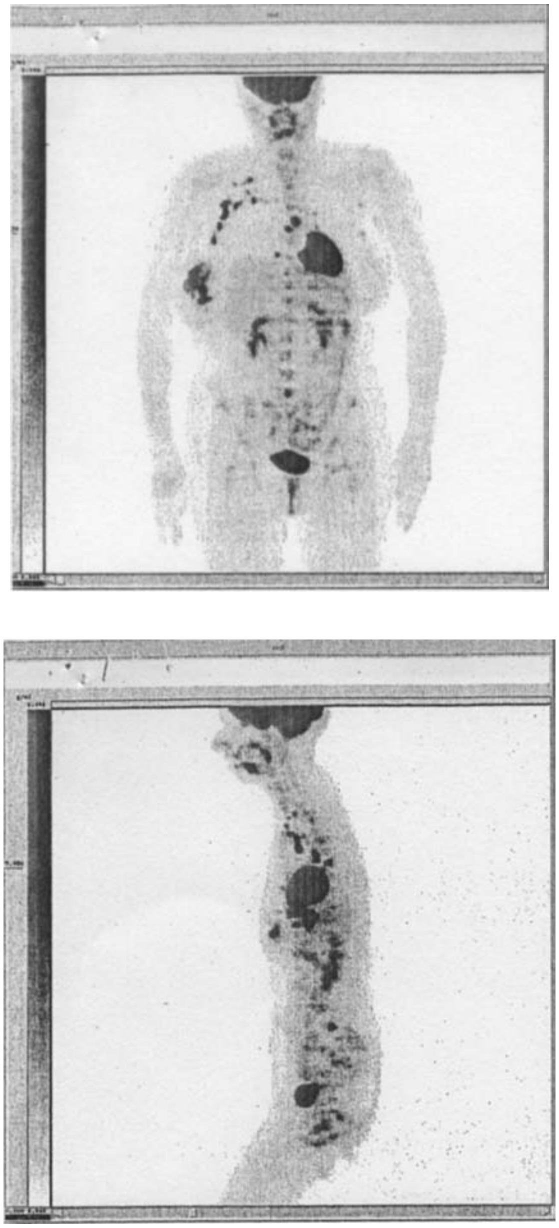


Figure 14. A PET image of a breast cancer patient following injection of 10 mCi of FDG. A MIPS projection is shown with areas of elevated FDG appearing as dark foci. Note accumulation in regional lymph nodes near the breast primary.

Lack of anatomic correlation has been one of the most difficult issues in the history of nuclear imaging. Physiological data determined with nuclear techniques are considered complementary to anatomical information separately obtained by other imaging modalities such as

CT or magnetic resonance imaging (MRI). The radiologist or referring clinician will frequently have to conceptually fuse disparate data sets to help identify the specific organ or tissue where a nuclear tracer uptake zone occurs. Using DICOM and PACS technologies, one may also attempt to digitally overlay nuclear and anatomic images. In this case, however, magnification, rotation, and translation of one image relative to the other must be accounted for with appropriate software and adjustable parameters. Using commercial programs, CT and MRI digital images may be fused to nuclear imaging results using least-squares techniques and an external workstation.

In order to remove this conceptual and computational bottleneck, recent developments in nuclear medicine have included manufacture of hybrid physiologic/anatomic imagers. In this strategy, both devices share a common patient bed so that two types of images are spatially registered and, although successive, nonetheless obtained within a few minutes of each other. Note that the PM tubes of a typical gamma camera or PET system are sensitive to magnetic field effects at the level of the earth's value; that is, at ~0.5 G. Yet clinical MRI scanners operate in the range of 1.5–3.0 T (1.5×10^4 to 3.0×10^4 G) so that hybrids of MRI and nuclear devices would be problematic. Thus, essentially all of the hybrid systems have involved combinations of nuclear and CT imagers.

SPECT/CT Hybrid Imagers

A logical approach to the issue of radionuclide localization is to have two scanners, one nuclear and one based on X-ray attenuation, located on attached gantries. This pair of devices shares the same patient couch. Because the distances of bed movement can be known within 1 mm or less, the user can identify an uptake volume in the nuclear SPECT image with a geometrically corresponding part of the anatomy as seen via CT scan. Additionally, attenuation corrections may be made more effectively using the CT data to improve SPECT sectional images. Some difficulties remain: (*1*) the breathing motion of the patient, and (*2*) possible changes in posture from one sequence to the other during the double imaging procedure. Complementary nature of the two images makes the interpretation of either somewhat clearer.

PET–CT Hybrid Imagers

Analogous to the gamma camera, a PET detector ring imager can be mounted adjacent to a CT scanner to provide registration of images from two modalities. As in the case of SPECT–CT devices, disparities in the speed of the two data acquisitions leads to some remaining ambiguity involving organs that move with respiration such as liver or lungs. While it is possible to hold one's breath for a CT scan, the PET whole body nuclear imaging time remains on the order of 20–30 min to preclude such possibilities for the emission segment of the study. A set of hybrid images and their superimposition are given in Fig. 15.

Radiation therapy treatment planning has been one of the primary beneficiaries of hybrid imaging devices. It may be that some mass lesions visible via CT or other anatomic imagers are necrotic or at least not active metabolically. This result can most clearly be seen in the fused image so that the more physiologically active sites may be treated with higher external beam doses. Likewise, with appropriate resolution, the radiation oncologist may elect to treat part of a lesion that has heterogeneous tracer uptake in an effort to spare contiguous normal (albeit sensitive) sites, such as in the lung, spinal cord, or brain. Those segmental regions of a tumor mass that are metabolically active may be targeted with external beam therapy using a number of linear accelerator strategies including conformal therapy, intensity modulated radiation therapy (IMRT) and tomotherapy using a rotating radiation source.

ANIMAL IMAGING DEVICES

As indicated previously, the growth of nuclear medicine is limited by availability of specific radiopharmaceuticals. Historically, useful agents were often discovered (sometimes by accident) and were almost never invented. This strategy is inefficient and modern molecular biologists and pharmacists attempt to directly engineer improved tracers for a given clinical objective; that is, imaging or therapy of a particular tissue or tumor type. A specific molecule or cellular organelle is generally the target in these efforts. Molecular imaging has become an alternative name for nuclear medicine. After initial protein or nanostructure development is completed, the next task is the determination of the relative usefulness of the prototype in an animal study. Usually, this work involves mouse or rat radiotracer biodistributions involving sacrifice of 5–10 animals at each of a number of serial times. If multiple time points and comparison of various similar radiotracers are involved, numbers of mice may approach thousands for the development of a single radiopharmaceutical.

It is more analogous to clinical procedure if serial images of the same animal are obtained during the course of the research study. Far fewer animals are required and the data are more homogenous internally. Imaging with standard-sized nuclear technology is generally unsatisfactory due to poor spatial resolution associated with typical gamma cameras (1 cm) or PET scanners (0.5 cm). Early investigators had utilized a suitably small pinhole collimator and gamma camera combination on mouse and rat imaging studies. By collimator magnification, the image can be made large enough that the internal structures can be resolved. As noted previously, magnification and sensitivity depend on distance from the pinhole so that quantitative interpretation of these images was difficult. Sensitivity of pinhole imaging was likewise low so that relatively large amounts of activity were required for the study. It is more effective if a dedicated, high efficiency, animal-size imaging device is designed for the experimental species. Such instruments have been developed for planar and SPECT gamma camera as well as PET imager systems.

Animal Gamma Cameras

Imaging a 10 cm mouse is best done with a gamma camera having approximately that sized crystal. Rather than employing a hexagonal array of multiple, miniaturized PM tubes to locate the scintillation, an animal camera

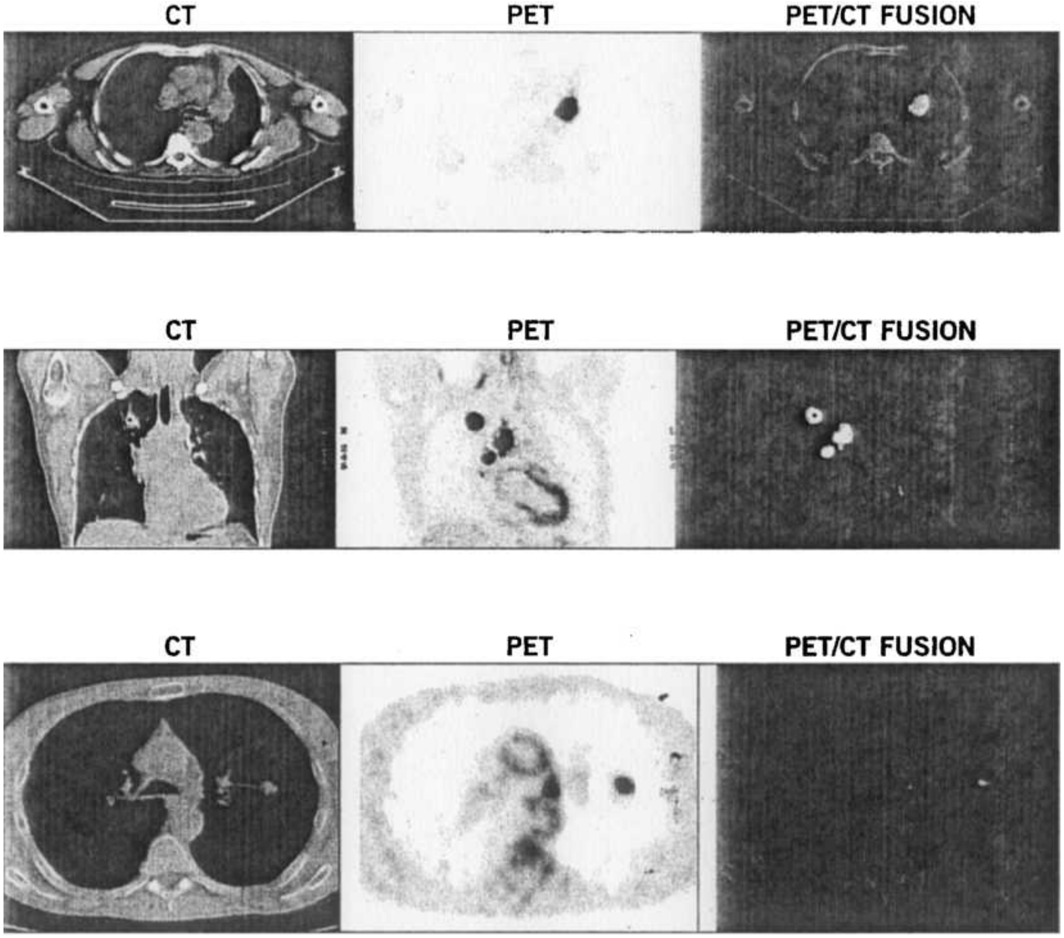


Figure 15. The CT–PET hybrid image showing respective CT, PET, and combined images. Clarity of location follows from the last of these results.

relies on the use of a single spatially sensitive PM tube. This device sends both x and y coordinates and the energy of the scintillation to a dedicated computer. Otherwise, the murine camera is operated essentially identically to the full-size version. Parallel-hole collimation is most common, although pinholes may be used to form highly magnified images of murine organs, such as the liver, kidneys, or even the thyroid. Figure 16 illustrates the last of these targets for a mouse receiving a tracer injection of ^{125}I to enable imaging of the murine thyroid. SPECT imaging is also possible; it is accomplished by rotating a rigorously constrained mouse or other small animal within the field of view of the camera. The usual projections, coronal, sagittal and transaxial are then available.

Animal PET Imagers

Miniature PET scanners have become of importance to the development of new radiopharmaceuticals. Here, a ring of BGO or LSO crystals is installed in a continuous cylinder extending over the entire length of the mouse. Spatial resolution is on the order of 2 mm or less over the 12 cm axial dimension. A sample image is given in Fig. 17 where a number of coronal sections are superimposed to improve the image statistics. Both ^{18}F-FDG and ^{64}Cu labeled to a modified antibody protein called the minibody were the positron emitters used in this study. Again, as in the clinical case, the PET images are intrinsically tomographic unlike the gamma camera results. Therefore, the PET animal imagers have a theoretical advantage in biodistribution

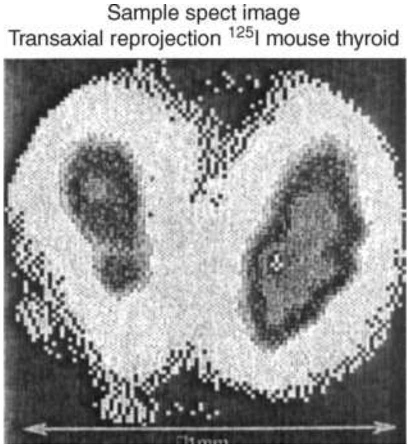


Figure 16. Animal gamma camera image of a mouse thyroid. Iodine-123 was used as the tracer with a pinhole collimator to obtain an image of the normal organ.

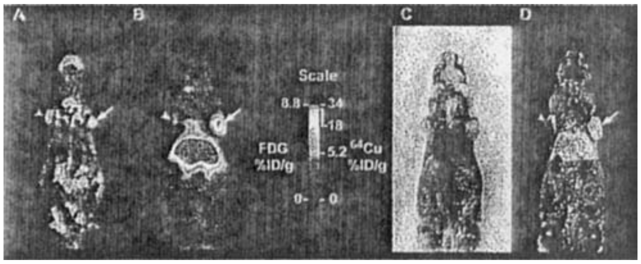


Figure 17. A PET animal scanner murine image. Both FDG (Image a) and ^{64}Cu-minibody (Image b) were used as radiotracers. Images c and d show pathology and autoradiographic results, respectively.

assays. As mentioned there remains the difficulty of finding a suitable positron emitter and method of attachment for a particular imaging experiment.

Hybrid Animal Imaging Devices

Quantitation of radioactivity at sites within the mouse's body is more easily done with an animal gamma camera than in the comparable clinical situation. This follows since attenuation of photons is relatively slight for a creature only a few cm thick in most cross-sections. Because of this simplicity, the output of the small animal gamma camera imaging systems can be modified to yield percent-injected dose (%ID). In order to correct for organ perfusion, it is historically conventional in biodistribution work using sacrificed animals to obtain uptake in %ID/g of tissue. Given the organ %ID, this last parameter may be obtained if the total mass of the target organ can be determined. Two avenues are available; one may employ miniaturized CT or a reference table of organ sizes for the particular strain of animal being imaged. We should note that suitably sized CT scanners are produced commercially and may be used to estimate organ mass. Hybrid SPECT/CT, PET–CT and SPECT–PET–CT animal imagers are now available for mouse-sized test animals.

One caveat regarding the small-scale imaging devices should be added; these systems cannot give entirely comparable results to biodistribution experiments. In animal sacrifice techniques, essentially any tissue may be dissected for radioactivity assay in a well counter. Miniature cameras and PET systems will show preferentially the highest regions of tracer accumulation. Many tissues may not be observable as their activity levels are not above blood pool or other background levels. Hybrid animal scanners can reduce this limitation, but not eliminate it entirely. Those developing new pharmaceuticals may not be concerned about marginal tissues showing relatively low accumulation, but regulatory bodies, such as the U.S. Food and Drug Administration (FDA), may require their measurement by direct biodistribution assays.

BIBLIOGRAPHY

Reading List

Aktolun C, Tauxe WN, editors. Nuclear Oncology, New York: Springer; 1999. Multiple images, many in full color, are presented of clinical studies in oncology using nuclear imaging methods.

Cherry SR, Sorenson JA, Phelps ME. Physics in Nuclear Medicine, 3rd ed. Philadelphia: Saunders; 2003. A standard physics text that describes SPECT and PET aspects in very great detail. This book is most suitable for those with a physical science background; extensive mathematical knowledge is important to the understanding of some sections.

Christian PE, Bernier D, Langan JK, editors. Nuclear Medicine and PET, Technology and Techniques, 5th ed. St. Louis: Mo. Mosby; 2004 The authors present a thorough description of the methodology and physical principles from a technologist's standpoint.

Conti PS, Cham DK, editors. PET-CT, A Case-Based Approach, New York: Springer; 2005. The authors present multiple hybrid (PET/CT) scan case reports on a variety of disease states. The text is structured in terms of organ system and describes the limitations of each paired image set.

Sandler MP, et al. editors. Diagnostic Nuclear Medicine, 4th ed. Baltimore: Williams and Wilkins; 2002. A more recent exposition that is a useful compilation of imaging methods and study types involved in diagnosis. No description of radionuclide therapy is included.

Saha GP, Basics of PET Imaging. Physics, Chemistry and Regulations, New York: Springer; 2005. This text is a useful for technical issues and is written at a general level for technologists. Animal imaging is described in some detail and several of the commercial instruments are described.

Wagner HN, editor. Principles of Nuclear Medicine, Philadelphia: Saunders; 1995. The Father of Nuclear Medicine is the editor of this reasonably recent review of the concepts behind the field. A rather complete but somewhat dated exposition of the entire technology of nuclear medicine operations in a medical context.

Wahl RL, editor. Principles and Practice of Positron Emission Tomography. Philadelphia: Lippincott Williams and Wilkins; 2002. A solid review of PET clinical principles and practical results. Logical flow is evident throughout and the reader is helped to understand the diagnostic process in clinical practice.

See also COMPUTED TOMOGRAPHY, SINGLE PHOTON EMISSION; NUCLEAR MEDICINE, COMPUTERS IN; POSITRON EMISSION TOMOGRAPHY; RADIATION PROTECTION INSTRUMENTATION.

NUCLEAR MEDICINE, COMPUTERS IN

PHILIPPE P. BRUYANT
MICHAEL A. KING
University of Massachusetts
North Worcester, Massachusetts

INTRODUCTION

Nuclear medicine (NM) is a medical specialty where radioactive agents are used to obtain medical images for diagnostic purposes, and to a lesser extent treat diseases (e.g., cancer). Since imaging is where computers find their most significant application in NM, imaging will be the focus of this article.

Radioactive imaging agents employed to probe patient pathophysiology in NM consist of two components. The first is the pharmaceutical that dictates the *in vivo* kinetics

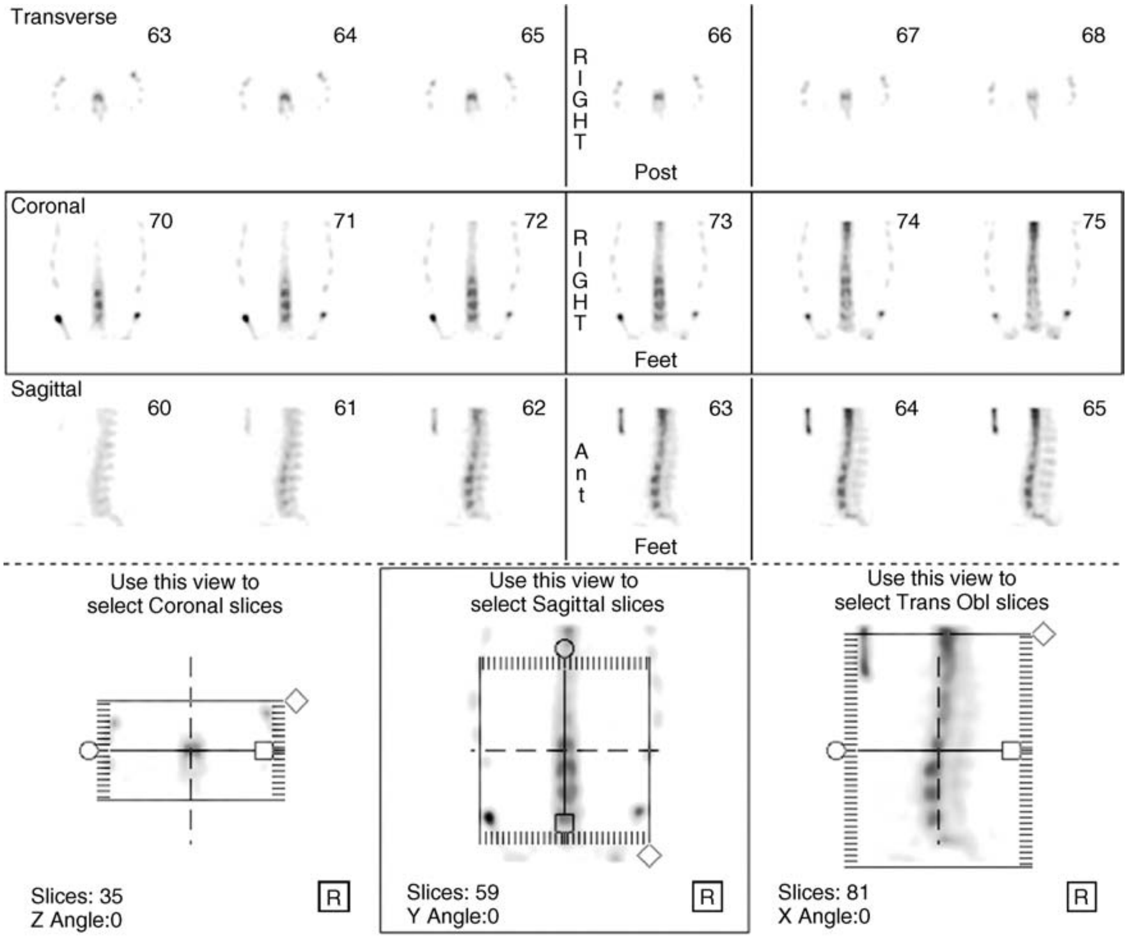


Figure 1. Images after a bone scan.

or distribution of the agent as a function of time. The pharmaceutical is selected based on the physiological function it is desired to image. The second component is the radionuclide that is labeled to the pharmaceutical and emits radiation that allows the site of the disintegration to be imaged by a specifically designed detector system (1). An example of an imaging agent is technetium-99 m labeled diphosphate, which is used to image the skeleton. The diphosphate is localized selectively on bone surfaces by 3 h postinjection and the technetium-99 m is a radionuclide that emits a high energy photon when is decays. A normal set of patient bone images of the mid-section is shown in Fig. 1. Another imaging agent example is thallium-201 chloride, which is localizes in the heart wall in proportion to local blood flow. In this case, thallium-201 is both the radiopharmaceutical and radionuclide. A normal thallium-201 cardiac study is shown in Fig. 2. A final example is the use of an imaging agent called fluorodeoxyglucose (FDG), which is labeled by the positron emitting fluorine-18. As a glucose analog, FDG is concentrated in metabolically active tissue such as tumors. Figure 3 shows a patient study with FDG uptake in a patient with lung cancer. Dozens of tracers are available to study a variety of pathologies for almost all organs (heart, bones, brain, liver, thyroid, lungs, kidneys, etc.).

Because the amount of radioactivity and the imaging duration are kept at a minimum, NM images are typically noisy and lack detail, compared to images obtained with other modalities, such as X-ray computerized tomography (CT), and magnetic resonance imaging (MRI). However, CT and MRI provide mainly anatomical information. They provide less functional information (i.e., information regarding the way organs work) in the part because these techniques are based on physical properties (such as tissue density. . .) that are not strikingly different between normal and abnormal tissues. Actually, after recognizing the differences between the anatomically and physiologically based imaging techniques, the current trend in the diagnostic imaging strategies is, as seen below, to combine anatomical information (especially from CT) and functional information provided by NM techniques (2,3).

As seen below, computers play a number of fundamental roles in nuclear medicine (4). First, they are an integral part of the imaging devices where they perform a crucial role in correcting for imaging system limitations during data acquisition. If the acquired data is to be turned from two-dimensional (2D) pictures into a set of three-dimensional (3D) slices, then it is the computer that runs the reconstruction algorithm whereby this is performed.

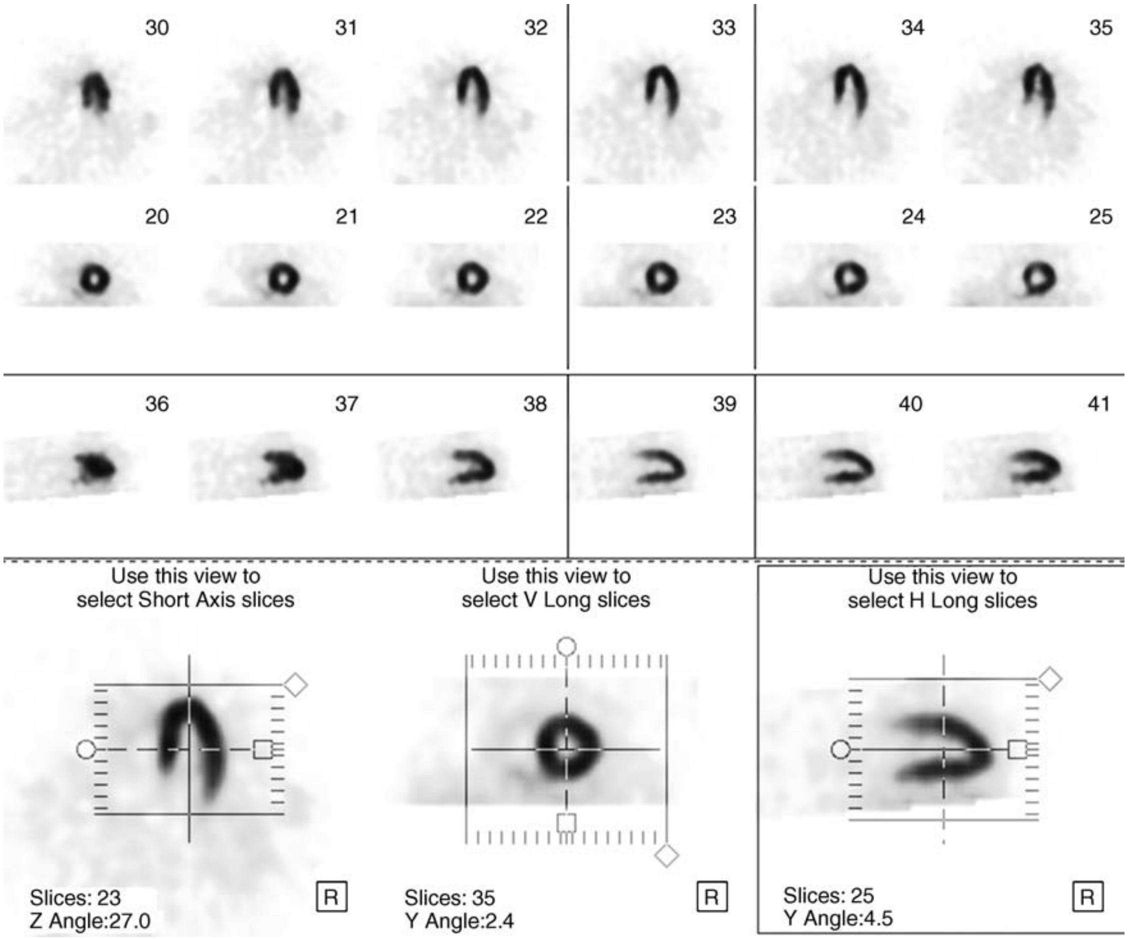

30
31
32
33
34
35
20
21
22
23
24
25
36
37
38
39
40
41
Use this view to select Short Axis slices
Use this view to select V Long slices
Use this view to select H Long slices
Slices: 23
Z Angle:27.0
R
Slices: 35
Y Angle:2.4
R
Slices: 25
Y Angle:4.5
R

Figure 2. Normal thallium-201 cardiac study. The first three rows show six slices of the left ventricle in three different axes (vertical long axis, short axis, horizontal long axis) of the heart. The fourth row shows how the images can be reoriented along each axis.

Once the final set of pictures are ready for clinical use, then it is the computer that is used for image display and analysis. The computer is also used for storage of the clinical studies and to allow their use by multiple readers at various sites and time points during patient care as required for optimal usage of the diagnostic information they provide. They are also very useful in research aimed at optimizing imaging strategies and systems, and in the education and training of medical personnel.

NM IMAGING

Computers play an essential role in NM as an integral part of the most common imaging device used in NM, which is a gamma camera, and in the obtention of slices through the body made in emission computerized tomography (ECT). Emission CT is the general term referring to the computer-based technique by which the 3D distribution of a radioactive tracer in the human body is obtained and presented as a stack of 2D slices. The acronym ECT should not be confused with CT (for computerized tomography), which refers to imaging using X rays. Historically, the use of two different kinds of radioactive tracers has led to the parallel evolution of two types of ECT techniques: Single-photon emission computerized tomography (SPECT) and positron emission tomography (PET). The SPECT technology is used with gamma emitters, that is, unstable nuclei whose disintegration led to the emission of high energy photons, called γ rays. Gamma rays are just like X rays except that X rays are emitted when electrons loose a good amount of energy and γ rays are emitted when energy is given off as a photon or photons during a nuclear disintegration, or when matter and antimatter annihilate. As the name implies, the gamma camera is used with gamma emitters in planar imaging (scintigraphy) where 2D pictures of the distribution of activity within a patients body are made. An illustration of a three-headed SPECT system is shown in Fig. 4. The evolution of SPECT systems has led to a configuration with one, then two and three detectors that are gamma-camera heads. The positron emission tomography is used with emitters whose disintegration results in the emission of a positron (a particle similar to an electron, but with the opposite charge making it the antiparticle to the electron). When a positron that has lost all of its kinetic energy hits an electron, the two annihilate and two photons are emitted from the annihilation. These two photons have the same energy (511 keV) and opposite directions. To

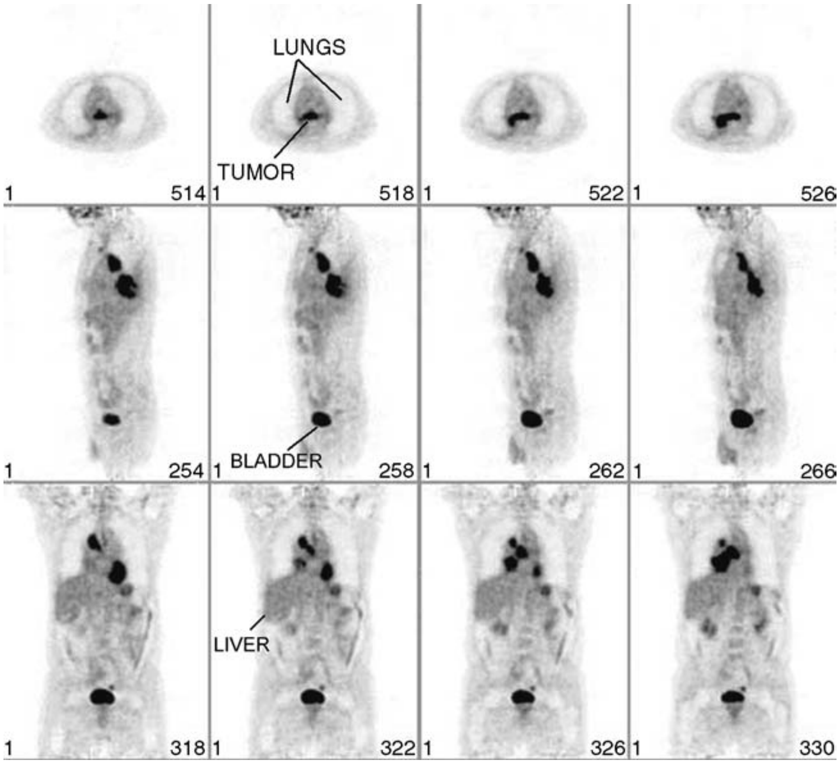


Figure 3. The FDG study for a patient with lung cancer. Upper to lower rows: transverse, sagittal, and coronal slices.

detect these two photons, a natural configuration for a PET system is a set of rings of detectors. The two points of detection of the opposite detectors form a line, called the line of response (LOR). A state-of-the-art PET system combined with an X-ray CT system (PET/CT) is shown in Fig. 5.

A gamma camera has three main parts: the scintillating crystal, the collimator, and the photomultiplier tubes (PMT) (Fig. 6). When the crystal (usually thallium-activated sodium iodide) is struck by a high energy photon (γ or X ray), it emits light (it scintillates). This light is detected by an array of PMT located at the back of the camera. The sum of the currents emitted by all the PMT after one scintillation is proportional to the energy of the incoming photon, so the γ rays can be sorted according to their energy based upon the electrical signal they generate. Because, for geometrical reasons, the PMT closer to the scintillation see more light than the PMT located farther away, the relative amounts of current of the PMT are used to locate the scintillation. This location alone would be of little use if the direction of the incoming photon was not known. The current way to know the direction is by using a collimator. The collimator is a piece of lead with one or more holes, placed in front of the scintillating crystal, facing the patient. Although different kinds of collimators exist, they are all used to determine, for each incoming photon, its direction before its impact on the crystal. To understand the role of the collimator, one can use the analogy of the gamma-camera with a camera that takes photographs. The

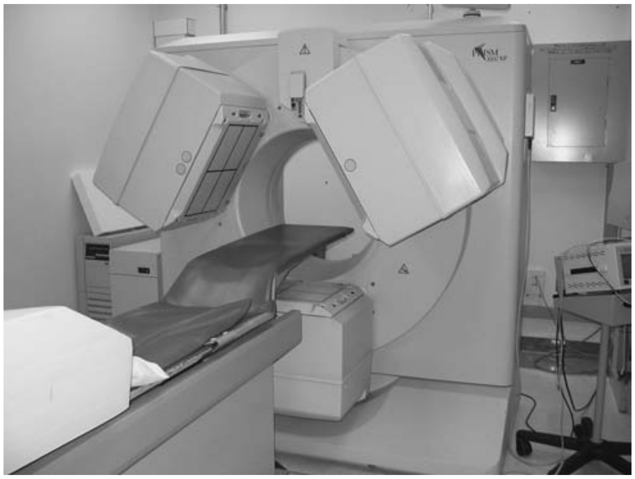


Figure 4. A three-headed SPECT system with the third head below the imaging bed the patient lays on.

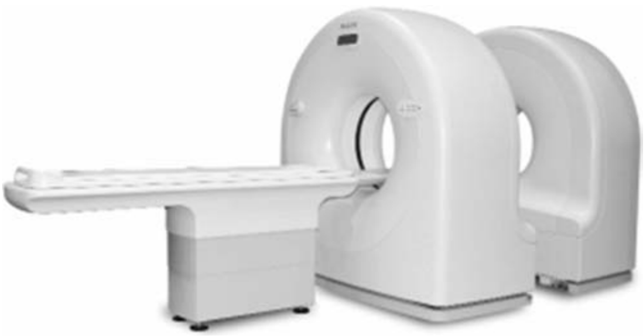


Figure 5. Illustration of a state-of-the-art PET/CT system (Philips Medical Systems) with patient bed. The CT system is the first ring-shaped gantry and the PET system is the second ring-shaped gantry. (Reproduced with permission of Philips Medical Systems.)

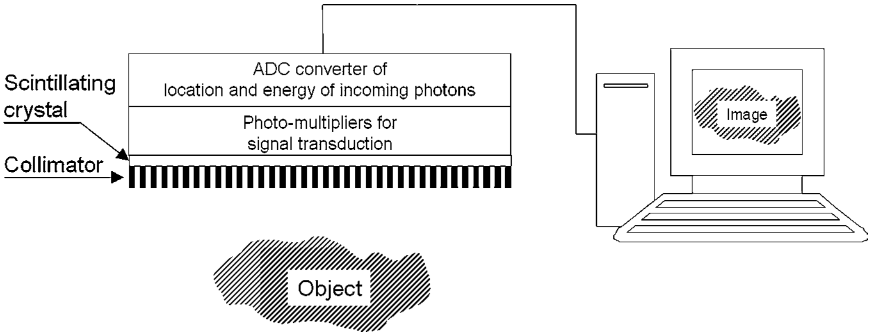

Scintillating crystal
Collimator
ADC converter of location and energy of incoming photons
Photo-multipliers for signal transduction
Image
Object

Figure 6. Main parts of a SPECT camera and basic principle. A gamma photon going through the holes of the collimator strikes the crystal. The crystal restitutes the energy of the gamma photon by scintillation, that is, by emitting some transient ultraviolet (UV) light. Some of the UV light is collected by photomultiplier tubes, whose function is to ensure the transduction of the signal (i.e., the conversion of light into electricity). The location of the scintillation and the energy of the photon are estimated, digitized and sent to the computer.

collimator plays the role of the objective lens in a camera. An image acquired without a collimator would be totally blurry, as would be a photograph taken with a camera with no lens. This is because γ rays are emitted in all directions with equal probabilities, and without a collimator, the photons emitted from a radioactive point source would strike the detector almost uniformly. With a collimator, only the photons whose direction is parallel to the axis of the holes may be potentially detected, while others are stopped by the lead. As a result, the image of a source is (ideally) the projection of the source distribution onto the crystal plane (Fig. 7). Gamma rays can be stopped or scattered, but due to their high penetrating power, it is very difficult to bend them like light rays with lenses, and this is the reason why collimators are used instead of lenses. Photons emitted at different distances from the camera, but along the same direction parallel to a hole, are detected at the same location in the crystal. Thus, the image obtained is the projection of the 3D distribution of the tracer onto the 2D plane of the detector. In that sense, a projection is similar to a chest X ray, in which the images of all the organs (ribs, heart, spine, etc.) are overlaid on the film even though organs do not spatially overlap in the body. The overlay might not be a problem for relatively thin parts of the body, such as a hand, or when tracer-avid structures do not overlap, such as the skeleton. In that case, only one projection is obtained from the best angle of view for the gamma-camera head. As stated above, this technique is called planar imaging, or scintigraphy. However, for other thicker organs like the myocardium and the brain, for which one is interested in measuring the 3D tracer inner distribution, more information is gathered by rotating the heads to acquire projections from multiple angles of view (tomographic acquisition, presented later in this article).

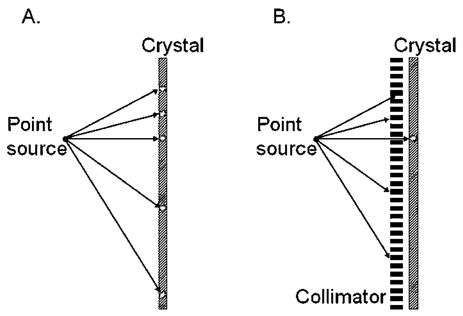


Figure 7. The role of the collimator in a SPECT system.

In PET scanners, hundreds of small crystals arranged in rings are used so that the data can be simultaneously acquired along multiple LOR (Fig. 8). Thousands of photons hit the crystals every second, so how to know which two photons are the result of the same electron-positron annihilation? If two photons are detected almost simultaneously, chances are that they are of the same pair (it is called a true coincidence), so an electronic circuitry checks whether one photon is detected $\sim$ 10 ns (the time window) at most after the previous one. It may happen that, although two photons are detected within that time window, they are not of the same pair, and such an event is called a random coincidence. Because in PET the direction of the photons is known (it is the LOR), collimators are not needed; however, because of the limited counting rate capabilities of older systems, septa made of lead may be used to limit the acquisition to the LORs roughly perpendicular to the axial direction, inside the same ring (2D acquisition). With modern PET systems having a high couting rate capability, a 3D acquisition is possible by detecting LORs even when the two photons hit crystals of different rings.

The computer plays an important role in the formation of the image coming from the gamma camera. As described above, the crystal is viewed by an array of 37 to $>$ 100, depending on the model, of PMT. These are analog devices that can drift with time. Also the positioning in the image of the location of the flash of light when a γ ray is absorbed in the crystal depends to some extent on where the ray interacts relative to the array of PMT. Such local variations lead to nonuniformity (uneven apparent sensitivity) and

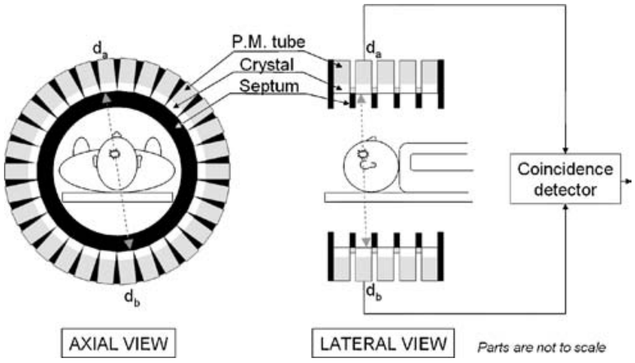


Figure 8. Main parts of a PET system. Pairs of photons are detected in coincidence following the annihilation of a positron with an electron.

nonlinearity (straight lines of activity are bent in the image). Prior to the incorporation of computers into the gamma camera electronics, the impact of such local variations was minimized by allowing the light to spread out before reaching the PMT by passing it through a light guide. This resulted in more PMT receiving enough light to participate in determining the location of the interaction thus improving uniformity and linearity, but at the expense of spatial resolution (i. e., determination of where in the crystal the flash of light originated). Modern gamma cameras incorporate computers to correct for local variations in camera performance so that the light guide is virtually eliminated. This in turn has improved spatial resolution.

Computer correction of the camera image usually takes place in three steps (5). The first is energy correction. As we said, the total magnitude of the signal from all the PMT is related to the energy deposited in the crystal by the γ ray. However, if a large number of γ rays of exactly the same energy interact in the crystal, the magnitude of the electrical signal will vary due to the statistics of turning the light emitted into an electrical signal and local variation in camera performance. By placing a source that will uniformly irradiate the crystal, such as the commercial sheet source shown in Fig. 9, the local variation on average in the magnitude of the signal can be determined on a pixel by pixel basis by computer. The centering of the window employed to select electrical pulses for inclusion in image formation can then be adjusted to give a more uniform response.

Besides varying in the average size of the total electrical pulse detected from the PMT locally, gamma cameras vary in how well they map the true location of the flash of light into its perceived location in the image. Thus, in some regions, detected events are compressed together and in others they are spread apart. Correction of this nonlinear mapping constitutes the second step in computer correction of the gamma-camera image and is called linearity correction. Linearity correction is performed by placing an attenuating sheet with an exactly machined array of very small holes in a precise location between the gamma camera and the sheet source of Fig. 9. A high resolution image consisting of a large number of gamma-ray events is then acquired. The images of the holes do not match where they should appear. The vector displacement of the image of the hole back to its true location defines how the mapping from true to detected location is inaccurate at that location. By using the computer to interpolate between the array of measured distortions at the pixel level, a map is generated giving how each event detected at a location in the crystal should be displaced in the resulting image.

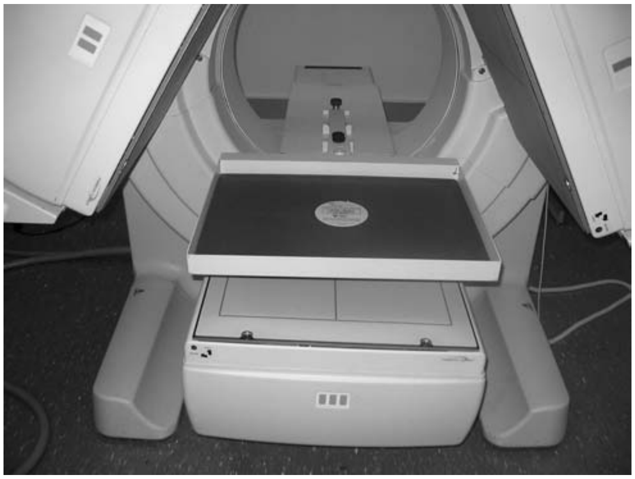
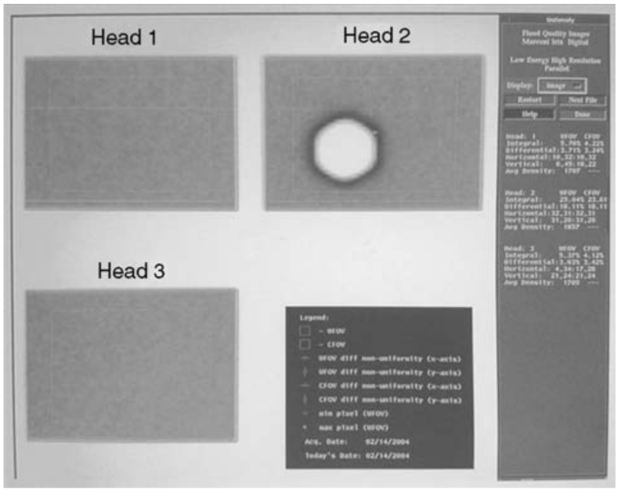


Figure 9. Radioactive sheet source in front of the third head of a three-headed SPECT system in position for checking uniformity and loading correction factors.



```
Head: 1          UFOV   CFOV
Integral:        5.70% 4.22%
Differential:3.71% 3.24%
Horizontal:10,32:10,32
Vertical:    6,45:10,22
Avg Density:  1707   ---

Head: 2          UFOV   CFOV
Integral:       25.04% 23.6
Differential:18.11% 18.1
Horizontal:32,31:32,31
Vertical:   31,28:31,28
Avg Density:  1857   ---

Head: 3          UFOV   CFOV
Integral:        5.37% 4.12%
Differential:3.83% 3.42%
Horizontal: 4,34:17,28
Vertical:   21,24:21,24
Avg Density:  1705   ---
```

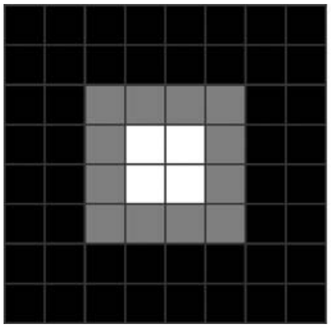
```
Legend:
☐  - UFOV
☐  - CFOV
—  UFOV diff non-uniformity (x-axis)
|  UFOV diff non-uniformity (y-axis)
—  CFOV diff non-uniformity (x-axis)
|  CFOV diff non-uniformity (y-axis)
-  min pixel (UFOV)
+  max pixel (UFOV)
Acq. Date:    02/14/2004
Today's Date: 02/14/2004
```

Figure 10. Output from checking camera uniformity when a single PM on head 2 of the three-headed SPECT system of Fig. 9 has failed.

The final step in image correction is called flood correction. If an image of a large number of events from a sheet source is acquired with energy and linearity correction enabled, then any residual nonuniformity is corrected by determining with computer a matrix that when multiplied by this image would result in a perfectly uniform image of the same number of counts. This matrix is then saved and used to correct all images acquired by the gamma camera before they are written to disk.

An example of testing camera uniformity is shown in Fig. 10. Here again, a sheet source of radioactivity is placed in front of the camera head as shown in Fig. 9. High count images of the sheet source are inspected numerically by computer and visually by the operator each day before the camera is employed clinically. Heads 1 and 3 in Fig. 10 show both numerically and visually good uniformity. A large defect is seen just below and to the left of center in the image from head 2. This is the result of the failure of a single PMT. A single PMT affects a region much larger than its size because it is the combined output of all the PMT close to the interation location of a gamma-ray that are used to determine its location.

Images can be classified in two types, mutually exclusive: analog or digital. A chest X ray on a film is a typical example of an analog image. Analog images are not divided into a finite number of elements, and the values in the image can vary continuously. An example of digital image is a photograph obtained with a digital camera. Much like roadmaps, a digital image is divided into rows and columns, so that it is possible to find a location given its row



0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	127	127	127	127	0	0
0	0	127	255	255	127	0	0
0	0	127	255	255	127	0	0
0	0	127	127	127	127	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

Figure 11. Left: example of an 8×8 image. Each square represents a pixel. The dark gray borders of each square have been added here for sake of clarity, but are not present in the image when stored on the computer. Each pixel has a level of gray attached to it. Right: the values in each pixel. In its simplest form, the image is stored in a computer as a series of lexicographically ordered values. The rank of each value in the series defines the pixel location in the image (e.g., the 10th value of the series refers to the 10th pixel of the image). There is a 1:1 relationship between the brightness and the value. The correspondence between the color and the value is defined in a table called look-up table (LUT) or a color map. An image can be a shade of grays (black and white) or in color.

and column. The intersection of a row and a column defines one picture element, or pixel, of the image. Each pixel has a value that usually defines its brightness, or its color. A digital image can be seen as a rectangular array of values (Fig. 11), and thus be considered, from a mathematical point of view, as a matrix. Computers cannot deal with analog values, so whenever an analog measurement (here, the current pulse generated after the impact of a gamma photon in the crystal) is made, among the first steps is the analog-to-digital conversion (ADC), also called digitization. The ADC is the process during that an infinite number of possible values is reduced to a limited (discrete) number of values, by defining a range (i.e., minimum and maximum values), and dividing the range into intervals, or bins (Fig. 12). The performance of an ADC is defined by its ability to yield a digital signal as close as possible to the analog input. It is clear from Fig. 12 that the digitized data are closer to the analog signal when the cells are smaller. The width of the cells is defined by the sampling rate, that is, the number of measurements the ADC can convert per unit of time. The height of the cells is defined by the resolution of the ADC. A 12-bit resolution means that the ADC sorts the amplitude of the analog values among one of $2^{12} = 4096$ possible values. The ADC has also a range, that is, the minimum and maximum analog amplitudes it can handle. For example, using a 12-bit analog-to-digital converter with a -10 to $+10$ V range (i.e., a 20 V range), the height of each cell is 20/4096 (i.e., ~ 0.005 V). Even if the analog signal is recorded with a 0.001 V accuracy, after ADC the digital signal accuracy will be at best 0.005 V. The point here is that any ADC is characterized by its sampling rate, its resolution and its range. Both the SPECT and PET systems measure the location and the energy of the photons hitting their detectors. The measurements are initially analog, and are digitized as described above before being stored on a computer. In SPECT, prior to the beginning of the acquisition, the operator chooses the width and height (in pixels) of the digital image to be acquired. Common sizes are 64 by 64 pixels (noted 64×64, width by height), 128×128 or 256×256. Dividing the size of the field of view (in cm) by the number of pixels yields the pixel size (in cm). For example, if the usable size for the detector is 40×40 cm, the pixel size is $40/64 = 0.66$ cm for a 64×64 image. Because all devices are imperfect, a point source is seen as a blurry spot on the image. If two radioactive point sources in the field of view are close enough, their spots overlap each other. When the two sources get closer, at some point the spots cannot be visually separated in the image. The smallest distance between the sources that allows us to see one spot for each source is called the spatial resolution. The pixel size is not to be confused with the spatial resolution. While the pixel size is chosen by the operator, the spatial resolution is imposed by the camera characteristics, and most notably by the collimator now that thick light guides are no longer employed. The pixel size is chosen to be smaller than the resolution, so that we can get as much detail as the resolution allows us to get, but it is important to understand that using a pixel size much smaller than the resolution does not increase the image quality. If the pixel size is small (i.e., when the number of pixels in the field of view is large), then the spot spills over many pixels, but with no improvement to image resolution.

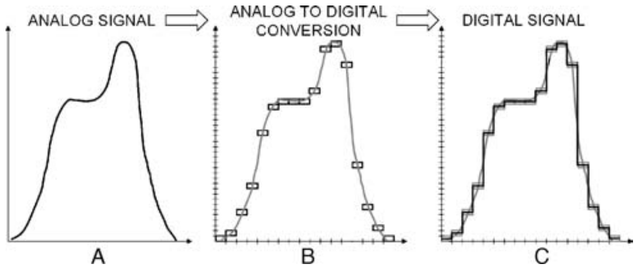


Figure 12. (a) Example of an analog signal, for example the intensity of gray (vertical axis) along a line (horizontal axis) drawn on a photograph taken with a nondigital camera. (b) The process of ADC, for example when the photograph is scanned to be archived as an image file on a computer. The 2D space is divided into a limited number of rectangular cells as indicated by the tick marks on both axes. For sake of clarity, only the cells whose center is close to the analog signal are drawn in this figure, and this set of cells is the result of the ADC. (c) The digital signal is drawn by joining the center of the cells, so that one can compare the two signals.

The energy resolution (i.e., the smallest change in energy the detector can measure) is limited, and its value has a great impact on image quality, as explained below. Between the points of emission and detection, photons with an energy < 1 MeV frequently interact with electrons by scattering, during which their direction changes and some of their energy is lost. Because their direction changes, an error is made on their origin. However, it is possible to know that a photon is scattered because it has lost some energy, so an energy window, called the photopeak window defined around the energy of nonscattered (primary) photons (the photopeak), is defined prior to the acquisition, and the scattered photons whose energy falls outside the photopeak window can be identified by energy discrimination and ignored. Unfortunately, photons in the photopeak window can either be scattered photons, or a primary photon whose energy has been underestimated (due to the limited energy resolution of the detectors, an error can be made regarding the actual energy of the photons). If the photopeak window is wide, many scattered photons are accepted, and the image has a lot of scattered activity that reduces the contrast; if the energy window is narrow, many primary photons are rejected, and the image quality is poor because of a lack of signal. As the energy resolution increases, the energy window can be narrowed, so that most scattered photons can be rejected while most primary photons are kept.

As mentioned above, the 3D distribution of the tracer in the field of view is projected onto the 2D plane of the camera heads. As opposed to the list-mode format (presented later in this article), the projection format refers to the process of keeping track of the total number of photons detected (the events, or counts) for each pixel of the projection image. Each time a count is detected for a given pixel, a value of 1 is added to the current number of counts for that pixel. In that sense, a projection represents the accumulation of the counts on the detector for a given period of time. If no event is recorded for any given pixel, which is not uncommon especially in the most peripheral parts of the image, then the value for that pixel is 0. Usually, 16 bits (2 bytes) are allocated to represent the number of counts per pixel, so the range for the number of events is 0 to $2^{16}-1 = 65{,}535$ counts per pixel. In case the maximal value is reached for a pixel (e.g., for a highly active source and a long acquisition time), then the computer possibly stops incrementing the counter, or reinitializes the pixel value to 0 and restarts counting from that point on. This yields images in which the most radioactive areas in the image may paradoxically have a lower number of counts than surrounding, less active, areas.

Different acquisitions are possible with a gamma camera:

Planar (or static): The gamma-camera head is stationary. One projection is obtained by recording the location of the events during a given period from a single angle of view. This is equivalent to taking a photograph with a camera. The image is usually acquired when the tracer uptake in the organ of interest has reached a stable level. One is interested in finding the quantity of radiopharmaceutical that accumulated in the region of interest. Planar images are usually adequate for thin or small structures (relative to the resolution of the images), such as the bones, the kidneys, or the thyroid.

Whole body: This acquisition is similar to the planar acquisition, in the sense that one projection is obtained per detector head, but is designed, as the name implies, to obtain an image of the whole body. Since the human body is taller than the size of the detector (~ 40 × 40 cm), the detector slowly moves from head to toes. This exam is especially indicated when looking for metastases. When a cancer starts developing at a primary location, it may happen that cancer cells, called metastases, disseminate in the whole body, and end up in various locations, especially bones. There, they may start proliferating and a new cancer may be initiated at that location. A whole-body scintigraphy is extremely useful when the physician wants to know whether one or more secondary tumors start developing, without knowing exactly where to look at.

Dynamic: Many projections are successively taken, and each of them is typically acquired over a short period (a few seconds). This is equivalent to recording of a movie. Analyzing the variations as a function of time allows us to compute parameters, such as the uptake rate, which can be a useful clinical index of normality.

Gated: The typical application of a gated acquisition is the cardiac scintigraphy. Electrodes are placed on the patient's chest to record the electrocardiogram (ECG or EKG). The acquisition starts at the beginning of the cardiac cycle, and a dynamic sequence of 8 or 16 images is acquired over the cardiac cycle (~ 1 s), so that a movie of the beating heart is obtained. However, the image quality is very poor when the acquisition is so brief. So, the acquisition process is repeated many times (i.e., over many heart beats), and the first image of all cardiac cycles are summed together, the second image of all cardiac cycles are summed together, and so on.

Tomographic: The detector heads are rotating around the patient. Projections are obtained under multiple angles of view. Through a process called tomographic reconstruction (presented in the next section), the set of 2D projections is used to find the 3D distribution of the tracer in the body, as a stack of 2D slices. The set of 1D projections of one slice for all projection angles is called a sinogram.

Tomographic gated: As the name implies, this acquisition is a tomographic one with gating information. The ECG is recorded during the tomographic acquisition, and for each angle of view, projections are acquired over many cardiac cycles, just as with a gated acquisition (see above). Thus, a set of projections is obtained for each point of the cardiac cycle. Each set is reconstructed, and tomographic images are obtained for each point of the cardiac cycle.

In contrast with the types of acquisition above in which the data are accumulated in the projection matrix for several seconds or minutes (frame-mode acquisition), a much less frequent type of acquisition called list-mode acquisition, can also be useful, because more information is available in this mode. As the name implies, the information for each individual event is listed in the list-mode file, and are not arranged in a matrix array. In addition to the coordinates of the scintillations, additional data are stored in the file. Figure 13 illustrates the typical list-mode format for a SPECT system. List-mode information is similar with a PET system, except that the heads location and *X*–*Y* coordinates are replaced with the location of the event on the detector rings. The list-mode file (~ 50 Mb

```
Gantry Angle 123 degrees
  Timestamp: 0 ms
    X:1002, Y:1270, Energy:1640
    X:1044, Y:1211, Energy:1767
    X:1077, Y:741, Energy:1788
    X:570, Y:819, Energy:1674
  Timestamp: 10 ms
    X:1280, Y:1595, Energy:1603
    X:576, Y:1181, Energy:1768
  Timestamp: 20 ms
    X:919, Y:1162, Energy:1828
    X:973, Y:1078, Energy:1765
    X:1023, Y:1045, Energy:1708
  Timestamp: 30 ms
    X:955, Y:773, Energy:1717
    X:989, Y:702, Energy:1732
    X:1060, Y:1145, Energy:1853
          ...
  Timestamp: 19990 ms
    X:577, Y:862, Energy:1818
    X:556, Y:766, Energy:1682
Gantry Angle 126 degrees
  Timestamp: 0 ms
          ...
```

Figure 13. Example of data stored on-the-fly in a list-mode file data in a SPECT system. Gantry angle defines the location of the detectors. The X and Y coordinates are given in a 2048×2048 matrix. The energy is a 12-bit value (i.e., between 0 and 4,095), and a calibration is required to convert these values in usual units (i.e., kiloelectron volt, keV).

in size in SPECT) can be quite large relative to a projection file. The list-mode format is far less common than the projection format, because it contains information, such as timing, that would usually not be used for a routine clinical exam. The list-mode data can be transformed into projection data through a process called rebinning (Fig. 14). Since the timing is known, multiple projections can be created as a function of time, thus allowing the creation of "movies" whose rate can be defined postacquisition. A renewed interest in the list-mode format has been fueled these past years by the temporal information it contains, which is adequate for the temporal correlation of the acquisition with patient, cardiac, or respiratory motions through the synchronized acquisition of signal from motion detectors.

TOMOGRAPHIC RECONSTRUCTION

Tomographic reconstruction has played a central role in NM, and has heavily relied on computers (6). In addition to data acquisition control, tomographic reconstruction is the other main reason for which computers have been early introduced in NM. Among all uses of computers in NM, tomographic reconstruction is probably the one that symbolizes most the crunching power of computers. Tomographic reconstruction is the process by which slices of the 3D distribution of tracers are obtained based upon the projections obtained under different angles of view. Because the radioactivity emitted in the 3D space is projected on the 2D detectors, the contrast is usually low. Tomographic reconstruction greatly restores the contrast, by estimating the 3D tracer distribution. Reconstruction is possible using list-mode data (SPECT or PET), but mainly

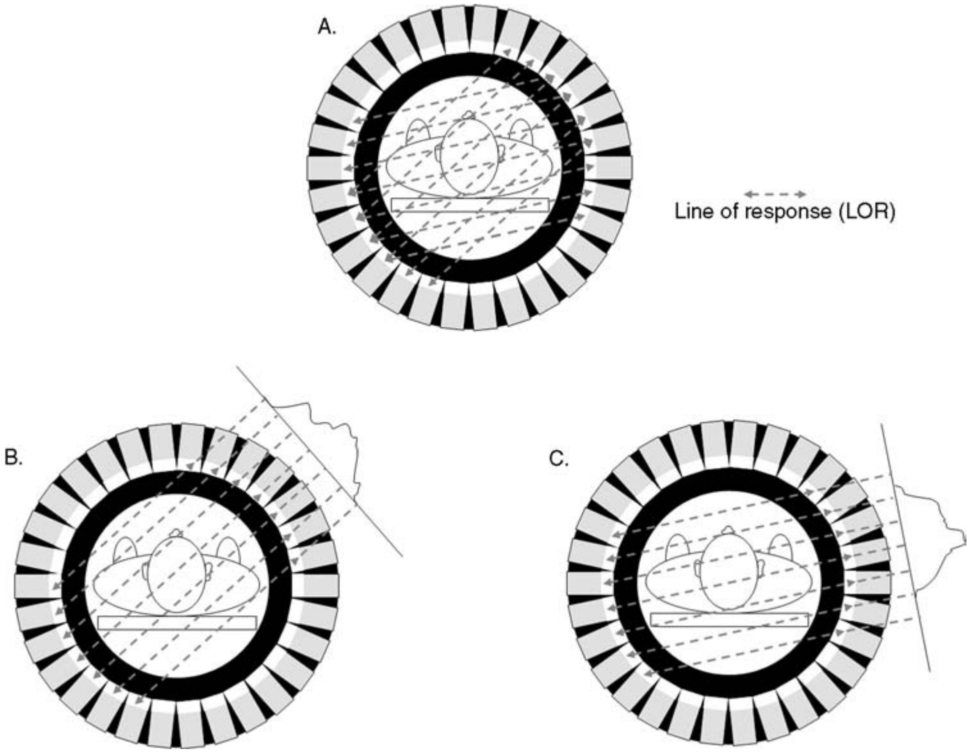


Figure 14. Rebinning process. (a) The line joining pairs of photons detected in coincidence is called a line of response (LOR). (b and c) LOR are sorted so that parallel LOR are grouped, for a given angle.

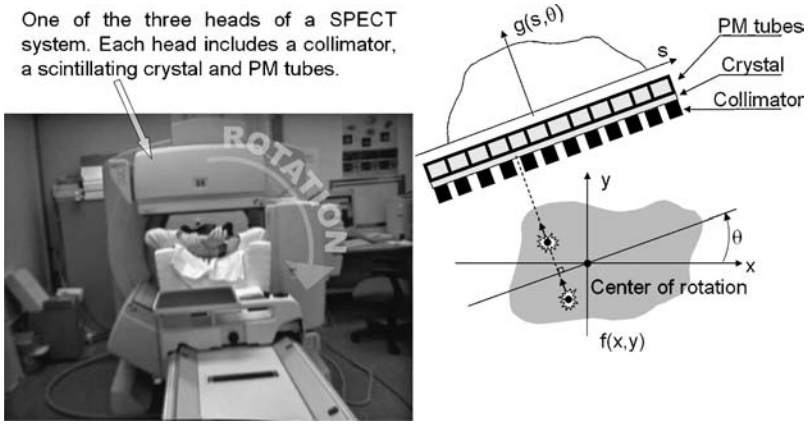

One of the three heads of a SPECT system. Each head includes a collimator, a scintillating crystal and PM tubes.
ROTATION
g(s,θ)
s
PM tubes
Crystal
Collimator
y
θ
x
Center of rotation
f(x,y)

Figure 15. The SPECT acquisition. Left. A three-head IRIX SPECT system (Philips Medical Systems). A subject is in the field of view while the camera heads are slowly rotating around him. Right. Physical model and geometric considerations. The 2D distribution of the radioactivity *f*(*x*,*y*) in one slice of the body is projected and accumulated onto the corresponding 1D line *g*(*s*,θ) of detector bins.

for research purposes. The PET data, although initially acquired in list-mode format, are usually reformatted to form projections, so that the algorithms developped in SPECT can also be used with PET data. There are many different algorithms, mainly the filtered back-projection (FBP) and the iterative algorithms, that shall be summarized below (7–9).

In the following, focuses on tomographic reconstruction when input data are projections, which is almost always the case on SPECT systems. During a SPECT acquisition, the detecting heads rotate around the subject to gather projections from different angles of views (Fig. 15). Figure 16 presents the model used to express the simplified projection process in mathematical terms. Associated with the projection is the backprojection (Fig. 17). With backprojection, the activity in each detector bin *g* is added to all the voxels which project onto bin *g*. It can be shown (10) that backprojecting projections filtered with a special filter called a ramp filter (filtered backprojection, or FBP) is a way to reconstruct slices. However, the ramp filter is known to increase the high frequency noise, so it is usually combined with a low pass filter (e.g., Butterworth filter) to form a band-pass filter. Alternatively, reconstruction can be performed with the ramp filter only, and then the reconstructed images can be smoothed with a 3D low pass filter. The FBP technique yields surprisingly good results considering the simplicity of the method and its approximations, and is still widely used today. However, this rather crude approach is more and more frequently replaced by the more sophisticated iterative algorithms, in which many corrections can be easily introduced to yield more accurate results. An example of an iterative reconstruction algorithm includes the following steps:

1. An initial estimate of the reconstructed is created, by attributing to all voxels the same arbitrary value (e.g. 0 or 1).
2. The projections of this initial estimate are computed.
3. The estimated projections are compared to the measured projections, either by computing their difference or their ratio.
4. The difference (resp. the ratio) is added (resp. multiplied) to the initial estimate to get a new estimate.
5. Steps 2–4 are repeated until the projections of the current estimate are close to the measured projections.

Figure 18 illustrates a simplified version of the multiplicative version of the algorithm. This example has been voluntarily oversimplified for sake of clarity. Indeed, image reconstruction in the real world is much more complex for several reasons: (*1*) images are much larger; typically, the 3D volume is made of 128 × 128 × 128 voxels, (*2*) geometric considerations are included to take into account the volume of each volume element (voxel) that effectively project onto each bin at each angle of view, (*3*) camera characteristics, and in particular the spatial resolution, mainly defined by the collimator characteristics, are introduced in the algorithm, and (*4*) corrections presented below are applied during the iterative process. The huge number of operations made iterative reconstruction a slow process and prevented its routine use until recently, and FBP was

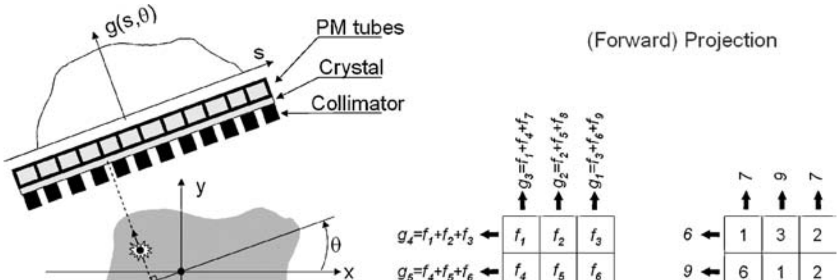


Figure 16. Projection. Each plane in the FOV (left) is seen as a set of values *f* (center). The collimator is the device that defines the geometry of the projection. The values in the projections are the sum of the values in the slices. An example is presented (right). (Reproduced from Ref. 8 with modifications with permission of the Society of Nuclear Medicine Inc.)

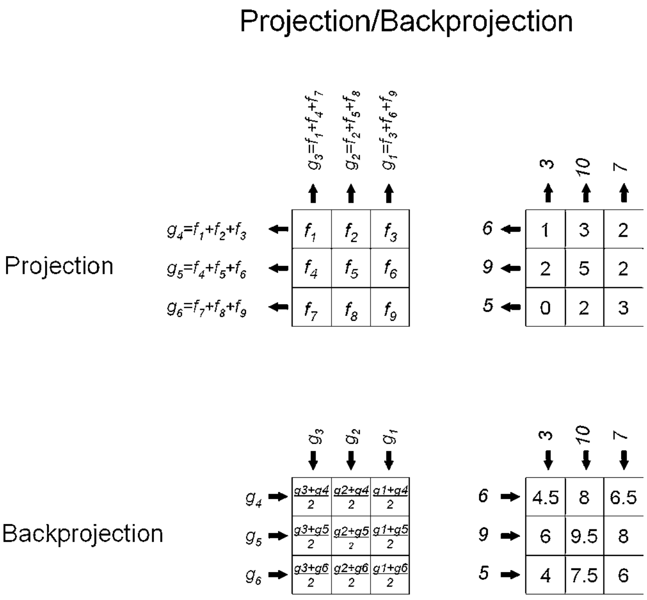


Figure 17. Projection and backprojection. Notice that backprojection is not the invert of projection. (Reproduced from Ref. 8 with modifications with permission of the Society of Nuclear Medicine Inc.)

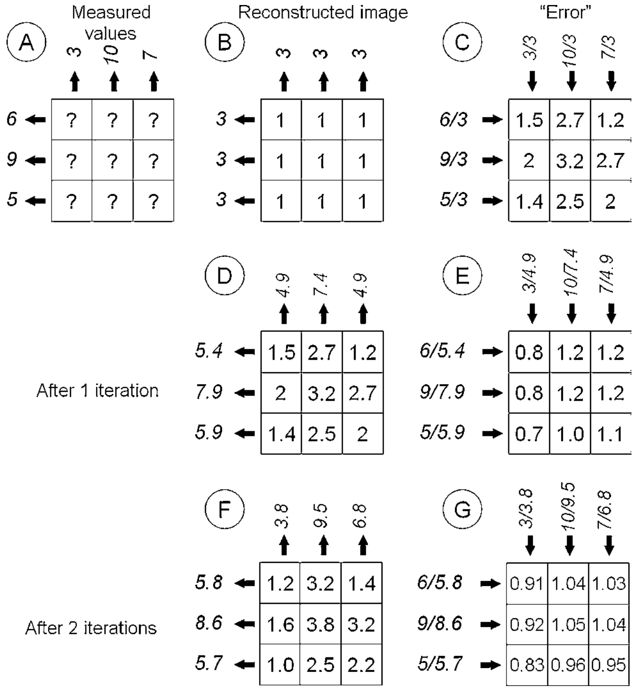


Figure 18. A simplified illustration of tomographic reconstruction with an iterative algorithm. (a) The goal is to find the values in a slice (question marks) given the measured projection values 7, 10, 3, 6, 9, 5. (b) Voxels in the initial estimate have a value of 1, and projections are computed as described in Fig. 17. (c) The error in the projections is estimated by dividing the actual values by the estimated values. The ratios are then backprojected to get a slice of the "error". (d) Multiplying the error obtained in c by the estimate in b yields a second estimate, and projections are computed again (g). After an arbitrary number of iterations (e, f), an image whose projections are close to the measured projections is obtained. This image is the result of the iterative tomographic reconstruction process. Such a process is repeated for the stack of 2D slices.

preferred. Modern computers are now fast enough for iterative algorithms, and since these algorithms have many advantages over the FBP, they are more and more widely used.

A number of corrections usually need to be applied to the data during the iterative reconstruction to correct them for various well-known errors caused by processes associated with the physics of the detection, among which the more important are attenuation (11–13), Compton scattering (11,12), depth-dependent resolution (in SPECT) (11,12), random coincidences (in PET) (14), and partial volume effect (15). These sources of error below are briefly presented:

Attenuation occurs when photons are stopped (mostly in the body), and increases with the thickness and the density of the medium. Thus, the inner parts of the body usually appear less active than the more superficial parts (except the lungs, whose low density makes them almost transparent to gamma photons and appear more active than the surrounding soft tissues). Attenuation can be compensated by multiplying the activity in each voxel by a factor whose value depends upon the length and the density of the tissues encountered along the photons path. The correction factor can be estimated (e.g., by assuming a uniform attenuation map) or measured using an external radioactive source irradiating the subject. A third possibility, which is especially attractive with the advent of SPECT/CT and PET/CT systems (presented below), is to use the CT images to estimate the attenuation maps.

Photons may be scattered when passing through soft tissues and bones, and scattered photons are deflected from their original path. Because of the error in the estimated origin of the scattered photons, images are slightly blurred and contrast decreases. As mentioned in the previous section, the effects of scattering can be better limited by using detectors with a high energy resolution. Scatter can also be estimated by acquiring projection data in several energy windows during the same acquisition. Prior to the acquisition, the user defines usually two or three windows per photopeak (the photopeak window plus two adjacent windows, called the scatter windows). As mentioned, photons can be sorted based upon their energy, so they can be assigned to one of the windows. The amount of scattering is estimated in the photopeak window using projection data acquired in the scatter windows, and assuming a known relationship between the amount of scattering and the energy. Another approach to Compton scattering compensation uses the reconstructed radioactive distribution and attenuation maps to determine the amount of scatter using the principles of scattering interactions.

In SPECT, collimators introduce a blur (i.e., even an infinitely small radioactive source would be seen as a spot of several mm in diameter) for geometrical reasons. In addition, for parallel collimators (the most commonly used, in which the holes are parallel), the blur increases as the distance between the source and the collimator increases. Depth-dependent resolution can be corrected either by filtering the sinogram in the Fourier domain using a filter whose characteristics vary as a function of the distance to the collimator (frequency–distance relationship, FDR) or by modelling the blur in iterative reconstruction algorithms.

In PET, a coincidence is defined as the detection of two photons (by different detectors) in a narrow temporal window of $\sim$ 10 ns. As mentioned, a coincidence is true when the two photons are of the same pair, and random when the photons are from two different annihilations. The amount of random coincidences can be estimated by defining a delayed time window, such that no true coincidence can be detected. The estimation of the random coincidences can then be subtracted from the data including both true and random, to extract the true coincidences.

Partial volume effect (PVE) is directly related to the finite spatial resolution of the images: structures that are small (about the voxel size and smaller) see their concentration underestimated (the tracer in the structure appears as being diluted in the voxel). Spillover is observed at the edges of active structures: some activity spreads outside the voxels, so that although it stems from the structure, it is actually detected in neighboring voxels. Although several techniques exist, the most accurate can be implemented when the anatomical boundaries of the structures are known. Thus, as presented below, anatomical images from CT scanners are especially useful for PVE and spillover corrections, if they can be correctly registered with the SPECT or PET data.

IMAGE PROCESSING, ANALYSIS AND DISPLAY

Computers are essential in NM not only for their ability to control the gamma cameras and to acquire images, but also because of their extreme ability to process, analyze and display the data. Computers are essential in this respect because (*1*) the amount of data can be large (millions of pixel values), and computers are extremely well suited to handle images in their multimegabytes memory, (*2*) repetitive tasks are often needed and central processor units (CPUs) and array processors can repeat tasks quickly, (*3*) efficient algorithms have been implemented as computer programs to carry on complex mathematical processing, and (*4*) computer monitors are extremely convenient to display images in a flexible way.

Both the PET and SPECT computers come with a dedicated, user-friendly graphical environment, for acquisition control, patient database management, and a set of programs for tomographic reconstruction, filtering and image manipulation. These programs are usually written in the C language or in Fortran, and compiled (i.e., translated in a binary form a CPU can understand) for a given processor and a given operating system (OS), usually the Unix OS (The Open Group, San Francisco CA) or the Windows OS (Microsoft Corporation, Redmond WA). As an alternative to these machine-dependent programs, Java (Sun Microsystems Inc.) based programs have been proposed (see next section).

As seen in the first section, an image can be seen as a rectangular array of values, which is called, from a mathematical point of view, a matrix. A large part of image processing in NM is thus based upon linear algebra (16), which is the branch of mathematics that deals with matrices. One of the problems encountered in NM imaging is the noise (random variations due to the probabilistic nature of the radioactive processes and to the limited accuracy of the measurements). A number of methods are available to reduce the noise after the acquisition, by smoothing the minor irregularities or speckles in the images (10). The most common way to filter images is by convolution (a pixel value is replaced by a weighted average of its neighbors) or by Fourier methods. Computers are extremely efficient at computing discrete Fourier transforms thanks to a famous algorithm called the fast Fourier transform (FFT) developed by Cooley and Tukey (17).

The NM images can be displayed or printed in black and white (gray levels) or in color. Pixel values can be visually estimated based on the level of gray or based on the color. Color has no special meaning in NM images, and there is no consensus about the best color map to use. Most often, a pixel value represents a number of counts, or events. However, units can be something else (e.g., flow units), especially after some image processing. So, for proper interpretation, color map and units should always accompany an image.

Regions of interest (ROIs) are defined by line segments drawn to set limits in images and can have any shape or be drawn by hand. The computer is then able to determine which pixels of the image are out and which are in the ROI, and thus computations can be restricted to the inside or to the outside of the ROI. ROIs are usually drawn with the mouse, based on the visual inspection of the image. Drawing a ROI is often a tricky task, due to the low resolution of the images and to the lack of anatomical information regarding the edges of the organs. In an attempt to speed up the process, and to reduce the variability among users, ROIs can also be drawn automatically (18). When a dynamic acquisition is available, a ROI can be drawn on one image and reported on the other images of the series, the counts in the ROI are summed and displayed as a time–activity curve (TAC), so that one gets an idea of the kinetics of the tracer in the ROI. The TAC are useful because with the appropriate model, physiological parameters such as pharmacological constants or blood flow can be determined based on the shape of the curve. An example of dynamic studies with ROI and TAC is the renal scintigraphy, whose goal is to investigate the renal function through the vascularization of the kidneys, their ability to filter the blood, and the excretion in the urine. A tracer is administered with the patient lying on the bed of the camera, and a two-stage dynamic acquisition is initiated: many brief images are acquired (e.g., 1 image per second over the first 60 s). Then, the dynamic images are acquired at a lower rate (e.g., 1 image per minute for 30 min). After the acquisition, ROIs are drawn over the aorta, the cortical part of the kidneys, and the background (around the kidneys), and the corresponding TAC are generated for analysis (Fig. 19). The TAC obtained during the first stage of the acquisition reflect the arrival of the tracer in the renal arteries (vascular phase). The rate at which the tracer invades the renal vascularization, relative to the speed at which it arrives in the aorta above indicates whether the kidneys

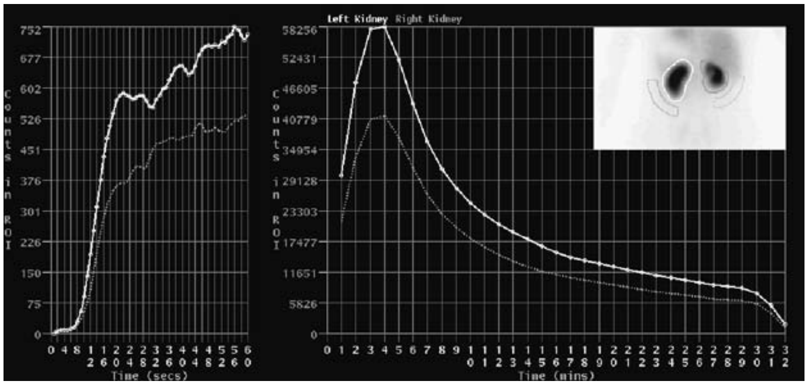


Figure 19. Output of a typical renal scintigraphy. Left: the TAC for both kidneys in the first minute after injection of the tracer. The slope of the TAC gives an indication of the state of the renal vascularization. Center: One-minute images acquired over 32 min after injection. The ascending part evidences the active tracer uptake by the kidneys, while the descending part shows the excretion rate. Right: Insert shows the accumulation of the tracer in the bean-shaped kidneys. The ROI are drawn over the kidneys and the background.

are normally vascularized. The renal TAC obtained in the second stage of the acquisition (filtration phase) show the amount of tracer captured by the kidneys, so that the role of the kidneys as filters can be assessed. When the tracer is no longer delivered to the kidneys, and as it passes down the ureters (excretion phase), the latest part of the renal curves normally displays a declining phase, and the excretion rate can be estimated by computing the time of half-excretion (i.e., the time it takes for the activity to decreases from its peak to half the peak), usually assuming the decrease is an exponential function of time.

The corrections presented at the end of the previous section are required to obtain tomographic images in which pixel values (in counts per pixel) are proportional to the tracer concentration with the same proportion factor (relative quantitation), so that different areas in the same volume can be compared. When the calibration of the SPECT or PET system is available (e.g., after using a phantom whose radioactive concentration is known), the images can be expressed in terms of activity per volume unit, (e.g., in becquerels per milliliters, $\mathrm{Bq{\cdot}mL^{-1}}$; absolute quantitation). Absolute quantitation is required in the estimation of a widely used parameter, the standardized uptake value (SUV) (19), which is an index of the FDG uptake that takes into account the amount of injected activity and the dilution of the tracer in the body. The SUV in the region of interest is computed as SUV = (uptake in the ROI in $\mathrm{Bq{\cdot}mL^{-1}}$)/(injected activity in Bq/body volume in mL). Another example of quantitation is the determination of the blood flow (in $\mathrm{mL{\cdot}g^{-1}{\cdot}min^{-1}}$), based upon the pixel values and an appropriate model for the tracer kinetics in the area of interest. For example, the absolute regional cerebral blood flow (rCBF) is of interest in a number of neurological pathologies (e.g., ischemia, haemorrage, degenerative diseases, epilepsy). It can be determined with xenon ^{133}Xe, a gas that has the interesting property of being washed out from the brain after its inhalation as a simple exponential function of the rCBF. Thus, the rCBF can be assessed after at least two fast tomographic acquisitions (evidencing the differential decrease in activity in the various parts of the brain), for example, using the Tomomatic 64 SPECT system (Medimatic, Copenhagen, Denmark) (20).

An example of image processing in NM is the equilibrium radionuclide angiography (ERNA) (21), also called multiple gated acquisition (MUGA) scan, for assessment of the left ventricle ejection fraction (LVEF) of the heart. After a blood sample is taken, the erythrocytes are labelled with $^{99\ \mathrm{m}}$Tc and injected to the patient. Because the technetium is retained in the erythrocytes, the blood pool can be visualized in the images. After a planar cardiac gated acquisition, 8 or 16 images of the blood in the cardiac cavities (especially in the left ventricle) are obtained during an average cardiac cycle. A ROI is drawn, manually or automatically, over the left ventricle, in the end-diastolic (ED) and end-systolic (ES) frames, that is, at maximum contraction and at maximum dilatation of the left ventricle respectively. Another ROI is also drawn outside the heart for background activity subtraction. The number of counts nED and nES in ED and ES images, respectively, allows the calculation of the LVEF as LVEF = (nED-nES)/nED. Acquisitions for the LVEF assessment can also be tomographic in order to improve the delineation of the ROI over the left ventricle, and several commercial softwares are available (22) for largely automated processing, among which the most widely used are the Quantitative Gated SPECT (QGS) from the Cedars-Sinai Medical Center, Los Angeles, and the Emory Cardiac Tool box (ECTb) from the Emory University Hospital, Atlanta.

INFORMATION TECHNOLOGY

An image file typically contains, in addition to the image data, information to identify the images (patient name, hospital patient identification, exam date, exam type, etc.), and to know how to read the image data (e.g., the number of bytes used to store one pixel value). File format refers to the way information is stored in computer files. A file format can be seen as a template that tell the computer how and where in the file data are stored. Originally, each gamma-camera manufacturer had its own file format, called proprietary format, and for some manufacturers the proprietary format was confidential and not meant to be widely disclosed. To facilitate the exchange of images between different computers, the Interfile format (23) was proposed in the late 1980s. Specifically designed for NM images, it was intended to be a common file format that anyone could understand and use to share files. At the same period, the American College of Radiology (ACR) and the National Electrical Manufacturers Association (NEMA) developed their standard for NM, radiology, MRI and ultrasound images: the ACR-NEMA file format, version 1.0 (in 1985) and 2.0 (in 1988). In the early 1990s, local area networks (LANs) connecting NM, radiology and MRI departments started to be installed. Because Interfile was not designed to deal with modalities other than NM, and because ACR-NEMA 2.0 was “only” a file format, and was not able to handle robust network communications to exchange images over a LAN, both became obsolete and a new standard was developed by the ACR and the NEMA, ACR-NEMA 3.0, known as Digital Imaging and Communications in Medicine (DICOM) (24). Although quite complex (the documentation requires literally thousands of pages), DICOM is powerful, general in its scope and designed to be used by virtually any profession using digital medical images. Freely available on the Internet, DICOM has become a standard among the manufacturers, and it is to be noted that DICOM is more than a file format. It also includes methods (programs) for storing and exchanging image information; in particular, DICOM servers are programs designed to process requests for handling DICOM images over a LAN.

DICOM is now an essential part of what is known as Picture Archiving and Communications Systems (PACS). Many modern hospitals use a PACS to manage the images and to integrate them into the hospital information system (HIS). The role of the PACS is to identify, store, protect (from unauthorized access) and retrieve digital images and ancillary information. A web server can be used as an interface between the client and the PACS (25). Access to the images does not necessarily require a dedicated software on the client. A simple connection to the Internet and a web browser can be sufficient, so that the images can be seen from the interpreting or prescribing physician’s office. In that case, the web server is responsible for submitting the user’s request to the PACS, and for sending the image data provided by the PACS to the client, if the proper authorization is granted. However, in practice, the integration of NM in a DICOM-based PACS is difficult, mainly because PACS evolved for CT and MR images (26,27), that is, as mostly static, 2D, black and white images. The NM is much richer from this point of view, with different kinds of format (list-mode, projections, whole-body, dynamic, gated, tomographic, tomographic gated, etc.) and specific postacquisition processing techniques and dynamic displays. The information regarding the colormaps can also be a problem for a PACS when dealing with PET or SPECT images fused with CT images (see next section) because two different colormaps are used (one color, one grayscale) with different degrees of image blending.

In the spirit of the free availability of programs symbolized by the Linux operating system, programs have been developed for NM image processing and reconstruction as plug-ins to the freely available ImageJ program developed at the U.S. National Institutes of Health (28). ImageJ is a general purpose program, written with Java, for image display and processing. Dedicated Java modules (plugs-in) can be developed by anyone and added as needed to perform specific tasks, and a number of them are available for ImageJ (29). Java is a platform-independent language, so that the same version of a Java program can run on different computers, provided that another program, the Java virtual machine (JVM), which is platform-dependent, has been installed beforehand. In the real world, however, different versions of the JVM may cause the Java programs to crash or to cause instabilities when the programs require capabilities the JVM cannot provide (30). The advantage of the Java programs is that they can be used inside most Internet browsers, so that the user has no program to install (except the JVM). A Java-based program called JaRVis (standing for Java-based remote viewing station) has been proposed in that spirit for viewing and reporting of nuclear medicine images (31).

It is very interesting to observe how, as the time goes by, higher levels of integration have been reached: with the early scintigraphy systems, such as rectilinear scanners (in the 1970s), images were analog, and the outputs were film or paper hard copies. In the 1980s images were largely digital, but computers were mainly stand alone machines. One decade later, computers were commonly interconnected through LANs, and standard formats were available, permitting digital image exchange and image fusion (see next section). Since the mid-1990s, PACS and the worldwide web make images remotely available, thus allowing telemedecine.

HYBRID SPECT/CT AND PET/CT SYSTEMS

Multimodality imaging (SPECT/CT and PET/CT) combines the excellent anatomical resolution of CT with SPECT or PET functional information (2,3). Other advantages of multimodality are (*1*) the use of CT images to estimate attenuation and to correct for PVE in emission images, (*2*) the potential improvement of emission data reconstruction by inserting in the iterative reconstruction program prior information regarding the locations of organ and/or tumor boundaries, and (*3*) the possible comparison of both sets of images for diagnostic purposes, if the CT images are of diagnostic quality. The idea of combining information provided by two imaging modalities is not new, and a lot of work has been devoted to multimodality. Multimodality initially required that the data acquired from the same patient, but on different systems and on different occasions, be grouped on the same computer, usually using tapes to physically transfer the data. This was, $\sim$ 20 years ago, a slow and tedious process. The development of hospital computer network in the 1990s greatly facilitated the transfer of data, and the problem of proprietary image formats to be decoded was eased when a common format (DICOM) began to spread. However, since the exams were still carried out in different times and locations, the data needed to be registered. Registration can be difficult, especially because emission data sometimes contain very little or no anatomical landmarks, and external fiducial markers were often needed. Given the huge potential of dual-modality systems, especially in oncology, a great amount of energy has been devoted in the past few years to make it available in clinical routine. Today, several manufacturers propose combined PET/CT and SPECT/CT hybrid systems (Fig. 20): The scanners are in the same room, and the table on which the patient lies can slide from one scanner to the other (Fig. 21). An illustration of PET/CT images is presented in Fig. 22.

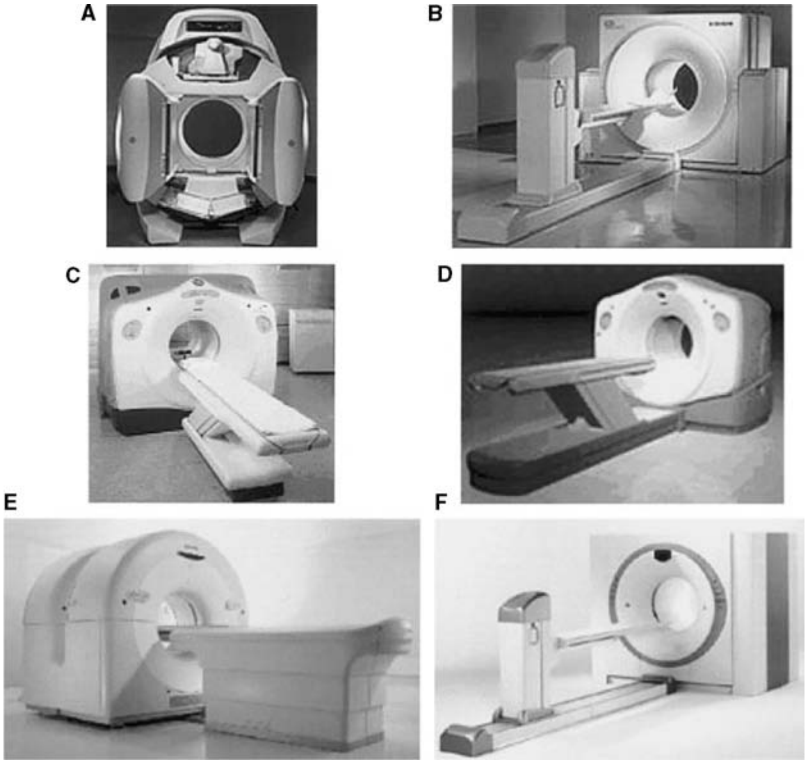


Figure 20. Current commercial PET/CT scanners from 4 major vendors of PET imaging equipment: (a) Hawkeye (GE Medical Systems); (b) Biograph (Siemens Medical Solutions) or Reveal (CTI, Inc); (c) Discovery LS (GE Medical Systems); (d) Discovery ST (GE Medical Systems); (e) Gemini (Philips Medical Systems); (f) Biograph Sensation 16 (Siemens Medical Solutions) or Reveal XVI (CTI, Inc.). (Reproduced from Ref. 2, with permission of the Society of Nuclear Medicine Inc.)

Although patient motion is minimized with hybrid systems, images from both modalities are acquired sequentially, and the patient may move between the acquisitions, so that some sort of registration may be required before PET images can be overlaid over CT images. Again, computer programs play an essential role in finding the best correction to apply to one dataset so that it matches the other dataset. Registration may be not too difficult with relatively rigid structures, such as the brain, but tricky for chest imaging for which nonrigid transformations are needed. Also, respiratory motion introduces in CT images mushroom-like artifacts that can be limited by asking the patients to hold their breath at mid-respiratory cycle during CT acquisition, so that it best matches the average images obtained in emission tomography with no respiratory gating.

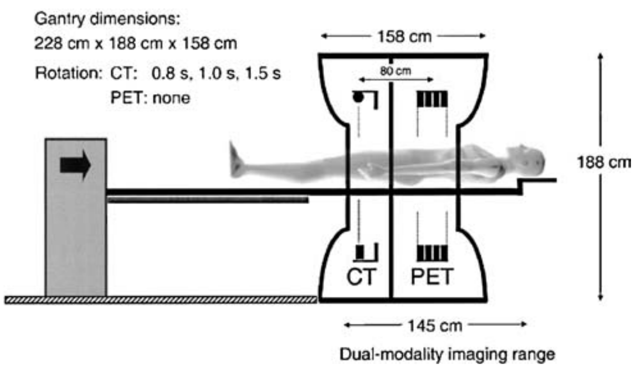


Figure 21. Schematic of PET/CT developped by CPS Innovations. Axial separation of two imaging fields is 80 cm. The coscan range for acquiring both PET and CT has maximum of 145 cm. (Reproduced from Ref. 2, with permission of the Society of Nuclear Medicine Inc.)

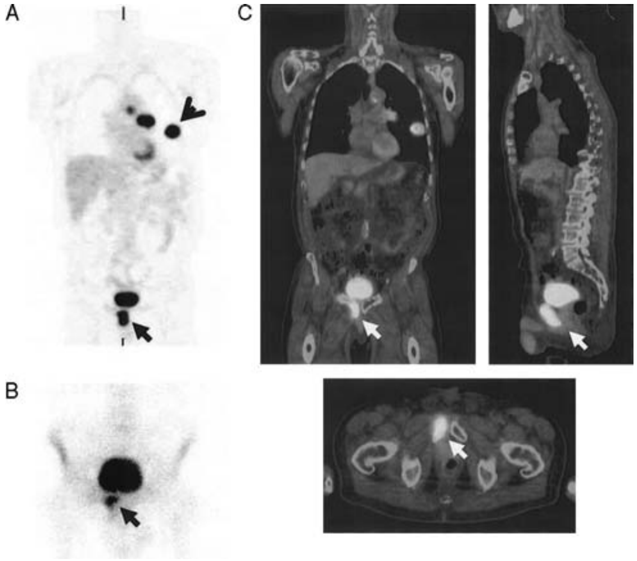


Figure 22. Image of a 66 years old male patient with history of head-and-neck cancer. In addition to ^{18}FDG uptake in lung malignancy, intense uptake is seen on PET scan (a) in midline, anterior and inferior to bladder. Note also presence of lung lesion (arrowhead) due to primary lung cancer. ^{99m}Tc bone scan (b) subsequently confirmed that uptake was due to metastatic bone disease. PET/CT scan (c) directly localized uptake to pubic ramus (arrowed). (Reproduced from Ref. 2, with permission of the Society of Nuclear Medicine Inc.)

Dedicated programs are required for multimodality image display (*1*) to match the images (resolution, size, orientation); (*2*) to display superimposed images from both modalities with different color maps (CT data are typically displayed with a gray scale, while a color map is used to display the tracer uptake); (*3*) to adjust the degree of transparency of each modality relative to the other in the overlaid images; and (*4*) to select the intensity scale that defines the visibility of the bones, soft tissues and lungs in the CT images. For the interpretation of SPECT/CT or PET/CT data, the visualization program has to be optimized, for so much information is available (dozens of slices for each modality, plus the overlaid images, each of them in three perpendicular planes) and several settings (slice selection, shades of gray, color map, the degree of blending of the two modalities in the superimposed images) are to be set. Powerful computers are required to be able to handle all the data in the computer random access memory (RAM) and to display them in real time, especially when the CT images are used at their full quality (512 × 512 pixel per slice, 16-bit shades of gray). Finally, these new systems significantly increase the amount of data to be archived (one hundred to several hundreds megabytes per study), and some trade-off may have to be found between storing all the information available for later use and minimizing the storage space required. An excellent review of the current software techniques for merging anatomic and functional information is available (32).

COMPUTER SIMULATION

Simulation is a very important application of computers in NM research, as it is in many technical fields today. The advantage of simulation is that a radioactive source and a SPECT or PET system are not required to get images similar to the ones obtained if there were real sources and systems. Simulation is cheaper, faster and more efficient to evaluate acquisition hardware or software before they are manufactured and to change its design for optimization. Thus, one of the applications is the prediction of the performance of a SPECT or PET system by computer simulation. Another application is to test programs on simulated images that have been created in perfectly known conditions, so that the ouput of the programs can be predicted and compared to the actual results to search for possible errors.

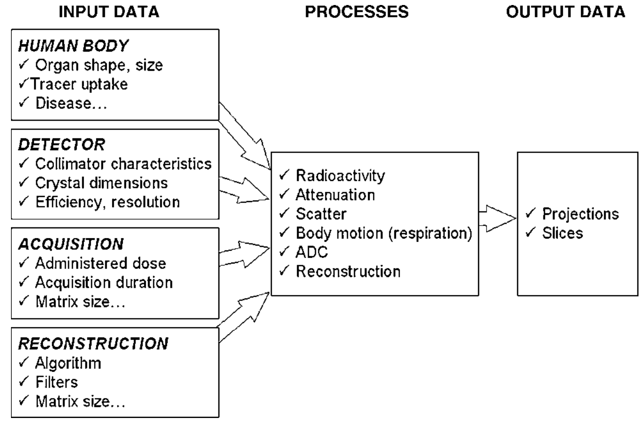


Figure 23. Principle of computer simulation in NM. Parameters and all available information are defined and used by the simulation processes (i.e., computer programs), which in turn generate output data that constitute the result of the simulation.

Computer simulation is the art of creating information about data or processes without real measurement devices, and the basic idea is the following: if enough information is provided to the computer regarding the object of study (e.g., the 3D distribution of radioactivity in the body, the attenuation), the imaging system (the characteristics of the collimator, the crystal, etc.), and the knowledge we have about the interactions of gamma photons with matter, then it is possible to generate the corresponding projection data (Fig. 23). Although this may seem at first very complex, it is tractable when an acquisition can be modeled with a reasonable accuracy using a limited number of relevant mathematical equations. For example, a radioactive source can be modeled as a material emitting particles in all directions, at a certain decay rate. Once the characteristics of the source: activity, decay scheme, half-life, and spatial distribution of the isotope are given, then basically everything needed for simulation purposes is known. Radioactive disintegrations and interactions between gamma photons and matter are random processes: one cannot do predictions about a specific photon (its initial direction, where it will be stopped, etc.), but the probabilities of any event can be computed. Random number generators are used to determine the fate of a given photon.

Let us say that we want to evaluate the resolution of a given SPECT system as a function of the distance from the radioactive source to the surface of the detector. The first solution is to do a real experiment: prepare a radioactive source, acquire images with the SPECT system, and analyze the images. This requires (*1*) a radioactive source, whose access is restricted, and (*2*) the SPECT system, which is costly. Because the resolution is mainly defined by the characteristics of the collimator, a second way to evaluate the resolution is by estimation (analytical approach): apply the mathematical formula that yield the resolution as a function of the collimator characteristics (diameter of the holes, collimator thickness, etc.). This approach may become tricky as more complex processes have to be taken into account. A third solution is to simulate the source, the gamma camera, and the physical interactions. It is an intermediate solution, between real acquisition and estimation. Simulation yields more realistic results than estimations, but does not require the use of real source or SPECT system. Simulation is also powerful because when uncertainties are introduced in the model (e.g., some noise), then it is possible to see their impact on the projection data.

Simulation refers either to the simulation of input data (e.g., simulation of the human body characteristics), or to the simulation of processes (e.g., the processes like the interaction between photons and matter). For simulation of input data, a very useful resource in nuclear cardiology is the program for the mathematical cardiac torso (MCAT) digital phantom developed at the University of North Carolina (33). The MCAT program models the shape, size, and location of the organs and structures of the human chest using mathematical functions called non-uniform rational B-splines (NURBS). The input of this program is a list of many parameters, among them: the amount of radioactivity to be assigned to each organ (heart, liver, spleen, lungs, etc.), the attenuation coefficients of these organs, their size, the heart rate, and so on. This phantom used the data provided by the Visible Human Project and is accurate enough from an anatomical point of view for NM, and it can be easily customized and dynamic sets of slices can be obtained that simulate the effects of respiration and a beating heart. The output is the 3D distribution (as slices) of radioactivity (called emission data) and 3D maps of the attenuation (attenuation data) in the human torso. This output can then be used as an input for a Monte Carlo program to generate projection data. Monte Carlo programs (34,35) use computer programs called random number generators to generate data (for example, the projection data) based upon the probability assigned to any possible event, such as the scattering of a photon, or its annihilation in the collimator. The programs were named Monte Carlo after the city on the French Riviera, famous for its casinos and games based upon probabilities. Among the Monte Carlo simulation programs used in NM are: Geant (36), Simind (37), SimSET (38), and EGS4 (39). Programs, such as Geant and its graphical environment Gate (40), allow the definitions of both the input data (the body attenuation maps, the collimator characteristics) and the interactions to be modeled (photoelectric effect, scatter, attenuation, etc.).

EDUCATION AND TEACHING

As almost any other technical field, NM has benefited from the Internet as a prodigious way to share information. Clinical cases in NM are now available, and one advantage of computers over books is that an image on a computer can be manipulated: the color map can be changed (scale, window, etc.) and, in addition to images, movies (e.g., showing tracer uptake or 3D images) can be displayed. Websites presenting clinical NM images can be more easily updated with more patient cases and some on-line processing is also possible. One disadvantage is the sometimes transitory existence of the web pages, that makes (in the author's opinion) the Internet an unreliable source of information from this point of view. NM professionals can also share their experience and expertise on list servers (see Ref. 41 for a list of servers). The list servers are programs to which e-mails can be sent, and that distribute these e-mails to every registered person.

The Society of Nuclear Medicine (SNM) website hosts its Virtual Library (42), in which > 90 h of videos of presentations given during SNM meetings are available for a fee. The Joint Program in Nuclear Medicine is an example of a program including on-line education by presenting clinical cases (43) for > 10 years. More than 150 cases are included, and new cases are added; each case includes presentation, imaging technique, images, diagnosis, and discussion. Other clinical cases are also available on the Internet (44). Another impressive on-line resource is the Whole Brain Atlas (45) presenting PET images of the brain, coregistered with MRI images. For each case, a set of slices spanning over the brain is available, along with the presentation of the clinical case. The user can interactively select the transverse slices of interest on a sagittal slice. As the last example, a website (46) hosts a presentation of normal and pathologic FDG uptake in PET and PET/CT images.

CONCLUSION

Computers are used in NM for a surprisingly large variety of applications: data acquisition, display, processing, analysis, management, simulation, and teaching/training. As many other fields, NM has benefited these past years from the ever growing power of computers, and from the colossal development of computer networks. Iterative reconstruction algorithms, which have been known for a long time, have tremendously benefited from the increase of computers crunching power. Due to the increased speed of CPUs and to larger amounts of RAM and permanent storage, more and more accurate corrections (attenuation, scatter, patient motion) can be achieved during the reconstruction process in a reasonable amount of time. New computer applications are also being developed to deal with multimodality imaging such as SPECT/CT and PET/CT, and remote image viewing.

ACRONYMS

ACR	American College of Radiology
ADC	Analog to Digital Conversion (or Converter)
CPU	Central Processing Unit
CT	(X-ray) Computerized Tomography
DICOM	Digital Imaging and COmmunications in Medicine
ECT	Emission Computerized Tomography
FBP	Filtered BackProjection
FDG	Fluoro-Deoxy Glucose
FDR	Frequency-Distance Relationship
FFT	Fast Fourier Transform
HIS	Hospital Information System
keV	kiloelectron Volt
LAN	Local Area Network
LOR	Line of Response
MCAT	Mathematical Cardiac Torso
MRI	Magnetic Resonance Imaging
NEMA	National Electrical Manufacturers Association
NM	Nuclear Medicine
NURBS	Nonuniform Rational B-Splines
PACS	Picture Archiving and Communication System
PET	Positron Emission Tomography
PMT	Photomultiplier
PVE	Partial Volume Effect
RAM	Random Access Memory
rCBF	regional Cerebral Blood Flow
ROI	Region of Interest
SPECT	Single-Photon Emission Computerized Tomography
SNM	Society of Nuclear Medicine
SUV	Standardized Uptake Value
TAC	Time–Activity Curve

BIBLIOGRAPHY

Cited References

1. Wagner HN, Szabo Z, Buchanan JW, editors. Principles of Nuclear Medicine. 2nd ed. New York: W.B. Saunders; 1995.
2. Townsend DW, Carney JPJ, Yap JT, Hall NC. PET/CT today and tomorrow. J Nucl Med 2004;45:4S–14S.
3. Ratib O. PET/CT navigation and communication. J Nucl Med 2004;45:46S–55S.
4. Lee K. Computers in NM: a Practical Approach. New York: The Society of Nuclear Medicine Inc.; 1991.
5. Simmons GH. On-line corrections for factors that affect uniformity and linearity. J Nucl Med Tech 1988;2:82–89.
6. Rowland SW. Computer implementation of image reconstruction formulas. In: Herman GT, editor. Topics in Applied Physics: Image Reconstruction from Projections. vol. 32. Heidelberg, Germany: Springer-Verlag; 1979. pp. 29–79.
7. Zeng GL. Image reconstruction: a tutorial. Comput Med Imaging Graph 2001;25:97–103.
8. Bruyant PP. Analytical and iterative algorithms in SPECT. J Nucl Med 2002;43:1343–1358.
9. Brook RA, DiChiro G. Principles of computer assisted tomography in radiographic and radioisotope imaging. Phys Med Biol 1976;21:689–732.
10. Jain AK. Fundamentals of digital image processing. In: Kailath Th, editor. Englewood Cliffs (NJ): Prentice Hall; 1989.
11. King MA, et al. Attenuation, scatter, and spatial resolution compensation in SPECT. Emission Tomography: The Fundamentals of PET and SPECT. Chapt. 22. In: Wernick MN, Aarsvold JN, editors. San Diego: Elsevier Academic Press; 2004.
12. King MA, Tsui BMW, Pretorius PH. Attenuation/scatter/resolution correction: Physics Aspects. In: Zaret BL, Beller GA, editors. Clinical Nuclear Cardiology: State of the Art and Future Directions. 3d ed. Philadelphia: Elsevier Science; 2004.
13. Zaidi H, Hasegawa B. Determination of the attenuation map in emission tomography. J Nucl Med 2003;44(2):291–315.
14. Brasse D, et al. Correction methods for random coincidences in fully 3D whole-body PET: impact on data and image quality. J Nucl Med 2005;46(5):859–867.
15. Rousset OG, Ma Y, Evans AC. Correction for partial volume effect in PET: Principles and validation. J Nucl Med 1998;39:904–911.
16. Anton H. Elementary Linear Algebra. New York: John Wiley & Sons, Inc.; 2000.
17. Cooley JW, Tukey JW. An algorithm for the machine calculations of complex Fourier series. Math Comput 1965;19:297–301.
18. Dai X, et al. Left-ventricle boundary detection from nuclear medicine images. J Digit Imaging 1998;11(1):10–20.
19. Huang SC. Anatomy of SUV. Nucl Med Biol 2000;27:643–646.
20. Celsis P, Goldman T, Henriksen L, Lassen NA. A method for calculating regional cerebral blood flow from emission computed tomography of inert gas concentrations. J Comput Assist Tomogr 1981;5:641–645.
21. Bacharach SL, Green MV, Borer JS. Instrumentation and data processing in cardiovascular nuclear medicine: evaluation of ventricular function. Semin Nucl Med 1979;9(4):257–274.
22. Nakajima K, et al. Accuracy of Ventricular Volume and Ejection Fraction Measured by Gated Myocardial SPECT: Comparison of 4 Software Programs. J Nucl Med 2001;42(10): 1571–1578.
23. Todd-Pokropek A, Cradduck TD. (1998). Interfile resources. [Online]. The Keston Group. Available at http://www.keston.com/Interfile/interfile.htm. Accessed 2005 Sept 27.
24. Clark H. (2004). NEMA-Digital Imaging and Communications in Medicine (DICOM) Part 1 :Introduction and Overview. [Online]. National Electrical Manufacturers Association. Available at http://www.nema.org/stds/ps3-1.cfm. Accessed 2005 Sept 27.
25. Barbaras L, Parker JA, Donohoe KJ, Kolodny GM. (1996) Clinical data on the World Wide Web.
26. Surface D. Making PACS work with nuclear medicine. Radiol Today 2004;12:22.
27. Laβmann M, Reiners C. A DICOM-based PACS for nuclear medicine. Electromedica 2002;70:21–30.
28. Anonymous (2004) Image J. [Online]. National Institutes of Health. Available at http://rsb.info.nih.gov/ij. Accessed 2005 Sept 27.
29. Parker JA. (2004). Parker Plugins. [Online]. The Harvard Medical School. Available at http://www.med.harvard.edu/JPNM/ij/plugins/. Accessed 2005 Sept 27.
30. Wallis JW. Java and teleradiology. J Nucl Med 2000;41(1): 119–122.
31. Slomka PJ, Elliott E, Driedger AA. Java-based remote viewing and processing of nuclear medicine images: towards "the imaging deprtment without walls." J Nucl Med 2000;41(1): 111–118.
32. Slomka PJ. Software approach to merging molecular with anatomic information. J Nucl Med 2004;45(1 Suppl):36S–45S.
33. Lacroix KL. (1998) Home page for MCAT phantomo [Online] University of North Carolina. Available at http://www.bme.unc.edu/mirg/mcat/. Accessed 2005 Sept 27.
34. Buvat I, Castiglioni I. Monte Carlo simulations in SPET and PET. Q J Nucl Med 2002;46:48–61.
35. Ljungberg M, Strand S-E, King MA, editors. Monte Carlo Calculations in Nuclear Medicine. London: Institute of Physics Publishing; 1998.
36. Agostinelli S. (2005). Geant4 home page. [Online]. CERN. Available at wwwasd.web.cern.ch/wwwasd/geant4/geant4. html. Accessed 2005 Sept 27.
37. Ljungberg M. (2003). simind Monte Carlo home page. [Online]. Lund University. Available at http://www.radfys.lu.se/simind/. Accessed 2005 Sept 27.
38. Harrison R. (2004). Simulation system for emission tomography (SimSET) home page. [Online]. University of Washington. Available at http://depts.washington.edu/~simset/html/simset_main.html. Accessed 2005 Sept 27.
39. Liu JC. (2001). Electron gamma shower (EGS) Monte Carlo radiation transport code. [Online]. Stanford Linear Accelerator Center. Available at http://www.slac.stanford.edu/egs/index.html. Accessed 2005 Sept 27.
40. Morel C. (2004). GATE. [Online]. Ecole Federale Polytechnique de Lausanne. Available at http://www-lphe.epfl.ch/~PET/research/gate/index.html. Accessed 2005 Sept 27.
41. Zaidi H. (1997). Medical Physics on the Internet and the world-wide web. [Online]. Geneva University Hospital. Available at http://dmnu-pet5.hcuge.ch/Habib/Medphys.html. Accessed 2005 Sept 27.
42. Anonymous (2005) The Society of Nuclear Medicine Virtual Library. [Online]. The Society of Nuclear Medicine Inc. Available at http://snm.digiscript.com/. Accessed 2005 Sept 27.
43. Anthony Parker J. (2005). Joint Program in Nuclear Medicine: Electronic Learning Resources. [Online]. The Joint Program in Nuclear Medicine. Available at http://www.jpnm.org/elr.html. Accessed 2005 Sept 27.
44. Wallis J. (2005). MIR Nuclear medicine network access page. [Online]. The Mallinckrodt Institute of Radiology. Available at http://gamma.wustl.edu/home.html. Accessed 2005 Sept 27.
45. Johnson KA, Becker JA. (1999). The Whole Brain Atlas. [Online]. Harvard Medical School. Available at http://www.med.harvard.edu/AANLIB/home.html. Accessed 2005 Sept 27.
46. Rajadhyaksha CD, Parker JA, Barbaras L, Gerbaudo VH. (2004). Findings in 18FDG-PET & PET-CT. [Online]. Harvard Medical School and The Joint Program in Nuclear Medicine.

Available at http://www.med.harvard.edu/JPNM/chetan/. Accessed 2005 Sept 27.

Reading List

Lee K. Computers in NM: a Practical Approach. New York: The Society of Nuclear Medicine Inc; 1991.

Wernick MN, Aarsvold JN, editors. Emission Tomography: The Fundamentals of PET and SPECT. San Diego: Elsevier Academic Press; 2004.

See also NUCLEAR MEDICINE INSTRUMENTATION; RADIATION DOSE PLANNING, COMPUTER-AIDED; RADIATION THERAPY SIMULATOR; RADIOPHARMACEUTICAL DOSIMETRY.

NUTRITION, PARENTERAL

HOWARD SILBERMAN
University of Southern
California
Los Angeles, California

INTRODUCTION

Parenteral nutrition, total parenteral nutrition (TPN), and hyperalimentation are terms referring to a variety of methods by which all required nutrients are provided intravenously, independent of alimentary tract function. Effective parenteral nutrition represents one of the major advances in medicine in the last 50 years and has led to intense interest in the incidence, consequences, and treatment of malnutrition, an area of study largely heretofore neglected because of a lack of effective therapy. The rationale for prescribing nutrients intravenously for patients unable to eat or receive tube feedings is based on the fact that malnutrition ultimately has an adverse impact on all organs and systems including the heart, lungs, gastrointestinal tract, and the immune system. The degree of morbidity and the rapidity of its onset depend on the magnitude and duration of nutritional deprivation. In addition, there are subtle, but perhaps quantitatively more important, indirect effects of nutritional deficits. Thus, nutritional derangements often have an adverse effect on the prognosis and treatment of concurrent illnesses. For example, malnourished patients who undergo elective surgery have a higher rate of postoperative sepsis and mortality. In addition, significant nutritional deficits may preclude the safe administration of optimal neoplastic chemotherapy because its inherent toxicity may be intolerable in the face of malnutrition.

DEVELOPMENT OF PARENTERAL NUTRITION

The adverse consequences of malnutrition became apparent to gastrointestinal (GI) surgeons during the first half of the twentieth century. Their patients with GI diseases were unable to eat, and the ensuing weight loss was associated with poor wound healing, anastomotic dysfunction and leakage, and poor surgical results. Such observations stimulated experimentation with intravenous nutrition. The early workers identified the nutrients required for human beings and further were able to prepare these nutrients in solutions that could be safely administered intravenously (1,2). However, clinical application was hampered by the fact that available nutrient solutions prepared in isotonic concentration could not meet energy and protein requirements when given in physiologic volumes. Experimentally, massive infusions of isotonic solutions of glucose and protein hydrolysates were indeed capable of supporting the anabolic state, but the intensive care required to monitor fluid balance made it a clinically impractical method. Early attempts at concentrating these solutions in order to provide all the required nutrients in acceptable volumes failed because of the thrombophlebitis that inevitably developed in the peripheral veins through which the solutions were infused. This was the state of the art until the mid-1960s when Rhoads, Dudrick, Wilmore, Vars, and their associates at the University of Pennsylvania undertook experiments in which very concentrated solutions were infused directly into the superior vena cava or the right atrium where instantaneous dilution of the solution occurred (3,4). By using hypertonic solutions of glucose and protein hydrolysates, these investigators demonstrated for the first time, initially in animals and then in humans, that intravenous "hyperalimentation" could support normal growth and development (2).

ESSENTIAL COMPONENTS OF NUTRIENT SOLUTIONS

The essential ingredients of solutions designed to meet all known nutritional requirements include nonprotein calories, utilizable nitrogen for protein synthesis, minerals, essential fatty acids, trace elements, and vitamins (5). In addition, it is likely that other micronutrients are necessary for optimal human nutrition. Normally, these as yet unidentified factors are automatically provided in a well-rounded diet derived from natural foodstuffs. Dudrick et al. (2,3) used glucose exclusively for nonprotein energy. Now fat emulsions are available as an additional clinically useful caloric source. Protein hydrolysates, derived from the enzymatic degradation of fibrin or casein, provided the nitrogen source in the early studies of parenteral nutrition. Currently, synthetic amino acid solutions are used to supply nitrogen.

Meeting the therapeutic goal of homeostasis (nitrogen equilibrium) or growth or nutritional repletion (positive nitrogen balance) is dependent on a variety of factors, most important among which are the levels of energy and nitrogen consumption. At any given level of protein or nitrogen intake, nitrogen balance progressively improves to some maximum level as caloric intake increases from levels below requirements to levels exceeding requirements (6). Maximum protein sparing and optimal utilization of dietary protein are achieved when the energy sources include at least 100–150 g of carbohydrate daily. The remaining energy requirements of most individuals can be met equally effectively by carbohydrate, fat, or a combination of these two. The requirement for this minimum amount of carbohydrate is based on its unique ability to satisfy the energy requirements of the glycolytic tissues, including the central nervous system, erythrocytes, leukocytes, active fibroblasts, and certain phagocytes.

At any given level of energy intake, nitrogen balance improves as nitrogen consumption increases. This dose-response relationship is curvilinear, and the nitrogen balance plateaus at higher dosages of nitrogen intake (6). To avoid the limiting effects of calories on nitrogen or of nitrogen on calories, parenteral nutrition solutions are prepared so that the nitrogen provided bears a fixed relationship to the nonprotein calories provided. In studies of normal, active young men fed orally, optimal efficiency was achieved at a calorie/nitrogen ratio of ~ 300–350 kcal to 1 g of nitrogen (6,7). However, protein economy decreases during most serious illnesses, and nitrogen losses increase; therefore, dietary protein requirements rise. Nitrogen equilibrium or retention can usually be achieved, however, by approximately doubling the quantity of nitrogen required by a normal man at any given level of caloric intake. Thus, a calorie/nitrogen ratio of 150:1 is thought optimal for seriously ill patients, although the ratio actually may range between 100 : 1 and 200 : 1 (6).

Minerals required in amounts exceeding 200 mg day^{-1} include sodium, potassium, calcium, magnesium, chloride, and phosphate. These macronutrients are essential for the maintenance of water balance; cardiac function; mineralization of the skeleton; function of nerve, muscle, and enzyme systems; and energy transformation. In addition, protein utilization is affected by the availability of sodium, potassium, and phosphorus in the diet; nitrogen accretion is impaired when any of these mineral nutrients is withdrawn from the diet. Nutritional repletion evidently involves the formation of tissue units containing protoplasm and extracellular fluid in fixed proportion and with fixed elemental composition (8). Thus, the retention of 1 g of nitrogen is characteristically associated with the retention of fixed amounts of phosphorus, potassium, sodium, and chloride.

Linoleic acid is the primary essential fatty acid for humans, and consequently it must be provided in order to avoid chemical and clinical evidence of deficiency. Linoleic acid requirements are generally met when a fat emulsion is used to provide at least 4% of calories as linoleic acid.

Micronutrients, or trace elements, presently recognized as essential for humans include iron, iodine, cobalt, zinc, copper, chromium, manganese, and possibly selenium. Cobalt is supplied as vitamin B12, and iron is generally withheld because of poor marrow utilization in critical or chronic illness. The remaining trace elements are routinely supplied in the nutrient solution (Table 1).

The remaining essential components are the water and fat soluble vitamins. Guidelines for parenteral vitamin administration have recently been revised by the Food and Drug Administration to include vitamin K, heretofore withheld, and increased amounts of vitamins B1, B6, C and folic acid (13–15) (Table 2).

Table 1. Trace Element Requirements[a–c]

Chromium	10–15 mcg
Copper	0.5–1.5 mg
Iodine	1–2 mcg kg^{-1}
Manganese	60–100 mcg
Zinc	2.5–4.0 mg

[a] Daily intravenous maintenance allowances for adults with normal initial values.

[b] Based on recommendations of Refs. 9–11.

[c] Selenium, 40–80 mcg day^{-1}, recommended for patients requiring long-term TPN, or those with burns, acquired immunodeficiency syndrome or liver failure. Ref. 12. p 125.

Table 2. Vitamin Requirements[a]

Thiamine (B_1)	6 mg
Riboflavin (B_2)	3.6 mg
Niacin (B_3)	40 mg
Folic acid	600 mcg
Pantothenic acid (B_5)	15 mg
Pyridoxine (B_6)	6 mg
Cyanocobalamin (B_{12})	5 mcg
Biotin	60 mcg
Ascorbic acid (C)	200 mg
Vitamin A	3300 IU
Vitamin D	200 IU
Vitamin E	10 IU
Vitamin K	150 mcg

[a] Daily intravenous allowances for adults. Based on Food and Drug Administration requirements Ref. 15.

The specific nutrient requirements for a given individual depend on the initial nutritional and metabolic status of the patient and his/her underlying disease process. Although the precise requirements for each nutrient can be determined by metabolic balance studies and direct or indirect calorimetry, such techniques are generally not employed in routine clinical practice. Instead, on the basis of data from clinical investigations applying such balance studies to patients with various diseases, injuries, and degrees of stress, requirements can be accurately estimated (16). As a result of these estimates, it is possible to formulate basic nutrient solutions of essentially fixed composition that can be used to meet the needs of most patients by varying only the volume of the basic formulation and by making simple adjustments in the electrolyte content of the solution. Thus, in clinical practice the caloric requirement is usually estimated from simple formulas that are adjusted for activity and stress. Although nitrogen requirements can likewise be determined or estimated, nitrogen needs are usually met as a consequence of the fixed calorie/nitrogen ratio of the nutrient solution. Thus, as the volume of infusate is increased to meet increased caloric demands, the additional nitrogen requirements, which generally parallel the rising caloric needs, are likewise met. Although such standard preparations can be used to satisfy the needs of most individuals, fluid-restricted patients, severely hypermetabolic patients, or those with renal or hepatic failure may require special nutrient formulations.

FORMULATING NUTRIENT SOLUTIONS

In current practice, nutrient solutions designed for parenteral administration are formulated to provide nonprotein calories as carbohydrate or a combination of carbohydrate and lipid. Glucose is the carbohydrate of choice since it is the normal physiologic substrate; it naturally occurs in blood; and it is abundant, inexpensive, and readily purified for intravenous administration. Glucose can be given in high concentrations and in large amounts that are well tolerated by most patients after a period of adaptation. Other carbohydrates such as fructose, sorbitol, xylitol, and maltose have been evaluated experimentally, but each has disadvantages that preclude clinical application at the present time. Glucose for parenteral infusion is commercially available in concentrations from 5 to 70% and is provided as glucose monohydrate with a caloric density of 3.4 $\text{kcal}\cdot\text{g}^{-1}$. Although isotonic (5%) solutions of glucose are available, concentrated glucose solutions are necessary in parenteral nutrition protocols in order to provide required calories in physiologic volumes of fluid.

Lipid is the alternative clinically useful nonprotein caloric source. Fat emulsions derived from soybean and safflower oil were approved for use in the United States in 1975 and 1979, respectively. The soybean oil emulsions had been used in Europe for nearly 20 years prior to their introduction in the United States. Currently available fat emulsions are derived from soybean oil or are mixtures of soybean and safflower oil emulsions. The use of lipid emulsions in intravenous feeding regimens is attractive because of the high caloric density of fat (9 $\text{kcal}\cdot\text{g}^{-1}$), and because they are isotonic solutions that can, therefore, provide many calories in relatively small volumes via peripheral veins. Although early experience with lipids using the cottonseed oil emulsion Lipomul was unsatisfactory because of the toxicity of that preparation, fat emulsions derived from soybean and safflower oil have proven safe for clinical use. These newer preparations are purified plant oils emulsified in water. Egg phospholipids are added to regulate the size of the fat particles, stabilize the emulsion, and prevent fusion of the oil drops. Glycerol is added to make the emulsion isotonic, since oil and water emulsions have no osmolal effect. The resulting fat droplets have characteristics that are similar to those of naturally occurring chylomicrons found in the circulation after absorption of dietary lipid from the small intestine. Thus the particle size and the plasma elimination characteristics of these fat emulsions appear comparable to those of chylomicrons (17,18).

Studies of the elimination kinetics of soybean emulsion triglycerides indicate that at very low concentrations the rate of removal from the plasma is dependent on the triglyceride concentration. Above a certain critical concentration representing saturation of binding sites of lipoprotein-lipase enzymes, a maximum elimination capacity is reached that is independent of concentration. This maximum elimination capacity is influenced by the clinical state of the patient. It is increased during periods of starvation, after trauma, and in severely catabolic states (18,19). The infusion of fats is associated with an increase in heat production and oxygen consumption, a decrease in respiratory quotient, and the appearance of carbon-14 (^{14}C) in the expired air of patients receiving ^{14}C-labeled fat. These observations indicate that the infused fats are in fact used for energy. Soybean and soybean–safflower oil emulsions are available in 10, 20, and 30% concentrations and are mixtures of neutral triglycerides of predominantly unsaturated fatty acids. The major component fatty acids are linoleic, oleic, palmitic, and linolenic. The total caloric value of the 10% emulsions, including triglyceride, phospholipid, and glycerol, is 1.1 $\text{kcal}\cdot\text{mL}^{-1}$. The corresponding values for the 20 and 30% emulsions are 2 and 3 $\text{kcal}\cdot\text{mL}^{-1}$, respectively. In each of these preparations, ~0.1 $\text{kcal}\cdot\text{mL}^{-1}$ of the total caloric value is derived from the added glycerol.

All nutrient solutions must provide at least 100–150 g of glucose day^{-1} to meet the needs of the glycolytic tissues, as described above. The proportional distribution of glucose and fat to provide the remaining required nonprotein calories apparently can be widely variable with the expectation of achieving the same nutritional goals. Commonly cited guidelines suggest that fat should not provide >30% of nonprotein calories and that the daily dosage of fat not exceed 2.5 g kg^{-1} in adults (11). However, evidence-based reports indicate infusions providing >80% of nonprotein calories as fat and daily dosages of as much a 12 g fat kg^{-1} day^{-1} have been tolerated with clinical benefit and no adverse effects (20–22).

Consequently, protocols for solution preparation vary from institution to institution. Practical considerations in choosing the amount of glucose and the amount of fat relate primarily to the route of administration and to the fluid status of the patient. Parenteral nutrition solutions providing all nonprotein calories as glucose are highly concentrated and require central venous administration (see below). As an increasing proportion of the caloric content of the nutrient solution is provided by an isotonic fat emulsion, the content of glucose is thereby reduced, and, consequently, the concentration of the resultant solution falls. Nutrient solutions with an osmolarity not exceeding approximately three times normal serum levels can be successfully infused through peripheral veins (23); more concentrated solutions must be infused centrally. Parenteral nutrition solutions are formulated in accordance with one of three commonly used protocols: the glucose-based system, the lipid-based system, and a three-in-one system of variable composition.

The nutrient solution prescribed for a 24 h period usually is prepared in single container in a pharmacy under strict aseptic conditions. Manufacturing pharmacies responsible for preparing solutions for many patients often employ automated, computerized compounding apparatus that is programmed to add the specified nutrient components to the infusion container.

THE GLUCOSE SYSTEM

This is the original carbohydrate-based system developed by Dudrick and his associates at the University of

Table 3. The Glucose System[a–d]

50% glucose	500 mL
8.5% Amino acids	500 mL
Sodium (as acetate)	25 meq
Sodium (as chloride)	5 meq
Sodium (as phosphate)	14.8 meq
Potassium (as chloride)	40 meq
Phosphate (as sodium salt)	11.1 m*M*
Magnesium sulfate	8 meq
Calcium gluconate	5 meq

[a] Composition per liter.

[b] Provides 850 nonprotein kilocalories and 7.1 g nitrogen/L.

[c] Trace elements and vitamins are provided daily (see Tables 1 and 2).

[d] Electrolyte additives based on Travasol as the amino acid source.

Pennsylvania (3). The nutrient solution is prepared by the admixture of equal volumes of 50% glucose and 8.5% crystalline amino acids, and the addition of appropriate electrolytes, vitamins, and trace elements (Tables 1–3). A 1 L solution provides 850 nonprotein kilocalories, and ~7 g of nitrogen, equivalent to 44 g protein. Consequently, the solution has a calorie/nitrogen ratio of ~ 120 : 1. The nitrogen content and the calorie/nitrogen ratio will vary slightly depending on the brand of amino acid solution used. Due to the osmolar contribution of each of the constituents, this nutrient solution has a final concentration of ~ 2000 $\mathrm{mOsm \cdot L^{-1}}$. Such a solution can never be safely infused through peripheral veins; consequently, the glucose system must be delivered into a central vein where the infusate is immediately diluted. Vascular access is usually through a percutaneously placed subclavian venous catheter or, for short-term use, a peripherally inserted central venous catheter (PICC line). Other routes (eg, via jugular, saphenous, or femoral veins) are used occasionally with variable success. The incidence of morbidity associated with establishing and maintaining central venous access is influenced by the site and technique of insertion of the venous cannula and the diligence with which the apparatus is managed during the course of nutrition therapy.

Because of the high concentration of glucose in this form of parenteral nutrition, therapy should begin gradually to allow adaption and thereby avoid hyperglycemia. Generally, on the first day a patient receives 1 L of nutrient solution, which is infused at a constant rate over the full 24 h period. Blood and urine glucose levels are monitored frequently, and if this initial rate of infusion is well tolerated, the volume prescribed is increased from day-to-day until the volume infused meets the caloric requirement of the individual patient. For the average patient, the nutritional requirements as well as the requirements for fluid and electrolytes are usually met by 2.5 $\mathrm{L \cdot day^{-1}}$ of the nutrient solution, providing 2125 nonprotein kilocalories.

Infusion of the glucose system at a constant rate is a critical feature of safe practice since abrupt changes in the rate of delivery may be associated with marked fluctuations in blood sugar levels. The constant rate of infusion is most efficiently achieved by using an infusion pump. At the conclusion of therapy, the rate of infusion should be tapered gradually over several hours to avoid hypoglycemia. When infusion must be abruptly terminated, a solution of 10% glucose is substituted for the nutrient solution.

THE LIPID SYSTEM

The system of total parenteral nutrition based on glucose as the major caloric source is simple in concept and the least expensive, but patients' glucose metabolism must be closely monitored, and administration of the infusate requires technical expertise to achieve and maintain the central venous access necessary for safe treatment. The use of lipid emulsions as the major caloric source is attractive because of the high caloric density and isotonicity of these products. These considerations have logically led to the preparation of nutrient solutions based on fat as the major caloric source, with the goal of providing all required nutrients by peripheral vein. An example of such a lipid-based system of total parenteral nutrition is presented in Table 4. This nutrient solution was devised with the aim of maximizing caloric and amino acid content without producing a solution with a concentration that would preclude safe peripheral venous administration (23). Each liter provides 720-nonprotein kilocalories and 5.0 g of nitrogen, equivalent to 31 g of protein. The calorie/nitrogen ratio is 144: 1. The nitrogen content of the solution will vary slightly, depending on the amino acid product used in its preparation. As with the glucose system, sufficient volume is given to meet measured or estimated caloric requirements. The safety and efficacy of nutrient solutions utilizing lipid as the major caloric source were established by Jeejeebhoy and associates in their landmark investigation of lipid-based TPN in which 83% of nonprotein calories were supplied as fat (22). As many as 5 L daily of the lipid system described here have been infused for periods of weeks to months without apparent adverse effect.

Table 4. The Lipid System[a–e]

10% Fat emulsion	500 mL
50% Glucose	100 mL
8.5% Amino acids	350 mL
Sodium (as acetate)	35 meq
Sodium (as chloride)	5 meq
Sodium (as phosphate)	6 meq
Potassium (as chloride)	40 meq
Phosphate (as sodium salt)[e]	4.5 m*M*
Magnesium sulfate	8 meq
Calcium gluconate	5 meq
Heparin sodium	1000 U
Distilled water	q.s.ad 1000 mL

[a] Composition per liter.
[b] Provides 720 nonprotein kilocalories and 5.0 g nitrogen/L.
[c] Trace elements and vitamins are provided daily (see Tables 1 and 2).
[d] Electrolye additives are based on Travasol as amino acid source.
[e] Aproximately 7 m*M* additional phosphorus derived from fat emulsion.

THE THREE-IN-ONE SYSTEM

Innumerable nutrient solutions can be prepared with a distribution of glucose and lipid calories that differs from the two systems described above. Although there is no established biological advantage of differing proportions of fat and glucose as long as the minimum 100–150 g of carbohydrate are supplied, many clinicians prefer a profile of nutrients that mimics the optimal oral diet. Such a system calls for the admixture of amino acids, glucose, and lipids in which carbohydrate provides 65–85% of nonprotein calories and lipid 15–35% (11,12) (Table 5).

However, altering the lipid system described here by increasing the glucose content would have certain nonnutritional effects. Thus, a solution with a higher proportion of glucose calories could be produced by replacing some of the fat emulsion with isotonic glucose. The concentration of the final solution would remain unchanged, but a much greater total volume would be required to provide the same number of calories. If the substitution were made with hypertonic glucose, as called for in the three-in-one system

Table 5. A Three-in-One System[a–d]

50% Glucose	300 mL
10% Fat emulsion	300 mL
10% Amino acids	400 mL
Sodium (as acetate)	25 meq
Sodium (as chloride)	5 meq
Sodium (as phosphate)	14.8 meq
Potassium (as chloride)	40 meq
Phosphate (as sodium salt)	1.1 m*M*
Magnesium sulfate	8 meq
Calcium gluconate	5 meq

[a] Composition per liter.
[b] Provides 840 nonprotein kilocalories and 6.8 g nitrogen/L.
[c] Trace elements and vitamins are provided daily (Tables 4 and 5).
[d] Electrolyte additives based on Travasol as the amino acid source.

described in Table 5, the final concentration of the nutrient solution would be so increased as to require central venous administration, thereby losing the advantage of peripheral venous delivery.

COMPLICATIONS OF PARENTERAL NUTRITION

Morbidity associated with intravenous feedings may be related to drug toxicity, difficulties with vascular access, sepsis, or metabolic derangements.

Drug Toxicity

Adverse reactions to the components of parenteral nutrition solutions are uncommon. Although glucose is virtually nontoxic, the hypertonic solutions employed in the glucose system of TPN may be associated with potentially serious complications usually related to alterations in blood glucose levels. Currently used solutions of synthetic amino acids provide all of the nitrogen in the form of free l-amino acids and, in contrast to previously used protein hydrolysates, no potentially toxic ammonia or peptide products are present. Toxicity associated with the intravenous infusion of the currently available fat emulsions also has been minimal. The most frequent acute adverse reactions are fever, sensations of warmth, chills, shivering, chest or back pain, anorexia, and vomiting. Similarly, adverse reactions associated with chronic infusions of fat emulsions are also quite uncommon. Anemia and alterations in blood coagulation have been observed during treatment, but the etiologic relationship to lipid infusions has been unconfirmed. The most serious adverse effects have been observed in infants and children. The "fat overload" syndrome associated with the older cottonseed emulsion has rarely been observed with the newer current preparations. Nevertheless, several reports have been published (24) in which children receiving fat emulsions have developed marked hyperlipidemia, GI disturbances, hepatosplenomegaly, impaired hepatic function, anemia, thrombocytopenia, prolonged clotting time, elevated prothrombin time, and spontaneous bleeding. These findings resolved when the fat emulsion was withdrawn.

Complications of Vascular Access

The lipid-based system of parenteral nutrition can be infused through the ordinary peripheral venous cannulae used for the administration of crystalloid solutions. Local phlebitis and inflammation from infiltration and cutaneous extravasation occur with about the same frequency as that associated with the infusion of nonnutrient solutions. In contrast, a central venous catheter is required for infusion of the highly concentrated glucose and three-in-one systems. Insertion and maintenance of such catheters may be associated with a variety of complications. Complications that may occur during the placement of the catheter include improper advancement of the catheter tip into one of the jugular veins or the contralateral innominate vein, instead of into the superior vena cava. In addition, air embolization or cardiac arrhythmias may occur. Percutaneous jugular or subclavian cannulation may rarely result in an injury to an adjacent anatomic structure, such as the brachial plexus, great vessels, or thoracic duct. Pneumothorax, usually resulting from inadvertent entrance into the pleural cavity, is probably the most common complication of attempted subclavian catheterization and has been reported to occur in ~ 2–3% of attempts in large series. Late complications after successful central catheterization may include air embolism, catheter occlusion, central vein thrombophlebitis, and catheter-related sepsis.

Systemic Sepsis

Sepsis attributable primarily to the administration of parenteral nutrition should be an infrequent complication in modern practice. A variety of factors may contribute to the development of this complication. Hyperglycemia, which may be induced or aggravated by nutrient infusions (see below), has been associated with sepsis in critically ill patients. Maintaining blood glucose levels between 80 and 110 $\text{mg}\cdot\text{dL}^{-1}$ has been shown to significantly reduce the incidence of septicemia (25). In addition, patients requiring TPN are often inordinately susceptible to infection because of serious illness, malnutrition, and chronic debilitation—all conditions associated with impaired immune responses. Patients receiving immunosuppressive therapy, cytotoxic drugs, or corticosteroids are likewise susceptible to infection. These drugs as well as prolonged administration of broad-spectrum antibiotics may subject patients to sepsis from unusual, ordinarily saprophytic, microorganisms.

In addition to these patient-related factors, several specific TPN-related factors contribute to the pathogenesis of sepsis. The various components of the nutrient solution can become contaminated during manufacture or at the time of component admixture in the hospital pharmacy. The ability of the nutrient solution to support microbial growth is well established, but with present techniques of solution preparation sepsis from contamination should be rare. The vascular access apparatus appears to be the most common source of TPN-associated sepsis. Contamination may take place when the infusion catheter is inserted; when containers of the nutrient solution are changed; when intravenous tubing is replaced; when in-line filters are inserted; or when the intravenous cannula is used for measurement of central venous pressure, blood sampling, or the infusion of medication or blood products. In addition, to-and-fro motion of a subclavian catheter due to inadequate fixation will allow exposed portions of the catheter to enter the subcutaneous tract leading to the vein, which may result in infection. Hematogenous contamination of the infusion catheter may occasionally occur following bacteremia secondary to a distant focus of infection. More commonly, however, catheter-related sepsis is due to contamination of the catheter by organisms colonizing the skin surrounding the catheter insertion site. The incidence of sepsis varies greatly in reported series, but in recent years TPN has been administered with very low rates of infection. This improving trend is evidently due to adherence to rigid protocols of practice, and the employment in many hospitals of a dedicated, multidisciplinary team to manage the nutritional therapy. With this approach, TPN-related sepsis occurs in $\sim 3\%$ of patients receiving the glucose system. This complication is much less common among patients receiving the lipid-based system of parenteral nutrition through a peripheral vein (16).

Metabolic Complications

A variety of metabolic derangements have been observed during the course of total parenteral nutrition. These derangements may reflect preexisting deficiencies, or they may develop during the course of parenteral nutrition as a result of an excess or deficiency of a specific component in the nutrient solution. As would be expected, the standard solutions may not contain the ideal combination of ingredients for a given individual. In fact, adverse effects from an excess or deficiency of nearly every component of nutrient solutions have been described. Consequently, patients must be carefully monitored so that the content of the nutritional solution can be adjusted during the course of therapy. For example, minor alterations in electrolyte content are often necessary.

Abnormalities of blood sugar are the most common metabolic complications observed in patients receiving total parenteral nutrition. Hyperglycemia may be associated with critical illness independent of nutrient infusions. However, patients receiving the glucose-rich glucose and the three-in-one systems are particularly susceptible to elevated blood sugar levels. In addition, hyperglycemia may be manifest when the full caloric dosage of the glucose and three-in-one systems is inappropriately given initially and later if rates of infusion are abruptly increased. In addition, glucose intolerance may be a manifestation of overt or latent diabetes mellitus, or it may reflect reduced pancreatic insulin response to a glucose load, a situation commonly observed during starvation, stress, pain, major trauma, infection, and shock. Hyperglycemia also may be a reflection of the peripheral insulin resistance observed during sepsis, acute stress, or other conditions that are accompanied by high levels of circulating catecholamines and glucocorticoids. Decreased tissue sensitivity to insulin is also associated with hypophosphatemia, and hyperglycemia has been observed in patients with a deficiency of chromium. The latter trace metal probably acts as a cofactor for insulin. The incidence of hyperglycemia can be minimized by initiating therapy gradually with either of the two glucose-rich systems. Full dosage should be achieved over a 3 day period, during which time adaption to the glucose load takes place. In addition, careful metabolic monitoring during this period will disclose any tendency to hyperglycemia. Subsequently, a constant rate of infusion is maintained. An inadvertent decrease in the rate of the infusion should not be compensated by abrupt increases in rate; such "catching up" is not allowed. When hyperglycemia supervenes despite these precautions, the etiology is sought. The common cause of hyperglycemia after a period of stability is emerging sepsis, the overt manifestations of which may not appear for 18–24 h after development of elevated glucose levels.

Recent evidence indicates that maintenance of blood sugars levels between 80 and 110 $\text{mg}\cdot\text{dL}^{-1}$ in critically ill patients is associated with a significant reduction in mortality and the incidence of septicemia (25). Uncomplicated, moderate hyperglycemia is controlled initially by subcutaneous or intravenous administration of insulin; the TPN infusion is continued at the usual rate. Subsequently, the appropriate amount of insulin is added to the TPN solution during its aseptic preparation in the pharmacy. Providing insulin in the TPN solution has the advantage that inadvertent alterations in the rate of glucose delivery are automatically accompanied by appropriate adjustments in the amount of insulin administered. Patients with hyperglycemia complicated by massive diuresis, dehydration, neurologic manifestations, or the syndrome of hyperosmolar nonketotic coma are managed by immediate termination of the TPN infusion, fluid resuscitation, and insulin administration.

In contrast to the problem of hyperglycemia, blood sugar levels decrease when the rate of infusion of the glucose system is abruptly reduced. Symptomatic hypoglycemia is most likely to occur when the reduction of the infusion rate had been preceded by an increased rate. When the glucose system is to be discontinued electively, the rate of delivery should be tapered gradually over several hours. Patients who are hemodynamically unstable or who are undergoing surgery should not receive TPN, since fluid resuscitation may be inadvertently carried out using the TPN solution. Therefore, the TPN infusion is discontinued abruptly in such patients, and hypoglycemia is averted by infusing a solution of 10% glucose. Hypoglycemia may also reflect an excessive dosage of exogenous insulin. This most commonly occurs as a result of failure to recognize the resolution of peripheral insulin resistance and the associated decreased insulin requirement when the provoking condition responds to therapy.

Serum lipid profiles, which are routinely monitored during treatment with the lipid system, commonly reveal elevations of free fatty acids, cholesterol, and triglycerides. However, adverse clinical effects are uncommon (22,26). Nevertheless, triglyceride levels $> 400\ \text{mg}\cdot\text{dL}^{-1}$ should be avoided since hypertrigylceridemia of this magnitude may be associated with an increased risk of pancreatitis, immunosuppression, and altered pulmonary hemodynamics (11).

Deficiencies of the major intracellular ions may occur in the catabolic state, since the protein structure of cells is metabolized as an energy source, intracellular ions are lost, and the total body concentration of these ions, including potassium, magnesium, and phosphate, are decreased. Furthermore, during nutritional repletion, these ions, derived from the serum, are deposited or incorporated in newly synthesized cells. When supplementation of these ions in nutrient solutions is insufficient, hypokalemia, hypomagnesemia, and hypophosphatemia ensue. Serum levels of these substances should be measured regularly during TPN since such monitoring will disclose deficiencies before the clinical manifestations develop. Symptoms of hypokalemia are unusual when serum levels of potassium $> 3.0\ \mathrm{meq{\cdot}L^{-1}}$. Asymptomatic hypokalemia can be managed by increasing the potassium supplement added to the nutrient solution at the time of preparation. When cardiac arrhythmias or other significant symptoms develop, the rate of TPN infusion should be tapered promptly while serum glucose levels are monitored closely, and an intravenous infusion of potassium chloride is begun.

Intracellular consumption of inorganic phosphate during the synthesis of proteins, membrane phospholipids, DNA, and adenosine triphosphate (ATP) may produce a striking deficit in the serum phosphate level after only several days of intravenous feedings devoid of or deficient in phosphate. Symptoms of hypophosphatemia may occur when serum phosphate levels fall to $2\ \mathrm{mg{\cdot}dL^{-1}}$. However, severe manifestations are particularly apt to occur as levels fall to $< 1\ \mathrm{mg{\cdot}dL^{-1}}$. These include acute respiratory failure, marked muscle weakness, impaired myocardial contractility, severe congestive cardiomyopathy, acute hemolytic anemia, coma, and death. Hypophosphatemic patients who are asymptomatic can be managed by increasing the phosphate supplement in the nutrient solution. Symptomatic patients or those with serum phosphate levels $< 1\ \mathrm{mg\ dL^{-1}}$ should be repleted intravenously through a separate infusion line. Parenteral nutrition should be stopped, and a 10% glucose solution should be infused to avert hypoglycemia. Since intracellular phosphate consumption is dependent on caloric intake, withdrawing TPN alone often results in an increased serum phosphate level within 24 h.

A variety of adverse effects comprising the *refeeding syndrome* has been associated with the rapid induction of the anabolic state in severely malnourished, cachectic patients using standard nutrient solutions (16). Cardiac decompensation, the most serious feature of the syndrome, may be due to overhydration and salt retention in the face of starvation-induced low cardiac reserve. Hypophosphatemia, consequent to rapid refeeding, is another important contributing factor to cardiac failure. Rapid nutritional repletion also is implicated in producing deficits of the other major intracellular ions, magnesium and potassium, as well as acute deficiencies of vitamin A (associated with night blindness), thiamine (associated with the high output cardiac failure of beriberi, Wenicke's encephalopathy, and lactic acidosis), and zinc (associated with diarrhea, cerebellar dysfunction, dermatitis, impaired wound healing, and depressed immunity). Refeeding alkalosis also has been described. To avoid the refeeding syndrome in the chronically starved patient, parenteral nutrition should be introduced more gradually than usual, perhaps reaching the full caloric and protein requirements over the course of 5–7 days (16).

Healthy or malnourished individuals who receive a constant parenteral infusion of a fat-free, but otherwise complete diet eventually develop clinical and biochemical manifestations that are completely reversed by the administration of linoleic acid. Thus, the syndrome of essential fatty acid deficiency in humans is due principally, if not exclusively, to a lack of linoleic acid. Exogenous linolenic acid is required by some species, but its essentiality for humans is unproven. The most commonly recognized manifestation of linoleic acid deficiency is an eczematous desquamative dermatitis largely, but not always, confined to the body folds. Other clinical findings may include hepatic dysfuntion, anemia, thrombocytopenia, hair loss, and possibly impaired wound healing. Growth retardation has been observed in infants. Fatty acid deficiency is treated by the administration of linoleic acid, usually by infusing one of the currently available fat emulsions. Patients receiving the glucose-based system of parenteral nutrition should be treated prophylactically by providing 4% of calories as linoleic acid. This requirement is usually met by infusing $1\ \mathrm{L{\cdot}week^{-1}}$ of a 10% fat emulsion.

Abnormalities in bone metabolism have been observed in patients receiving parenteral nutrition for prolonged periods, especially in home treatment programs. Such metabolic bone disease includes the common disorders of osteoporosis and osteomalacia and is characterized by hypercalciuria, intermittent hypercalcemia, reduced skeletal calcium, and low circulating parathormone levels. The clinical features have included intense periarticular and lower extremity pain. The pathogenesis of this syndrome is obscure, but hypotheses include an abnormality of vitamin D metabolism and aluminum toxicity (27–29). Most recently, vitamin K deficiency has been considered an etiologic factor since it has been recognized that this condition increases the risk of osteoporosis and fractures and that these risks can be reduced with vitamin K therapy. Vitamin K also appears to be necessary for the synthesis of a diverse group of proteins involved in calcium homeostasis (13,14,30,31). These findings have lead to the recent recommendation to routinely add vitamin K to TPN solutions, as discussed above.

NON-NUTRITIONAL EFFECTS OF PARENTERAL NUTRITION

Effects on the Stomach

Gastric acid secretion is significantly increased during the initial period of treatment with the glucose system, but the duration of this effect is unknown. The acid secretory response observed is due primarily to the infusion of crystalline amino acids, and this effect of amino acids on gastric secretion is virtually abolished by the concurrent intravenous infusion of a fat emulsion. The effect of chronic TPN on gastric secretory function is less clear. Chronic parenteral nutrition in animals has been associated with decreased antral gastrin levels and atrophy of the parietal cell mass. This observation is consistent with anecdotal clinical reports in which gastric hyposecretion has been observed in patients receiving long-term parenteral nutrition at home (32).

Effects on the Intestinal Tract

Morphologic and functional changes occur in the small intestine and the colon when nutrition is maintained exclusively by vein. A significant reduction in the mass of the small and large intestine occurs, and there is a marked decrease in mucosal enzyme activity. Enzymes affected include maltase, sucrase, lactase, and peroxidase. These changes are not in response to intravenous nutrition per se, but reflect the need for luminal nutrients for maintenance of normal intestinal mass and function. The mechanism by which food exerts a trophic effect is at least in part endocrine in that intraluminal contents stimulate the release of enterotrophic hormones such as gastrin (16,33,34).

Effects on the Pancreas

Similar morphologic and functional atrophy of the pancreas is observed during the course of parenteral nutrition. In contrast to the effect of fat consumed orally, intravenous lipids do not stimulate pancreatic secretion (16,35).

Effects on the Liver

Transient derangements of liver function indexes occur in the majority of patients receiving parenteral nutrition regardless of the proportion of glucose and lipid (36,37). Similar abnormalities also have been observed in patients receiving enteral nutrition (tube feedings) (38,39). The etiology of these changes is uncertain and probably multifactorial. One hypothesis is that glucose and protein infusions in amounts exceeding requirements may contribute to these changes. In addition, an infectious etiology, perhaps related to the underlying condition requiring nutritional support, has been suggested, since oral metronidazole has been reported to reverse the changes in some patients. Administration of ursodesoxycholic acid also has been associated with improvement of TPN-related cholestasis (40). In any case, the clinical course associated with the liver changes is nearly always benign so that TPN need not be discontinued. Nevertheless, patients receiving TPN for several years or more are at greater risk for developing severe or chronic liver disease, but again the etiologic relationship is unclear (36).

Effects on the Respiratory System

Fuel oxidation is associated with oxygen consumption and carbon dioxide production. Oxygenation and carbon dioxide elimination are normal pulmonary functions. Consequently, patients with respiratory failure receiving aggressive nutritional support may not be able to meet these demands of fuel metabolism. It is particularly important to avoid infusing calories in amounts exceeding requirements, since this aggravates the problem, increases tidal volume, respiratory rate, and $P\text{CO}_2$, and offers no nutritional benefit (41).

INDICATIONS FOR PARENTERAL NUTRITION

Although the clinical benefits derived from nutritional substrates infused intravenously appear equivalent to those derived from substrates absorbed from the alimentary tract, feeding through the alimentary tract is preferable when feasible because this route of administration is less expensive, less invasive, and, most importantly, is associated with a significantly lower incidence of infectious complications (42). Nevertheless, many hospitalized patients have conditions in which alimentary tract nutrition either by mouth or tube feeding is inadequate, inadvisable, or would require an operative procedure (e.g., gastrostomy or jejunostomy) to establish access. It is for these patients that parenteral feeding should be considered. Normally nourished patients unable to eat for as long as 7–10 days generally do not require parenteral nutrition. The protein-sparing effect of 100–150 g of glucose provided in a 5% solution is sufficient. Patients in this category include those undergoing GI surgery in whom only several days of ileus are anticipated postoperatively. However, if the resumption of adequate intake is not imminent after 7–10 days, parenteral feedings are recommended. In contrast, normally nourished patients should receive TPN promptly when initial evaluation discloses gastrointestinal dysfunction that is expected to persist beyond 7–10 days. In addition, malnourished or markedly hypercatabolic patients (e.g., those with severe burns, sepsis, or multiple trauma) with GI dysfunction are given parenteral nutrition immediately.

In some patients, parenteral feedings have benefits in addition to improved nutrition. When all nutrients are provided intravenously, a state of bowel rest can be achieved in which the mechanical and secretory activity of the alimentary tract declines to basal levels (see earlier). These nonnutritional effects may be beneficial in the management of GI fistulas and acute inflammatory diseases, such as pancreatitis and regional enteritis. Parenteral nutrition may also be useful as a “medical colostomy”. Thus, the reduction or elimination of the fecal stream associated with intravenous feedings may benefit patients with inflammation or decubitis ulcers adjacent to the anus or an intestinal stoma or fistula.

Any preexisting acute metabolic derangement should be treated before parenteral nutrition is begun. In addition, TPN should not be used during periods of acute hemodynamic instability or during surgical operations since the nutrient solution may be used inadvertently for fluid resuscitation. Parenteral nutrition is not indicated for patients with malnutrition due to a rapidly progressive disease that is not amenable to curative or palliative therapy.

COMPARING METHODS OF TOTAL PARENTERAL NUTRITION

Factors to be considered in comparing the glucose, the lipid, and the three-in-one systems of parenteral nutrition include the composition and nutrient value of the three systems, the relative efficacy of glucose and lipid calories, and the ease and safety of administration.

Comparative Composition of Parenteral Nutrition Systems

As outlined in Table 6, the lipid system provides fewer calories and less nitrogen per unit volume than the glucose and three-in-one systems. Thus, greater volumes of the lipid system are required to provide an isocaloric and isonitrogenous regimen. On the other hand, the lower osmolarity of the lipid system permits safe peripheral venous administraton of all required nutrients, whereas

Table 6. Comparison of Parenteral Nutrition Systems

	Glucose System	Lipid System	Three-in-One System
Carbohydrate calorie	850 $\text{kcal}\cdot\text{L}^{-1}$	220 $\text{kcal}\cdot\text{L}^{-1}$	540 $\text{kcal}\cdot\text{L}^{-1}$
Lipid calories	0 $\text{kcal}\cdot\text{L}^{-1}$	500 $\text{kcal}\cdot\text{L}^{-1}$	300 $\text{kcal}\cdot\text{L}^{-1}$
Caloric density	0.85 $\text{kcal}\cdot\text{mL}^{-1}$	0.72 $\text{kcal}\cdot\text{mL}^{-1}$	0.84 $\text{kcal}\cdot\text{mL}^{-1}$
Nitrogen provided	7.1 $\text{g}\cdot\text{L}^{-1}$	5.0 $\text{g}\cdot\text{L}^{-1}$	6.8 $\text{g}\cdot\text{L}^{-1}$
Protein equivalent	44 $\text{g}\cdot\text{L}^{-1}$	31 $\text{g}\cdot\text{L}^{-1}$	42.5 $\text{g}\cdot\text{L}^{-1}$
Calorie/nitrogen ratio	120:1	144:1	124:1
Concentration (approximate)	2000 $\text{mOsm}\cdot\text{L}^{-1}$	900 $\text{mOsm}\cdot\text{L}^{-1}$	1500 $\text{mOsm}\cdot\text{L}^{-1}$

the higher concentration of the other two systems mandates central venous infusion.

Glucose Versus Lipid as a Caloric Source

The relative impact of glucose and lipid calories on nitrogen retention or body composition has been the subject of extensive investigation often with disparate conclusions, depending on the subset of patients studied (16). However, the preponderance of evidence supports the conclusion that the two caloric sources are of comparable value in their effect on nitrogen retention in normal persons or in chronically ill, malnourished patients. The major study supporting this conclusion is that of Jeejeebhoy and associates (22), who observed that optimal nitrogen retention with the lipid system requires a period of ~ 4 days to establish equilibrium, after which nitrogen balance is positive to a comparable degree with both the glucose and lipid systems. More recent data now attest to the equivalent efficacy of lipid as a major caloric source in critically illness and sepsis (43–45).

Ease and Safety of Administration

The glucose and three-in-one systems require central venous administration. Percutaneously inserted central venous catheters must be placed by physicians and peripherally inserted central venous catheters (PICC) by physicians or specially trained nurses under sterile conditions. Insertion and use of central catheters may be associated with certain complications discussed previously that are not seen with the peripherally administered lipid system.

The ordinary venous cannulae used for infusion of the lipid system can be easily inserted at the bedside and maintained by ward personnel. Whereas a central venous catheter requires special care and attention to prevent catheter sepsis, best provided by a dedicated team, the cannulae used in the lipid system require the same simple care as those used in the peripheral venous administration of crystalloid solutions. The peripherally-infused lipid system is rarely associated with systemic sepsis (16,23).

SELECTING THE TPN REGIMEN

For many patients, the nutritional requirements can be met equally well by any of the TPN systems discussed. The selection in these cases is often based on nonnutritional factors, such as the experience of the physician, ease of administration, and anticipated duration of therapy. On the other hand, there are subsets of patients requiring intravenous nutritional support who have associated or concurrent medical conditions that influence the choice of treatment.

Fluid Restriction

The lipid system described here has the lowest caloric and nitrogen content per unit volume of the three standard regimens (Table 6). Thus, a greater volume has to be infused to provide the same nutrients. Fluid restriction is facilitated, therefore, by prescribing the more concentrated glucose or three-in-one system. For patients who must be severely fluid restricted, these two systems may be modified by substituting 70% glucose and 10–15% amino acids for the 50 and 8.5% preparations, respectively, in order to supply equivalent nutrient content in a smaller volume. The recently available 30% fat emulsion, providing 3 $\text{kcal}\cdot\text{mL}^{-1}$, may prove useful in designing additional TPN regimens for fluid-restricted patients.

Acute Myocardial Ischemia

In some studies, lipid infusions have been associated with elevated circulating free fatty acid levels. The effect of the latter on patients with acute myocardial ischemia is controversial, but there is evidence that arrhythmias may be precipitated and the area of ischemic damage may be extended in patients with acute myocardial infarctions (46–48). In view of these data, the glucose system is recommended in this group of patients.

Glucose Intolerance

It appears that hyperglycemia due to stress or diabetes mellitus is more easily managed if less glucose is infused, as in the three-in-one and lipid systems (49–51).

HYPERLIPIDEMIA

The lipid infusions are contraindicated in patients with conditions in which the metabolism of endogenous lipids is abnormal. Here the glucose system is prescribed.

Pulmonary Disease

In patients with pulmonary insufficiency it is particularly important that lipogeneis induced by excess glucose be avoided because it results in an increase in total CO_2

production, which may in turn lead to elevated PCO_2 values. In addition, significantly less CO_2 is produced during the metabolism of isocaloric amounts of lipid compared to glucose. Thus, increasing the proportion of lipid calories in the nutrient solution, as in the three-in-one and lipid systems, may facilitate the clinical management of patients with chronic pulmonary insufficiency and hypercarbia (52–54). In contrast, impaired pulmonary function has been observed when patients with acute respiratory distress syndrome receive lipids infusions. The adverse effects reported include decreased PO_2 and compliance and increased pulmonary vascular resistance (55).

HOME PARENTERAL NUTRITION

Methods of TPN have become sufficiently standardized and simplified that such care can now be safely and effectively provided at home on an ambulatory basis. Candidates for such homecare include those in whom the acute underlying medical condition requiring initial hospitalization has resolved, but who still require intravenous nutrition for a prolonged or indefinite period or even permanently. Patients with anorexia nervosa, Crohn's disease, short bowel syndrome, or severe hyperemesis gravidarum are among those who have been successfully managed with ambulatory TPN. Other candidates for home therapy include cancer patients with anorexia associated with chemotherapy or radiation therapy and patients with controlled enterocutaneous fistulas, radiation enteritis, or partial intestinal obstruction.

While the general principles of TPN outlined previously are applicable here, there are certain specific considerations in homecare necessary to make this method safe, convenient, and practical. Home patients should receive their nutrient solution through a tunneled, cuffed, silicone rubber or polyurethane central venous catheter (Hickman-type catheter). Such catheters are of low thrombogenicity, and passing the cuffed catheter through a subcutaneous tunnel reduces the incidence of ascending infection. Central placement, usually through the subclavian vein or internal jugular vein, frees the patient's extremities from any apparatus.

Whereas inpatient TPN is infused around the clock, home TPN is often infused in cyclic fashion, usually during sleeping hours, so that patients may be free of the infusion apparatus for part of the day. Patients must adapt to the more rapid hourly rates of infusion necessary to provide the required volume in a shorter period. Alterations of blood sugar, the commonest acute metabolic abnormalities, are best prevented by gradually increasing the rate of delivery over 1–2 h at the beginning of therapy and tapering the rate over several hours at the conclusion of the daily treatment.

Finally, chronic TPN for months or years appears to be unmasking requirements for additional nutrients that are stored in significant quantities or that are required in minute amounts. For example, further investigation may indicate requirements for selenium, molybdenum, taurine, and probably other micronutrients.

BIBLIOGRAPHY

1. Vinnars E, Wilmore D. Jonathan Roads Symposium Papers. History of parenteral nutrition. JPEN J Parenter Enteral Nutr 2003;27:225–231.
2. Dudrick SJ. Early developments and clinical applications of total parenteral nutrition. JPEN J Parenter Enteral Nutr 2003;27:291–299.
3. Dudrick SJ, Wilmore DW, Vars HM, Rhoads JE. Long-term total parenteral nutrition with growth, development, and positive nitrogen balance. Surgery 1968;64:134–142.
4. Rhoads JE. The history of nutrition. In: Ballinger WF, Collins JA, Drucker WR, Dudrick SJ, Zeppa R, editors. Manual of Surgical Nutrition. Philadelphia: W. B. Saunders; 1975. pp. 1–9.
5. Silberman H. Nutritional requirements. In: Silberman H, editor. Parenteral and Enteral Nutrition. 2nd ed. Norwalk, (CT): Appleton & Lange; 1989. pp. 85–116.
6. Wilmore DW. Energy requirements for maximum nitrogen retention. In: Greene HL, Holliday MA, Munro HN, editors. Clinical Nutrition Update: Amino Acids. Chicago: American Medical Association; 1977. pp. 47–57.
7. Calloway DH, Spector H. Nitrogen balance as related to caloric and protein intake in active young men. Am J Clin Nutr 1954;2:405–412.
8. Rudman D, Millikan WJ, Richardson TJ, Bixler TJ, II, Stackhouse J, McGarrity WC. Elemental balances during intravenous hyperalimentation of underweight adult subjects. J Clin Invest 1975;55:94–104.
9. Guidelines for essential trace element preparations for parenteral use: A statement by an expert panel, AMA Department of Foods and Nutrition. JAMA 1979;241:2051–2054.
10. Shils ME.. Minerals in total parenteral nutrition. Proceedings of the AMA Symposium on Total Parenteral Nutrition Nashville (TN): January 17–19 1972. pp. 92–114.
11. Mirtallo J, Canada T, Johnson D, Kumpf V, Petersen C, Sacks G, Seres D, Guenter P. Safe practices for parenteral nutrition. JPEN J Parenter Enteral Nutr 2004;28:S39–S70.
12. Mirtallo J. Parenteral formulas. In: Rombeau JL, Rolandelli RH, editors. Clinical Nutrition: Parenteral Nutrition. 3rd ed. Philadelphia: W.B. Saunders Company; 2001. pp. 118–139.
13. Bern M. Observations on possible effects of daily vitamin K replacement, especially upon warfarin therapy. JPEN J Parenter Enteral Nutr 2004;28:388–398.
14. Helphingstine CJ, Bistrian BR. New Food and Drug Administration requirements for inclusion of vitamin K in adult parenteral multivitamins. JPEN J Parenter Enteral Nutr 2003;27:220–224.
15. Fed Reg 2000;65:21200–212010.
16. Silberman H. Parenteral and Enteral Nutrition. 2nd ed. Norwalk, (CT): Appleton & Lange; 1989.
17. Hallberg D. Therapy with fat emulsion. Acta Anaesthesiol Scand Suppl 1974;55:131–136.
18. McNiff BL. Clinical use of 10% soybean oil emulsion. Am J Hosp Pharm 1977;34:1080–1086.
19. Hallberg D. Studies on the elimination of exogenous lipids from the blood stream. The effect of fasting and surgical trauma in man on the elimination rate of a fat emulsion injected intravenously. Acta Physiol Scand 1965;65:153–163.
20. Blanchard R, Gillespie D. Some comparisons between fat emulsion and glucose for parenteral nutrition in adults at the Winnipeg Health Sciences Center. In: Meng H, Wilmore D, editors. Fat Emulsions in Parenteral Nutrition. Chicago: American Medical Association; 1976. pp. 63–64.

21. Hadfield J. High calorie intravenous feeding in surgical patients. Clin Med 1966;73:25–30.
22. Jeejeebhoy KN, Anderson GH, Nakhooda AF, Greenberg GR, Sanderson I, Marliss EB. Metabolic studies in total parenteral nutrition with lipid in man. Comparison with glucose. J Clin Invest 1976;57:125–136.
23. Silberman H, Freehauf M, Fong G, Rosenblatt N. Parenteral nutrition with lipids. JAMA 1977;238:1380–1382.
24. Hansen LM, Hardie BS, Hidalgo J. Fat emulsion for intravenous administration: clinical experience with intralipid 10%. Ann Surg 1976;184:80–88.
25. van den Berghe G, Wouters P, Weekers F, Verwaest C, Bruyninckx F, Schetz M, Vlasselaers D, Ferdinande P, Lauwers P, Bouillon R. Intensive insulin therapy in the critically ill patients. N Engl J Med 2001;345:1359–1367.
26. Eisenberg D, Schmidt B, Silberman H. Safety and efficacy of lipid-based TPN: I Effects of 20% fat emulsion on serum lipids and respiratory functions. JPEN J Parenter Enteral Nutr 1982;6: 586.
27. Fuhrman MP. Complication management in parenteral nutrition. In: Matarese LE, Gottschlich MM, editors. Contemporary Nutrition Support Practice: A Clinical Guide. 2nd ed. Philadelphia: W.B. Saunders; 2003.
28. Klein GL, Targoff CM, Ament ME, Sherrard DJ, Bluestone R, Young JH, Norman AW, Coburn JW. Bone disease associated with total parenteral nutrition. Lancet 1980;2:1041–1044.
29. Shike M, Harrison JE, Sturtridge WC, Tam CS, Bobechko PE, Jones G, Murray TM, Jeejeebhoy KN. Metabolic bone disease in patients receiving long-term total parenteral nutrition. Ann Intern Med 1980;92:343–350.
30. Buchman AL, Moukarzel A. Metabolic bone disease associated with total parenteral nutrition. Clin Nutr 2000; 19:217–231.
31. Hamilton C, Seidner DL. Metabolic bone disease and parenteral nutrition. Curr Gastroenterol Rep 2004;6:335–341.
32. Kotler DP, Levine GM. Reversible gastric and pancreatic hyposecretion after long-term total parenteral nutrition. N Engl J Med 1979;300:241–242.
33. Magnotti LJ, Deitch EA. Mechanisms and significance of gut barrier function and failure. In: Rolandelli RH, Bankhead R, Boullata JI, Compher CW, editors. Clinical Nutrition: Enteral and Tube Feeding. 4th ed. Philadelphia: Elsevier-Saunders; 2005. pp. 23–31.
34. Tilson MD. Pathophysiology and treatment of short bowel syndrome. Surg Clin North Am 1980;60:1273–1284.
35. Grundfest S, Steiger E, Selinkoff P, Fletcher J. The effect of intravenous fat emulsions in patients with pancreatic fistula. JPEN J Parenter Enteral Nutr 1980;4:27–31.
36. Shattuck KE, Klein GL. Hepatobiliary complications of parenteral nutrition. In: Rombeau JL, Rolandelli RH, editors. Clinical Nutrition: Parenteral Nutrition. 3rd ed. Philadelphia: W.B. Saunders; 2001. pp. 140–156.
37. Wagner WH, Lowry AC, Silberman H. Similar liver function abnormalities occur in patients receiving glucose-based and lipid-based parenteral nutrition. Am J Gastroenterol 1983;78:199–202.
38. Kwan V, George J. Liver disease due to parenteral and enteral nutrition. Clin Liver Dis 2004;8:ix–x, 893–913.
39. Silk DBA. Nutritional Support in Hospital Practice. Oxford: Blackwell Scientific Publications; 1983.
40. Krawinkel MB. Parenteral nutrition-associated cholestasis—what do we know, what can we do? Eur J Pediatr Surg 2004; 14:230–234.
41. Askanazi J, Rosenbaum SH, Hyman AI, Silverberg PA, Milic-Emili J, Kinney JM. Respiratory changes induced by the large glucose loads of total parenteral nutrition. JAMA 1980;243: 1444–1447.
42. Gramlich L, Kichian K, Pinilla J, Rodych NJ, Dhaliwal R, Heyland DK. Does enteral nutrition compared to parenteral nutrition result in better outcomes in critically ill adult patients? A systematic review of the literature. Nutrition 2004;20:843–848.
43. de Chalain TM, Michell WL, O'Keefe SJ, Ogden JM. The effect of fuel source on amino acid metabolism in critically ill patients. J Surg Res 1992;52:167–176.
44. Druml W, Fischer M, Ratheiser K. Use of intravenous lipids in critically ill patients with sepsis without and with hepatic failure. JPEN J Parenter Enteral Nutr 1998;22:217–223.
45. Garcia-de-Lorenzo A, Lopez-Martinez J, Planas M, Chacon P, Montejo JC, Bonet A, Ortiz-Leyba C, Sanchez-Segura JM, Ordonez J, Acosta J, Grau T, Jimenez FJ. Safety and metabolic tolerance of a concentrated long-chain triglyceride lipid emulsion in critically ill septic and trauma patients. JPEN J Parenter Enteral Nutr 2003;27:208–215.
46. Editorial: Free fatty acids and arrhythmias after acute myocardial infarction. Lancet 1975;1:313–314.
47. Jones JW, Tibbs D, McDonald LK, Lowe RF, Hewitt RL. 10% Soybean oil emulsion as a myocardial energy substrate after ischemic arrest. Surg Forum 1977;28:284–285.
48. Opie LH, Tansey M, Kennelly BM. Proposed metabolic vicious circle in patients with large myocardial infarcts and high plasma-free-fatty-acid concentrations. Lancet 1977;2:890–892.
49. Baker JP, Detsky AS, Stewart S, Whitwell J, Marliss EB, Jeejeebhoy KN. Randomized trial of total parenteral nutrition in critically ill patients: metabolic effects of varying glucose-lipid ratios as the energy source. Gastroenterology 1984;87: 53–59.
50. Meguid MM, Schimmel E, Johnson WC, Meguid V, Lowell BC, Bourinski J, Nabseth DC. Reduced metabolic complications in total parenteral nutrition: pilot study using fat to replace one-third of glucose calories. JPEN J Parenter Enteral Nutr 1982;6:304–307.
51. Watanabe Y, Sato M, Abe Y, Nakata Y, Lee T, Kimura S. Fat emulsions as an ideal nonprotein energy source under surgical stress for diabetic patients. Nutrition 1995;11:734–738.
52. Silberman H, Silberman AW. Parenteral nutrition, biochemistry and respiratory gas exchange. JPEN J Parenter Enteral Nutr 1986;10:151–154.
53. Askanazi J, Nordenstrom J, Rosenbaum SH, Elwyn DH, Hyman AI, Carpentier YA, Kinney JM. Nutrition for the patient with respiratory failure: glucose vs. fat. Anesthesiology 1981;54:373–377.
54. Sherman SM. Parenteral nutrition and cardiopulmonary disease. In: Rombeau JL, Rolandelli RH, editors. Clinical Nutrition: Parenteral Nutrition. 3rd ed. Philadelphia: W.B. Saunders; 2001. pp. 335–352.
55. Lekka ME, Liokatis S, Nathanail C, Galani V, Nakos G. The impact of intravenous fat emulsion administration in acute lung injury. Am J Respir Crit Care Med 2004;169:638–644.

See also DRUG INFUSION SYSTEMS; GLUCOSE SENSORS; HOME HEALTH CARE DEVICES.

NYSTAGMOGRAPHY. See OCULAR MOTILITY RECORDING AND NYSTAGMUS.

O

OCULAR FUNDUS REFLECTOMETRY

Amol D. Kulkarni
Amruta M. Dattawadkar
University of Wisconsin,
Madison, Wisconsin

INTRODUCTION

Ophthalmology involves a study of diagnosis and management of eye diseases. The retina, also called the fundus oculi, constitutes a major component of the posterior segment of the eye. The central area of the fundus that is responsible for vision is called macula. The macula has a large number of photoreceptors that are specialized neurons containing colored light sensitive dyes (visual pigments). The center of macula is known as the fovea. The absorption of light initiates a cascade of events that bleaches these visual pigments and generates an electrochemical signal that is responsible for vision. This sequence of events is known as the visual cycle. Ocular fundus reflectometry is a noninvasive technique for an *in vivo* study of the visual cycle (1). It provides an objective and quantitative assessment of the kinetics of visual pigments.

Fundus reflectometry involves measuring the intensity of light of different wavelengths reflected by the ocular fundus (1). It has been primarily used to study reflectance properties of various structures in fundus, but also can be used for the study of oximetry and blood flow (2,3). However, the interpretation of measurement shows a lot of variation due to the effect of different types of photoreceptors with its own type of pigment, and spectral distortions. It is used in practice to characterize eye disorders with abnormalities in visual pigments, to detect autofluorescence of retinal lesions, and for various research studies in animals and humans (1).

HISTORICAL ASPECT

In 1851, a qualitative method for observation of the light reflected at the fundus was developed by Helmhotz. In 1952, the absorption spectrum of macular pigment was measured by Brindley and Willmer (4,5). In 1954–1971, the density and spectral properties of human rhodopsin was established by Rushton (6,7). He further developed a densitometer in 1971. A spectrophotographic technique that projects the entire spectrum of light was developed by Weale in 1953. This principle was used by Weale (8,9) to measure the density of cone pigments.

PHYSIOLOGIC BASIS AND PRINCIPLES

There are two types of photoreceptors in the human retina, namely, the rods and the cones. The rods are concerned with scotopic vision (dim lighting conditions) and the cones are responsible for photopic (daytime vision) and color vision. The cones are present in large numbers in the macula. Ultrastructure of both the types of photoreceptors as studied by electronmicroscopy consists of an outer and inner segment. The outer segments are made of stack of disks containing the visual pigment. The inner segment is responsible for pigment production and regeneration of outer segments. The pigment in rods is an aldehyde of vitamin A and in combination with protein (opsin) forms a compound known as rhodopsin. Of the various isomeric forms, the 11-cis form is a vital component of visual cycle and is converted to all-trans state on absorption of a photon (8). This sequence of events is responsible for vision and a similar process occurs in the cones (9).

The ability of visual pigments to absorb certain wavelength of the projected light can be determined by spectrophotographic techniques. This forms the basis of fundus reflectometry. In *in vitro* conditions, a monochromatic light can be projected on a sample of pigment and the intensity of the emergent beam is measured. It is possible to deduce the absorbing effect of the pigment by again projecting the light without the pigment. By using various wavelengths, an absorbance spectrum can be calculated. However, this is not possible *in vivo*, and hence an alternative technique has been devised. It consists of measuring the intensity of the emergent beam before and after bleaching of the pigment *in vivo*. This constitutes the physiologic basis of fundus reflectometry (1).

METHODS OF FUNDUS REFLECTOMETRY

The reflectometers can be classified as either spectral or imaging reflectometers. The various types of spectral reflectometers are Utrecht, Boston 1, Jena, Boston 2, and Utrecht 2^1. These instruments measure the absolute spectral reflectance. The Utrecht reflectometer measures the foveal reflectance and determines the absorption characteristics of the cone visual pigments. The Utrecht 2 is a newer device and measures cone-photoreceptor directionality along with foveal reflectometry. The Boston 1 reflectometer was devised for oximetry and could simultaneously measure the reflectance at six wavelengths between 400 and 800 nm. The Boston 2 consisted of a modified Zeiss fundus camera, which could assess the orientation of foveal photoreceptors, their directionality, and the ratio of directional to diffuse flux (10–13). The Jena was a combination of a xenon lamp with a monochromator, and measured reflectance by photon-counting techniques.

The fundus imaging systems can also be modified to measure reflectance. This is called as imaging densitometry and has inferior resolution as compared to spectral densitometry. The various techniques used include fundus camera, video-based systems, scanning laser ophthalmoscope, and a charge-coupled device (CCD) camera. The fundus camera and the video-based system generate maps of the visual pigment (14). However, there is an error in measuring reflectance due to stray light. Therefore a scanning laser ophthalmoscope (SLO) was developed in which a laser beam is moved in a raster pattern over the retina. In spite of high contrast and large dynamic range, the SLO did not provide a quantitative determination of fundus reflection. Around the same time, the CCD camera came into vogue and measured fundus reflectance spectra in 400–710 nm wavelength range.

The above-mentioned techniques of reflectometry have been used to develop models to quantify the spectral distribution of light pathways in human fundus. The various structures in the eye with reflectance properties include the cornea, lens, internal limiting membrane, nerve fiber layer, photoreceptors, retinal pigment epithelium, and sclera (15). On the contrary, the various structures that absorb light include lens, macular pigments, visual pigment, lipofuscin, melanin, and hemoglobin. Taking into consideration these various structures and their reflectance properties, numerous attempts have been made to study the visual cycle. These can be as simple as measuring light transmission by the retina to determining foveal fundus reflectance using spectral, directional (16–19) and bleaching effects.

CLINICAL APPLICATIONS OF FUNDUS REFLECTOMETRY

Fundus reflectometry is primarily used to estimate the optical density of various pigments in the eye (19,20). This includes the lens, macular, visual, and melanin pigments. In the lens and macular pigments, optical density measurement helps in determining the effects of aging. Visual pigments are vital component of the light cycle and densitometry can be used to classify photoreceptors on the basis of wavelength sensitive pigments. The extent of melanin pigmentation can be characterized by reflectometry and an index of pigmentation can be established (21).

Apart from measuring the pigment density, reflectometry is also used to study oxygen saturation, and orientation of foveal photoreceptors. These have applications in studying various congenital and acquired disorders of the retina including nutritional deficiency, infections, and degenerations (22–26). They can be used not only to characterize diseases, but also to study the effects of various treatment modalities (25). Moreover, sometimes it also contributes to early detection of particular diseases.

FUTURE DIRECTIONS

Fundus reflectometry has been principally used to measure the optical density of visual pigments and study the function of a normal and diseased retina. However, it is not used in routine patient care. The reason is because it is time consuming and requires complicated equipments. In addition the specificity is low, and so its use is limited only to research purposes. Therefore it has wide applications in epidemiologic studies, such as aging related ocular disorders (27,28).

Due to the ability to measure directionality and spatial distribution, fundus reflectometry is being tested in determining nerve fiber layer thickness, measurement of oxygen saturation, and to monitor the effects of laser therapy. As we move into a new era of prolonged longevity, due to advances in medicine, fundus reflectometry will be used to test new hypothesis and treatments.

BIBLIOGRAPHY

Cited References

1. Killlbride PE, Ripps H. Fundus reflectrometry. In: Martens BR, editor. Noninvasive Diagnostic Techniques in Ophthalmology. New York: Spinger; 1990. pp 479–498.
2. Beach JM, et al. Oximetry of retinal vessels by dual-wavelength imaging: calibration and influence of pigmentation. J Appl Physiol 1999;86:748–758.
3. Delori FC. Noninvasive technique for oximetry of blood in retinal vessels. Appl Opt 1998;27:1113–1125.
4. Killlbride PE, Alexander KR, Fishman GA. Human macular pigment assessed by imaging fundus reflectrometry. Vision Res 1989;29:663–674.
5. Chen SF, Chang Wu JC. The spatial distribution of macular pigment in humans. Curr Eye Res 2001;23:422–434.
6. Rushton WAH, Campbell FW. Measurement of rhodopsin in the living human eye. Nature (London) 1954;174:1096–1097.
7. Rushton WAH. Physical measurement of cone pigment in the living human eye. Nature (London) 1957;179:571–573.
8. Weale RA. Observations on photochemical reactions in living eyes. Br J Ophthalmol 1957;41:461–474.
9. Weale RA. Photochemical reactions in the living cat's retina. J Physiol 1953;121:322–331.
10. Van Blokland GJ. Directionality and alignment of the foveal receptors, assessed with light scattered from the human fundus *in vivo*. Vision Res 1986;26:495–500.
11. Zagers NPA, van de Karrats J, Berendschot TTJM, van Norren D. Simultaneous measurement of foveal spectral reflectance and cone photoreceptor directionality. Appl Opt 2002;41:4686–4696.
12. De Lint PJ, Berendschot TTJM, van Norren D. A Comparison of the optical stiles-crawford effect and retinal densitometry in a clinical setting. Invest Opthalmol Vis Sci 1998;39:1519–1523.
13. Macros S, Burns SA, He JC. Model for cone directionality reflectrometric measurements based on scattering. J Opt Soc Am A 1998;15:2012–2022.
14. Delori FC, Gragoudas ES, Francisco R, Pruett RC. Monochromatic ophthalmoscopy and fundus photography. The normal fundus. Arch Ophthalmol 1977;95:861–868.
15. Hammer M, Roggan A, Schweitzer D, Muller G. Optical properties of ocular fundus tissues -an in vitro study using the double-integrating-sphere technique and inverse Monte Carlo simulation. Phys Med Biol 1995;40:963–978.
16. Delori FC, Burns SA. Fundus reflectance and the measurement of crystalline lens density. J Opt Soc Am A 1996;13: 215–226.
17. Delori FC, Pflibsen KP. Spectral reflectance of the human ocular fundus. Appl Opt 1989;28:1061–1077.
18. Gorrand JM, Delori FC. A reflectometric technique for assessing photoreceptor alignment. Vision Res 1995;35:999–1010.
19. Wooten BR, Hammond BR Jr, Land RI, Snodderly DM. A practical method of measuring macular pigment optical density. Invest Ophthalmol Vis Sci 1999;40:2481–2489.
20. Savage GL, Johnson CA, Howard DL. A comparison of noninvasive objective and subjective measurement of the optical

density of human ocular media. Optom Vis Sci 2001;78:386–395.
21. Hunold W, Malessa P. Spectrophotometric determination of melanin pigmentation of the human fundus oculi. Ophthalmic Res 1974;6:355–362.
22. Highman VN, Weale RA. Rhodopsin density and visual threshold in retinitis pigmentosa. Am J Ophthalmol 1973; 75:822–832.
23. Carr RE, Ripps H, Siegel IM, Weale RA. Rhodopsin and the electrical activity of the retina in congenital night blindness. Invest Ophthalmol Vis Sci 1966;5:497–507.
24. Liem AT, Keunen JE, van Norren D. Clinical applications of fundus reflection densitometry. Surv Ophthalmol 1996;41: 37–50.
25. Augsten R, Konigsdorffer E, Schweitzer D, Strobel J. Nonproliferative diabetic retinopathy-reflection spectra of the macula before and after laser photocoagulation. Ophthalmologica 1998;212:105–111.
26. Landrum JT, Bone RA, Kilburn MD. The macular pigment: a possible role in protection from age-related macular degeneration. Adv Pharmacol 1997;38:537–556.
27. Delori FC, Goger DG, Dorey CK. Age-related accumulation and spatial distribution of lipofuscin in rpe of normal subjects. Invest Ophthalmol Vis Sci 2001;42:1855–1866.
28. Berendschot TTJM, et al. Influence of lutein supplementation on macular pigment, assessed with two objective techniques. Invest Ophthalmol Vis Sci 2000;41:3322–3326.

OCULAR MOTILITY RECORDING AND NYSTAGMUS

LOUIS F. DELL'OSSO
Case Western Reserve
University Cleveland, Ohio
L. A. ABEL
University of Melbourne
Melbourne, Australia

INTRODUCTION

This chapter will discuss the different types of eye movements generated by the ocular motor system, the advantages and disadvantages of commonly used recording systems, the requirements for accurate calibration of those systems, and the use of eye-movement recordings in research.

What Can We Record and Why? A Brief Introduction to Types of Eye Movements and Why We Record Them

Humans are highly visually driven animals. Our hearing may be inferior to that of the owl and our sense of smell far poorer than a dog's, but our visual acuity is excelled by few other species. High resolution vision, however, creates a bandwidth problem—if we processed our entire visual field simultaneously at maximal resolution, we would need so many optic nerve fibers to carry visual information back to the brain that our eyes might not fit into our heads. The solution that has evolved is to make the resolution of the retina—the light-sensitive neural layer of the eye—inhomogeneous. Visual acuity in the central 1° of the visual field is maximal, but it falls off rapidly as one moves toward the periphery. What keeps us from ever being aware of this fact is the nearly incessant motion of our eyes, which use a number of interconnected control systems to direct our gaze to an object of interest and to keep it fixated in the face of target and body movement. Considerable processing in the visual areas of the brain is needed to integrate the discontinuous flow of visual images into the clear, stable perception of the world that we usually experience.

EYE MOVEMENTS

The types of eye movements to be discussed here all play a part in the maintenance of vision. There are only 6 muscles per eye, arranged in opposing pairs and moving in a relatively constrained way by virtue of the anatomy of the orbit. Although each type of eye movement serves a specific purpose and is generated by partially distinct brain mechanisms, they nonetheless interact in the course of normal life. Examination and recording of eye movements has a surprisingly long history, going back to the pioneering work of Dodge and Cline (1). Eye-movement recording has enjoyed a number of advantages over the analysis of other motor control mechanisms. The following sections will briefly describe each type of eye movement, what purpose it serves, and why one might wish to record it.

Version and Vergence

The ocular motor system may be divided into two major subsystems: one that controls version (conjugate or conjunctive) eye movements, and one that controls vergence (disconjugate or disjunctive) eye movements. Saccades, pursuit, vestibuloocular, and optokinetic eye movements are types of version movements, and convergent and divergent refixations and pursuits are types of vergence eye movements. Patients may exhibit eye-movement abnormalities stemming from disorders in the version or vergence subsystem and both nystagmus and saccadic intrusions may be disconjugate, even uniocular. Recording systems used for all eye movements should be capable of independently recording data from both eyes, regardless of whether they are presumed to be conjugate, which is especially important when recording patients but is also applicable to normal individuals because conjugacy is not absolute. It is a common misconception that one can record "conjugate" movements from one eye only and presume the other eye is moving in exactly the same manner. In this chapter, only methods that fulfill this requirement are considered, regardless of either the experimental paradigm (version or vergence) or the subject population (normal or patient).

Saccades

Saccades are the fastest eye movements made, with velocities at times approaching 1000°/s. We make them nearly incessantly during our waking hours and during rapid eye movement sleep. Although at the end of each saccade only the most central area around the fixation location is seen with maximal acuity, our brains are able to integrate the rapidly acquired series of such images into a single, unified perception of the world. Saccades may be horizontal, vertical, or oblique, which has implications for their recording, as will be discussed below. Evaluation of saccades may be grouped broadly into assessment of the saccades themselves and analysis of where they go as an individual views a scene or an image. Some eye trackers are more suitable for one sort of study than another. In particular, some methods are poorly suited to vertical and completely unsuited to torsional eye movements, where as others may have insufficient temporal resolution for assessment of latency or accuracy but excel at mapping sequences of fixations in two dimensions. In this discussion, a somewhat arbitrary distinction will be drawn between the detailed evaluation of individual saccades (as is often done clinically) and the assessment of scanpaths (as is sometimes used in a clinical setting but more often used in studies of man-machine interaction).

Inherent Saccadic Characteristics.

Accuracy. Saccadic accuracy is usually expressed in terms of gain (eye position/target position). Most commonly, if a refixation were comprised of multiple steps toward the target, the gain would be based only on the first step. Gain may be either abnormally high or abnormally low in different neurological conditions.

Latency. Latency is the time between stimulus onset and onset of eye movement. In humans, latency may range from 80 to several hundred ms, depending on the task and the age and health of the patient. Normally, saccades made in anticipation of target motion are excluded, unless stimuli with predictable location and timing are used. To be measured accurately, data must be acquired at a rate permitting the precise resolution of saccade timing (e.g., 500 Hz). Thus, a 25 or 30 Hz video-based system would be useless for this application.

Peak Velocity. Peak velocity can be measured either using analog electronics or, more commonly now, by off-line differentiation using software. Peak velocity is affected by fatigue, sedating drugs, and diseases that affect the cells in the brainstem that generate the fast, phasic component of saccadic innervation. Again, very low frame rates will make accurate calculation of peak velocity impossible, as it would not be possible to measure the rate of change in eye position. Indeed, if the sampling rate is too low, small saccades may be lost altogether, as they could be completed between samples (or video frames).

Scanpaths. Scanpaths can be divided into the descriptions of how individuals view a scene and nystagmus scanpaths that describe the eye trajectories about a fixation point in an individual with nystagmus. The former contain refixation saccades and periods of fixation whereas the latter contain the oscillatory nystagmus movements, braking, and foveating saccades, plus intervals of relatively stable fixation, if present.

Clinical Applications. Demonstration of how individuals (including patients) view a scene is probably the most familiar application of eye movement recording. In these applications, the “fine structure” of each saccade is of less interest than knowledge of where the saccades take the eyes and in what sequence. The classic work of Yarbus demonstrated the stereotyped way in which individuals view faces (2). As these investigations are focused on how cognitive processes control gaze, such work can be used to examine how patients with Alzheimer’s disease (3) examine a novel scene or how individuals with schizophrenia attempt to judge the emotions expressed in a face. For scanpath analyses, high temporal resolution is unnecessary and spatial resolution on the order of a degree, not minute of arc, is acceptable. A wide linear range for vertical and horizontal eye movements is essential, however. Unobtrusiveness and minimal obstruction of the visual field are highly desirable when behavior is to be interfered with as little as possible.

Commercial Applications (Usability Studies, Man-Machine Interactions). Commercial applications are probably one of the most rapidly growing areas of eye movement research; it involves evaluating how humans interact with human displays. Here, the goal may be to see how a web page is examined or where a pilot is looking in a cockpit. The technical requirements for the eye tracker are essentially the same as for clinical applications. An exception is the area of gaze-contingent displays, where the endpoint of a saccade is predicted from recording its beginning, and the display is updated in high resolution only at that point. Such applications impose stricter temporal and spatial resolution criteria.

Nystagmus Scanpaths. Plots of the horizontal vs. vertical motion of nystagmus patients’ eye movements during fixation of a stationary target provide insight into their ability to foveate the target in a stable (i.e., low retinal-slip velocity) and repeatable (i.e., low variance in the mean positions of target foveation intervals) manner. Nystagmus phase-plane (eye position vs. eye velocity) and scanpath plots were developed to study the foveation periods present in many of the waveforms seen in infantile nystagmus (4–8). They are important methods that provide insight into how individuals with such oscillations can achieve high visual acuity. The recording equipment for nystagmus scanpaths and phase-planes needs to be both accurate and of sufficient bandwidth to record the small saccades imbedded in nystagmus waveforms.

Smooth Pursuit. A correlate of having only a small part of the retina—the fovea—with high spatial resolution is that if a moving object is to be seen clearly, it must be tracked precisely, so that its image remains on the fovea, which is the function of the smooth pursuit system. The brain substrates underlying smooth pursuit are, to a degree, separable from those of the saccadic system, but, as a recent review has noted (9), a high degree of parallelism exists. Given that the two systems must work together for successful tracking, this fact is not surprising. For example, if you hear a bird call in the sky and decide to follow it, you must first locate it with a saccade (and possibly a head movement). Your pursuit system then keeps your gaze on the bird, but if it moves too swiftly for this system, it can be reacquired by a saccade and tracking can the recommence. If it is lost again, the pattern repeats. Indeed, if pursuit gain (eye velocity/target velocity) is low or even zero, objects can still be tracked by repeated saccades, giving rise to the clinical observation of "cogwheel pursuit."

In contrast to the many roles that saccades serve, pursuit eye movements are rather specialized for tracking. We all can generate saccades at will, even in the absence of targets, but voluntary generation of smooth pursuit is extremely rare and of poor quality. When recorded, it may be examined qualitatively for the presence of saccades or the smooth tracking segments can be separated out and their gain analyzed. As a result of the bilateral organization of motor control in the brain, it is possible to have a unidirectional pursuit abnormality, which may be of diagnostic value. However, bilaterally reduced smooth pursuit is nonspecific, as it may result from boredom, inattention, alcohol, fatigue, as well as pathology. As the pursuit system cannot track targets moving at greater than approximately 2 Hz, the requirements for its recording are not very demanding. If pursuit velocity is to be derived, however, then a low-noise system with appropriate low-pass filtering is essential to prevent the velocity signal from being swamped by noise. A low-noise system is also crucial in computer analysis of smooth pursuit because the algorithm used to identify saccades must ensure that none of the saccade is included in the data segment being analyzed as pursuit. If the pursuit component of the eye movement is only $5^\circ/\mathrm{s}$ and portions of saccades with velocities $\leq 30^\circ/\mathrm{s}$ are included, pursuit gain calculations may be highly inaccurate, which is a concern when commercial systems incorporating proprietary algorithms are being used in clinical settings where this possibility has not been anticipated. See Calibration (below) for more information.

Vestibulo-Ocular Response (VOR)

The VOR is a fast reflex whose purpose is to negate the effects of head or body movement on gaze direction. Acceleration sensors in the semicircular canals provide a head-velocity input to the ocular motor system that is used to generate an eye-velocity signal in the opposite direction. The sum of head and eye velocity cancel to maintain steady gaze in space. The VOR is tuned to negate fast head movements and works in concert with the optokinetic reflex (see below), which responds to lower frequency background motion.

Rotational Testing. For vision to be maximally effective, it must continue to work properly as humans move around in the environment. Consider what would happen if the eyes were fixed in the head as one walked about—the image falling on the retina would oscillate with every step. Every turn of the head would cause the point of regard to sweep away from the fovea. Relying on visual input to compensate would be far too slow to generate an accurate compensatory input. Therefore, humans possess the semicircular canals, three approximately (but not precisely) orthogonal transducers of rotational motion, as part of each inner ear. Only three neurons separate the canals from the extraocular muscles that move the eyes. The canals are filled with fluid and, as the head moves, the inertia of the fluid causes it to lag behind, stimulating displacement-sensitive hair cells at the base of each canal. With only two synapses between sensory transducer and motor effector, the core of the VOR pathway can act very rapidly. Note, however, that constant velocity rotation elicits a signal that eventually decays to baseline, as the fluid eventually ceases to lag behind the canals (i.e., it moves with the same rotational velocity as the canals). Of course, prolonged constant velocity rotations are not part of our evolutionary history and are rarely encountered in daily life.

As the function of the VOR is to facilitate the maintenance of stable gaze as we move around in the environment, it makes intuitive sense to assess it in a moving subject. The most common way to make this assessment is to measure the horizontal component of the VOR as the patient is rotated while in the seated position. Spring-loaded Barany chairs were eventually superseded by electrically driven chairs, which could be driven with velocity steps, sinusoidally, or with more complex inputs. Step inputs may be used to quantify the time constant of decay of the VOR, whereas the other inputs can be used to generate gain and phase plots. Directional asymmetries or abnormal gains can be readily detected with such testing. Such tests are also carried out not only under baseline conditions, with the patient in complete darkness, but also with the VOR suppressed (patients fixate a target rotating with them) or enhanced (patients fixate an earth-fixed target).

Rotary chair testing has several shortcomings, particularly for low frequencies (e.g., 0.05 Hz). It takes a long time to obtain several cycles of data, during which time the patient may be lulled to sleep by the slow rotation in the dark. Alerting tasks (e.g., mental arithmetic) can be used to overcome this shortcoming, but the overall testing time may be quite long. Stimuli such as pseudo-random binary sequences have been used, with data analyzed by cross-correlation (10) in order to obtain results across a wide range of frequencies more rapidly. Another limitation, however, is that in order to obtain VOR data with a chair, the entire patient must be rotated, which limits the frequency range of the technique, because rotations of, for example, a 100 kg individual at 2 Hz would require very high forces. In addition, high frequency rotations increase the likelihood that because of inertia, the patient would not rotate precisely in phase or with the same amplitude as the chair. Systems are available that record eye movement and sense head movement during patient-initiated head shaking, which allows for testing at more physiological frequencies, but it requires a cooperative patient.

A fundamental problem with rotary chair testing is that although directional differences can be detected, localizing pathology to one ear is difficult. Obviously, rotating only one side of the head is impossible, and the "push-pull" nature of the vestibular system (due to the juxtaposed semicircular canals in the ears) makes lateralization difficult. For this reason, the next test remains valuable, in spite of its shortcomings.

Caloric Testing. Introduced by Barany in 1903, caloric testing is probably the most widely used of all vestibular tests. When carried out using EOG, it is still often referred to as "electronystagmography" (ENG), a term that is sometimes mistakenly applied to all forms of eye-movement recording. It involves the irrigation of the ear canal with either warm water or cold water, which alters the behavior of the horizontal semicircular canal on the side being irrigated. Cold water simulates reduced ipsilateral activity and warm water simulates an irritative lesion; the temperature of the water thus determines the direction of the resulting induced nystagmus fast phase in the way described by the acronym COWS: cold, opposite; warm, same. Vestibular nystagmus frequency and amplitude can readily be assessed for each ear at various temperature levels, which remains the only practical way to assess each side of the vestibular system independently. However, caloric stimulation has the appreciable shortcoming that it is a dc input to the vestibular system. It thus assesses the function of the system far from its physiological frequency range of several Hertz.

Optokinetic Response (OKR)

The OKR is a slow reflex whose purpose is to negate the effects of retinal image movement on gaze direction. Velocity sensors in the retina provide an input to the ocular motor system that is used to generate an eye-velocity signal in the same direction, maintaining gaze on the moving background. The OKR is tuned to respond to slow retinal image movement and works in concert with the VOR (see above), which responds to high frequency head motion.

Full-Field. The optokinetic nystagmus (OKN) response, like the VOR, may be induced in healthy individuals with appropriate visual stimuli. The fundamental form of the optokinetic response is induced by motion of all (or most) of the visual field, which elicits a slow eye movement in the direction of the stimulus, with a fast phase bringing the eyes back toward their initial position. This response continues as long as the stimulus continues. If one considers how the VOR decays with continuous motion and has low gain at very low frequencies, then it can be seen that the OKR and the VOR are functionally additive. Indeed, the relationship between OKR and VOR can be readily observed by anyone who has sat gazing out of a train window and felt himself moving, only to discover that it was the adjacent train which was pulling out of the station. The optokinetic stimulus evokes activity in the vestibular nuclei of the brain, and this activity elicits a sense of motion—the most common way to activate the vestibular system. This visually-induced motion percept is known as linearvection if the motion is linear and as circularvection if the stimulus is rotational. The nature of OKN differs depending on whether the stimulus is actively followed or passively viewed.

Small-Field (Hand-Held Drum, Tape). Although "train nystagmus" may be relatively easy to induce in the real world, presentation of a full-field OKN stimulus in a clinical setting requires a stimulus that essentially surrounds the patient. For this reason, OKN is more often tested using either a small patterned drum or a striped tape, both of which can be easily held in the examiner's hands. Although the OKN induced in this way looks no different than that deriving from a full-field stimulus, it is primarily a smooth pursuit response, whereas the full-field OKN includes both pursuit components as well as responses deriving from subcortical pathways, including the lateral geniculate body, accessory optic system, nucleus of the optic tract, and the brain stem and cerebellar circuitry governing eye movements.

Spontaneous Nystagmus & Saccadic Intrusions or Oscillations

Diagnostic Classification. In addition to induced nystagmus, some subjects exhibit either spontaneous or gaze-evoked nystagmus or saccadic intrusions or oscillations. The waveforms and other characteristics of these movements often have diagnostic value; accurate calibration of the data is necessary to extract diagnostic information or to deduce the mechanisms underlying the genesis of an intrusion or oscillation (nystagmus or saccadic).

Nystagmus Versus Saccadic Oscillations. The first distinction to be made when spontaneous oscillations are present is to distinguish between the many types of nystagmus and saccadic intrusions and oscillations. Both the slow-phase waveforms and their relationships to target foveation (placement of the target image on the small, high resolution portion of the retina) help in making this determination. Although the details of these determinations are beyond the scope of this chapter, the basic difference is that nystagmus is generated and sustained by the slow phases, whereas saccadic intrusions and oscillations are initiated by saccades that take the eyes off-target.

Congenital Versus Acquired Nystagmus. If nystagmus is present, determination of its origin is necessary (i.e., is it congenital or acquired?). Again, this field is complex and cannot be fully discussed here. Suffice is to say, certain nystagmus waveforms exist that are pathognomonic of congenital nystagmus; they, along with characteristic variations with gaze angle, convergence angle, or fixating eye, help to determine whether a nystagmus is congenital or acquired.

OCULAR MOTOR RECORDING SYSTEMS

Overview of Major Eye-Movement Recording Technologies

The following are descriptions of the more common technologies used to record the eye movements of both normals and patients. Technical descriptions, engineering, and physics of these and other methods may be found elsewhere in this volume (see "Eye Movement Measurement Techniques"). Emphasis in this chapter will be on the abilities of different types of systems and the calibration requirements to provide accurate eye-movement data in the basic and clinical research settings.

Electrooculography. ***Theory of Operation.*** Electrooculography (EOG) is the only eye-movement recording method that relies on a biopotential, in this case, the field potential generated between the inner retina and the pigment epithelium. This signal may approach 0.5 mV or more in amplitude. If two electrodes are placed on either side of, and two more above and below, the orbit (along with a reference electrode on the forehead or ear), then as the eye rotates in the orbit, a voltage proportional to the eye movement may be recorded, because one electrode becomes more positive and the other more negative with respect to the reference electrode. The technique is one of the oldest and most widespread and has been the standard for assessment of eye movements related to vestibular function. When the term ENG is seen, it is generally EOG that is used.

Characteristics. EOG has the considerable advantage that it requires only a high impedance, low noise instrumentation amplifier for its recording and that the voltage is linearly proportional to eye movement over most of its range. Such amplifiers are relatively inexpensive in comparison with many other eye-tracking technologies. As the electrodes are placed on the skin adjacent to the eye, no contact occurs with the eye itself and no obstruction of any part of the visual field exists. It also is unaffected by head motion, because the electrodes move with the head.

Applications. In theory, the EOG can be used anywhere eye movements are to be recorded. However, as the following section will show, it has a number of inherent limitations that practically eliminate it from many applications. Its widest use remains in the assessment of vestibular function and for the recording of caloric nystagmus and the vestibulo-ocular reflex. It is unsuited for use in environments with changing levels of illumination, as normal physiological processes will change the resting potential of the EOG and thus alter its relationship with amplitude of eye movement. EOG can be used in the assessment of saccades and smooth pursuit, but the low-pass filtering generally required will lead to artificially lowered saccade peak velocities. EOG has occasionally been used in scanpath studies, but its instability and fluctuating gain make it undesirable for this application, because if scenes differing in mean luminance are presented, the EOG will gradually change amplitude.

Limitations. Although conceptually simple and easy to implement, EOG has many shortcomings. One is that because the electrodes are placed on the surface of the facial skin, the EOG is not the only signal they detect. If the patient is nervous or clenches his or her teeth, the resulting electromyographic (EMG) activity in the facial muscles will be recorded as well, with the result that the signal actually recorded is the sum of the desired EOG and the unwanted EMG. As the spectra of the two signals overlap, no amount of filtering can completely separate them.

Another significant problem with EOG is the fact that, like many biopotential recordings, it is prone to drift. Some of this droft may reflect electrochemical changes at the electrode, causing a shift in baseline, which was particularly a problem when polarizable electrodes were used in the early days of the technique. Even nonpolarizable electrodes such as the commonly used Ag-AgCl button electrodes may still yield a varying baseline when first applied. Furthermore, the potential also shifts with changes in illumination. Indeed, assessment of this response to light is itself a clinical tool. This baseline variability can lead to the temptation to use an ac-coupled amplifier in the recording of the EOG, which has frequently been done, particularly in the ENG literature. Although not a problem if the only data required is nystagmus frequency, significant distortion occurs when ac-coupling is used to record saccades. The apparent drift back toward the center closely resembles a saccade whose tonic innervational component is inadequate. Noise and drift limit the resolution of EOG to eye movements of no less than 1°; this threshold may be even higher in a nervous patient or an elderly patient with slack, dry skin. An additional limitation undercuts the EOG's otherwise significant advantage in being able to record vertical eye movements, which is the overshooting seen on vertical saccades. It has long been suggested that the lids, moving somewhat independently of the globe, act as electrodes on the surface of the globe, conducting current in parallel to the other current path between globe and electrodes (11).

Another more practical drawback to the use of EOG when used for recording the movements of both eyes horizontally and vertically is that a total of nine electrodes are required (see Fig. 1). Each must be individually adhered to the patient and must be carefully aligned if spurious crosstalk between horizontal and vertical motion is to be avoided. Even if only horizontal motion is to be recorded, five accurately placed electrodes are still needed. A common but unfortunate clinical shortcut has been to use only three—two at either outer canthus of the eye and one

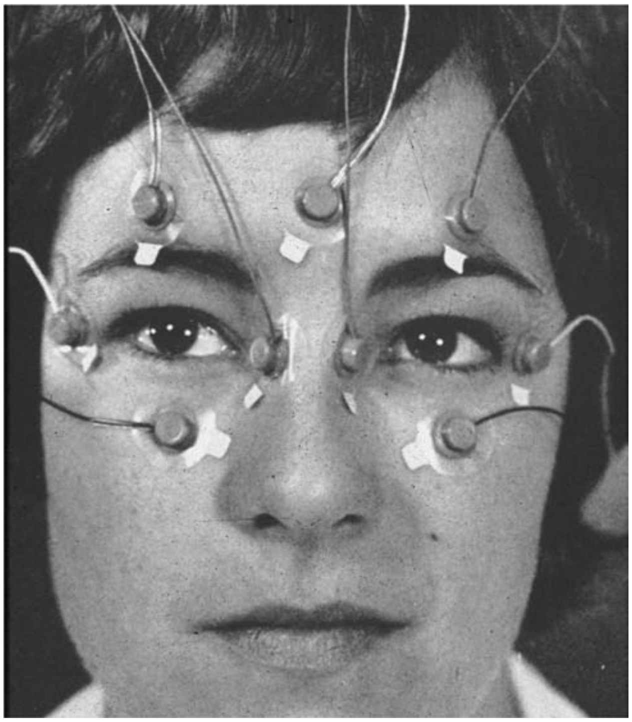


Figure 1. EOG electrodes arranged to record the horizontal and vertical eye movements of both eyes. Reference electrode is in the center of the forehead.

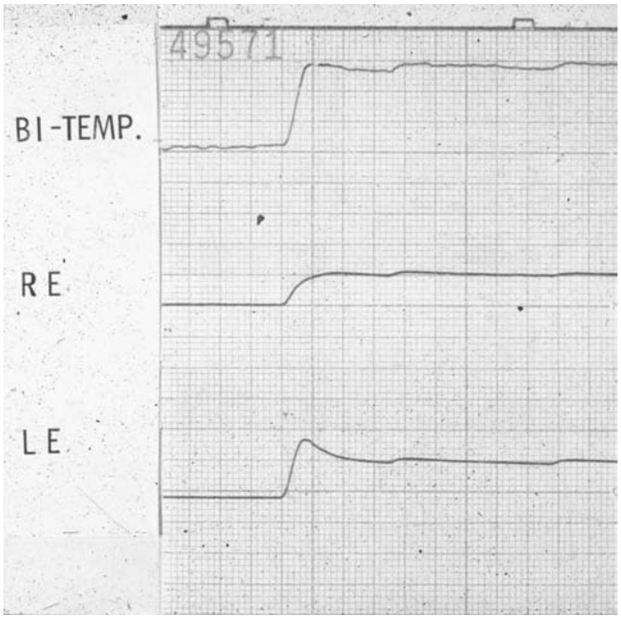


Figure 2. False saccadic trajectory from bitemporal EOG electrodes resulting from the summation of the individual saccadic trajectories shown below.

for reference. This shortcut effectively records a "cyclopean" eye by summing the potentials obtained from each eye. Although eye movements other than vergence are conjugate in normal individuals, it is not generally normal individuals who are seen for clinical evaluation. Figure 2 illustrates how an overshooting and an undershooting eye movement may be combined to give the appearance of a perfect saccade. For this reason, both ac-coupling and bitemporal electrode placement should be avoided when anything other than the crudest information about eye movement is desired.

Infrared Reflectance.

Theory of Operation. Although photographic recording of eye movements dates back to 1901 (1), such methods remained cumbersome to use, especially when they required frame-by-frame analysis of the location of some marker on the eye. Optical levers, where a beam of light was reflected from a mirror attached by a stalk to a scleral contact lens, offered the opportunity for precise registration of eye position, but occluded the view of the eye being recorded. As might be imagined, they were also unpleasant to wear. An alternative recording method that also makes use of reflected light relies on the differential reflectivity of the iris and sclera of the eye to track the limbus—the boundary between these structures. The earliest versions of this system were developed by Torok et al. (12) and refined by several investigators over the years (13–15). Although the number of emitters and detectors vary between designs, they share the same fundamental principle; that is, the eye is illuminated by chopped, low intensity infrared light (to eliminate the effects of variable ambient lighting). Photodetectors are aimed at the limbus on either side of the iris. As the eye moves, the amount of light reflected back onto some detectors increases and onto others decreases. The difference between the two signals provides the output signal. As would be expected, these signals are analog systems, so that the output of the photodetectors is electronically converted into a voltage that corresponds to eye position. Figure 3 shows an IR system mounted on an earth-fixed frame (a), spectacle

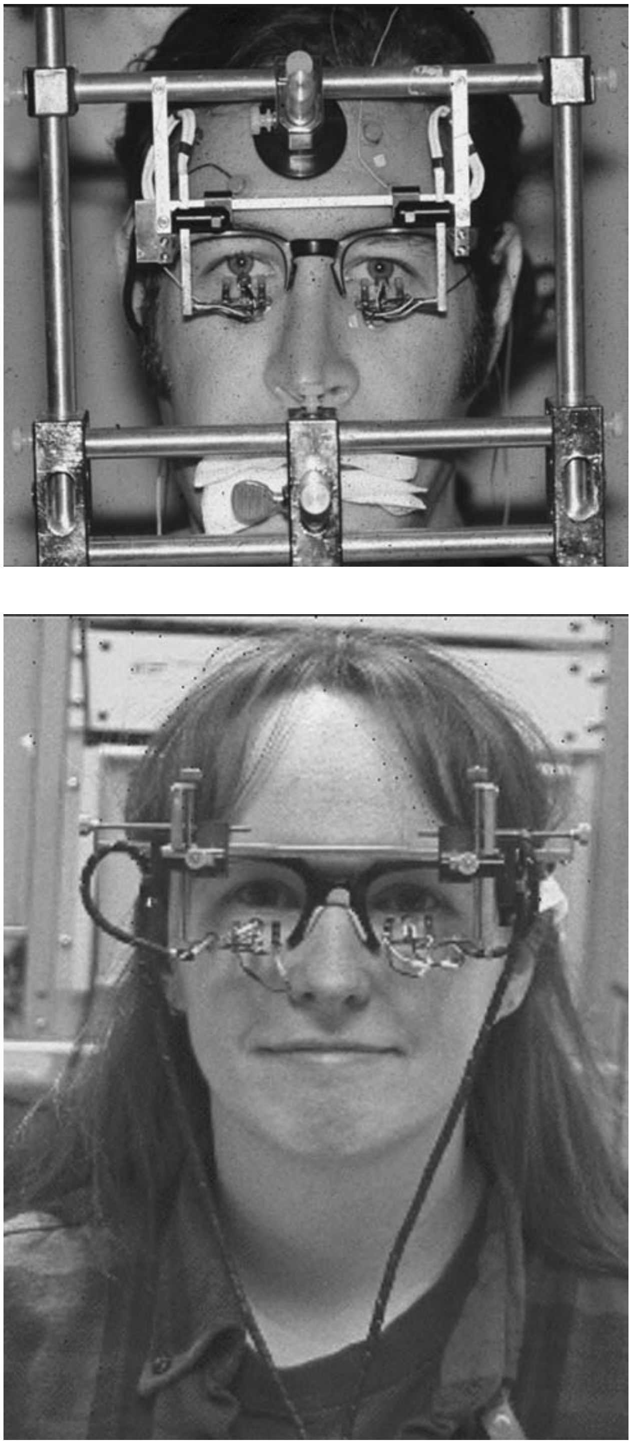


Figure 3. IR system to measure the horizontal eye movements of both eyes shown mounted on an earth-fixed frame (a) and spectacle frame (b) for human subjects and on a spectacle frame for a canine subject (c).

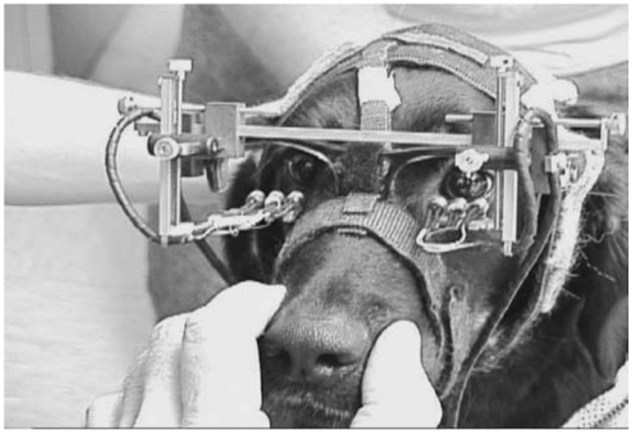


Figure 3. (*Continued*)

frame (b), and spectacle frame on a dog (c). Figure 4 shows an IR system mounted in goggles on a child (a) and a dog (b).

Characteristics. These systems offer a number of advantages over EOG, at least for the examination of horizontal eye movements. As the signal is not a biopotential, it is free of the instability found in the EOG; it is also immune to interference from muscle artifact and changes in electrode potentials. Unlike some earlier photographic methods, the device does not occlude the eyes, as the sensors and emitters are positioned above or below the eye. The field of view is somewhat obstructed by the emitter/detector, in contrast to EOG. Resolution is of the order of minutes of arc. Assuming that nothing disturbs the sensors, a shaken head or a rubbed eye, for example, stability is excellent. Thus, the question of using ac-coupling, as in many electronystagmographic applications of the EOG, never occurs. System bandwidth is generally on the order of 100 Hz, which is sufficient to capture fine details of saccades.

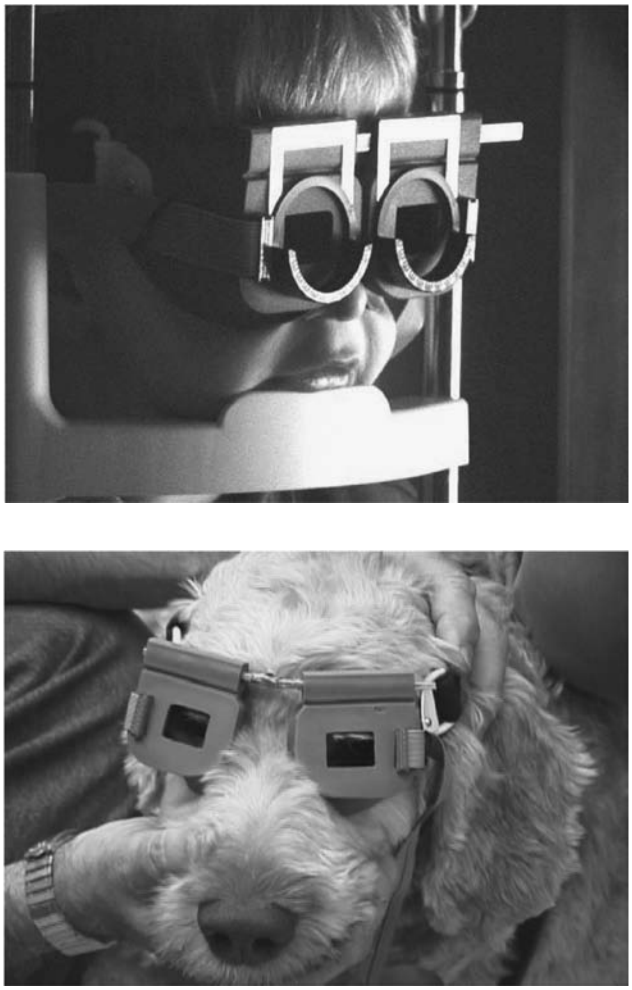


Figure 4. IR system to measure the horizontal and vertical eye movements of both eyes shown mounted in goggles for a human subject (a) and a canine subject (b).

The linear range of these systems generally is between $\pm 15^\circ$ and 20° in the horizontal plane and half this amount or less in the vertical plane (which requires vertical orientation of the detectors or summation of the signals from horizontally-oriented detectors).

Applications. IR limbus trackers are probably second only to EOG in their range of applications. Their ability to resolve fine detail with low noise makes them excellent for conditions where subtle features of the eye movement are important; examples include analyses of saccadic trajectories or analysis of small corrective saccades within a nystagmus waveform. An important advantage over EOG is that if eye velocity is to be calculated, the resulting signal is far less noisy than the derivative of an EOG recording, especially where broadband EMG noise has contaminated the signal developing from the eye. These systems are well suited to studies of any sort of eye movement that falls within their linear operating range in the horizontal plane. As they are generally head-mounted, they will tolerate modest head movement, but if the stimuli are fixed in the environment, such movement will certainly cause a loss of baseline and may move the tracker outside its linear range, which makes head stabilization highly desirable, especially when stimuli are presented at gaze angles where subjects would normally make both a head movement and an eye movement to acquire the target. Finally, IR systems are noninvasive, a major advantage for many patients and for children.

Limitations. One of the biggest shortcomings of these systems is their poor performance for vertical eye movement, their near-uselessness for oblique eye movements, and their complete lack of value for torsional eye movements. Although the limbus is clearly visible over a wide range of eye positions in the horizontal plane, the eyelids obscure its top and bottom margins. Although a degree of vertical tracking can be obtained by virtue of the differential reflectivity of the iris and pupil, the range over which this is possible is limited, again in part because of occlusion of the lids. Oblique movement suffers from inherent crosstalk because, as eye position changes in one plane, the sensitivity to motion in the other plane will vary, which is a hindrance to using these systems for studies of reading, scanpath analysis, or other applications where 2D eye movements are important. The use of the systems in rotational testing is also limited by the range of allowable gaze angles and by the possible slippage of the head mounting on the head if accelerations are sufficiently high. Their suitability for small children also varies; some of the systems do not fit small heads well, although if precise calibration is not important, one can generally record patients as young as 3 years. These systems are not generally appropriate for use with infants. The one exception is for diagnosing nystagmus from its waveform by simply holding the sensors in front of the eyes, which can be done for even the smallest infants (e.g., a premature infant still in an incubator).

Scleral Search Coil.

Theory of Operation. Robinson developed the Scleral Search Coil technique in 1963 (16). It relies on the principle that a coil of wire in an alternating magnetic field induces a voltage proportional to the area of the coil, the number of turns, and the number of field lines. This latter measure will vary with the sine of the angle the coil makes with the magnetic field. In the basic configuration, two orthogonal pairs of field coils are used, each modulated by phase-locked square wave sources either operating in quadrature (i.e., one signal 90° phase-shifted relative to the other) (16) or at a different frequency (e.g., 50 and 75 kHz) (17). An annular contact lens with a very fine coil of wire is placed on the eye, so that it surrounds the cornea (or in animals, is surgically implanted under the conjunctiva). Figure 5 shows an annular search-coil contact lens on the eye of a subject. Components of the induced voltage generated by the horizontal and vertical signals can be separated via phase-sensitive detectors. Note that this method of recording horizontal and vertical components of eye movement eliminates the crosstalk present in 2D recordings made by limbus trackers. With an appropriately wound coil added to the lens, torsional eye movements may also be recorded. This technique is the only one able to record torsion with high bandwidth.

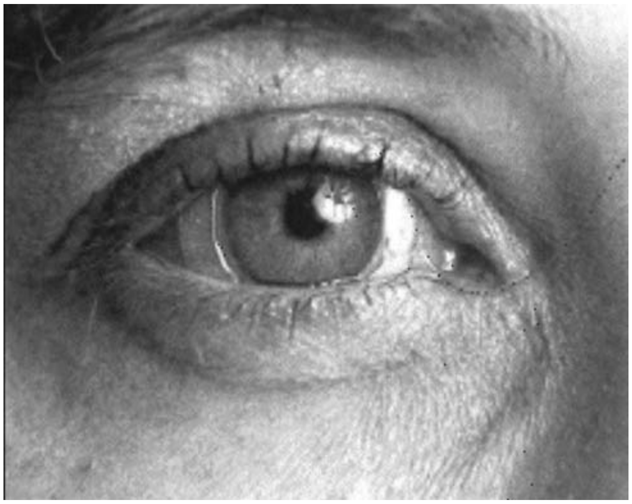


Figure 5. An annular search-coil contact lens used to measure the horizontal and vertical eye movements of a human subject. The fine wire from the imbedded coil exits at the nasal canthus.

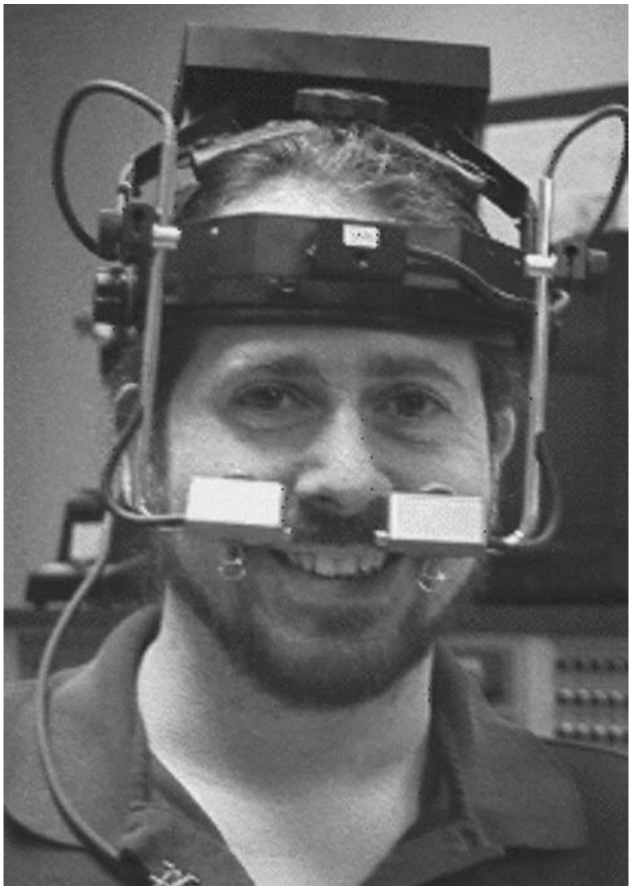
Characteristics. This technology serves as the "gold standard" for eye-movement recording. Resolution is in seconds of arc and the linear range ± 20°, with linearization possible outside this range, because the nonlinearity follows the sine function. The signals are extremely stable, because their source is determined by the geometry of coil and magnetic field alone. In the usual configuration, the maximum angle that can be measured is 90°. Although the eyes cannot rotate this far in the head, if the head is also allowed to turn (and its position recorded by a head coil), a net change of eye position $> 90^\circ$ is possible. A solution to this problem was developed whereby all the field coils were oriented vertically, generating a magnetic field whose vector rotates around 360°. Now, the phase of the field coil varies linearly over 360° of rotation (18,19), which is most often used for horizontal eye movements, with vertical and torsional eye movements recorded using the original Robinson design.

Applications. As the search-coil system provides such high quality data, it can be used in nearly any application where stability, bandwidth, and resolution are paramount and free motion by the subject is not essential. However, recent evidence suggests that the coils themselves may alter the eye movements being measured (20). Nonetheless, the low noise level and ability to independently record horizontal, vertical, and torsional movements at high bandwidth and high resolution still make this the gold standard of eye-movement recording techniques.

Limitations. As a result of their size, search-coil systems are clearly not suited for ambulatory studies or those carried out in other real-world settings such as a vehicle. The system also cannot be adapted to use in fMRI scanners, unlike IR limbus trackers or video-based systems. Search coils are invasive, making them unsuitable for some adult patients and for most children. A small risk of corneal abrasion exists when the coil is removed, but this risk is generally minor. Use of the coil in infants or small children would be undesirable, because they could not be instructed not to rub their eyes while the coil was in place. Another practical issue associated with the technology is the cost of the coils, which have a single supplier, have a limited lifetime, and are relatively expensive (> US$100 each). As recommended duration of testing with the coils is 30 minutes or less, long duration studies are also precluded.

Digital Video.

Theory of Operation. Although electronic systems that locate and store the location of the center of the pupil in a video image of the eye were developed in the 1960s, often in combination with pupil diameter measurement (21,22), video-based eye trackers became a major force in eye-tracking technology when digital rather than analog image analysis was implemented. If the camera is rigidly fixed to the head, then simply tracking this centroid is sufficient to identify the location of the eye in its orbit. However, if there is even slight translational movement of the camera with respect to the eye, a large error is introduced: 1 mm of translation equals 10° of angular rotation in the image. For this reason, video systems also track the specular reflection of a light source in the image in addition to the pupil centroid. As this first Purkinje image does not change with rotation but does change with translation, whereas the pupil center changes with eye rotation as well as translation, their relative positions can be used to compensate for errors induced by relative motion occurring between the head and camera. Figure 6 shows a digital video system in use on a human subject (a) and on dogs (b and c).



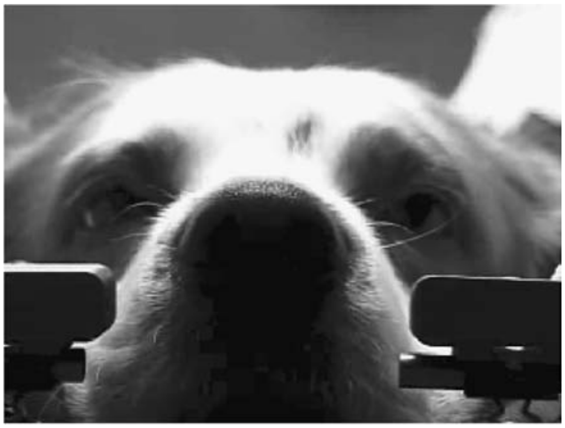


Figure 6. A high-speed digital video system to measure the horizontal and vertical eye movements of both eyes for a human subject (a) and canine subjects (b and c).

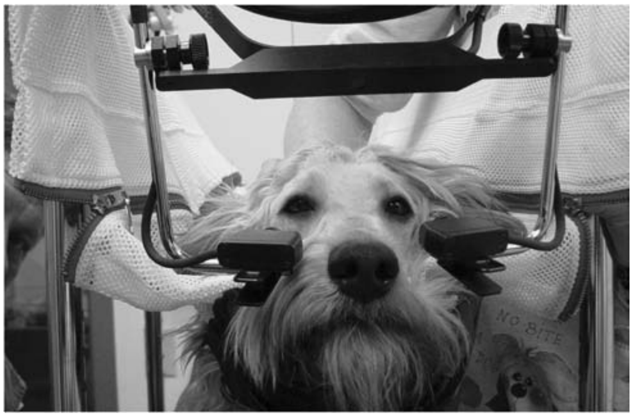


Figure 6. (*Continued*)

Characteristics. Assuming that the axes of the head and camera are aligned, then video-based systems are capable of recording both horizontal and vertical eye movements over a relatively wide range (often ± 30° horizontally, somewhat less vertically). Resolution is better than EOG but generally somewhat less than for IR or search-coil systems, often in the range of 0.5°. As analog video systems use a raster scan to represent an image, spatial resolution is limited by the nature of the video system used (e.g., PAL or NTSC). Bandwidth is limited by the frame rate of the video system. If conventional analog video is used, then frame rates are 50 Hz for PAL and 60 Hz for NTSC. These rates impose a maximum bandwidth of 25 and 30 Hz, respectively. Although adequate for examination of slow eye movements, these frame rates are inadequate for assessment of saccades; indeed, very small saccades could be completed within the inter-frame interval. Systems using digital video are free from the constraints imposed by broadcast TV standards and can make use of higher frame rate cameras—several now operate at 250 or 500 Hz. Generally, a frame rate versus resolution trade-off exists—higher frame rates imply lower image resolution. However, continued improvement in digital video technology and ever faster and cheaper computers continue to improve performance.

Although older video tracking systems often required a good deal of “tweaking” of brightness and contrast settings in an effort to obtain a reliable image of the pupil, many recent systems have more streamlined set-up protocols. In the past, some systems internally monitored fixation on calibration targets and rejected data that were unstable, thereby making the systems unsuitable for use with patients with nystagmus. However, default calibration settings generally permit data to be taken and the nystagmus records can then be retrospectively calibrated.

Applications. In principle, digital video is the most flexible of all eye-movement recording technologies. Some systems use cameras mounted on the head, using either helmets or some other relatively stable mounting system. Other systems use remote cameras, often mounted adjacent to or within a computer stimulus display. Systems used in vehicles may use either remote cameras or helmet-mount cameras. In addition to conventional clinical eye-movement testing, video systems, especially remote camera models, are increasingly being used in commercial applications such as advertising studies and usability analyses of websites. For such applications, the unobtrusiveness of the technology and the need to only monitor fixations rather than to study saccade dynamics makes even relatively low-frame-rate video ideal. Such systems are also excellent for use with infants and small children, who may be induced to look at some attractive display on a screen but who generally respond poorly to head-mounted apparatus. Remote systems that track more than one first Purkinje image can cope with a wider range of head movements, making the systems even less restrictive for the subjects. Some video systems can also analyze torsional eye movements by identifying some feature on the iris and then tracking changes in its orientation from frame to frame. High-speed (500 Hz) digital video systems are seeing increased use in basic and clinical laboratories, challenging magnetic search coils as the method of choice.

Limitations. The problems associated with calibrating patients whose eyes are never still have already been discussed. As noted before, the other serious limitations of some of these systems are their somewhat limited spatial resolution and bandwidth. Both parameters can be optimized, but doing so leads to marked increases in price. However, unlike other eye-tracking technologies, the limiting factors for high-speed, digital video eye-movement recording systems are the cameras and computing power. As the enormous general consumer market rather than the quite small eye-movement recording market drives improvements in both technologies, improvements can be anticipated to occur much faster than they would otherwise. Even within the eye-tracking field, the development of commercial uses for the technology will facilitate its advance faster than the smaller and less prosperous academic research community.

OCULAR MOTOR RECORDING TECHNIQUES

How Do We Record and Later Calibrate and Analyze Subjects' Eye Movements?

The initial recording and *post-hoc* calibration and analysis of eye movements require following a protocol conducive to both accurate calibration and obtaining the data specific to a particular study. Decades of experience have resulted in the following recording procedures and caveats and in the development of software that allows accurate calibration and linearization of the data.

Real-Time Monitoring. When recording subjects (especially patients), it is necessary to monitor the eye channels in real-time to ensure that the subject is following instructions, which is also imperative when calibrating subjects (see below). Unlike highly dedicated and motivated graduate students, most subjects quickly become bored by the task at hand or distracted and fail to fixate or pursue the stimuli; others may have difficulty doing so. Real-time monitoring via a strip chart or computer display allows the experimenter to detect and correct such failures with a simple verbal instruction encouraging the subject (e.g., "follow the target" or "look at the target").

Monocular Calibration. The key to obtaining accurate eye-movement data that will allow meaningful analysis is monocular calibration; that is, calibration of each eye independently while the other is behind cover. Too often, potentially accurate, commercially available recording systems are seriously compromised by built-in calibration techniques that erroneously presume conjugacy, even for so-called normal subjects. Just as bitemporal EOG makes it impossible to determine the position of either eye individually (see Fig. 2), so do calibration techniques carried out during binocular viewing of the stimuli. Most commercially available software calibration paradigms suffer from this fatal flaw, rendering them totally inappropriate for most clinical research and seriously compromising studies of presumed normal subjects. For methods that depend on subject responses to known target positions (e.g., IR or digital video), both the zero-position adjustment and gains at different gaze amplitudes in each direction must be calibrated for each eye during short intervals of imposed monocular fixation (i.e., the other eye occluded); for methods where precalibration is possible (e.g., magnetic search coils), the zero adjustment for each eye in each plane must also be made during imposed monocular fixation.

Linearization and Crosstalk Minimization. In addition to monocular calibration, linearization is required of most systems, even within the stated "linear" regions of those systems. As a result of different facial geometries and the inability to position the sensors in the precisely optimal positions for linearity, these systems are usually not linear over the range of gaze angles needed for many studies. System responses may be linearized by taking data during short intervals (5 s) of monocular fixation at all gaze angles of interest (e.g., 0°, ± 15°, ± 20°, ± 25°, and ± 30°) and applying post-recording linearization software. Even Robinson-type search coils need an arcsine correction for a linear response. For IR and video-based systems measuring eye motion in both the horizontal and vertical planes, crosstalk is a major problem due to sensor placement. Crosstalk can also be minimized post recording, using software written for that purpose. However, IR systems suffer from the additional problem that, as vertical eye position changes, a change may occur in the sensors' aim regions at the left and right limbal borders, which means that for a diagonal eye movement, the horizontal gain is an unknown function of vertical eye position, making IR systems essentially unsuitable for the recording of oblique eye movements.

All of the problems discussed above are accentuated when recording subjects with ocular motor oscillations, such as nystagmus. In these cases, the experimenter must be familiar with the type of nystagmus the subject has and be able to identify the portions of their waveforms that are used for target foveation. It is the "foveation periods" that are used to set the zero-position and gains at each target position; without them, accurate calibration is impossible.

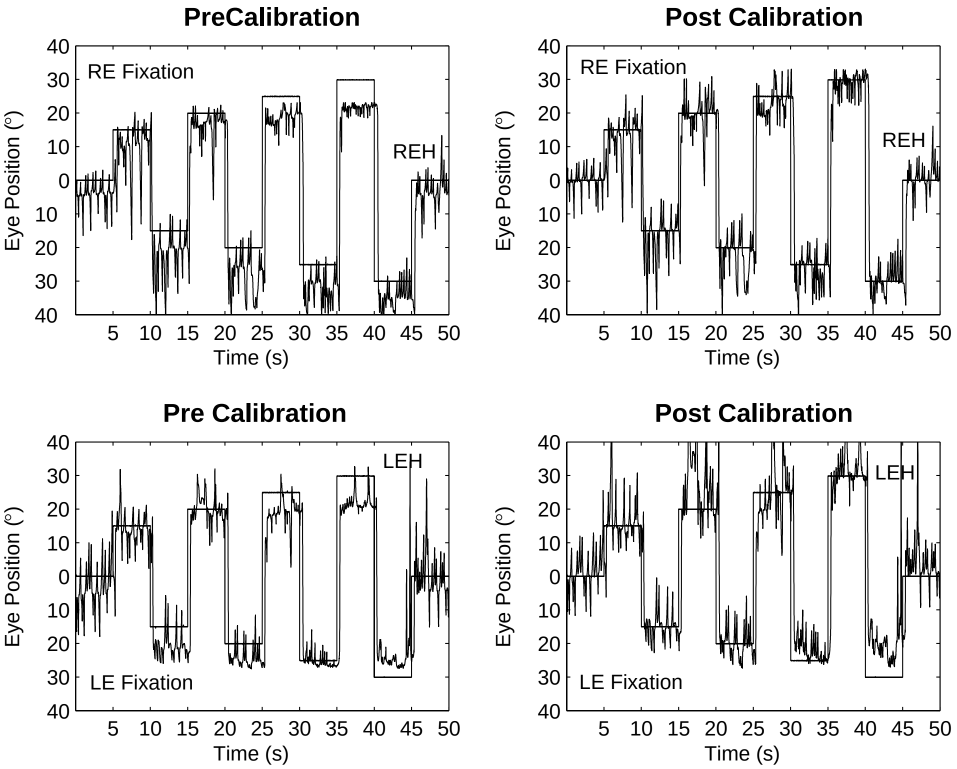


Figure 7. Monocular fixation precalibration and postcalibration (horizontal) records for the right (REH) and left (LEH) eye. Compare the offsets and nonlinear precalibration responses to the bias-adjusted, calibrated, and linearized postcalibration responses. Note the failure of the subject to look at the −30° target during LE fixation. In this figure and Fig. 8, target position is shown by the alternating direction, increasing offset, solid line.

The rest of the nystagmus waveform is irrelevant to target foveation and should be ignored during calibration. With a little practice, investigators can easily determine exactly where the subject with nystagmus is looking, which eye is fixating, and where the other eye is located with respect to the target; they can also determine periods of inattention by the associated waveform changes. Figure 7 demonstrates precalibration and postcalibration (horizontal) records of each eye made under imposed monocular fixation, and Fig. 8 shows the results of applying those calibration factors to a record made during binocular "viewing" of the targets. Note that the fixating eye is easily determined as well as the angle/position of the strabismic eye. Unfortunately, investigators with little or no experience in recording subjects with nystagmus are often reduced to using the average eye position during long periods of presumed binocular fixation to approximate calibration of subjects with nystagmus (and probably strabismus). Averaging anathema to accurate calibration and renders most potentially accurate recording systems (e.g., search coils) no better than bitemporal EOG. Needless to say, the results and conclusions of such studies must be highly suspect and are often incorrect; they exemplify how even the most sophisticated hardware and software can be misused, and prove the old adage, "garbage in, garbage out."

CONCLUSIONS

During the past 40 years, advances in eye-movement recording systems, coupled with the control-systems approach brought to the field by biomedical engineers, have resulted in an explosion of basic and clinical ocular motor research, at the systems as well as single-cell levels. Using the measurement systems and recording and calibration techniques described above, great strides have been made in our understanding of the ocular motor system. Animal studies have provided understanding at the single-cell and cell-network (bottom-up) levels, giving rise to computer models of small portions of the ocular motor system with neuroanatomical correlations. Normal human studies have allowed characterization of ocular motor behavior under a variety of stimulus conditions, giving rise to functional, top-down computer models of ocular motor behavior. Finally, studies of patients with many congenital and acquired ocular motor disorders have provided insights into the functional structure of the ocular motor system, which was not forthcoming from studies of normals (23,24). These latter studies have resulted in robust, behavioral models of the ocular motor system that are able to simulate normal responses and patient responses to a variety of ocular motor stimuli (25–27).

At present, accurate eye-movement recordings are an integral part of the diagnosis of both congenital and acquired forms of nystagmus, and of saccadic intrusions and oscillations. In addition, they provide objective measures of therapeutic efficacy that are related to visual function in patients afflicted with disorders producing ocular motor dysfunction. Indeed, ocular motor studies of the effects of a specific surgical procedure for congenital nystagmus produced an entirely new type of "nystagmus" surgery for both congenital and acquired nystagmus (28–31). This surgery (named "tenotomy") simply requires

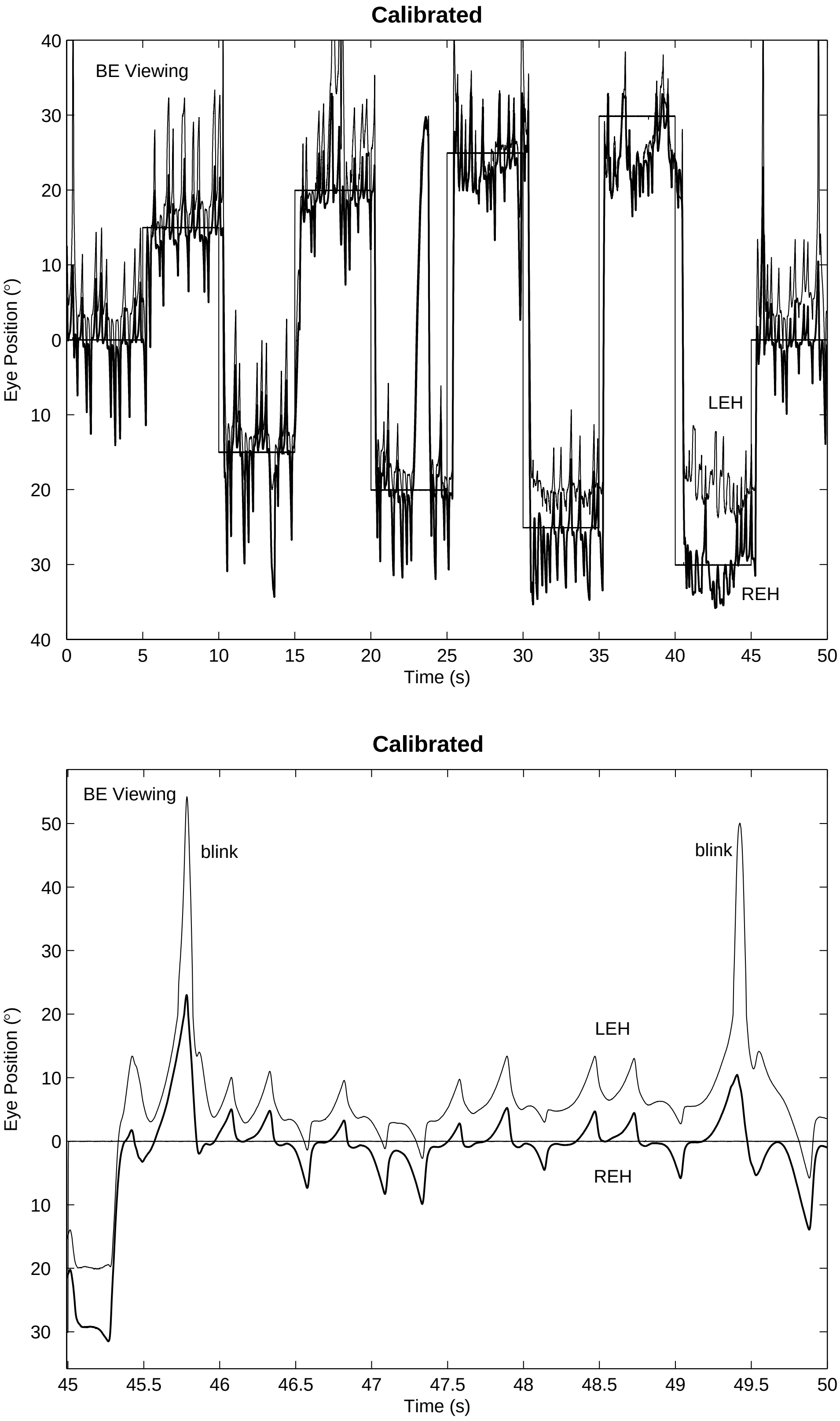


Figure 8. Calibrated binocular viewing records of both eyes (a) and final primary-position segment (b). Note the preference for RE fixation in left gaze and in the final primary-position segment, with the LE 3–5° esotropic. Note also how well the flat foveation periods of the RE line up on the 0° target despite the alternating direction of the nystagmus.

removal and reattaching, at their original insertion points, each of the four extraocular muscles in the plane of the nystagmus. Tenotomy represents a radical paradigm change from the "strabismus" surgeries that preceded it and has resulted in new insights into the anatomic structures responsible for proprioceptive signals from the extraocular muscles and their neurophysiologic role in the control of eye movements (32–34).

BIBLIOGRAPHY

1. Dodge R, Cline TS. The angle velocity of eye movements. Psychol Rev 1901;8:145–157.
2. Yarbus AL. Eye Movements and Vision. New York: Plenum Press; 1967.
3. Daffner KR, Scinto LF, Weintraub S, Guinessey JE, Mesulam MM. Diminished curiosity in patients with probable Alzheimer's disease as measured by exploratory eye movements. Neurology 1992;42:320–328.
4. Dell'Osso LF, Van der Steen J, Steinman RM, Collewijn H. Foveation dynamics in congenital nystagmus I: Fixation. Doc Ophthalmol 1992;79:1–23.
5. Dell'Osso LF, Van der Steen J, Steinman RM, Collewijn H. Foveation dynamics in congenital nystagmus II: Smooth pursuit. Doc Ophthalmol 1992;79:25–49.
6. Dell'Osso LF, Van der Steen J, Steinman RM, Collewijn H. Foveation dynamics in congenital nystagmus III: Vestibulo-ocular reflex. Doc Ophthalmol 1992;79:51–70.
7. Dell'Osso LF, Leigh RJ. Foveation period stability and oscillopsia suppression in congenital nystagmus. An hypothesis. Neuro Ophthalmol 1992;12:169–183.
8. Dell'Osso LF, Leigh RJ. Ocular motor stability of foveation periods. Required conditions for suppression of oscillopsia. Neuro Ophthalmol 1992;12:303–326.
9. Krauzlis RJ. Recasting the smooth pursuit eye movement system. J Neuriphysiol 2004;91:591–603.
10. Furman JM, O'Leary DP, Wolfe JW. Application of linear system analysis to the horizontal vestibulo-ocular reflex of the alert rhesus monkey using pseudorandom binary sequence frequency sinusoidal stimulation. Biol Cyber 1979;33:159–165.
11. Barry W, Jones GM. Influence of eyelid movement upon electro-oculographic recording of vertical eye movements. Aerospace Med 1965;36:855–858.
12. Torok N, Guillemin VJ, Barnothy JM. Photoelectric nystagmography. Ann Otol Rhinol Laryngol 1951;60:917–926.
13. Young LR. Measuring eye movements. Am J Med Electr 1963;2:300–307.
14. Kumar A, Krol G. Binocular infrared oculography. Laryngoscope 1992;102:367–378.
15. Reulen JP, Marcus JT, Koops D, de Vries FR, Tiesinga G, Boshuizen K, et al. Precise recording of eye movement: The IRIS technique. Part 1. Med Biol Eng Comput 1988;26: 20–26.
16. Robinson DA. A method of measuring eye movement using a scleral search coil in a magnetic field. IEEE Trans Bio Med Electron 1963;BME(10):137–145.
17. Remmel RS. An inexpensive eye movement monitor using the scleral search coil technique. IEEE Trans Biomed Eng 1984;31:388–390.
18. Hartmann R, Klinke R. A method for measuring the angle of rotation (movements of body, head, eye in human subjects and experimental animals). Pflügers Archiv (Suppl); 362:R52.
19. Collewijn H. Eye movement recording. In: Carpenter RHS, Robson JG, eds. Vision Research: A Practical Guide to Laboratory Methods. Oxford: Oxford University Press; 1999.
20. Frens MA, van der Geest JN. Scleral search coils influence saccade dynamics. J Neurophysiol 2002;88:676–691.
21. Young LR. Recording eye position. In: Clynes M, Milsum JH, eds. Biomedical Engineering Systems. New York: McGraw-Hill; 1970. pp 1–2.
22. Marchant J. The oculometer: NASA, 1967. Report No. CR-805.
23. Abel LA, Dell'Osso LF, Daroff RB. Analog model for gaze-evoked nystagmus. IEEE Trans Biomed Eng 1978;BME(25): 71–75.
24. Abel LA, Dell'Osso LF, Schmidt D, Daroff RB. Myasthenia gravis: Analogue computer model. Exp Neurol 1980;68:378–389.
25. Dell'Osso LF, Jacobs JB. A normal ocular motor system model that simulates the dual-mode fast phases of latent/manifest latent nystagmus. Biolog Cybernet 2001;85:459–471.
26. Dell'Osso LF. Nystagmus basics. Normal models that simulate dysfunction. In: Hung GK, Ciuffreda KJ, eds. Models of the Visual System. New York: Kluwer Academic/Plenum Publishers; 2002. pp 711–739.
27. Jacobs JB, Dell'Osso LF. Congenital nystagmus: Hypothesis for its genesis and complex waveforms within a behavioral ocular motor system model. JOV 2004;4(7):604–625.
28. Dell'Osso LF. Extraocular muscle tenotomy, dissection, and suture: A hypothetical therapy for congenital nystagmus. J Pediatr Ophthalmol Strab 1998;35:232–233.
29. Dell'Osso LF, Hertle RW, Williams RW, Jacobs JB. A new surgery for congenital nystagmus: Effects of tenotomy on an achiasmatic canine and the role of extraocular proprioception. JAAPOS 1999;3:166–182.
30. Hertle RW, Dell'Osso LF, FitzGibbon EJ, Yang D, Mellow SD. Horizontal rectus muscle tenotomy in patients with infantile nystagmus syndrome: A pilot study. JAAPOS 2004;8:539–548.
31. Hertle RW, Dell'Osso LF, FitzGibbon EJ, Thompson D, Yang D, Mellow SD. Horizontal rectus tenotomy in patients with congenital nystagmus. Results in 10 adults. Ophthalmology 2003;110:2097–2105.
32. Büttner-Ennever JA, Horn AKE, Scherberger H, D'Ascanio P. Motoneurons of twitch and non-twitch extraocular fibres in the abducens, trochlear and oculomotor nuclei of monkeys. J Comp Neurol 2001;438:318–335.
33. Büttner-Ennever JA, Horn AKE, Graf W, Ugolini G. Modern concepts of brainstem anatomy. From extraocular motoneurons to proprioceptive pathways. In: Kaminski HJ, Leigh RJ, eds. Neurobiology of Eye Movements. From Molecules to Behavior—Ann NY Acad Sci 956. New York: NYAS; 2002. pp 75–84.
34. Hertle RW, Chan C, Galita DA, Maybodi M, Crawford MA. Neuroanatomy of the extraocular muscle tendon enthesis in macaque, normal human and patients with congenital nystagmus. JAAPOS 2002;6:319–327.

See also Electroretinography; eye movement, measurement techniques for.

OCULOGRAPHY. See Ocular motility recording and nystagmus.

OFFICE AUTOMATION SYSTEMS

Jorge Cardoso
University of Madeira
Funchal, Portugal

INTRODUCTION

The purpose of this article is to help people in fields, such as healthcare, engineering, sales, manufacturing, consulting, and accounting to understand office automation systems from the viewpoint of a business professional. This is important because personal association with office automation systems is almost unavoidable in today's business world. The widespread adoption of personal computers in conjunction with the development of graphically driven operating systems gave people a more natural and intuitive way of visualizing and manipulating information. The applications that were developed, from word processors to spreadsheets, to take benefit of these new operating systems, led to a growth in the use and acceptance of personal computers that significantly altered the manner organizations conduct their daily business.

Healthcare enterprises involve complex processes that span diverse groups and organizations. These processes involve clinical and administrative tasks, large volumes of data, and large numbers of patients and personnel. The tasks can be performed either by humans or by automated systems. In the latter case, the tasks are supported by a variety of software applications and information systems that are very often heterogeneous, autonomous, and distributed. The development of systems to manage and automate these processes has increasingly played an important role in improving the efficiency of healthcare enterprises.

Office Automation Systems (OAS) are computer-based automated information systems that are used to execute a variety of office operations, such as word processing, electronic spreadsheet, e-mail, and video conferencing. These different office automation systems allow the automation of much of the administrative work in the office and typically focuses on the more repeatable and predictable aspects of individual and group work. They are more and more frequently used by managers, engineers, and clerical employees to increase efficiency and productivity. They support the general activities of workers and underlie the automation of document-centric tasks performed by production office workers.

The OAS encompass a broad set of capabilities, and provide much of the technological basis for the electronic workplace. The focus of OAS have typically been used in supporting the information and communication needs of office workers, and its use by organizations supporting the white-collar work force has revealed itself crucial.

HISTORICAL PERSPECTIVE

In its early days, office automation systems focused on needs generally found in all offices, such as reading and writing. Before the 1950s, electromechanical and electronic devices were used to carry out financial and other numerical record-keeping tasks. During the evolution of OAS solutions, manual typewriters have been replaced by the electric typewriter and the electronic typewriter.

The electronic typewriter, introduced in the early 1970s, was the first of the automated office systems. It could store and retrieve information from memory providing automated functions such as center, bold, underline, and spell check.

The advances in the development of mainframes have caused electromechanical devices to be increasingly replaced by computers. In the 1970s, integrated circuit technology made the production of small and relatively inexpensive personal computers possible. Yet, even with this available technology, many computer companies chose not to adopt personal computers. They could not imagine why anyone would want a computer when typewriters and calculators were sufficient.

In the mid-1970s, computers began to support offices and organizations in more complex ways. The rapid growth of computers furnished the market with sophisticated office automation devices.

In the late 1970s, several researchers started to describe the needs of office automation systems. Computer terminals had replaced electronic typewriters. With the rapid evolution of electronic technology, office information systems were developed to provide for the storage, manipulation, computation, and transmission of large amounts of information. The first sophisticated OAS prototypes included the SCOOP project (1), which was oriented to the automation of office procedures, and Officetalk (2), which provided a visual electronic desktop metaphor, a set of personal productivity tools for manipulating information, and a network environment for sharing information.

In 1981, IBM introduced the IBM PC (Personal Computer). The PC was a milestone and proved that the computer industry was more than a trend, and that the computer was in fact a necessary tool for the business community. Computers, designed solely for word processing and financial tasks, became common. At first, the PC was utilized to replace traditional typewriters and calculators, but persistent technological advances and innovation over the past two decades have put powerful PCs at the center of daily activities for people worldwide.

The growth and widespread adoption of PCs, networks, graphical user interfaces, and communications as allowed the development of complete OAS package suites. For example, in 1985 the Lotus Notes (3) groupware platform was introduced. The term groupware refers to applications that enhance communication, collaboration, and coordination among groups of people. This system included online discussion, e-mail, phone books, and document databases. Throughout the years, continuous improvements were made to Lotus Notes. Nowadays, this system includes new features, such enterprise-class instant messaging, calendaring, and scheduling capabilities with a strong platform for collaborative applications.

In 1992, Microsoft lunched its new operating system (OS), Microsoft for Workgroups (4). This OS allowed the sending of electronic mail and provided advanced networking capabilities to be used as a client on existing networks. This was an important stage in the vast evolution of the world's most popular operating system since it enabled the collaboration of groups of people. Microsoft has also invested in the development of full OAS suites, which are commonly available nowadays. The most well-known and widespread productivity software suite is Microsoft Office (5). Microsoft Office helps workers to complete common business tasks, including word processing, e-mail, presentations, data management and analysis.

ORGANIZATIONAL INFORMATION SYSTEMS AND OAS

While we are interested in studying office automation systems (OAS), it is important to relate this type of systems with other information systems (IS) commonly used inside an organization. An information system can be defined as a set of interrelated components that retrieve, process, store and distribute information to support decision making and control in an organization. The main role of IS is to assist workers in managing resources and information at each level in the organization.

This article, is primarily concerned with OAS and how they can be used in the medical community. For completeness, some other types of information systems commonly used by organizations are also mentioned. There will be no description of how such systems are developed, however, a brief description of their objectives will be given. Organizational information systems (OIS) are systems that support several functions in an organization and can be classified by the activity that they support. The OIS are usually split into six major types: Transaction Processing System, Knowledge Work Systems, Office Automation System, Management Information System, Decision Support System, and Executive Information System. These systems are illustrated in Fig. 1.

It is important to be able to distinguish the objectives and the level in the organization where a particular application or system can be used. For example, Transaction Processing Systems are employed to records daily routine transactions to produce information for other systems, while Office Automation Systems are oriented to increase that productivity of data workers using applications such as word processing and electronic mail applications.

Transaction Processing System (TPS): Is useful for daily transactions that are essential to the organization such as order processing, payroll, accounting, manufacturing, and record keeping.

Office Automation System (OAS): Aids office workers in the handling and management of documents, schedules, e-mails, conferences and communications. Data workers process information rather than create information and are primarily involved in information use, manipulation or dissemination.

Knowledge Work System (KWS): Promotes the creation of new information and knowledge and its dissemination and integration within the organization. In general, knowledge workers hold professional qualifications (e.g., engineers, managers, lawyers, analysts).

Management Information System (MIS): Provides middle-level managers with reports containing the basic operations of the organization which are generated by the underlying TPS. Typically, these systems focus on internal events, providing the information for short-term planning and decision making.

Decision Support System (DSS): Focuses on helping managers to make decisions from semistructured or unstructured information. These systems use internal information from TPS and MIS, but also information from external data sources, providing tools to support 'what-if' scenarios.

Executive Information System (EIS): Supports senior and top-level managers. They incorporate data from internal and external events, such as new legislation, tax laws, and summarized information from the

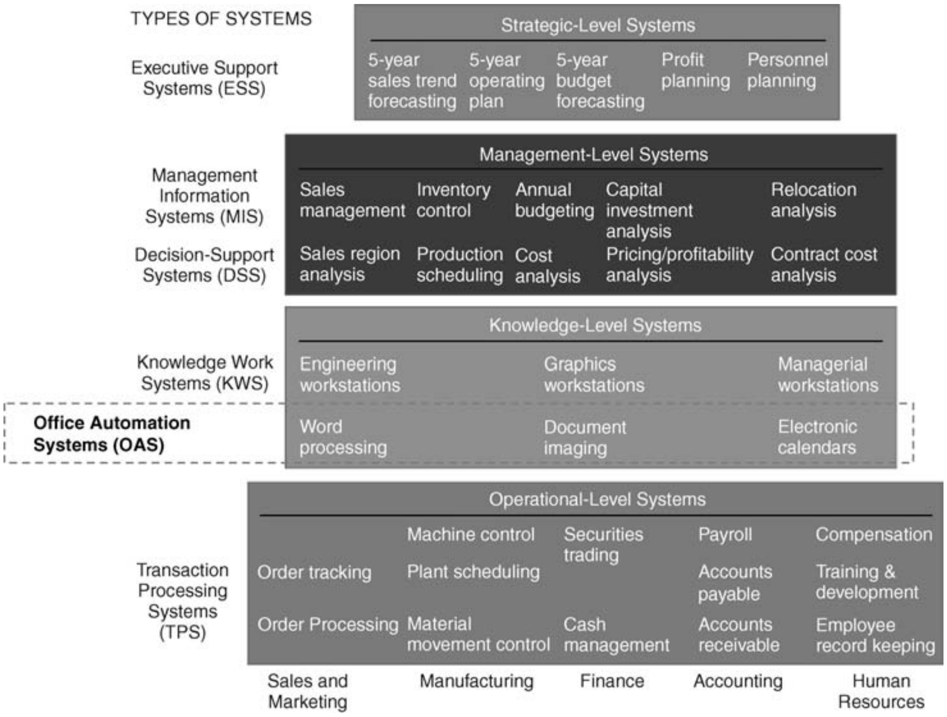


Figure 1. Types of information systems (6).

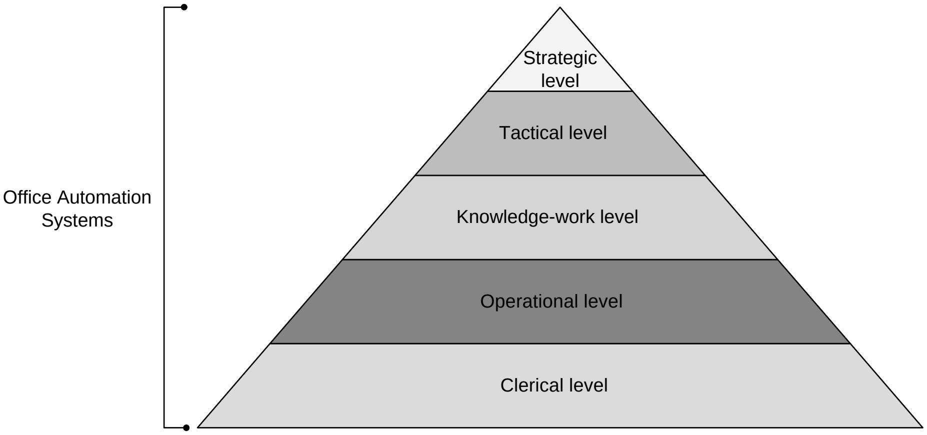


Figure 2. Support of OIS to the different organizational levels.

internal MIS and DSS. EIS software displays data graphically to provide easy-to-use representations of complex information.

Office information systems can also be classified by the organizational level they support. The human resources of an organization work in different areas and levels of operations, are in charge of different functions, and use different OIS. Any organization can be viewed as a multilevel entity with each level representing a different level of control. The levels of an organization can be arranged in a pyramid (Fig. 2).

The pyramid is divided into five horizontal sections:

Clerical level: Employees who support managers at all levels of the organization.

Operational level: First-line managers who make routine decisions and deal with the day-to-day operations of the organization.

Knowledge-work level: Advisors to both top and middle management who are often experts in a particular area.

Tactical level: Middle managers who deal with planning, organizing and the control of the organization.

Strategic level: Strategic managers who make decisions that guide the manner in which business is done.

Each successively lower level has different OIS requirements and a different, and less extensive, view of the organization. Obviously, the higher the level, the more interrelated the business functions become until, at the very top, they are viewed as one homogeneous organization with one continuous data flow.

Office automation systems can be effectively utilized in all the clerical, operational, knowledge-work, tactical, and strategic levels, as illustrated in Fig. 2. They can assist workers who work with word processors, electronic mail, and spreadsheets to use, manipulate, disseminate information, and help managers in planning, organizing, control and taking decisions.

OFFICE AUTOMATION SYSTEMS

Typical office automation systems handle and manage documents through word processing, desktop publishing, document imaging, and digital filing, scheduling through electronic calendars, and communication through electronic mail, voice mail, or video conferencing. In this section, 18 different types of OIS are discussed and described that are classify into four categories: productivity tools, digital communication systems, groupware applications, and teleconferencing systems (Fig. 3).

Productivity Tools

Productivity tools are software programs used to create an end product, such as letters, e-mails, brochures, or images. The most easily recognized tool is a word processing program, such as Microsoft Word (7) or Corel WordPerfect (8).

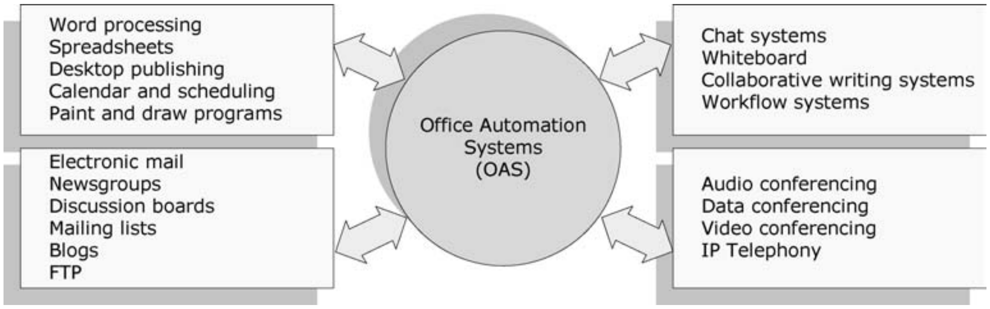


Figure 3. Office information systems.

Other tools help you view, create and modify general office documents such as letters, spreadsheets, memos, presentations, and images.

Word Processing. Of all computer applications, word processing is the most common. Almost every computer has a word processing program of some kind: whether it came free with the operating system or whether it was purchased separately.

In order to perform word processing, it is necessary to acquire a computer, a word processor, and a printer. A word processor enables you to create a document, store it, display it on the computer screen, modify it, and print it using a printer. There are many different word processing programs available, each offering different tools that make it easier to write everything from letters and term papers to theses and Web pages.

Most people use a word processor rather than a typewriter because it allows greater flexibility and control. It is possible to make changes without retyping the entire document. If mistakes are made while typing a text, the cursor can be used to correct errors. Word processors allow text rearranging, changing the layout, formatting the text, and inserting pictures, tables, and charts.

Most word processors available today allow more than just creating and editing documents. They have a wide range of other tools and functions, which are used in formatting documents. The following are the main features of word processors:

Insert, delete, copy, cut, and paste text: Allow to insert, erase, and copy text anywhere in the document. Cut and paste allow removing (cut) a section of text from one place and inserting (paste) it somewhere else in the document.

Search and replace: Allow searching for a particular word and also replacing groups of characters.

Font specifications: Allow to change fonts within a document. For example, you can specify bold, italics, font size and underlining.

Graphics: Allow adding pictures into a document.

Captions and cross-references: Allow placing captions to describe tables and pictures and creating references to them anywhere in the document.

Page setup, headers, and footers: Margins and page length can be adjusted as desired. Allow to specify customized headers and footers that the word processor will put at the top and bottom of every page.

Layout: Allows specifying different margins within a single document and to specify various methods for indenting paragraphs.

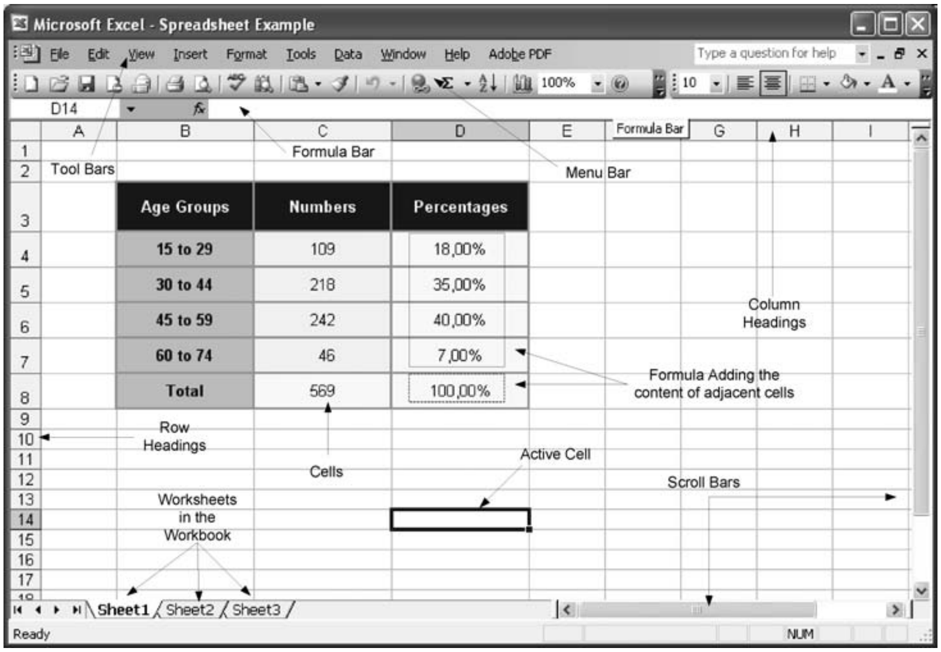
Spell checker and thesaurus: Spelling can be checked and modified through the spell check facility. The thesaurus allows the search for synonyms.

Tables of contents and indexes: Allow creating table of contents and indexing.

Print: Allows sending a document to a printer to get a hardcopy.

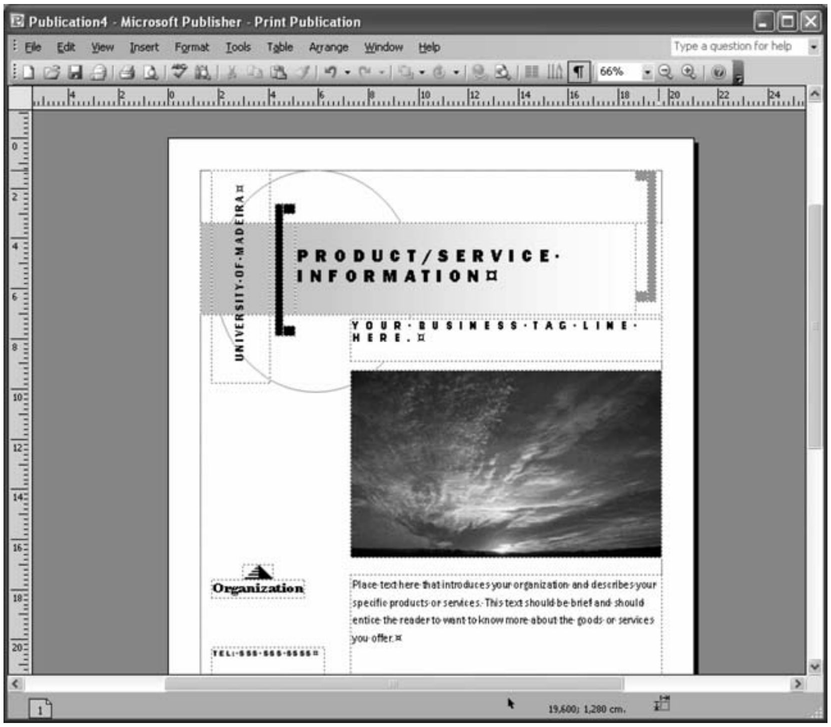
Spreadsheet. A spreadsheet is a computer program that presents data, such as numbers and text, in a grid of rows and columns. This grid is referred to as a worksheet. You can define what type of data is in each cell and how different cells depend on one another. The relationships between cells are called formulas, and the names of the cells are called labels.

There are a number of spreadsheet applications on the market, Lotus 1-2-3 (9) and Microsoft Excel (10) being among the most famous. In Excel, spreadsheets are referred to as workbooks and a workbook can contain several worksheets. An example of an Excel worksheet is shown in Fig. 4.



Age Groups	Numbers	Percentages
15 to 29	109	18,00%
30 to 44	218	35,00%
45 to 59	242	40,00%
60 to 74	46	7,00%
Total	569	100,00%

Figure 4. Microsoft Excel spreadsheet program.


Publication4 - Microsoft Publisher - Print Publication
File Edit View Insert Format Tools Table Arrange Window Help
Type a question for help
66%
UNIVERSITY·OF·MADEIRA¤
Organization
TEL:·555·555·5555¤
19,600; 1,280 cm.

PRODUCT/SERVICE· INFORMATION¤

YOUR·BUSINESS·TAG·LINE· HERE.¤

Place·text·here·that·introduces·your·organization·and·describes·your specific·products·or·services.·This·text·should·be·brief·and·should entice·the·reader·to·want·to·know·more·about·the·goods·or·services you·offer.¤

Figure 5. Editing a document with Microsoft Publisher.

Desktop Publishing. Desktop publishing is the use of the computer and specialized software to create high quality documents for desktop or commercial printing. Desktop publishing is the process of editing and layout of printed material intended for publication, such as books, magazines, brochures, and flyers using a personal computer.

Desktop publishing started in 1985, with the commercialization of the software Aldus PageMaker (11) (now from Abode). Nowadays, there are many software programs available for desktop publishing. QuarkXPress (12), Adobe InDesign (13), Abobe PageMaker (11), and Microsoft Publisher (14) are the most widespread. Figure 5 shows a document being created and edited with Microsoft Publisher.

As word processing programs become more sophisticated, the line separating such programs from desktop publishing systems is becoming fuzzy. Cutting-edge word processing programs give you most of the features you could want in a desktop publishing program. Such programs do not generally replace word processors and graphic applications, but are used to aggregate the text and graphic content created in these programs. The most powerful desktop publishing systems enable the creation of illustrations; while less powerful systems let you insert illustrations created by other programs.

Initial desktop publishing solutions were expensive due to the cost of specialized computing systems and accessories, such as printers and scanners. The cost of computers and printers has fallen dramatically in recent years (e.g., inkjet printers are amazingly inexpensive and most can print in color), allowing most personal users to acquire desktop publishing systems.

Calendars and Schedulers. A calendar program enables us to record events and appointments on an electronic calendar. Calendars allow scheduling, project management, and coordination among many people, and may provide support for scheduling equipment as well. Typical features identify conflicts in schedules, find meeting times that will work for everyone, signal upcoming events, and automatically fill in entries for regular events.

A special type of calendar, called a scheduler, is a solution to manage the daily scheduling needs of a business, such as scheduling appointments, equipment, staff (technicians, professionals, healthcare workers, others), vehicles, resources, projects, and meeting rooms. Scheduling software is an important investment for any type of business that wants to improve its scheduling processes. Every employee can have instant access to whom or what is available at any time of the day, week, month, or year and print detailed list reports. It is also possible to export schedules that may be easily opened in a word processor or spreadsheet.

Paint and Draw Program. A paint program or a graphics program enables the creation of pictures, backgrounds, buttons, lines, and other creative art. Paint programs provide easy ways to draw common shapes, such as straight lines, rectangles, circles, and ovals. Some programs also have photoediting capabilities and are optimized for working with specific kinds of images, such as photographs, but most of the smaller paint programs do not have this option. Paint programs are pixel based. They use "raster" images made up of small dots called pixels. As each dot is an individual, it can be difficult to move shapes around the screen.

A draw program is different from a paint program. Draw programs are object based, where an object is a geometrical shape, such as a line, a circle, a curve, a rectangle, a polygon, or a Bezier curve (curves that have hooks along their length so you can alter the angle of the curve at any point.) With draw programs, images are stored as mathematical information in the form of vectors for the lines and curves of each shape. Sophisticated programs often blur the difference between draw and paint, so it is possible to find programs that are able to do both types of work.

Digital Communication Systems

Nowadays, more and more computers are not isolated but, instead, are connected into a computer network that is often connected to other computer networks in much the same way as telephone systems connect telephones. If a computer is connected to such a network, it is possible to communicate with people whose computers are connected to the same network.

Electronic Mail. Electronic mail, or e-mail for short (another common spelling for e-mail is email), is one of the most popular uses of the Internet. It is a simple tool for exchanging brief messages between individuals or among a larger audience. Most mainframes, minicomputers, and computer networks have an e-mail system.

An e-mail address identifies a person and the computer for purposes of exchanging electronic mail messages. It consists of two parts: user name and mail domain or domain name. The user name identifies a particular person. The mail domain identifies the place on the Internet to which the e-mail for that person should be sent. An e-mail address is read from left to right. An example is illustrated in Fig. 6.

With an e-mail account, it is possible to send a message to anyone with an e-mail account. Just as a written letter can be sent to multiple recipients, an electronic mail message can be sent to one or more e-mail addresses. An e-mail can be broken down into several basic fields that include ‘From’, ‘To’, and ‘Cc’. The ‘From’ field contains the address of the sender of the message. The ‘To’ field indicates the addresses of one or more recipients who are the primary audience. All recipients can see every address listed in this field. Finally, the ‘Cc’ field (Cc - Carbon Copy) contains the addresses of recipients how are not the primary audience for the e-mail.

An electronic mail message is not limited to text. Other types of files can be added to mail messages as attachments. Attachments can be text files or binary files such as word processed documents, spreadsheets, images, files of sound and video, and software. To see if you have any e-mail, you can check your electronic mailbox periodically, although many programs can be configured to alert users automatically when mail is received. After reading an e-mail, it may be stored, deleted, replied to, forwarded to others, or printed.

One of the serious problems with reading e-mail on a PC computer running Windows operating system is that the computer can become infected with an e-mail virus program. It is always advisable to install and use anti-virus software. Such software will offer protection against known malicious programs. A malicious program may be a virus, a worm, a trojan horse, or a spyware. Once it is on your system, a malicious program cause disorder by corrupting, erasing, attaching to, or overwriting other files. In some cases malicious program, such as spyware, have the solely intent of monitoring Internet usage and delivering targeted advertising to the affected system. Unexpected e-mail attachments should not be opened since they are one of the most common ways for computer viruses to spread.

Newsgroups and Discussion Boards. Newsgroups, also known as Usenet, are comparable in essence to e-mail systems except that they are intended to disseminate messages among large groups of people instead of one-to-one communication (Fig. 7).

A newsgroup is a collection of messages posted by individuals to a news server. The concept of newsgroups was started in 1979 at the University of North Carolina and Duke University to create a place, where anyone could post messages.

Although some newsgroups are moderated, most are not. Moderated newsgroups are monitored by an individual (the moderator) who has the authority to block messages considered inappropriate. Therefore, moderated newsgroups have less spam than unmoderated ones. Anyone who has access to the board of a newsgroup can read and reply to a message that, in turn, will be read and replied to by anyone else who accesses it. If you have an interest in a certain topic, chances are it has its own newsgroup. A few examples of newsgroups are shown in Table 1.

Discussion boards (also called message boards) and newsgroups in general both accomplish the same task. They each have general topics, and visitors can post messages about specific topics. Discussion boards are usually read through a web browser, while newsgroups are usually read through a special program called a newsgroup reader. Nowadays, most people prefer discussion boards on the Web to newsgroups because they are easier to use.

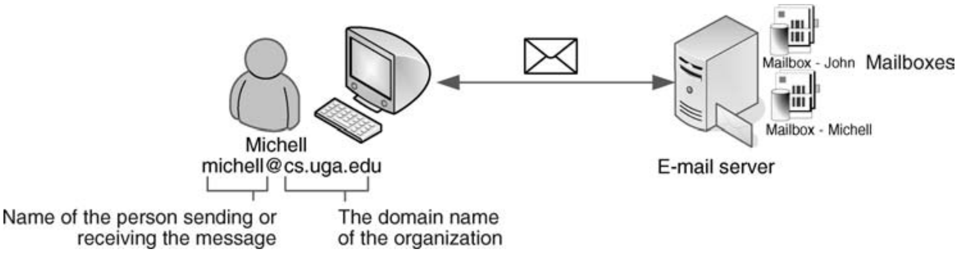

Mailbox - John
Mailboxes
Mailbox - Michell
Michell
michell@cs.uga.edu
E-mail server
Name of the person sending or receiving the message
The domain name of the organization

Figure 6. E-mail address structure.

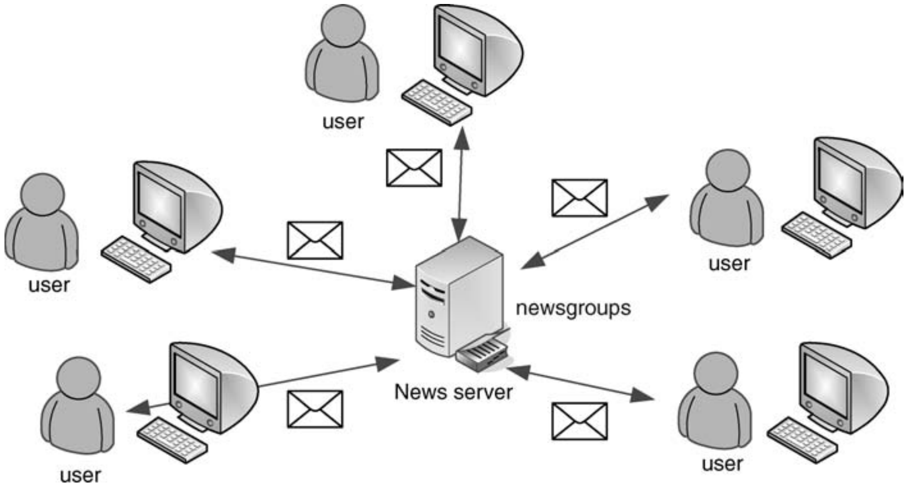


Figure 7. Newsgroup architecture.

Mailing Lists. The main difference between newsgroups and mailing lists is that newsgroups only show messages to a user when they are explicitly requested, while mailing lists deliver messages as they become available. Mailing lists are e-mail addresses that are used to distribute e-mail to many people. Typically, a list subscriber sends a message to the list address, and the message is then distributed to all the list subscribers for everyone to read.

Mailing lists are simple in operation. The first thing to do is to subscribe to a particular list; afterward the user can send messages to the mail server. The following steps are involved: (*1*) send a message (e-mail) to a mail server; (*2*) the mail server sends the message to everyone who is subscribed to the list; and (*3*) if someone replies to the message, then their reply goes to the mail server and is disseminated to everyone on the list.

Blogs. A weblog, or "blog", is a personal journal on the Web, although it can also be owned by a small group. The blog owner periodically writes entries and publishes them onto their blog. Weblogs cover as many different topics and express as many opinions, as there are people writing them.

A blog is used to show an up-to-date view of the owner's work, ideas, and activities. It provides a continuous record of activities, progress, and development. This type of systems can be effectively used by the healthcare community to discuss specific topics of interest. Examples of blog topics include product reviews, scientific endeavors, and any area of information where people have a deep expertise and a desire to express it. The power of blogs is that they are a fluid and dynamic medium that allow several people to easily publish their ideas and opinions, and allow other people to comment on them.

Table 1. Examples of Newsgroups

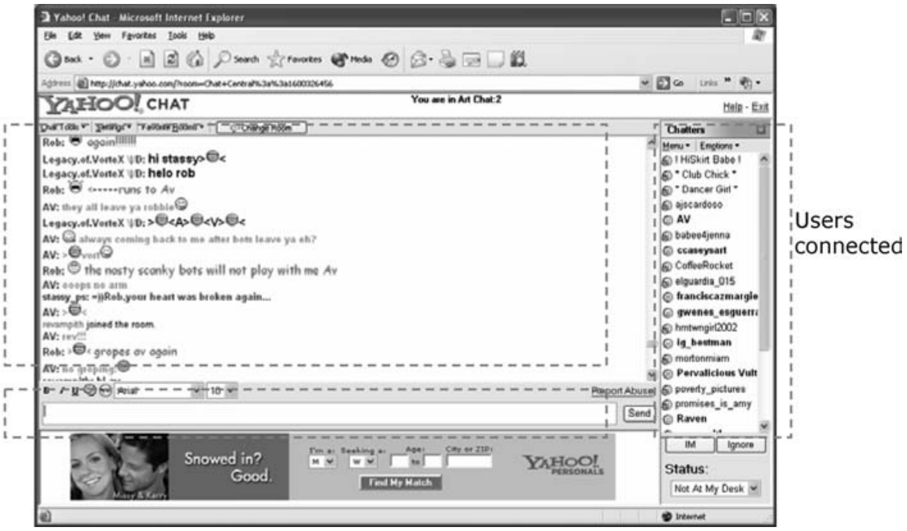
Newgroup name	Description
comp.ai	Artificial intelligence discussions
sci.cognitive	Perception, memory, judgment and reasoning
comp.groupware	Hardware & software for facilitating group interaction
comp.multimedia	Interactive multimedia technologies of all kinds
comp.infosystems	Any discussion about information systems
comp.graphics	Computer graphics, art, animation, image processing
alt.comp.blind-users	Discussion of the needs of blind users
comp.windows.misc	General issues regarding the use of windows

File Transfer Protocol. The ability to share information throughout organizations is essential in today's business environment. With the explosion of content creation and information in electronic formats, there is simply more electronic data today than ever before. File Transfer Protocol (FTP) is a standard method for sending files from one computer to another over networks, such as the Internet. Applications allow sharing and managing data between multiple remote, local, and home folders. It provides the ability to seamlessly work from a healthcare facility, a remote office, or home and is most commonly used to download a file from a server or to upload a file to a server.

Groupware Systems

Groupware refers to any computer-related tool that improves the effectiveness of person-to-person communication and collaboration. It is intended to create an environment that fosters the communication and coordination among a group of people. Where a traditional user interface generally focuses on the use of only one person, groupware relates to groups and understanding how people work and function in a group.

The groupware concept takes various applications and functionalities under the umbrella of communication and collaboration and integrates them together as a single client application. Groupware systems generally include some of the following systems: chat systems, whiteboarding, collaborative writing, workflow systems, and hypertext linking. Groupware packages are diverse in the



Rob: again!!!!!!!
Legacy.of.VorteX \|D: hi stassy>
Legacy.of.VorteX \|D: helo rob
Rob: <-----runs to Av
AV: they all leave ya robbie
Legacy.of.VorteX \|D: ><A><V><
AV: always coming back to me after bots leave ya eh?
AV: >yes!
Rob: the nasty sconky bots will not play with me Av
AV: ooops no arm
stassy_ps: =))Rob,your heart was broken again...
AV: ><
revampith joined the room.
AV: rev!!!
Rob: >< gropes av again
AV: no groping!

Figure 8. Web-based chat system.

functions they offer. Some include group writing, chat and/or e-mail. Sophisticated workgroup systems allow users to define workflows so that data is automatically forwarded to appropriate people at each stage of a process.

Chat Systems. Chat systems enable a type of group communication in which people located in different geographical locations get together in a virtual room and interact with each other by typing text. Chat systems make it possible for many people to write messages in a public space or virtual room. As each person submits a message, it appears on the screen of the other users located in the same virtual room. Chat groups are usually formed via listing chat rooms by name, location, number of people, topic of discussion, and so on.

Recently, systems accessible on the World Wide Web became widely spread among chat users. These types of chat systems are referred to as Web-based chat because they are accessible using a typical browser. One example of Web-based chat can be found at Yahoo.com (see Fig. 8).

Compared to e-mail, a chat system is a real-time synchronous system, while e-mail is neither real-time nor synchronous. When a user types a comment in a chat system, it is seen almost immediately by the others users present in the same virtual room. All the users are connected to the system at the same time. With e-mail, on the other hand, the two parties involved in the exchange of a message do not need to be connected to the system at the same time. For example, when reading an e-mail message the person who writes it may or may not be sitting in front of their computer at that time.

Whiteboard. A whiteboard provides real-time communication over the Internet and has a visual or graphical component in addition to text-based communication. Using a whiteboard, multiple users can simultaneously review, create, and update documents, images, graphs, equations, text, and information. All changes made by one user to the whiteboard area are displayed to all the other whiteboard users. The whiteboard allows participants to manipulate the contents by clicking and dragging with the mouse. In addition, they can use a remote pointer or highlighting tool to point out specific contents or sections of shared pages.

Most whiteboards are designed for informal conversation, but they may also serve structured communications or more sophisticated drawing tasks, such as collaborative graphic design, publishing, or engineering applications. For example, executives can meet and collaborate on slides for a presentation and architects can revise building plans.

Collaborative Writing Systems. Collaborative writing systems are applications that aim to help the joint editing of text documents by several authors. Coauthors, spread out across different network locations, can work together sharing common documents. When the interactions happen at the same time, they are called synchronous or real-time interactions. Otherwise, they are called asynchronous or non-real-time interactions.

Word processors may provide asynchronous support by showing authorship and by allowing users to track changes and make annotations to documents. It is possible to determine that only certain sections of documents may be modified by specific people to better protect how documents are modified and reduce the number of conflicting comments received. Reviewers can be prevented from making changes unless they turn revision marks on.

Workflow Systems. Workflow management systems (WfMS) appeared in the 1980s, but there is some consensus that the office information systems field is the predecessor of workflow systems (15). Advances in transaction processing and integrated office systems made workflow systems popular in the 1990s. They were innovative and had gained a high level of popularity. Commercial products include IBM MQSeries Workflow, Staffware, TIBCO InConcert, and COSA Workflow. General information on WfMSs can be found at the web sites of the Workflow and Reengineering International Association (16) and the Workflow Management Coalition (17).

A WfMS is implemented in accordance with a business process specification and execution paradigm. Under a WfMS, a workflow model is first created to specify organizational business processes, and then workflow instances are created to carry out the actual steps described in the workflow model. During the workflow execution, the workflow instances can access legacy systems, databases, applications, and can interact with users.

Workflow systems have been installed and deployed successfully in a wide spectrum of organizations. Most workflow management systems, both products and research prototypes, are rather monolithic and aim at providing fully fledged support for the widest possible application spectrum. The same workflow infrastructure can be deployed in various domains, such as bioinformatics, healthcare, telecommunications, military, and school administration.

In Fig. 9, a workflow process from the field of genomics exemplifies how workflow systems can be used to design business processes.

A major task in genomics is determining the complete set of instructions for making an organism. Genome projects are very demanding, and incur high costs of skilled manpower. There are many different types of tasks that must be performed, such as sequencing, sequence finishing, sequence processing, data annotation, and data submission. A single genomic workflow may be spread across multiple research centers, and the individual tasks in a workflow may be carried out at one or more of the participating centers. Many of the challenges of building an information system to manage a physically distributed genome project can be addressed by a workflow system.

The workflow model for such a workflow graphically specifies the control and data flow among tasks. For example, the workflow model in Fig. 9 is composed of several tasks and subworkflows. The tasks illustrated with machine gears represent automatic tasks, while the ones illustrated with boxes represent subworkflows.

At runtime, the workflow system reads the model specifications and transparently schedules task executions, providing the right data at the right time to the right worker. It manages distributed genomic tasks located at different research centers, such as DNA sequencing machines, matching algorithms, and human resources. Further, the workflow system provides a framework to easily reengineer a genomic workflow when new technological, biological, and chemical advances are made.

Teleconferencing

The term teleconferencing refers to a number of technologies that allow communication and collaboration among people located at different sites. At its simplest, a teleconference can be an audio conference with one or both ends of the conference sharing a speakerphone. With considerably more equipment and special arrangements, a teleconference can be a conference, called a videoconference, in which the participants can see still or motion video images of each other. Using teleconferencing systems, organizations can decrease costs and complexity, while increasing efficiency and productivity.

Audio Conferencing. Audio conferencing is the interaction between groups of people in two or more sites in real time using high quality, mobile, hands-free telephone technology. The interaction is possible with an audio connection via a telephone or network connection. It makes use of conventional communication networks such as POTS (Plain Old Telephone Service), ISDN (Integrated Services Digital Network), and the Internet.

Data Conferencing. Data conferencing is the connection of two or more computer systems, allowing remote groups to view, share, and collaborate on prepared documents or information. Data conferencing platforms make it possible to share applications and files with people in other locations. Everyone can see the same document at the same time and instantly view any changes made to it.

A user can share any program running on one computer with other participants in a conference. Participants can watch as the person sharing the program works, or the

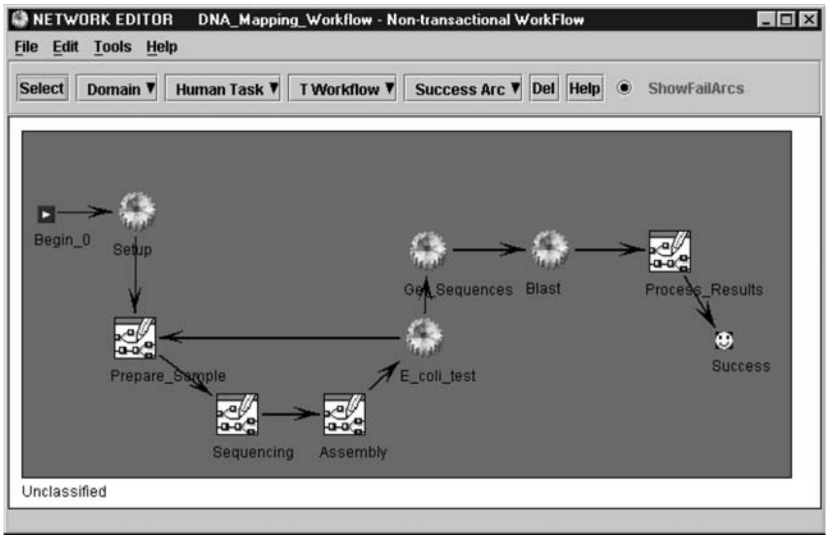
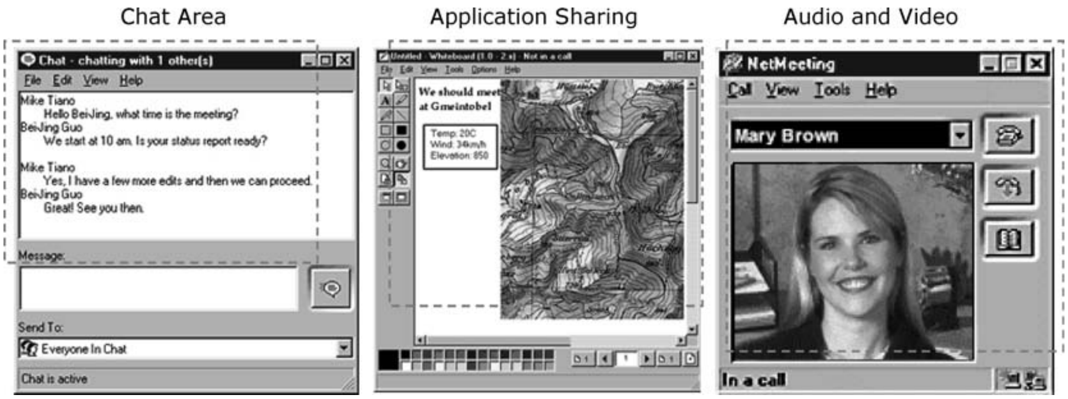


Figure 9. Genomic workflow example.



Mike Tiano
Hello BeiJing, what time is the meeting?
BeiJing Guo
We start at 10 am. Is your status report ready?

Mike Tiano
Yes, I have a few more edits and then we can proceed.
BeiJing Guo
Great! See you then.

We should meet at Gmeintobel

Temp: 20C
Wind: 34km/h
Elevation: 850

Figure 10. Microsoft NetMeeting with data, audio, and video conferencing (18).

person sharing the program can allow program control to other meeting participants.

Participants in a data conference can use a chat application to communicate in the absence of audio support. Chat can also be used to type text messages to share common ideas or topics with other conference participants or record meeting notes as part of a collaborative process.

Video Conferencing. Video conferencing uses telecommunications of audio and video to bring geographically dispersed people at different sites together for a meeting. Video conferencing is a valuable strategic tool for millions of individuals and small businesses for face-to-face meetings, team collaborations, brainstorming and training. There are two types of video conferencing: point-to-point and multipoint.

Point-to-point. A point-to-point conference is a connection between two video terminals. Each participant has a video camera, microphone, and speakers connected to a computer. As the two participants speak to one another, their voices are carried over the network and delivered to the other speakers, and images that appear in front of the video camera appear in a window on the other participants' monitor. Connecting two locations can be arranged simply by having one location dial the other, just as in a regular telephone call. No outside assistance is necessary.

Multipoint. A multipoint conference involves a connection among several video terminals connecting several sites with more than one person at different sites. This type of connection requires the assistance of a service to bridge the sites together into one conference. Until the mid-1990s, hardware costs made video conferencing prohibitively expensive for most organizations, but that situation is changing rapidly.

A video conference can involve just video, but some systems combine video, audio and data to provide a complete conferencing solution. One of the first and most popular video conferencing systems is NetMeeting (18). A product developed by Microsoft Corporation that enables groups to teleconference using the Internet as the transmission medium. NetMeeting (Fig. 10) supports video and audio conferencing, chat sessions, a whiteboard, and application sharing.

NetMeeting has been updated and extended with significant new capabilities designed to make it the most effective online meeting solution for integrated, interactive, and easy-to-use conferencing. The new version of this powerful application is now called Live Meeting (19).

Another well-known video conferencing program to transmit audio and video signals is CU-SeeMe. Originally developed by Cornell University, CU-SeeMe uses the standard bandwidth available on the Internet. Currently, CU-SeeMe is a low-cost software solution to the problem of electronic communication over the Internet.

IP Telephony. IP Telephony, also called 'Internet Telephony', allows voice and data to be transmitted over the same network using an open standards-based Internet Protocol (IP). It makes possible to exchange voice, fax, and other forms of information that have traditionally been carried over the dedicated circuit-switched connections of the public switched telephone network (PSTN). By combining different types of information on a single network connection, small and medium-sized businesses offices can decrease the costs of their voice and data networks.

IP Telephony is essential not just for its capability to reduce costs by combining voice and data communications, but also for its flexibility in supporting branch offices, mobile workers, and telecommuters that were not effective with PSTN. This technology allows an agile application deployment across the enterprise, increased personal and work group productivity, and permits a rapid return on investment.

CONCLUSIONS

Office Automation Systems specializes in allowing information workers to work, communicate, and collaborate. These systems are interactive and have the ability to allow workers to show and share documents or applications. These systems help workers worldwide to minimize the costs of business travel and streamline communications with co-workers, business partners, and customers.

Healthcare processes are very complex, involving both clinical and administrative tasks, large volumes of data, and a large number of patients and personnel. For example, an out-patient clinic visit involves administrative tasks performed by an assistant and clinical tasks performed by a doctor or by a nurse. For an in-patient hospital visit, this scenario involves more activities, and the process entails a duration that lasts at least as long as the duration of patient hospitalization. Healthcare processes are also very dynamic. As processes are instantiated, changes in healthcare treatments, drugs, and protocols may invalidate running instances, requiring reparative actions. Common problems reported by healthcare organizations include delays due to the lack of timely communication; time invested in completing and routing paper-based forms; errors due to illegible and incomplete patient information; frustration due to the amount of time spent on administrative tasks instead of patient interactions; long patient wait times caused by slow communication of patient information.

Office automation systems are a major asset to solve many of the problems identified by the healthcare community. For example, using Workflow management systems, paper forms can be easily converted into digital forms for use by caregivers. These electronic forms can be used throughout the patient care process from registration and triage to placing lab orders and charting treatment plans. These forms can be easily modified to accommodate changing business processes. By automating clinical forms processes and eliminating manual systems, caregivers can streamline patient information management and treatment flow. Workflow management systems can connect the data and processes in clinical forms with other systems, such as a lab or patient records system. As another example, using whiteboard technologies, caregivers and administrators can access a central location to view patient information and status including triage category, and lab order status. This level of access can help to quickly determine the next steps in each patient's care. Blogs can also be effectively used by healthcare professionals to discuss specific topics of interest, such as product reviews, scientific endeavors, patient's treatments, and any area of information where people have a deep expertise and a desire to express it.

BIBLIOGRAPHY

1. Zisman M. Representation, Specification and Automation of Office Procedures, Department of Business Administration, Wharton School. Philadelphia: University of Pennsylvania; 1977.
2. Ellis CA. Information Control Nets: A Mathematical Model of Office Information Flow. Conference on Simulation, Measurement and Modelling of Computer Systems. New York: ACM; 1979.
3. Notes. Lotus Notes. 2005. Available at http://www-130.ibm.com/developerworks/lotus/.
4. WFW. Windows for Workgroups. 2005. Available at http://www.microsoft.com/technet/archive/wfw/4_ch9.mspx. 2005.
5. Office. Microsoft Office. 2005. http://office.microsoft.com/. 2005.
6. Laudon JP, Laudon KC. Management Information Systems. 8th ed. New York: Prentice Hall; 2003.
7. Word. Microsoft Word. 2005. Available at http://office.microsoft.com/en-us/FX010857991033.aspx. 2005.
8. WordPerfect. Corel WordPerfect. 2005. http://www.corel.com/. 2005.
9. Lotus1-2-3. IBM Lotus 1-2-3. 2005. Available at http://lotus.com/products/product2.nsf/wdocs/123home. 2005.
10. Excel. Microsoft Excel. 2005. Available at http://office.microsoft.com/en-us/ FX010858001033.aspx. 2005.
11. PageMaker. Abode PageMaker. 2005. Available at http://www.adobe.com/products/pagemaker/main.html. 2005.
12. QuarkXPress. Quark. 2005. Available at http://www.quark.com/. 2005.
13. InDesign. Adobe InDesign. 2005. Available at http://www.adobe.com/products/indesign/main.html. 2005.
14. Publisher. Microsoft Publisher. 2005. Available at http://office.microsoft.com/en-us/ FX010857941033.aspx. 2005.
15. Stohr EA, Zhao JL. Workflow Automation: Overview and Research Issues. Information Systems Frontiers 2001;3(3): 281–296.
16. WARIA. Workflow and Reengineering International Association. 2002.
17. WfMC. Workflow Management Coalition. 2002.
18. NetMeeting, Microsoft NetMeeting. 2005. Available at http://www.microsoft.com/windows/netmeeting/. 2005.
19. LiveMeeting. Microsoft Live Meeting. 2005. Available at http://office.microsoft.com/en-us/FX010909711033.aspx 2005.

See also EQUIPMENT ACQUISITION; MEDICAL RECORDS, COMPUTERS IN.

OPTICAL FIBERS IN MEDICINE.

See FIBER OPTICS IN MEDICINE.

OPTICAL SENSORS

YITZHAK MENDELSON
Worcester Polytechnic Institute

INTRODUCTION

Optical sensing techniques have attracted extraordinary interest in recent years because of the key role they play in the development of medical diagnostic devices. Motivated by the expense and time constraints associated with traditional laboratory techniques, there is a growing need to continue and develop more cost-effective, simpler, and rapid methods for real-time clinical diagnostics of vital physiological parameters.

Optical sensors play a pivotal role in the development of highly sensitive and selective methods for biochemical analysis. The number of publications in the field of optical sensors used for biomedical and clinical applications has grown significantly during the past two decades. Numerous books, scientific reviews, historical perspectives, and conference proceedings have been published on biosensors including optical sensors and the reader interested in this rapidly growing field is advised to consult these excellent sources for additional reading (see Reading List). Some of these references discuss different optical sensors used in research applications and optical-based measurement techniques employed primarily in bench-top clinical analyzers. The emphasis of this article is on the basic concept employed in the development of optical sensors including specific applications highlighting how optical sensors are being utilized for real-time *in vivo* and *ex vivo* measurement of clinically significant biochemical variables, including some examples of optical sensor used for *in vitro* diagnosis. To narrow the scope, this article concentrates on those sensors that have generally progressed beyond the initial feasibility phase and have either reached or have a reasonable good potential of reaching the commercialization stage.

GENERAL PRINCIPLES OF OPTICAL BIOSENSING

The fundamental principle of optical sensors is based on the change in optical properties of a biological or physical medium. The change produced may be the result of the intrinsic changes in absorbance, reflectance, scattering, fluorescence, polarization, or refractive index of the biological medium. Optical sensors are usually based either on a simple light source–photodetector combination, optical fibers, or a planar waveguide. Some types of optical sensors measure changes in the intrinsic optical properties of a biological medium directly and others involve a specific indicator.

Biosensors are typically considered a separate subclassification of biomedical sensors. A biosensor, by definition, is a biomedical sensor consisting of an integrated biological component that provides the selectivity and a physical transducer to provide a solid support structure. Two major optical techniques are commonly available to sense optical changes at optical biosensor interfaces. These are usually based on evanescent wave, which was employed in the development of fiber optic sensors (see the section on Fiber Optic Sensors), and surface plasmon resonance principles, which played a pivotal role in the development and recent popularity of many optical biosensors. The basic principle of each measurement approach will be described first followed by examples arranged according to specific clinical applications.

Evanescent-Wave Spectroscopy

The propagation of light along a waveguide (e.g., a planar optical slab substrate or optical fiber) is not confined to the core region. Instead, when light travels through a waveguide at angles approaching the critical angle for total internal reflection, the light penetrates a characteristic short distance (on the order of one wavelength) beyond the core surface into the less optically dense (known as the cladding) medium as illustrated in Fig. 1. This effect causes the excitation of an electromagnetic field, called the "evanescent" wave, which depends on the angle of incidence and the incident wavelength. The intensity of the evanescent-wave decays exponentially with distance, starting at the interface and extending into the cladding medium.

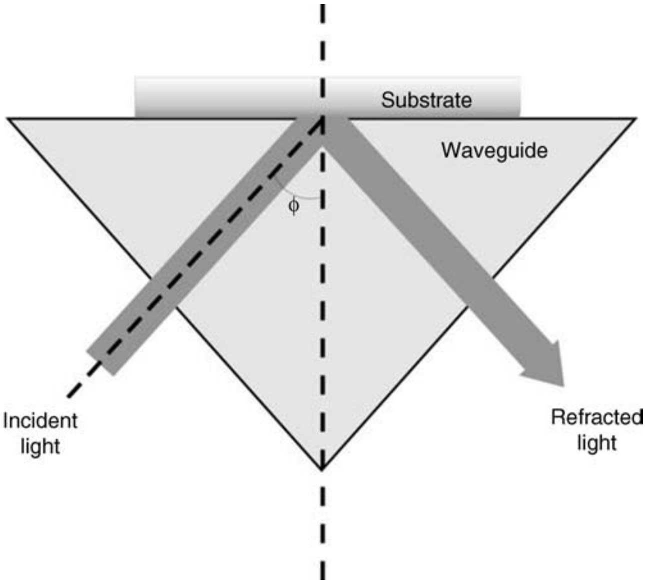


Figure 1. Principal diagram illustrating evanescence-wave spectroscopy sensing. Part of the incident light traveling through the waveguide at the critical angle (ϕ) penetrates a short distance into the substrate to be sensed and the remaining light is refracted.

The evanescent-wave can interact with molecules that are present within the penetration depth distance. This interaction causes attenuation of the incident light intensity and is related to the concentration of the molecules. This phenomenon can be exploited in the development of optical biosensors. For example, if the cladding is stripped and a substrate (such as a ligand) is immobilized on the core, the light will travel through this layer into the sample medium. Reactions close to the interface will perturb the evanescent field and the change in signal can be related to the amount of binding between the target and immobilized ligand at the interface. The measured parameter may be absorbance, fluorescence, or refractive index.

The method was first used as a means to study ultrathin films and coatings, and later was widely exploited to construct different types of optical sensors for biomedical applications. Because of the short penetration depth and the exponentially decaying intensity, the evanescent wave is absorbed by compounds that must be present very close to the surface. The principle can be utilized to characterize interactions between receptors that are attached to the surface of the optical sensor and ligands that are present in the solution probed by the sensor.

The key component in the successful implementation of evanescent-wave spectroscopy is the interface between the sensor surface and the biological medium. Receptors must retain their native conformation and binding activity and sufficient binding sites must be present for multiple interactions with the analyte. In the case of analytes having weakly optical absorbing properties, sensitivity can be enhanced by combining the evanescent-wave principle with multiple internal reflections along the sides of an unclad portion of a fiber optic tip. Alternatively, instead of an absorbing species, a fluorophore can be coated onto the uncladded fiber. Light propagating along the fiber core is partially absorbed by the fluorophore, emitting detectable fluorescent light at a higher wavelength and thus providing improved sensitivity.

Surface Plasmon Resonance

When monochromatic polarized light (e.g., from a laser source) impinges on a transparent medium having a conducting metalized surface (e.g., Ag or Au), there is a charge density oscillation at the interface. When light at an appropriate wavelength interacts with the dielectric-metal interface at a defined angle, called the resonance angle, there is a match of resonance between the energy of the photons and the electrons at the metal interface. As a result, the photon energy is transferred to the surface of the metal as packets of electrons, called plasmons, and the light reflection from the metal layer will be attenuated. This results in a phenomenon known as surface plasmon resonance (SPR) as illustrated schematically in Fig. 2. The resonance is observed as a sharp dip in the reflected light intensity when the incident angle is varied. The resonance angle depends on the incident wavelength, the type of metal, polarization state of the incident light, and the nature of the medium in contact with the surface. Any change in the refractive index of the medium will produce a shift in the resonance angle, thus providing a highly sensitive means of monitoring surface interactions.

The SPR is generally used for sensitive measurement of variations in the refractive index of the medium immediately surrounding the metal film. For example, if an antibody is bound to or adsorbed into the metal surface, a noticeable change in the resonance angle can be readily observed because of the change of the refraction index at the surface assuming all other parameters are kept constant. The advantage of this concept is the improved ability to detect the direct interaction between antibody and antigen as an interfacial measurement.

Optical Fibers

An optical fiber consists of two parts: a core (typically made of a thin glass rod) with a refractive index n_1, and an outer layer (cladding) with a refractive index n_2, where $n_1 > n_2$. The refractive indexes of the core and cladding depend on their material properties. Optical fibers are based on the principle of total internal reflection where incident light is transmitted through the core of the fiber with very little loss. The light strikes the cladding at an angle greater than the so-called critical angle, so that it is totally internally reflected at the core–cladding interface.

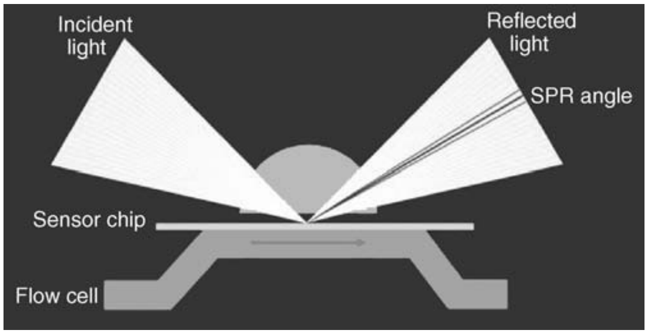


Figure 2. Principle of a SPR detection system. Courtesy of Biacore AB. An increased sample concentration causes a corresponding increase in refractive index that in turn alters the angle of incidence required to create the SPR angle.

Several types of biomedical measurements can be made using either a plain optical fiber as a remote device for detecting changes in the intrinsic spectral properties of tissue or blood, or optical fibers tightly coupled to various indicator-mediated transducers. The measurement relies either on direct illumination of a sample through the end-face of the fiber or by excitation of a coating on the sidewall surface through evanescent wave coupling. In both cases, sensing takes place in a region outside the optical fiber itself. Light emanating from the fiber end is scattered or fluoresced back into the fiber, allowing measurement of the returning light as an indication of the optical absorption or fluorescence of the sample at the fiber tip.

A block diagram of a generic instrument for an optical fiber-based sensor is illustrated in Fig. 3. The basic building blocks of such an instrument are the light source, various optical coupling elements, an optical fiber guide with or without the necessary sensing medium incorporated at the distal tip, and a photodetector.

Probe Configurations

A number of different methods can be used to implement fiber optic sensors. Most fiber optic chemical sensors employ either a single fiber configuration, where light travels to and from the sensing tip in one fiber, or a double-fiber configuration, where separate optical fibers are used for illumination and detection. A single fiber optic

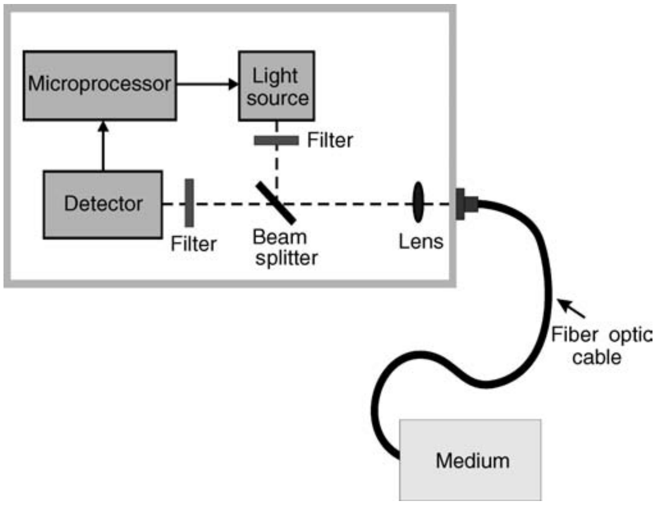


Figure 3. General building blocks of a fiber optic-based instrument for biomedical sensing applications. Typically, a microprocessor is used to control the light intensity and can be used to select either a fixed or a range of wavelengths for sample illumination. The microprocessor is also used to process the output of the photodetector.

configuration offers the most compact and potentially least expensive implementation. However, additional challenges in instrumentation are involved in separating the returning illuminating from the incident light illumination.

Intravascular fiber optic sensors are introduced into a vessel via a catheter. The design of intravascular fiber optic sensors requires additional considerations related to the sterility and biocompatibility of the catheter. For example, intravascular fiber optic sensors must be sterilizable and their material has to be nonthrombogenic and resistant to platelet and protein deposition. Therefore, these catheters are typically made of materials covalently bound with heparin or antiplatelet agents. The catheter is normally introduced via venous or arterial puncture and a slow heparin flush is maintained while the catheter remains in the body for short-term sensing, typically only for a few hours.

Indicator-Mediated Transducers

Indicator-mediated transducers are based on the coupling of light to a specific recognition element so that the sensor can respond selectively and reversibly to a change in the concentration of a particular analyte. The problem is that only a limited number of biochemical analytes have an intrinsic optical absorption that can be measured directly by spectroscopic methods with sufficient selectivity. Other species, particularly hydrogen ions and oxygen, which are of primary interest in diagnostic applications, do not have an intrinsic absorption and thus are not suitable analytes for direct photometry. Therefore, indicator-mediated transducers have been developed using specific reagents that can be immobilized on the surface of an optical sensor. These transducers may include indicators and ionophores (i.e., ion-binding compounds) as well as a wide variety of selective polymeric materials.

Figure 4 illustrates typical indicator-mediated fiber optic sensor configurations. In Fig. 4a, the indicator is immobilized directly on a membrane positioned at the end of a fiber. An indicator can be either physically retained in position at the end of the fiber by a special permeable membrane (Fig. 4b), or a hollow capillary tube (Fig. 4c). Polymers are sometimes used to enclose the indicator and selectively pass the species to be sensed.

Advantages and Disadvantages of Optical Fiber Sensors

Advantages of fiber optic sensors include their small size and low cost. In contrast to electrical measurements, fiber optic are self-contained, and therefore do not require an external reference signal from a second electrode. Because the signal that is transmitted is an optical signal, there is no electrical risk to the patient and the measurement is immune from interference caused by surrounding electric or magnetic fields. This makes fiber optic sensors very attractive for applications involving intense electromagnetic or radiofrequency fields, for example, near a magnetic resonance imaging (MRI) system or electrosurgical equipment. Chemical analysis can be performed in real-time with almost an instantaneous response. Furthermore, versatile sensors can be developed that respond to multiple analytes by utilizing multiwavelength measurements.

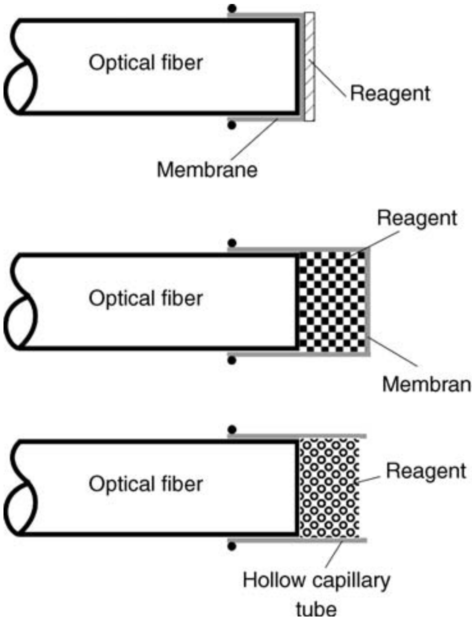


Figure 4. Typical configuration of different indicator-mediated fiber optic sensor tips showing different methods for placement of a reagent. Reprinted with permission from Fiber optic chemical sensors and biosensors, Vol. I, Wolfbeis OS, editor, CRC Press, Boca Raton, FL, c1991.

Despite unique advantages and promising feasibility studies, optical fiber sensors exhibit several shortcomings. Sensors with immobilized dyes and other indicators have limited long-term stability *in vivo* and their shelf-life degrades significantly over time.

INSTRUMENTATION

Instrumentation for optical measurements generally consist of (1) a light source to illuminate the medium, (2) optical components to generate a light beam with specific characteristics and to direct this light to some sensing agent or physical compartment, and (3) a photodetector for converting the optical signal to an electrical signal. The actual implementation of instrumentation designed to construct or interface with optical sensors vary greatly depending on the type of optical sensor used and its intended medical application.

Light Sources

Wide selections of light sources are available commercially for use in optical sensor applications. These include: narrow-band semiconductor diode lasers, broad spectral band incandescent lamps, and solid-state light emitting diodes (LEDs). The important requirement of a light source is obviously good stability. In certain applications, for example, in portable instrumentation, LEDs have significant advantages over other light sources since they are small, inexpensive, consume less power, produce selective wavelengths, and are easy to work with. In contrast, tungsten lamps produce a broader range of wavelengths, have higher intensities, but require a stable and sizable power supply. Furthermore, incandescent lamps produce significant heat that can degrade or destroy delicate biological samples so special provisions must be made for thermal dissipation of excessive heat.

Optical Elements

Different optical elements can be used to manipulate light in optical instrumentation. These may include mirrors, lenses, light-choppers, beam-splitters, and couplers for directing the light from the light source to the measurement site and back to the photodetector. If special waveguides are involved, additional components may be required to direct the light into the small aperture of an optical fiber or a specific area on a waveguide surface, and collecting the light from the sensor before it is processed by the photodetector. For a narrow wavelength selection when a broad bandwidth incandescent light source is utilized, optical filters, prisms, or diffraction gratings are the most common used components.

Photodetectors

Several specifications must be considered in choosing photodetectors for optical sensors. These include: detector sensitivity, noise level, spectral response, and response time. The most popular photodetectors employed in biomedical sensors are semiconductor quantum photodetectors, such as Silicon photodiodes. The choice, however, is somewhat dependent on the wavelength region of interest. Often, dual photodetectors are used in spectrophotometric instrumentation because it is frequently necessary to include a separate reference photodetector to track fluctuations in source intensity and temperature. By taking a ratio between the reference photodetector, which measures part of the light that is not affected by the measurement, and the primary photodetector, which measures the light intensity that interacts with the analyte, it is possible to obtain a more accurate and stable measurement.

Signal Processing

The signal obtained from a photodetector typically provides a current or voltage proportional to the measured light intensity. Therefore, simple analog circuitry (e.g., a current-to-voltage converter or connection to a programmable gain voltage amplifier) is required. Sometimes, the output from a photodetector is connected directly to a preamplifier before it is applied to a sampling and analog-to-digital conversion circuitry. More recently, advanced Sigma-Delta type analog-to-digital converters became available commercially that can accept input current directly from a photodiode, thus eliminating the need for a separate current-to-voltage converter stage.

Frequently, two different wavelengths of light are utilized to perform a specific measurement. One wavelength is usually sensitive to changes in the analyte being measured, while the other wavelength is unaffected by changes in the analyte concentration. In this manner, the unaffected wavelength is used as a reference to compensate for fluctuations in instrumentation properties over time. In other applications, additional discriminations, such as pulse excitation or electronic background subtraction utilizing a synchronized lock-in amplifier, are useful to gain significant improvement in selectivity and sensitivity.

IN VIVO APPLICATIONS

Glucose Sensors

Research has shown that tightly controlling blood sugar levels can prevent or slow down the development of problems related to diabetes. Presently, the conventional method for measuring blood glucose requires a drop of blood and relies on an electrochemical reaction of glucose with a glucose-specific enzyme such as glucose oxidase. During the past 20 years, numerous attempts have been made to develop optical sensors for continuous invasive and noninvasive measurement of blood glucose. The main driving forces for developing a blood glucose sensor is to enable the development of a closed-loop artificial endocrine pancreas for optimizing the treatment of diabetes. Continuous monitoring of blood glucose would provide the patient more useful information on daily fluctuations in glucose levels, will increase patient's motivation and compliance for daily self-monitoring, and would aid in the optimization of insulin therapy resulting in better metabolic control. If perfected, such a system could provide reliable early warning of large excursions in blood glucose levels that may lead to hypo- and hyperglycemic conditions. Therefore, it would be valuable in preventing long-term complications associated with diabetes, such as coronary artery disease, neuropathy, retinopathy, and hypertension.

A fiber optic sensor for measuring blood glucose *in vivo* utilizing the concept of competitive binding was described by Schultz et al. (1). The idea was based on an analyte (glucose) that competes for binding sites on a substrate (the Lectin Concanavalin A) with a fluorescent indicator-tagged polymer [fluorescein isothiocyanate (FITC)-dextran]. The sensor was arranged so that the substrate is fixed in a position out of the optical path of the fiber end. The substrate is bound to the inner wall of a glucose-permeable hollow fiber tubing and fastened to the end of an optical fiber. The hollow fiber acts as the container and is impermeable to the large molecules of the fluorescent indicator. The light beam that extends from the fiber "sees" only the unbound indicator in solution inside the hollow fiber, but not the indicator bound on the container wall. Excitation light passes through the fiber and into the solution, causing the unbound indicator to fluoresce, and the fluorescent light passes back along the same fiber to a measuring system. The fluorescent indicator and the glucose are in competitive binding equilibrium with the substrate. The interior glucose concentration equilibrates with its concentration exterior to the probe. If the glucose concentration increases, the indicator is driven off the substrate to increase the concentration of the indicator. Thus, fluorescence intensity as seen by the optical fiber follows changes in the glucose concentration. *In vivo* studies demonstrated fairly close correspondence between the sensor output and actual blood glucose levels. Another novel approach based on fluorescent molecules was suggested by Pickup et al.(2). The idea relied on the covalent binding of a fluorescent dye to glucose that results in a reduction of its fluorescence intensity when excited by light.

Although the measurement of glucose in plasma and whole blood *in vitro* is feasible (3), a more attractive approach for measuring blood glucose involves noninvasive measurement (4–6). The basic premise is to direct a light beam through the skin or laterally through the eye and analyze either the backscattered or transmitted light intensity. Three methods have been commonly attempted based either or changes in light absorption, polarization, or light scattering induced by variations in blood glucose. Although light in the visible and lower part of the near-infrared (NIR) region of the spectrum (700–2400 nm) can penetrate safely down to the vascular layer in the skin without significant attenuation there are major obstacles associated with noninvasive glucose measurement using spectrophotometry. Specifically, the concentration of glucose in tissue and blood is relatively low and light absorption in the NIR region is profoundly dependent on the concentration of water and temperature. Moreover, glucose has no unique absorption peaks in the visible or NIR region of the spectrum. Therefore, physiological variations in blood glucose induce only small and nonspecific changes in backscattered light intensity. Since measurements in the NIR region are due to low energy electronic vibrations, as well as high order overtones of multiple bands, NIR spectroscopy remains purely empirical. Thus, to extract accurate quantitative information related to variations in blood glucose, it is necessary to employ multivariate statistical calibration techniques and extensive chemometric analysis.

In principle, changes in light scattering caused by variations in blood glucose may offer another potential method for measuring glucose noninvasively. The fundamental principle exploited assumes that the refractive index of cellular structures within the skin remains unchanged while an increase in blood glucose leads to a subsequent rise in the refractive index of the blood and interstitial fluid. Limited studies involving glucose tolerance tests in humans (7,8) showed that changes in blood glucose could be measured from changes in light scattering. However, obtaining reliable and accurate measurement of blood glucose noninvasively using NIR spectroscopy remains challenging, mainly because other blood analytes (proteins, urea, cholesterol, etc.), as well as confounding physiological factors such as variations in blood flow, temperature, water content, and physical coupling of the sensor to the skin, are known to influence the measurement.

The implementation of an artificial pancreas would ultimately represent a major technological breakthrough in diabetes therapy. To date, however, inadequate specificity and insufficient accuracy within the clinically relevant range provide major obstacle in achieving this milestone with noninvasive blood glucose sensors using optical means.

Oximetry

Oximetry refers to the colorimetric measurement of the degree of oxygen saturation (SO_2), that is, the relative amount of oxygen carried by the hemoglobin in the erythrocytes. The measurement is based on the variation in the color of deoxyhemoglobin (Hb) and oxyhemoglobin (HbO_2). A quantitative method for measuring blood oxygenation is of great importance in assessing the circulatory and respiratory status of a patient.

Various optical methods for measuring the oxygen saturation in arterial (SaO_2) and mixed-venous (SvO_2) blood have been developed, all based on light transmission through, or reflection from, tissue and blood. The measurement is performed at two specific wavelengths: λ_1, where there is a large difference in light absorbance between Hb and HbO_2 (e.g., 660 nm red light), and a second wavelength, λ_2, which can be an isobestic wavelength (e.g., 805 nm IR light), where the absorbance of light is independent of blood oxygenation, or a higher wavelength in the near-IR region, typically between 805 and 960 nm, where the absorbance of Hb is slightly smaller than that of HbO_2.

The concept of oximetry is based on the simplified assumption that a hemolyzed blood sample consists of a two-component homogeneous mixture of Hb and HbO_2, and that light absorbance by the mixture of these two components is additive. Hence, a simple quantitative relationship can be derived for computing the oxygen saturation of blood based on the relationship:

$$SO_2 = K_1 - K_2[\mathrm{OD}(\lambda_1)/\mathrm{OD}(\lambda_2)]$$

where K_1 and K_2 are empirically derived coefficients that are functions of the specific absorptivities (also called optical extinction) of Hb and HbO_2, and OD (optical density) denotes the corresponding absorbance of the blood at a specific wavelength.

Since the original discovery of this phenomenon > 50 years ago (9), there has been progressive development in instrumentation to measure SO_2 along three different paths. Bench-top oximeters for clinical laboratories, which measure the concentration of Hb and HbO_2 from a small sample of arterial blood, fiber optic catheters for intravascular monitoring, and transcutaneous sensors, which are noninvasive devices placed on the skin.

Mixed-Venous Fiber Optic Catheters

Fiber optic oximeters for measuring mixed-venous oxygen saturation (SvO_2) were first described in the early 1960s by Polanyi and Hehir (10). They demonstrated that in a highly scattering medium, such as blood, it is feasible to use reflectance measurement to determine SO_2 in a flowing blood medium. Accordingly, they showed that a linear relationship exists between SO_2 and the ratio of the infrared-to-red (IR/R) light that is backscattered from blood:

$$SO_2 = A - B\,(\mathrm{IR/R})$$

where, A and B are empirically derived calibration coefficients.

The ability to rely on light reflectance for measurement of SO_2 *in vivo* subsequently led to the commercial development of fiber optic catheters for intravascular monitoring of SvO_2 inside the pulmonary artery.

Under normal conditions, oxygen consumption is less than or equal to the amount of oxygen delivered. However, in critically ill patients, oxygen delivery is often insufficient for the increased tissue demands, because many such patients have compromised compensatory mechanisms. If tissue oxygen demands increase and the body's compensatory mechanisms are overwhelmed, the venous oxygen reserve will be tapped, and that change will be reflected as a decreased SvO_2. Venous oxygen saturation in the pulmonary artery is normally ~ 75%. Although no specific SvO_2 level has been correlated with adverse physiological effects, an SvO_2 level of 53% has been linked to anaerobic metabolism and the production of lactic acid (11), while an SvO_2 level of 50% or less indicates that oxygen delivery is marginal for oxygen demands, and thus venous oxygen reserve is reduced. Thus, continuous SvO_2 monitoring can be used to track the available venous oxygen reserve.

A fiber optic pulmonary artery catheter, with its tip in the pulmonary artery, can be used to sample the outflow from all tissue beds. For this reason, SvO_2 is regarded as a reliable indicator of tissue oxygenation (12) and therefore is used to indicate the effectiveness of the cardiopulmonary system during cardiac surgery and in the ICU.

Fiber optic SvO_2 catheters consist of two separate optical fibers; one fiber is used for transmitting the light to the flowing blood and a second fiber directs the backscattered light to a photodetector. The catheter is introduced into the vena cava and further advanced through the heart into the pulmonary artery by inflating a small balloon located at the distal end. The flow-directed catheter also contains a small thermistor for measuring cardiac output by thermodilution.

Several problems limit the wide clinical application of intravascular fiber optic oximeters. These include the dependence of the optical readings on hematocrit and motion artifacts due to catheter tip "whipping" against the blood vessel wall. Additionally, the introduction of the catheter into the heart requires an invasive procedure and can sometimes cause arrhythmias.

Pulse Oximetry

Noninvasive monitoring of SaO_2 by pulse oximetry is a rapidly growing practice in many fields of clinical medicine (13). The most important advantage of this technique is the capability to provide continuous, safe, and effective monitoring of blood oxygenation.

Pulse oximetry, which was first suggested by Aoyagi et al. (14) and Yoshiya et al. (15), relies on the detection of time-variant photoplethysmographic (PPG) signals, caused by changes in arterial blood volume associated with cardiac contraction. The SaO_2 is derived by analyzing the time-variant changes in absorbance caused by the pulsating arterial blood at the same R and IR wavelength used in conventional invasive-type oximeters. A normalization process is commonly performed by which the pulsatile (ac) component at each wavelength, which results from the expansion and relaxation of the arterial bed, is divided by the corresponding nonpulsatile (dc) component of the PPG, which is composed of the light absorbed by the blood-less tissue and the nonpulsatile portion of the blood compartment. This effective scaling process results in a normalized R/IR ratio, which is dependent on SaO_2, but is largely independent of the incident light intensity, skin pigmentation, tissue thickness, and other nonpulsatile variables.

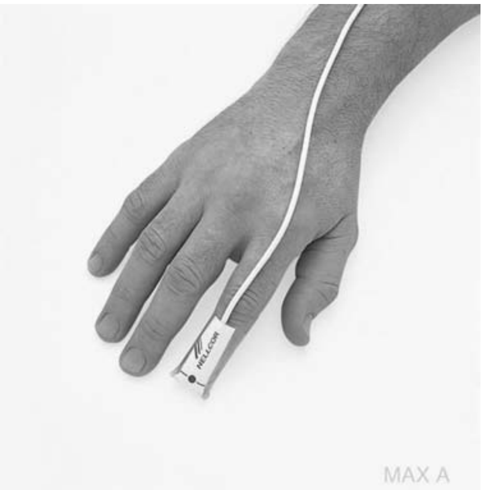


Figure 5. A disposable finger probe of a noninvasive transmission pulse oximeter. Reprinted by permission of Nellcor Puritan Bennett, Inc., Pleasanton, California. The sensor is wrapped around the fingertip using a self-adhesive tape backing.

Pulse oximeter sensors consist of a pair of small and inexpensive R and IR LEDs and a highly sensitive silicon photodiode. These components are mounted inside a reusable rigid spring-loaded clip, a flexible probe, or a disposable adhesive wrap (Fig. 5). The majority of the commercially available sensors are of the transmittance type in which the pulsatile arterial bed (e.g., ear lobe, fingertip, or toe) is positioned between the LEDs and the photodiode. Other probes are available for reflectance (backscatter) measurement, where both the LEDs and photodetector are mounted side-by-side facing the skin (16,17).

Numerous studies have evaluated and compared the accuracy of different pulse oximeters over a wide range of clinical conditions (18–21). Generally, the accuracy of most noninvasive pulse oximeters is acceptable for a wide range of clinical applications. Most pulse oximeters are accurate to within ± 2–3% in the SaO_2 range between 70 and 100%.

Besides SaO_2, most pulse oximeters also offer other display features, including pulse rate and analogue or bar graph displays indicating pulse waveform and relative pulse amplitude. These important features allow the user to assess in real time the quality and reliability of the measurement. For example, the shape and stability of the PPG waveform can be used as an indication of possible motion artifacts or low perfusion conditions. Similarly, if the patient's heart rate displayed by the pulse oximeter differs considerably from the actual heart rate, the displayed saturation value should be questioned.

Several locations on the body, such as the ear lobes, fingertips, and toes, are suitable for monitoring SaO_2 with transmission pulse oximeter sensors. The most popular sites are the fingertips since these locations are convenient to use and a good PPG signal can be quickly obtained.

Other locations on the skin that are not accessible to conventional transillumination techniques can be monitored using a reflection (backscatter) SaO_2 sensor. Reflection sensors are usually attached to the forehead or to the temples using a double-sided adhesive tape.

Certain clinical and technical situations may interfere with the proper acquisition of reliable data or the interpretation of pulse oximeter readings. Some of the more common problems are low peripheral perfusion associated, for example, with hypotension, vasoconstriction, or hypothermia conditions. Also, motion artifacts, the presence of significant amounts of dysfunctional hemoglobins (i.e., hemoglobin derivatives that are not capable of reversibly binding with oxygen), such as HbCO and methemoglobin, and intravenous dyes introduced into the blood stream (e.g., methylene blue).

Pulse oximetry is widely used in various clinical applications including anesthesia, surgery, critical care, hypoxemia screening, exercise, during transport from the operating room to the recovery room, in the emergency room, and in the field (22–26). The availability of small and lightweight optical sensors makes SaO_2 monitoring especially applicable for preterm neonates, pediatric, and ambulatory patients. In many applications, pulse oximetry has replaced transcutaneous oxygen tension monitoring in neonatal intensive care. The main utility of pulse oximeters in infants, especially during the administration of supplemental oxygen, is in preventing hyperoxia since high oxygen levels in premature neonates is associated with increased risk of blindness from retrolental fibroplasia.

During birth, knowing the blood oxygenation level of the fetus is of paramount importance to the obstetrician. Lack of oxygen in the baby's blood can result in irreversible brain damage or death. Traditionally, physicians assess the well being of the fetus by monitoring fetal heart rate and uterine contractions through the use of electronic fetal monitoring, which are sensitive, but generally not specific. Approximately one-third of all births in the United States are marked by a period in which a nonreassuring heart rate is present during labor. Without a reliable method to determine how well the fetus is tolerating labor and when dangerous changes in oxygen levels occur, many physicians turn to interventions, such as cesarean deliveries.

Recently, the FDA approved the use of new fetal oxygen monitoring technology, originally developed by Nellcor (OxiFirst, Tyco Healthcare). The pulse oximeter utilizes a disposable shoehorn-shaped sensor (Fig. 6), which is inserted through the birth canal during labor, after the amniotic membranes have ruptured and the cervix is dilated at least 2 cm. The sensor, which comprises the same optical components as other pulse oximeter sensors used for nonfetal applications, rests against the baby's cheek, forehead, or temple, and is held in place by the uterine wall. The fetus must be of at least 36 weeks gestation with the head in the normal vertex position for delivery. A controlled, randomized clinical trial, has demonstrated the safety and effectiveness of this new technology (27,28). While some clinical studies found that the rate of caesarean section for nonreassuring heart rate was significantly lower among the group of women monitored by the OxiFirst system, it is not yet clear whether the use of a fetal pulse oximeter may lead to a reduction in the number of Caesarean sections performed. Therefore, the American College of Obstetrics and Gynecology has not yet endorsed the use of this technology in clinical practice.

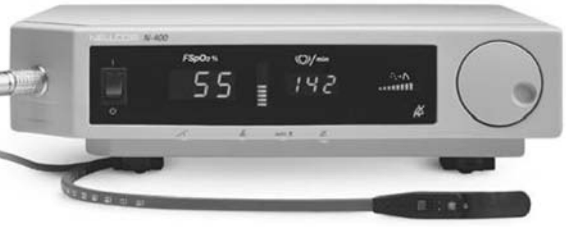


Figure 6. Optical probe and a fetal pulse oximeter. Reprinted by permission of Nellcor Puritan Bennett, Inc., Pleasanton, California. The shape of the optical sensor is contoured to enable proper placement of the sensor against the fetus face.

Noninvasive Cerebral Oximetry

Somanetics, Inc. (Troy, MI) has developed a noninvasive cerebral oximeter (INVOS) for monitoring of changes in brain oxygen saturation that can be used to alert clinicians to changes in the critical balance between arterial oxygen delivery and cerebral consumption (29–31). The method is based on NIR light photons injected by a light source into the skin over the forehead as illustrated in Fig. 7. After being scattered inside the scalp, skull, and brain, some fraction of the incident photons return and exit the skin. By measuring the backscattered light intensity as a function of two wavelengths, 730 and 810 nm, it is possible to

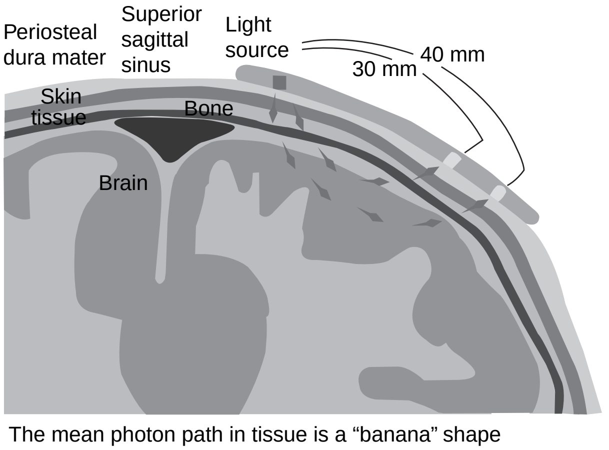


Figure 7. Principle of the INVOS cerebral oximeter. Courtesy of Somanetics, Corp, Troy, MI. Two photodetectors are used to capture the backscattered light from different depths in the brain.

measure changes in regional hemoglobin oxygen saturation (rSO_2 index).

The measurement is intuitively based on the fact that the greater the separation of source and detector, the greater the average depth of penetration. Photons that happen to meander close to the surface are very likely to be lost out of the skin before getting to a distant detector. Large source-detector separation is therefore biased against "shallow" photons except in the tissues directly under the optical sensor. On the other hand, geometry and absorption also make it statistically unlikely that very deeply penetrating photons will find their way back to the detector. Most of the photons reaching the detector tend to take some optimum middle course. This mean photon path is shaped approximately like a "banana" with ends located at the light source and photodetector.

To reduce extraneous spectroscopic interference that is dominated by light scattered by the surrounding bone, muscle, and other tissues, the INVOS SomaSensor® oximeter uses two source-detector separations: a "near" (shallow) spacing and a "far" (deep) spacing. The dual detectors sample about equally the shallow layers in the illuminated tissue volumes positioned directly under the light sources and photodetectors, but the far-spaced detector "sees" deeper than the near-spaced detector. By subtracting the two measurements, the instrument is able to suppress the influence of the tissues outside the brain to provide a measurement of changes in brain oxygen saturation.

Measurement of Blood Gases

Accurate measurement of arterial blood gases, that is, oxygen partial pressure (pO_2), carbon dioxide partial pressure (pCO_2), and pH, is one of the most frequently performed tests in the support of critically ill patients. The measurement is essential for clinical diagnosis and management of respiratory and metabolic acid–base problems in the operating room and the ICU. Traditionally, blood gases have been measured by invasive sampling, either through an indwelling arterial catheter or by arterial puncture, and analyzed in a clinical laboratory by a bench-top blood gas analyzer. However, this practice presents significant drawbacks: Sampling is typically performed after a deleterious event has happened, the measurement is obtained intermittently rather than continuously so physicians can not immediately detect significant changes in a patient's blood gas status, there could be a considerable delay between the time the blood sample is obtained and when the readings become available, there is an increased risk for patient infection, and there is discomfort for the patients associated with arterial blood sampling. Furthermore, frequent arterial blood gas sampling in neonates can also result in blood loss and may necessitate blood transfusions.

In view of the above drawbacks, considerable effort has been devoted over the last three decades to develop either disposable extracorporeal sensors (for *ex vivo* applications) or intraarterial fiber optic sensors that can be placed in the arterial line (for *in vivo* applications) to enable continuous trending that is vital for therapeutic interventions in ICU patients who may experience spontaneous and often unexpected changes in acid–base status. Thus, with the advent of continuous arterial blood gas monitoring, treatment modalities can be proactive rather than reactive. Although tremendous progress has been made in the miniaturization of intravascular blood gas and pH sensors, in order to be acceptable clinically, further progress must be achieved on several fronts. Specifically, there is a need to improve the accuracy and reliability of the measurement especially in reduced blood flow and hypotensive conditions. Additionally, sensors must be biocompatible and nonthrombogenic, the readings must be stable and respond rapidly to changes in physiological conditions, and the disposable sensors need to be inexpensive and more cost effective.

Intravascular Optical Blood Gas Catheters

In the early 1970s, Lubbers and Opitz (32) originated what they called "optodes" (from the Greek word '*optical path*') for measurements of important physiological gases in fluids and in gases. The principle upon which these sensors were designed originally was based on a closed compartment containing a fluorescent indicator in solution, with a membrane permeable to the analyte of interest (either ions or gases) constituting one of the compartment walls. The compartment was coupled by optical fibers to a system that measured the fluorescence inside the closed compartment. The solution equilibrates with the pO_2 or pCO_2 of the medium placed against it, and the fluorescence intensity of an indicator reagent in the solution was calibrated to the partial pressure of the measured gas.

Intraarterial blood gas optodes typically employ a single or a double fiber configuration. Typically, the matrix containing the indicator is attached to the end of the optical fiber. Since the solubility of O_2 and CO_2 gases, as well as the optical properties of the sensing chemistry itself, are affected by temperature variations, fiber optic intravascular sensors include a thermocouple or thermistor wire running alongside the fiber optic cable to monitor and correct for temperature fluctuations near the sensor tip. A nonlinear response is characteristic of most chemical indicator sensors, so they are designed to match the concentration region of the intended application. Also, the response time of optodes is somewhat slower compared to electrochemical sensors.

Intraarterial fiber optic blood gas sensors are normally placed inside a standard 20-gage arterial cannula, which is sufficiently small thus allowing adequate spacing between the sensor and the catheter wall. The resulting lumen is large enough to permit the withdrawal of blood samples, introduction of a continuous heparin flush, and the recording of a blood pressure waveform. In addition, the optical fibers are encased in a protective tubing to contain any fiber fragments in case they break off. The material in contact with the blood is typically treated with a covalently bonded layer of heparin, resulting in low susceptibility to fibrin deposition. Despite excellent accuracy of indwelling intraarterial catheters *in vitro* compared to blood gas analyzers, when these multiparameter probes were first introduced into the vascular system, it quickly became evident that the readings (primarily pO_2) varies frequently

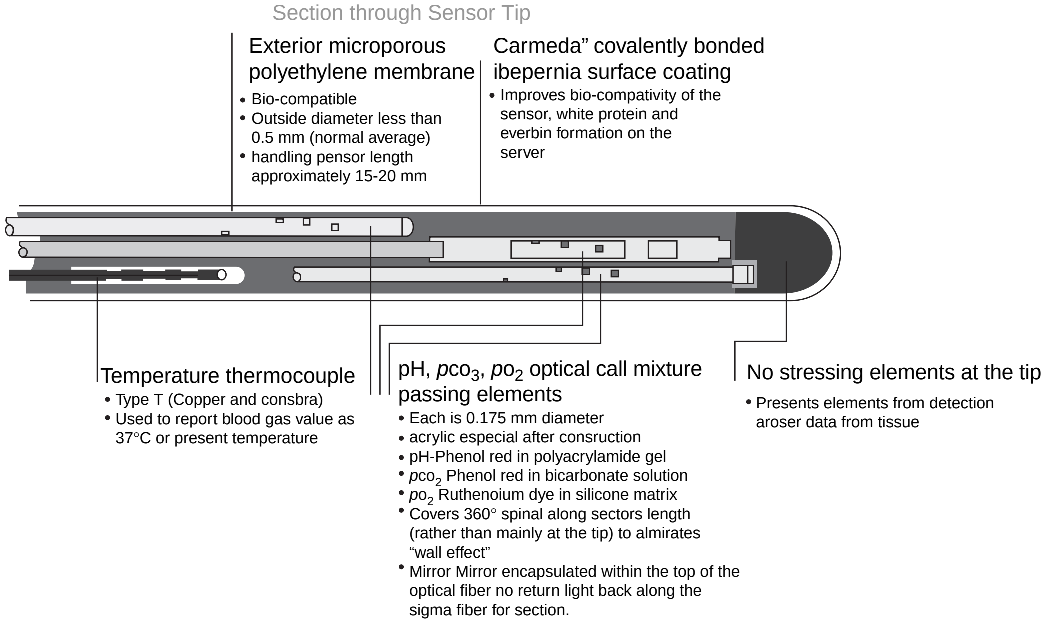

Section through Sensor Tip
Exterior microporous polyethylene membrane
• Bio-compatible
• Outside diameter less than 0.5 mm (normal average)
• handling pensor length approximately 15-20 mm
Carmeda" covalently bonded ibepernia surface coating
• Improves bio-compativity of the sensor, white protein and everbin formation on the server
Temperature thermocouple
• Type T (Copper and consbra)
• Used to report blood gas value as 37°C or present temperature
pH, pco3, po2 optical call mixture passing elements
• Each is 0.175 mm diameter
• acrylic especial after consruction
• pH-Phenol red in polyacrylamide gel
• pco2 Phenol red in bicarbonate solution
• po2 Ruthenoium dye in silicone matrix
• Covers 360° spinal along sectors length (rather than mainly at the tip) to almirates "wall effect"
• Mirror Mirror encapsulated within the top of the optical fiber no return light back along the sigma fiber for section.
No stressing elements at the tip
• Presents elements from detection aroser data from tissue

Figure 8. Principle of an indwelling arterial optical blood gas catheter. Courtesy of Diametrics, Inc, St Paul, MN. A heparin-coated porous polyethylene membrane encapsulates the optical fibers and a thermistor.

and unpredictably, mainly due to the sensor tip intermittently coming in contact with the wall of the arterial blood vessel and intermittent reductions in blood flow due to arterial vasospasm (33–35).

Recently, a more advanced multiparameter disposable probe (Fig. 8) consisting of a $p\text{O}_2$, $p\text{CO}_2$, and pH sensors was developed by Diametrics Medical, Inc. (36–38). Several technical issues discovered during clinical evaluations of earlier probes were successfully addressed by this new product, and multiple clinical studies confirmed that the system is adequate for trend monitoring. The device has been evaluated in neurosurgical patients for continuous monitoring in the brain and in critically ill pediatric patients (39–41).

Although significant progress has been made, most of the intravascular probes are adversely affected by patient–probe interfacing problems and high manufacturing cost. These problems must be fully overcome before continuous intravascular blood gas monitoring can be performed reliably and cost effectively in widely divergent patient groups. Continuous arterial blood gas monitoring will not completely replace traditional invasive blood sampling. Nevertheless, it may extend the clinician's ability to recognize and intervene more effectively to correct potentially life-threatening conditions.

$p\text{O}_2$ Sensors

The basic principle of measuring $p\text{O}_2$ optically relies on the fluorescence quenching effect of oxygen. Fluorescence quenching is a general property of aromatic molecules. In brief, when light is absorbed by a molecule, the absorbed energy is held as an excited electronic state of the molecule. It is then lost by coupling to the mechanical movement of the molecule (heat), reradiated from the molecule in a mean time of $\sim$ 10 ns (fluorescence), or converted into another excited state with much longer mean lifetime (phosphorescence). Quenching reduces the intensity of the emitted fluorescence light and is related to the concentration of the quenching molecules. The quenching of luminescence by oxygen was initially described by Kautsky in 1939 (42).

Opitz and Lubbers (43) and Peterson et al. (44) developed a fiber optic sensor for measuring $p\text{O}_2$ using the principle of fluorescence quenching. The dye is excited at $\sim$ 470 nm (blue) and fluoresces at $\sim$ 515 nm (green) with an intensity that depends on the $p\text{O}_2$. The optical information is derived from the ratio of light intensities measured from the green fluorescence and the blue excitation light, which serves as an internal reference signal. The original sensor contained a Perylene Dibutyrate dye contained in an oxygen-permeable porous polystyrene envelope (45). The ratio of green to blue intensity is processed according to the Stern–Volmer equation (46):

$$I_0/I = 1 + Kp\text{O}_2$$

where I and I_0 are the fluorescence emission intensities in the presence and absence (i.e., $p\text{O}_2 = 0$) of the quencher, respectively, and K is the Stern–Volmer quenching coefficient. The method provides a nearly linear readout of $p\text{O}_2$ over the range of 0–150 mmHg (0–20 kPa), with a precision of $\sim$ 1 mmHg (0.13 kPa). The original sensor was 0.5 mm in diameter its 90% response time in an aqueous medium was $\sim$ 1.5 min.

pH Sensors

Peterson and Goldstein et al.(47) developed the first fiber optic chemical sensor for physiological pH measurement by placing a reversible color-changing indicator at the end of a pair of optical fibers. In the original development, the popular indicator phenol red was used since this dye changes its absorption properties from the green to the blue part of the spectrum as the acidity is increased. The dye was covalently bound to a hydrophilic polymer in the form of water-permeable microbeads that stabilized the indicator concentration. The indicator beads were contained in a sealed hydrogen ion-permeable envelope made out of hollow cellulose tubing. In effect, this formed a miniature spectrophotometric cell at the end of the fibers.

The phenol red dye indicator is a weak organic acid, and its unionized acid and ionized base forms are present in a concentration ratio that is determined according to the familiar Henderson–Hasselbalch equation by the ionization constant of the acid and the pH of the medium. The two forms of the dye have different optical absorption spectra. Hence, the relative concentration of one of the forms, which varies as a function of pH, can be measured optically and related to variations in pH. In the pH sensor, green and red light emerging from the distal end of one fiber passes through the dye where it is backscattered into the other fiber by the light-scattering beads. The base form of the indicator absorbs the green light. The red light is not absorbed by the indicator and is therefore used as an optical reference. The ratio of green to red light is measured and is related to the pH of the medium.

A similar principle can also be used with a reversible fluorescent indicator, in which case the concentration of one of the indicator forms is measured by its fluorescence rather than by the absorbance intensity. Light, typically in the blue or ultraviolet (UV) wavelength region, excites the fluorescent dye to emit longer wavelength light. The concept is based on the fluorescence of weak acid dyes that have different excitation wavelengths for the basic and acidic forms but the same emitted fluorescent wavelength. The dye is encapsulated in a sample chamber that is permeable to hydrogen ions. When the dye is illuminated with the two different excitation wavelengths, the ratio of the emitted fluorescent intensities can be used to calculate the pH of the solution that is in contact with the encapsulated dye.

The prototype device for measuring pH consisted of a tungsten lamp for illuminating the optical fiber, a rotating filter wheel to select the red and green light returning from the fiber, and signal processing to provide a pH output based on the ratio of the green-to-red light intensity. This system was capable of measuring pH in the physiologic range between 7.0 and 7.4 with an accuracy and precision of 0.01 pH units. However, the sensor was susceptible to ionic strength variations.

Further development of the pH probe for practical use was continued by Markle et al. (48). They designed the fiber optic probe in the form of a 25-gage ($\phi = 0.5$ mm) hypodermic needle, with an ion-permeable side window, using 75-mm diameter plastic optical fibers. With improved instrumentation, and with a three-point calibration, the sensor had a 90% response time of 30 s and the range was extended to ±3 pH units with a precision of 0.001 pH units.

Different methods were suggested for fiber optic pH sensing (49–52). A classic problem with dye indicators is the sensitivity of their equilibrium constant to variations in ionic strength. To circumvent this problem, Wolfbeis and Offenbacher (53) and Opitz Lubber (54) demonstrated a system in which a dual sensor arrangement can measure ionic strength and pH while simultaneously correcting the pH measurement for ionic strength variations.

pCO$_2$ Sensors

The $p\text{CO}_2$ of a sample is typically determined by measuring changes in the pH of a bicarbonate solution. The bicarbonate solution is isolated from the sample by a CO_2-permeable membrane, but remains in equilibrium with the CO_2 gas. The bicarbonate and CO_2, as carbonic acid, form a pH buffer system and, by the Henderson–Hasselbalch equation,

$$\text{pH} = 6.1 + \log \frac{\text{HCO}_3^-}{p\text{CO}_2}$$

the hydrogen ion concentration is proportional to the $p\text{CO}_2$ of the sample. This measurement is done with either a pH electrode or a dye indicator. Both absorbance (55) and fluorescence (56) type $p\text{CO}_2$ sensors have been developed.

Vurek et al. (57) demonstrated that the same technique could be used also with a fiber optic sensor. In his design, one optical fiber carries light to the transducer, which is made of a silicone rubber tubing ~ 0.6 mm in diameter and 1.0 mm long, filled with a phenol red solution in a 35-mM bicarbonate. Ambient $p\text{CO}_2$ controls the pH of the solution that changes the optical absorption of the phenol red dye. The CO_2 permeates through the rubber to equilibrate with the indicator solution. A second optical fiber carries the transmitted signal to a photodetector for analysis. A different design by Zhujun and Seitz (58) used a $p\text{CO}_2$ sensor based on a pair of membranes separated from a bifurcated optical fiber by a cavity filled with bicarbonate buffer.

Extracorporeal Measurement

Several extracorporeal systems suitable for continuous online *ex vivo* monitoring of blood gases, base-access, and HCO_3 during cardiopulmonary bypass operations (Fig. 9) are available commercially (59–61). While this approach circumvents some of the advantages of continuous *in vivo* monitoring, this approach is useful in patients requiring frequent measurements of blood gases. The basic approach is similar to the optical method utilized in intravascular probes, but the sensors are located on a cassette that is placed within the arterial pressure line that is inserted into the patient's arm. The measurement is performed extravascularly by drawing a blood sample past the in-line sensor. After the analysis, which typically takes ~ 1 min, the sample can then be returned to the patient and the line is flushed. The sensor does not disrupt the pressure waveform or interfere with fluid delivery. Several clinical studies showed that in selected patient groups, the performance of these sensors is comparable to that of conventional *in vitro* blood gas analyzers (62–66).

Hematocrit Measurement

In recent years, an optical sensor has been developed by Hemametrics (Kaysville, UT) for monitoring hematocrit,

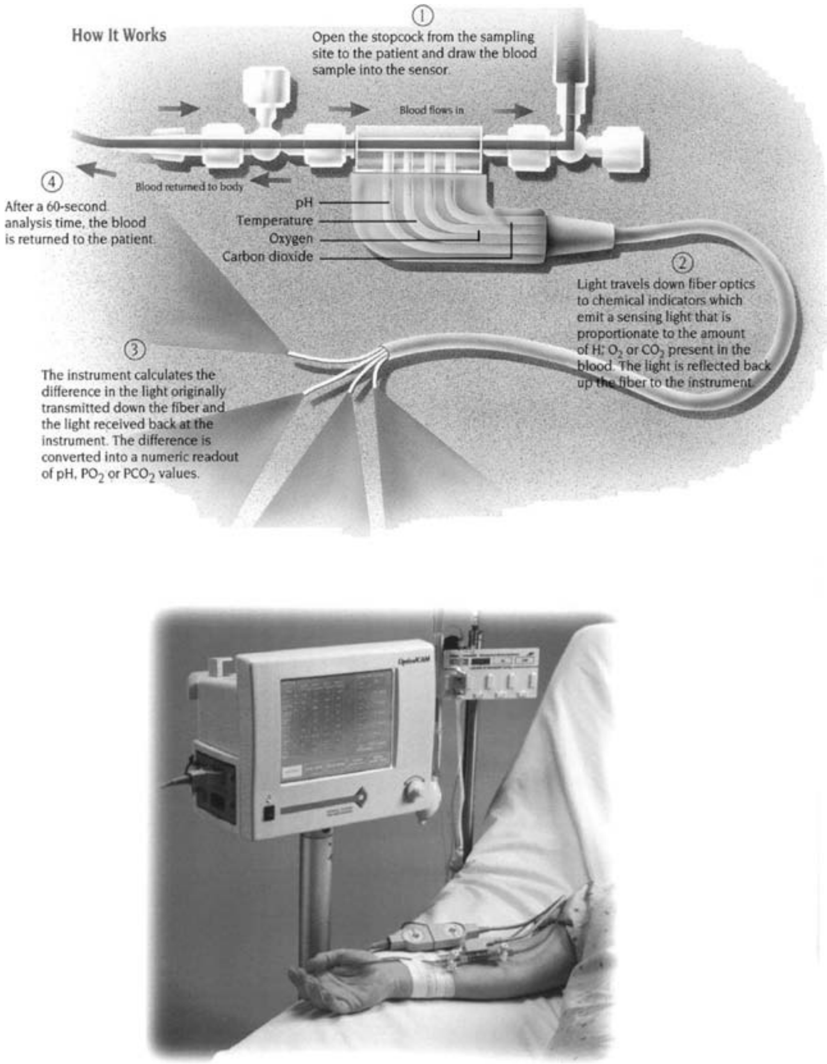

How It Works
①
Open the stopcock from the sampling site to the patient and draw the blood sample into the sensor.
Blood flows in
④
After a 60-second analysis time, the blood is returned to the patient.
Blood returned to body
pH
Temperature
Oxygen
Carbon dioxide
②
Light travels down fiber optics to chemical indicators which emit a sensing light that is proportionate to the amount of H; O_2 or CO_2 present in the blood. The light is reflected back up the fiber to the instrument.
③
The instrument calculates the difference in the light originally transmitted down the fiber and the light received back at the instrument. The difference is converted into a numeric readout of pH, PO_2 or PCO_2 values.

Figure 9. Principle of the GEM SensiCath fiber optic-based blood gas system. Courtesy of Optical Sensors, Inc., Minneapolis, MN.

oxygen saturation and blood volume *ex vivo* during hemodialysis (67–69). It provides vital and real time trending information that can be used to alert clinicians instantaneously to rapid changes during renal therapy (e.g., gastrointestinal bleeding or hemolysis). The optical measurements are based on detecting variations in scattering and absorption by blood flowing through a disposable cuvette that is connected to the arterial side of the dialyzer. A noninvasive version of this technology (CritScan) was recently cleared by the FDA for measuring absolute hematocrit and oxygen saturation by resting a fingertip against an optical sensor array of LEDs and photodetectors. One potential applications of this device, which was developed for spot checking, is for use in blood banking as an anemia screening tool to qualify blood donors, eliminating the need for performing hematocrit measurement based on the traditional and painful needle-stick approach.

Transcutaneous Bilirubin Measurement

Jaundice (hyperbilirubinemia) in newborn babies is evident in $> 50\%$ of newborns during the first week of birth. While hyperbilirubinemia (i.e., total serum bilirubin > 1.0 mg/dL) occurs in nearly all infants, moderate to significant hyperbilirubinemia usually peaks between 3 and 7 days of age and typically requires phototheraphy and/or exchange blood transfusions. If the condition is not recognized and treated properly, it can lead to potentially irreversible bilirubin-induced neurotoxicity and neurologic dysfunction. Visual recognition of jaundice is often inaccurate and unreliable, but excessive hyperbilirubinemia can be diagnosed by laboratory-based assay of total serum bilirubin obtained from a heel stick or an umbilical line.

To achieve a more objective measurement of jaundice, Yamanouchi et al. (70) developed a hand-held transcutaneous bilirubinometer developed by the Minolta Company. The meter used a dual optical filter design to measure a hemoglobin-corrected yellow color that is a distinctive skin color in jaundice patients. The measurement gives a reflectance value that must first be correlated to patient's serum bilirubin. Clinical testing of this device revealed significant inaccuracies and patient dependent variability compared with serum bilirubin determinations due to the presence of melanin pigmentation in the skin and variations in hemoglobin content.

More recently, several companies introduced more advance hand-held devices that uses multiwavelengths to measure bilirubin based on spectral analysis of light reflectance from the skin pioneered by Jacques (71) and others (72–74). The device consists of a light source, a fiberoptic probe, and a photodetector, essentially functioning as a microspectrophotometer. The unit is housed in a hand-held assembly that is positioned against the infant's forehead.

Pressure Sensors

In vivo pressure measurements provide important diagnostic information. For example, pressure measurements inside the heart, cranium, kidneys, and bladder can be used to diagnose abnormal physiological conditions that are otherwise not feasible to ascertain from imaging or other diagnostic modalities. In addition, intracranial hypertension resulting from injury or other causes can be monitored to assess the need for therapy and its efficacy. Likewise, dynamic changes of pressure measured inside the heart, uterus, and bladder cavities can help to assess the efficiency of these organs during muscular contractions.

There has long been an interest in developing fiber optic transducers for measuring pressure inside the cranium, vascular system, or the heart. Several approaches have been considered in the development of minimally invasive miniature pressure sensors. The most common technique involves the use of a fiber optic catheter. Fiber optic pressure sensors have been known and widely investigated since the early 1960s. The major challenge is to develop a small enough sensor with a high sensitivity, high fidelity, and adequate dynamic response that can be inserted either through a hypodermic needle or in the form of a catheter. Additionally, for routine clinical use, the device must be cost effective and disposable.

A variety of ideas have been exploited for varying a light signal in a fiber optic probe with pressure (75–77). Most designs utilize either an interferometer principle or measure changes in light intensity. In general, interferometric-based pressure sensors are known to have a high sensitivity, but involve complex calibration and require complicated fabrication. On the other hand, fiber optic pressure sensors based on light intensity modulation have a lower sensitivity, but involve simpler construction.

The basic operating principle of a fiber optic pressure sensor is based on light intensity modulation. Typically, white light or light produced by a LED is carried by an optical fiber to a flexible mirrored surface located inside a pressure-sensing element. The mirror is part of a movable membrane partition that separates the fiber end from the fluid chamber. Changes in the hydrostatic fluid pressure causes a proportional displacement of the membrane relative to the distal end of the optical fiber. This in turn modulates the amount of light coupled back into the optical fiber. The reflected light is measured by a sensitive photodetector and converted to a pressure reading.

A fiber optic pressure transducer for *in vivo* application based on optical interferometry using white light was recently developed by Fisco Technologies (Fig. 10). The sensing element is based on a Fabry–Perot principle. The Fabry–Perot cavity is defined on one end by a stainless steel diaphragm and on the opposite side by the tip of the optical fiber. When an external pressure is applied to the transducer, the deflection of the diaphragm causes variation of the cavity length that in turn is converted to a pressure reading.

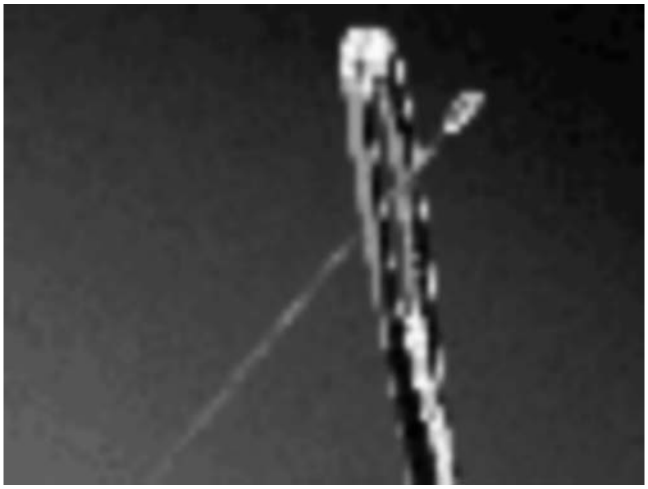


Figure 10. Fiber optic in vivo pressure sensor. Courtesy of Fiso Technologies, Quebec, Canada.

IN VITRO DIAGNOSTIC APPLICATIONS

Immunosensors

The development of immunosensors is based on the observation of ligand-binding reaction products between a target analyte and a highly specific binding reagent (78–80). The key component of an immunosensor is the biological recognition element typically consisting of antibodies or antibody fragments. Immunological techniques offer outstanding selectivity and sensitivity through the process of antibody–antigen interaction. This is the primary recognition mechanism by which the immune system detects and fights foreign matter and has therefore allowed the measurement of many important compounds at micromolar and even picomolar concentrations in complex biological samples.

In principle, it is possible to design competitive binding optical sensors utilizing immobilized antibodies as selective reagents and detecting the displacement of a labeled antigen by the analyte. In practice, however, the strong binding of antigens to antibodies and vice versa causes difficulties in constructing reversible sensors with fast dynamic responses. Other issues relate to the immobilization and specific properties related to the antibody-related reagents on the transducer surface.

Several immunological sensors based on fiber optic waveguides have been demonstrated for monitoring antibody–antigen reactions. Typically, several centimeters of cladding are removed along the fiber's distal end and the recognition antibodies are immobilized on the exposed core surface. These antibodies bind fluorophore–antigen

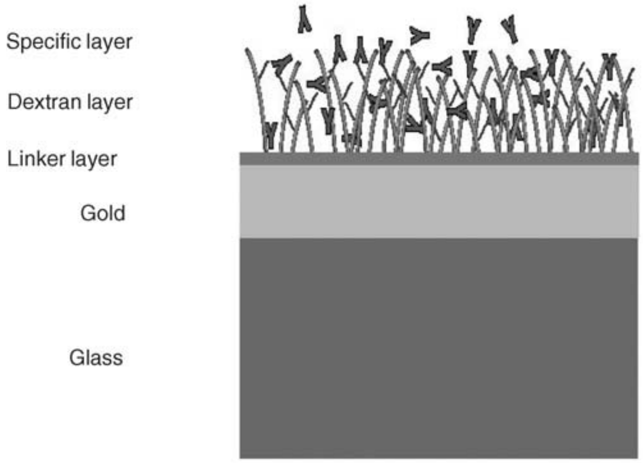


Figure 11. Typical principle of an immunodiagnostic sensor. Courtesy of Biacore AB. The sensor chip consists of a glass surface, coated with a thin layer of gold modified with a Dextran layer for binding of different biomolecules.

complexes within the evanescent wave. The fluorescent signal excited within the evanescent wave is then transmitted through the cladded fiber to a fluorimeter for processing.

Despite an overwhelming number of papers describing immunosensing techniques, there are still only few commercial instruments based upon immunosensors in clinical diagnostics today. Among them is the technology introduced by Biacore (Uppsala, Sweden) (81), which utilizes SPR as the principal probing method (Fig. 11) for real time monitoring of binding events without labeling or often purification of the substances involved. The sensor chip of the Biacore SPR system consists of a glass surface, coated with a thin layer of gold that is modified with a carboxymethylated dextran layer. This dextran hydrogel layer, which provides a hydrophilic environment for attached biomolecules, preserving them in a nondenatured state, forms the basis or a range of specialized surfaces designed to optimize the binding of a variety of molecules that occurs between active biomolecules. The IAsys from Affinity Sensors (Cambridge, UK) (82) also exploits the evanescence field principle, but the method is based on measuring the difference in the phase of a laser beam interacting with a waveguide structure. The method is utilized to study and measure the interactions of proteins binding to nucleic acids, lipids, and carbohydrates and can be used in molecular recognition experiments to study for example the DNA kinetics interactions.

Optical Biosensors in Drug Discovery

Optical biosensors exploit evanescent waves or SPR techniques, where binding of molecules in solution to surface-immobilized receptors causes an alteration of the refractive index of the medium near the surface. This optical change can be used to measure the amount of bound analyte using very low amounts of compound without the need for prior chemical labeling. Hence, evanescent waves and SPR techniques can be used in screening compounds for receptor binding affinity or to study the kinetic analysis of molecular interactions throughout the process of drug development (83). These methods also allow researchers to study *in vitro* the interaction between immobilized receptors and analytes and the general interrogation of intermolecular interactions (84–88). These sensors have become a standard method for characterizing biomolecular interactions and are useful tools to study, for example, if a certain analyte binds to a particular surface, how strong is the binding mechanism, and how much of the sample remains active. The recent development of highly multiplexed optical biosensor arrays has also accelerated the process of assay development and target identification that can speed up the tedious and time-consuming process involved in screening and discovery of new drugs.

BIBLIOGRAPHY

1. Schultz JS, Mansouri S, Goldstein IJ. Affinity sensor: A new technique for developing implantable sensors for glucose and other metabolites. Diabetes Care 1982;5:245.
2. Pickup J, McCartney L, Rolinski L, Birch D. In vivo glucose sensing for diabetes management: progress towards noninvasive monitoring. Br Med J 1999;319:1289–1293.
3. Heise HM. Noninvasive monitoring of metabolites using near infrared spectroscopy: state of the art. Horm Metab Res 1996;28:527–534.
4. Klonoff DC. Noninvasive blood glucose monitoring. Diabetes Care 1997;20:433–437.
5. Cote GL. Noninvasive optical glucose sensing-an overview. J Clin Eng 1997;22(4): 253–259.
6. Heise HM, Marbach R, Bittner A. Clinical chemistry and near infrared spectroscopy: technology for non-invasive glucose monitoring. J Near Infrared Spectrosc 1998;6:349–359.
7. Heinemann L, Schmelzeisen-Redeker G. Noninvasive continuous glucose monitoring in type I diabetic patients with optical glucose sensors. Diabetologia. 1998;41:848–854.
8. Heinemann L et al. Noninvasive glucose measurement by monitoring of scattering coefficient during oral glucose tolerance tests. Diabetes Technol Ther 2000;2:211–220.
9. Severinghaus JW. Historical development of oxygenation monitoring. In: Payne JP, Severinghaus JW, editors. Pulse Oximetry. Berlin: Springer-Verlag; 1986.
10. Polanyi ML, Heir RM. In vivo oximeter with fast dynamic response. Rev Sci Instrum 1962;33:1050.
11. Nelson LD. Mixed venous oximetry. In: Snyder JV, Pinsky JV, editors. Oxygen Transport in the Critically Ill. Chicago: Year Book Medical; 1987. pp. 235–248.
12. Enger EL, Holm K. Perspectives on the interpretation of continuous mixed venous oxygen saturation. Heart Lung 1990;19:578–580.
13. Mendelson Y. Pulse oximetry: theory and applications for noninvasive monitoring. Clin Chem 1992;38(9):1601–1607.
14. Aoyagi T, Kishi M, Yamaguchi K, Watanable S. Improvement of the earpiece oximeter. Jpn Soc Med Electron Biomed Eng 1974; 90–91.
15. Yoshiya I, Shimada Y, Tanaka K. Spectrophotometric monitoring of arterial oxygen saturation in the fingertip. Med Biol Eng Comput 1980;18:27.
16. Mendelson Y, Solomita MV. The feasibility of spectrophotometric measurements of arterial oxygen saturation from the

scalp utilizing noninvasive skin reflectance pulse oximetry. Biomed Instrum Technol 1992;26:215–224.

17. Mendelson Y, McGinn MJ. Skin reflectance pulse oximetry: In vivo measurements from the forearm and calf. J Clin Monit 1991;7:7–12.
18. Iacobelli L, Lucchini A, Asnaghi E, Nesci M. Oxygen saturation monitoring. Minerva Anesthesiol 2002;68(5): 488–491.
19. Sinex JE. Pulse oximetry: principles and limitations. Am J Emerg Med 1999;17(1): 59–67.
20. Wouters PF et al. Accuracy of pulse oximeters: The European multicenter trial. Anesthesiol Analg 2002;94:S13–S16.
21. Severinghaus JW, Naifeh KH. Accuracy of response of six pulse oximeters to profound hypoxia. Anesthesiology 1987;67:551–558.
22. Lee WW, Mayberry K, Crapo R, Jensen RL. The accuracy of pulse oximetry in the emergency department. Am J Emerg Med 2000;18(4):427–431.
23. Kopotic RJ, Lindner W. Assessing high-risk infants in the delivery room with pulse oximetry. Anesthesiol Analg 2002;94:S31–S36.
24. Moller JT et al. Randomized evaluation of pulse oximetry in 20,802 patients: I. Anesthesiology 1993;78:436–444.
25. Jubran A. Pulse Oximetry. In Principles and practice of intensive care monitoring. In: Tobin MJ, editor. New York: McGraw Hill; 1998.
26. Yamaya Y et al. Validity of pulse oximetry during maximal exercise in normoxia, hypoxia, and hyperoxia. J Appl Physiol 2002;92:162–168.
27. Yam J, Chua S, Arulkumaran S. Intrapartum fetal pulse oximetry. Part I: Principles and technical issues. Obstet Gynecol Surv 2000;55(3): 163–172.
28. Luttkus AK, Lubke M, Buscher U, Porath M, Dudenhausen JW. Accuracy of fetal pulse oximetry. Acta Obstet Gynecol Scand 2002;81(5):417–423.
29. Edmonds HL. Detection and treatment of cerebral hypoxia key to avoiding intraoperative brain injuries. APSF Newslett 1999;14(3):25–32.
30. Kim M, Ward D, Cartwright C, Kolano J, Chlebowski S, Henson L. Estimation of jugular venous O_2 saturation from cerebral oximetry or arterial O_2 saturation during isocapnic hypoxia. J Clin Monit 2001;16:191–199.
31. Samra SK, Stanley JC, Zelenock GB, Dorje P. An assessment of contributions made by extracranial tissues during cerebral oximetry. J Neurosurg Anest 1999;11(1):1–5.
32. Lubbers DW, Opitz N. The pCO_2/pO_2-optode: A new probe for measurement of pCO_2 or pO_2 in fluids and gases. Z Naturforsch, C: Biosci 1975;30C:532–533.
33. Hansmann DR, Gehrich JL. Practical perspectives on the in vitro and in vivo evaluation of a fiber optic blood gas sensor. Proc SPIE Opt Fibers Med III 1988;906:4–10.
34. Shapiro BA, Cane RD, Chomka CM, Bandala LE, Peruzzi WT. Preliminary evaluation of an intra-arterial blood gas system in dogs and humans. Crit Care Med 1989;17:455–460.
35. Mahutte CK et al. Progress in the development of a fluorescent intravascular blood gas system in man. J Clin Monit 1990;6:147–157.
36. Venkatesh B, Clutton-Brock TH, Hendry SP. A multiparameter sensor for continuous intraarterial blood gas monitoring: a prospective evaluation. Crit Care Med 1994;22:588–594.
37. Venkatesh B, Clutton-Brock TH, Hendry SP. Continuous measurement of blood gases using a combined electrochemical and spectrophotometric sensor. J Med Eng Technol 1994;18:165–168.
38. Abraham E, Gallagher TJ, Fink S. Clinical evaluation of a multiparameter intra-arterial blood-gas sensor. Intensive Care Med 1996;22:507–513.
39. Zauner A et al. Continuous monitoring of cerebral substrate delivery and clearance: initial experience in 24 patients with severe acute brain injuries. Neurosurgery 1997;40(2):294–300.
40. Coule LW, Truemper EJ, Steinhart CM, Lutin WA. Accuracy and utility of a continuous intra-arterial blood gas monitoring system in pediatric patients. Crit Care Med 2001;29(2):420–426.
41. 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.
42. Kautsky H. Quenching of luminescence by oxygen. Trans Faraday Soc 1939;35:216–219.
43. Opitz N, Lubbers DW. Theory and development of fluorescence-based optochemical oxygen sensors: oxygen optodes. Int Anaesthesiol Clin 1987;25:177–197.
44. Peterson JI, Fitzgerald RV, Buckhold DK. Fiber-optic probe for in vivo measurement of oxygen partial pressure. Anal Chem 1984;56:62.
45. Vaughan WM, Weber G. Oxygen quenching of pyrenebuteric acid fluorescence in water: a dynamic probe of the microenvironment. Biochemistry 1970;9:464–473.
46. Stern O, Volmer M. Uber die Abklingzeit der Fluorescenz. Z Phys 1919;20:183–188.
47. Peterson JI, Goldstein SR, Fitzgerald RV. Fiber optic pH probe for physiological use. Anal Chem 1980;52:864–869.
48. Markle DR, McGuire DA, Goldstein SR, Patterson RE, Watson RM. A pH measurement system for use in tissue and blood, employing miniature fiber optic probes. In: Viano DC, editor. Advances in Bioengineering. New York: American Society of Mechanical Engineers; 1981 p 123.
49. Wolfbeis OS, Furlinger E, Kroneis H, Marsoner H. Fluorimetric analysis. 1. A study on fluorescent indicators for measuring near neutral (physiological) pH values. Fresenius' Z Anal Chem 1983;314:119.
50. Gehrich JL et al. Optical fluorescence and its application to an intravascular blood-gas system. IEEE Trans Biomed Eng 1986;33:117–132.
51. Seitz WR. Chemical sensors based on fiberoptics. Anal Chem 1984;56:17A–34A.
52. Yafuso M et al. Optical pH measurements in blood. Proc SPIE Opt Fibers Med IV 1989;1067:37–43.
53. Wolfbeis OS, Offenbacher H. Fluorescence sensor for monitoring ionic strength and physiological pH values. Sens Actuators 1986;9:85.
54. Opitz N, Lubbers DW. New fluorescence photomatrical techniques for simultaneous and continuous measurements of ionic strength and hydrogen ion activities. Sens Actuators 1983;4:473.
55. Smith BE, King PH, Schlain L. Clinical evaluation-continuous real-time intra-arterial blood gas monitoring during anesthesia and surgery by fiberoptic sensor. Int J Clin Monit 1992;9:45.
56. Miller WW, Gehrich JL, Hansmann DR, Yafuso M. Continuous in vivo monitoring of blood gases. Lab Med 1988;19:629–635.
57. Vurek GG, Feustel PJ, Severinghaus JW. A fiber optic pCO_2 sensor. Ann Biomed Eng 1983;11:499.
58. Zhujun Z, Seitz WR. A carbon dioxide sensor based on fluorescence. Anal Chim Acta 1984;160:305.
59. Clark CL, O'Brien J, McCulloch J, Webster J, Gehrich J. Early clinical experience with GasStat. J Extra Corporeal Technol. 1986;18:185.
60. Mannebach PC, Sistino JJ. Monitoring aortic root effluent during retrograde cardioplegia delivery. Perfusion 1997; 12(5):317–323.

61. Siggaard-Andersen O, Gothgen IH, Wimberley PD, Rasmussen JP, Fogh-Andersen N. Evaluation of the GasStat fluorescence sensors for continuous measurement of pH, pCO_2 and pO_2 during CPB and hypothermia. Scand J Clin Lab Invest 1988;48 (Suppl. 189):77.
62. Shapiro BA, Mahutte CK, Cane RD, Gilmour IJ. Clinical performance of an arterial blood gas monitor. Crit Care Med 1993;21:487–494.
63. Mahutte CK. Clinical experience with optode-based systems: early in vivo attempts and present on-demand arterial blood gas systems, 12th IFCC Eur. Cong. Clin. Chem. Medlab. 1997;53.
64. Mahutte CK. On-line arterial blood gas analysis with optodes: current status. Clin Biochem 1998;31(3):119–130.
65. Emery RW et al. Clinical evaluation of the on-line Sensicath™ blood gas monitoring system. Am J Respir Crit Care Med 1996;153:A604.
66. Myklejord DJ, Pritzker MR. Clinical evaluation of the on-line Sensicath™ blood gas monitoring system. Heart Surg Forum 1998;1(1):60–64.
67. Steuer R et al. A new optical technique for monitoring hematocrit and circulating blood volume: Its application in renal dialysis. Dial Transplant 1993;22:260–265.
68. Steuer R et al. Reducing symptoms during hemodialysis by continuously monitoring the hematocrit. Am J Kidney Dis 1996;27:525–532.
69. Leypoldt JK et al. Determination of circulating blood volume during hemodialysis by continuously monitoring hematocrit. J Am Soc Nephrol 1995;6:214–219.
70. Yamanouchi I, Yamauchi Y, Igarashi I. Transcutaneous bilirubinometry: preliminary studies of noninvasive transcutaneous bilirubin meter in the Okayama national hospital. Pediatrics 1980;65:195–202.
71. Jacques SL. Reflectance spectroscopy with optimal fiber devices and transcutaneous bilirubinometers. Biomed Opt Instrum Laser Assisted Biotechnol 1996;84–94.
72. Robertson A, Kazmierczak S, Vos P. Improved transcutaneous bilirubinometry: comparison of SpectR$_x$, Bili*Check* and Minolta jaundice meter JM-102 for estimating total serum bilirubin in a normal newborn population. J Perinatol 2002;22: 12–14.
73. Rubaltelli FF et al. Transcutaneous bilirubin measurement: A multicenter evaluation of a new device. Pediatrics 2001;107(6): 1264–1271.
74. Bhutani VK et al. Noninvasive measurement of total serum bilirubin in a multiracial predischarge newborn population to assess the risk of severe hyperbilirubinemia. Pediatrics 2000;106(2):e17.
75. Ivan LP, Choo SH, Ventureyra ECG. Intracranial pressure monitoring with the fiber optic transducer in children. Child's Brain 1980;7:303.
76. Kobayashi K, Miyaji H, Yasuda T, Matsumoto H. Basic study of a catheter tip micromanometer utilizing a single polarization fiber. Jpn J Med Electron Biol Eng 1983;21:256.
77. Hansen TE. A fiberoptic micro-tip pressure transducer for medical applications, Sens. Actuators 1983;4:545.
78. Luppa PB, Sokoll LJ, Chan DW. Immunosensors-Principles and applications to clinical chemistry. Clin Chem Acta 2001;314:1–26.
79. Morgan CL, Newman DJ, Price CP. Immunosensors: technology and opportunities in laboratory medicine. Clin Chem 1996;42:193–209.
80. Leatherbarrow RJ, Edwards PR. Analysis of molecular recognition using optical biosensors. Curr Opin Chem Biol 1999;3:544–547.
81. Malmqvist M. BIACORE: an affinity biosensor system for characterization of biomolecular interactions. Biochem Soc Trans 1999;27:335–340.
82. Lowe P et al. New approaches for the analysis of molecular recognition using IAsys evanescent wave biosensor. J Mol Recogn 1998;11:194–199.
83. Cooper MA. Optical biosensors in drug discovery, Nature Reviews. Drug Discov 2002;1:515–528.
84. Ziegler C, Gopel W. Biosensor development. Curr Opin Chem Biol 1998;2:585–591.
85. Weimar T. Recent trends in the application of evanescent wave biosensors. Angew Chem Int Ed Engl 2000;39:1219–1221.
86. Meadows D. Recent developments with biosensing technology and applications in the pharmaceutical industry. Adv Drug Deliv Rev 1996;21:179–189.
87. Paddle BM. Biosensors for chemical and biological agents of defense interest. Biosens Bioelectro 1996;11:1079–1113.
88. Keusgen M. Biosensors: new approaches in drug discovery. Naturwissenschaften 2002;89:433–444.

Reading List

Barth FG, Humphrey JAC, Secomb TW, editors. Sensors and Sensing in Biology and Engineering. New York: Springer-Verlag; 2003.

Eggins BR, Chemical Sensors and Biosensors for Medical and Biological Applications. Hoboken, NJ: Wiley; 2002.

Fraser D, editor. Biosensors in the Body: Continuous In Vivo Monitoring. New York: Wiley; 1997.

Gauglitz G, Vo-Dinh T. Handbook of Spectroscopy. Hoboken, NJ: Wiley-VCH; 2003.

Kress-Rogers E, editor. Handbook of Biosensors and Electronic Noses: Medicine, Food, and the Environment. Boca Raton, FL: CRC Press; 1996.

Ligler FS, Rowe-Taitt CA, editors. Optical Biosensors: Present and Future. New York: Elsevier Science; 2002.

Mirabella FM, editor. Modern Techniques in Applied Molecular Spectroscopy. New York: Wiley; 1998.

Ramsay G, editor. Commercial Biosensors: Applications to Clinical, Bioprocess and Environmental Samples. Hoboken, NJ: Wiley; 1998.

Rich RL, Myszka DG. Survey of the year 2001 optical biosensor literature. *J Mol Recogn* 2002;15:352–376.

Vo-Dinh T, editor. Biomedical Photonics Handbook. Boca Raton, FL: CRC Press; 2002.

Webster JG, editor. Design of Pulse Oximeters. Bristol UK: IOP Publishing; 1997.

Yang VC, Ngo TT. Biosensors and Their Applications. Hingham, MA: Kluwer Academic Publishing; 2002.

See also Blood gas measurements; cutaneous blood flow, doppler measurement of; fiber optics in medicine; glucose sensors; monitoring, intracranial pressure.

OPTICAL TWEEZERS

Henry Schek III
Alan J. Hunt
University of Michigan
Ann Arbor, Michigan

INTRODUCTION

Since the invention of the microscope, scientists have peered down at the intricate workings of cellular machinery, laboring to infer how life is sustained. Doubtlessly, these investigators have frequently pondered What would happen if I could push on that, or pull on this? Today, these formerly rhetorical thought experiments can be accomplished using a single-beam optical gradient trap, more commonly known as "optical tweezers". This article surveys the history, theory and practical aspects of optical trapping, especially for studying biology. We start by reviewing the early demonstrations of optical force generation, and follow this with a discussion of the theoretical and practical concerns for constructing, calibrating, and applying an optical tweezers device. Finally, examples of important optical tweezers experiments and their results are reviewed.

OPTICAL TWEEZERS SYSTEMS

History

In 1970, Arthur Ashkin published the first demonstration of light pressure forces manipulating microscopic, transparent, uncharged particles, a finding that laid the groundwork for optical trapping (1). Significant application of optical forces to study biology would not occur until almost two decades later, after the first description of a single-beam optical gradient trap was presented in 1986 (2). Soon after, the ability to trap living cells was demonstrated (3,4) and by the late 1980s biophysicists were applying optical tweezers to understand diverse systems, such as bacterial flagella (5), sperm (6), and motor proteins, such as kinesin (7). Today, optical tweezers are a primary tool for studying the mechanics of cellular components and are rapidly being adapted to applications ranging from cell sorting to the construction of nanotechnology (8,9).

Trapping Theory

Ashkin and co-workers 1986 description of a single-beam trap presented the necessary requirements for stable trapping of dielectric particles in three dimensions (2). For laser light interacting with a particle of diameter much larger than the laser wavelength, $d \gg \lambda$, that is the so-called Mie regime, the force on the particle can be calculated using ray optics to determine the momentum transferred from the refracted light to the trapped particle. Figure 1a–c schematically shows the general principle; two representative rays are bent as they pass through the spherical particle producing the forces that push the trapped particle toward the laser focus. A tightly focused beam that is most intense in the center thus pulls the trapped object toward the beam waist. More rigorous treatment and calculations can be found in (2,10).

In the Rayleigh regime, $d \ll \lambda$, the system must be treated in accordance with wave optics, and again the tight focusing results in a net trapping force due to the fact that the particle is a dielectric in a non-uniform electric field. Figure 1d shows a simplified depiction of force generation in this regime. The electric field gradient from the laser light induces a dipole in the dielectric particle; this results in a force on the particle directed up the gradient toward the area of greatest electric field intensity, at the center of the laser focus. Detailed calculations can be found in Ref. 11.

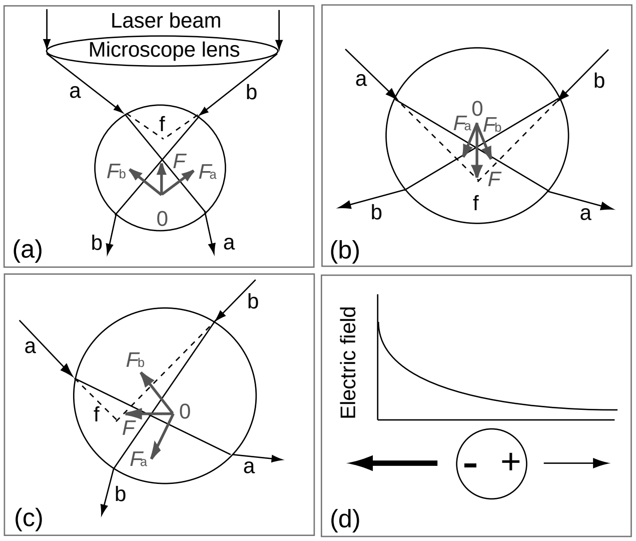


Figure 1. Optical forces on a dielectric sphere. (Adapted from Ref. 2.) Parts a–c show three possible geometries in the ray optics regime: particle above, particle below, and particle to the right of the laser focus, respectively. In each case, two representative rays are shown interacting with the particle. The two rays change direction upon entering and exiting the sphere causing a net transfer of momentum to the particle. Dotted lines show the focus point in the absence of the sphere. The forces resulting from each ray are shown in gray along with the summed force. (Please see online version for color in figure.) In each case the net force pushes the particle toward the focus. A highly simplified mechanism for the generation of force in the Rayleigh regime. An electric field profile is shown above a representative particle. The labels on the sphere show the net positioning of charges due to the formation of the dipole and the arrows show the forces. The induced dipole results in the bead being attracted to the area of most intense electric field, the laser focus.

In either regime, the principal challenge to producing a stable trap is overcoming the force produced by light scattered back in the direction of the oncoming laser, which imparts momentum that pushes the particle in the direction of beam propagation, and potentially out of the trap. When the trapping force that pulls the particle up the laser toward the beam waist is large enough to balance this scattering force, a stable trap results. From examination of Fig. 1a, it is apparent that the most important rays providing the force to balance the scattering force come at steep angles from the periphery of the focusing lens. For this reason, a tightly focusing lens is critical for forming an optical trap; typically oil-immersion lenses with a numerical aperture in excess of 1.2 are used. Furthermore, the beam must be expanded to slightly overfill the back aperture of the focusing lens so that sufficient laser power is carried in the most steeply focused periphery of the beam.

For biological experiments the preferred size of the trapped particle is rarely in the range appropriately treated in either the Mie or Raleigh regime; typically particles are on the order of 1 μm in diameter, or approximately equal to the laser wavelength. Theoretical treatment then requires generalized Lorenz–Mie theory (12,13). In

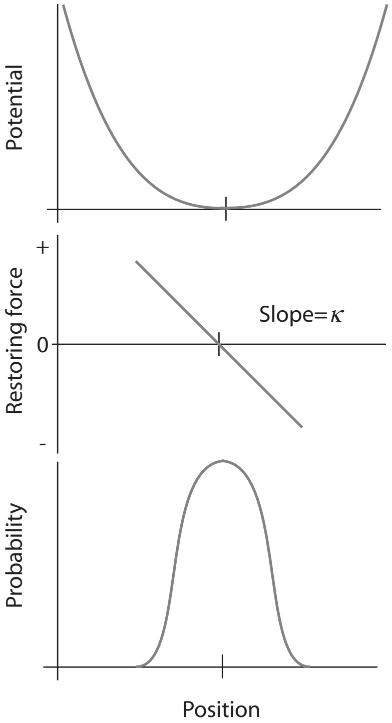


Figure 2. Relationship between potential, force, and bead position probability. The interaction of the tightly focused laser and dielectric results in a parabolic potential well, the consequence of which is a linear restoring force; the trap behaves as a linear radial spring. Here a positive force pushes the particle to the right. When driven by thermal events in solution a particle's position has a Gaussian distribution.

practice, theoretical analysis of specific trapping parameters is not necessary because several methods allow direct calibration of trapping forces. Regardless of particle size, a tightly focused laser with a Gaussian intensity profile (TEM$_{00}$ mode) interacting with a spherical particle, traps the particle in a parabolic potential well. The parabolic potential is convenient because it results in a trap that behaves like a Hookian spring: restoring force that pushes the particle back toward the focus increases linearly with the displacement from the focus. Figure 2 schematically illustrates the potential well, force profile, and distribution of particle positions for a hypothetical trapped sphere. This convenient relationship results in spherical particles being the natural choice for most trapping applications including gradient. These particles are often referred to as beads, microbeads, or microspheres interchangeably.

Optical Tweezers Systems

Implementations of optical tweezers as they are applied to biology are variations on a theme: most systems share several common features although details vary and systems are often optimized for the application. Figures 3 and 4 show a photograph and a diagram of an example system. Generally, the system begins with a collimated, plane-polarized laser with excellent power and pointing stability. The optimal laser light wavelength depends on the application, but near-infrared (IR) lasers with a wavelength ~1 μm are a typical choice for use with biological samples; such lasers are readily available and biological samples generally exhibit minimal absorption in the near-IR. A light attenuator allows for adjustment of laser power and therefore trap stiffness. In the example system this is accomplished with a variable half-wave plate that adjusts the polarity prior to the beam passing through a polarized beam splitting cube that redirects the unwanted laser power. Most systems then contain a device and lenses to actively steer the beam. In the case of the example system, a piezo-actuated mirror creates angular deflections of the beam that the telescope lenses translate into lateral position changes of the laser focus at the focal plane of the objective. This is accomplished by arranging the telescope so that a virtual image of the steering device is formed at the back aperture of the microscope objective lens. This arrangement also assures that the laser is not differentially clipped by the objective aperture as the trap is steered. Following the steering optics, the laser enters a microscope, is focused by a high numerical

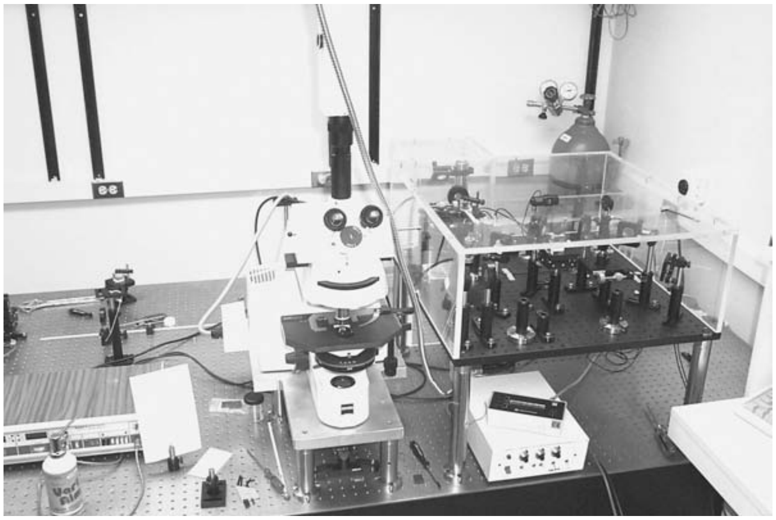


Figure 3. Photograph of an optical tweezers instrument. Note the vibration isolation table that prevents spurious vibrations from affecting experi- ments. The clear plastic cover minimizes ambient air convection that might affect the laser path in addition to limiting access of dust and preventing accidental contact with the optics.

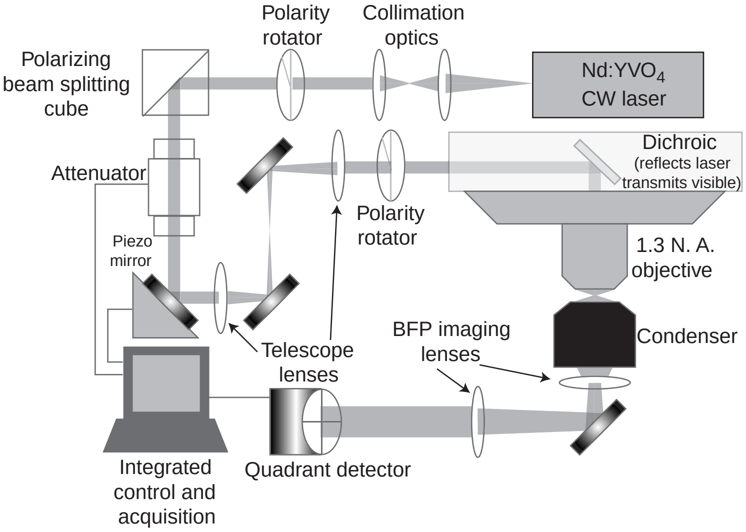


Figure 4. Optical tweezers schematic showing major system components.

aperture objective, and encounters the sample. After passing through the trapping/image plane, the laser usually exits the microscope and enters a detection system, which in this case employs back-focal-plane interferometry to measure the position of the bead relative to the trap (14–16).

There are many optical tweezers designs that can produce stable, reliable trapping. In most cases, the system is integrated into an inverted, high quality research microscope (17,18), although successful systems have been incorporated into upright (e.g., 15) or custom built microscopes (14). Some use a single laser to form the trap (7,15) while a more complicated arrangement uses two counterpropagating lasers (19,20).

Various implementations for detecting the bead position include image analysis (21,22), back focal plane interferometry (BFPI) (14,16), and a host of less frequently applied methods based on measurements of scattered or fluorescent light intensity (23–25) or interference between two beams produced using the Wollaston prisms associated with differential interference contrast microscopy (26). Trap steering can be accomplished with acoustooptic deflectors (15,27,28), steerable mirrors (23), or actuated lenses (14). Figure 5 shows a silica microsphere in a trap being moved in a circle at > 10 revolutions s^{-1} using acoustooptic deflectors, resulting in the comet tail in the image. Alternatively, the sample can be moved with a motorized or piezoactuated nanopositioning stage. In addition, splitting the laser into two orthogonally polarized beams, fast laser steering to effectively multiplex the beam by rapidly jumping the laser between multiple positions (29), or holographic technology (9) can be used to create arrays of multiple traps. Figure 6 shows an image of a 3×3 array of optical traps created using fast beam steering to multiplex a single beam. Five traps are holding beads while four, marked with white circles are empty. Specific optical tweezers designs can also allow incorporation of other advanced optical techniques often used in biology, including, but not limited to, total internal reflection microscopy and confocal microscopy (30).

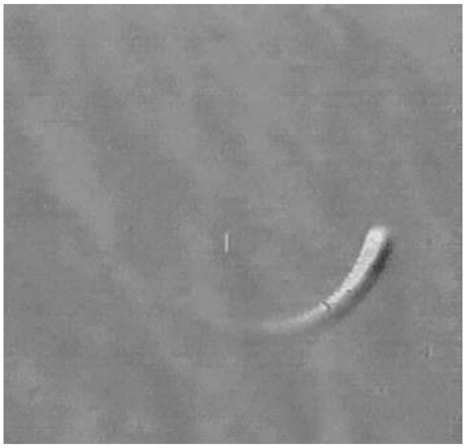


Figure 5. Micrograph of 1 μm bead being manipulated in a circle with an optical tweezers device. The comet tail is the result of the bead moving faster than the video frame rate.

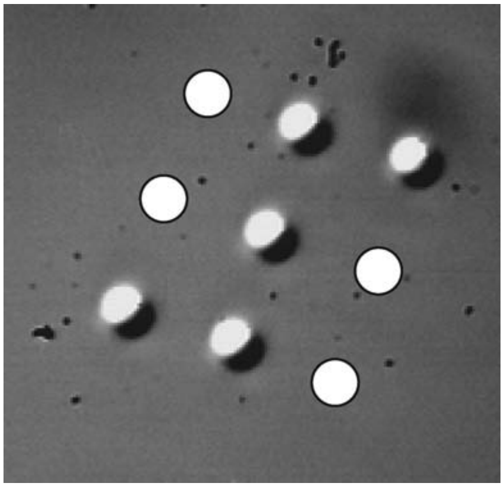


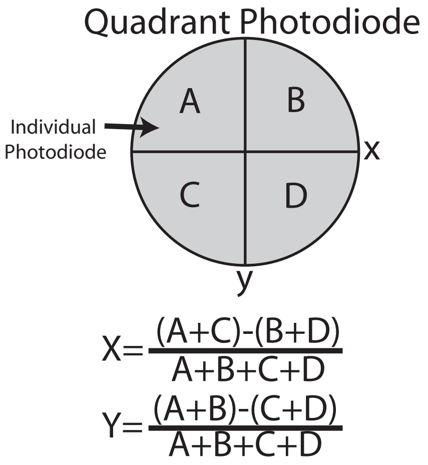
Figure 6. Micrograph of a 3×3 array of traps produced by time sharing a single laser between nine positions. The traps at the four locations marked with white circles are empty while the other five hold a 1 μm diameter silica microsphere.

Detection

Determining the force on a trapped sphere requires measurement of the displacement of the trapped particle from the center of the optical trap. Back focal plane interferometry (BFPI) is the most popular method, and allows high speed detection with nanometer resolution. Rather than following an image of the bead, which is also a viable detection technique, BFPI tracks the bead by examining how an interference pattern formed by the trapping laser is altered by the bead interacting with the laser further up the beam at the microscope focal plane. This interference pattern is in focus at the back focal plane of the condenser lens of the microscope, thus the name of the technique.

Laser light interacting with the bead is either transmitted or scattered. The transmitted and scattered light interfere with one another resulting in an interference pattern that is strongly dependent on the relative position of the bead within the trap. This interference pattern is easily imaged onto a quadrant photodiode (QPD) detector positioned at an image plane formed conjugate to the back focal plane of the condenser lenes by supplemental lenses (e.g., BFPI imaging lens in Fig. 4). A simple divider-amplifier circuit compares the relative intensity in each quadrant according to the equations shown in Fig. 7. This results in voltage signals that follows the position of the bead in the trap.

Back focal plane interferometry has several important advantages compared with the other commonly used bead tracking techniques. Typically, image analysis limits the data collection rate to video frame rate, 30 Hz, or slower if images must be averaged, while BFPI easily achieves sampling frequencies in the tens of kilohertz. Under ideal conditions, image analysis can detect particle positions with ~ 10 nm resolution, while BFPI achieves subnanometer resolution. Speeds and resolution similar to BFPI can be achieved by directly imaging a bead onto a QPD, however, it can be inconvenient in systems where trap steering is employed because the image of the particles will move across the detector as the trap moves in the image plane of the microscope, thus requiring frequent repositioning of the detector. With BFPI, the position of the intereference pattern at the back focal plane of the condenser does not move as the trap is steered, so the detector can remain stationary.



$$X=\frac{(A+C)-(B+D)}{A+B+C+D}$$

$$Y=\frac{(A+B)-(C+D)}{A+B+C+D}$$

Figure 7. Quadrant photodiode operation. The circular quadrant photodiode detector is composed of four individual detectors (A–D) each making up one quarter of the circle. Each diode measures the light intensity falling on the surface of that quadrant and the associated electronics compares the relative intensities according to the equations shown in the figure. The resulting voltages are calibrated to the intensity shifts caused by the bead as it moves to different positions in the trap.

Calibration

Although BFPI provides a signal corresponding to bead position, two important system parameters must be measured before forces and positions can be inferred. The first is the detection sensitivity, β, relating the detector signal to the particle position. The second is the Hookian spring constant, κ, where

$$F = -\kappa x \tag{1}$$

where x is the displacement of the particle from the center of the trap. Both sensitivity and stiffness depend on the diameter of the particle and the position of the trap relative to nearby surfaces. Stiffness is linear with laser power while sensitivity, when using BFPI, does not depend on laser power as long as the detector is operating in its linear range. Ideally, these parameters are determined in circumstances as near to the experimental conditions as possible. Indeed whenever reasonably possible, it is wise to account for bead and assay variation by determining these parameters for each bead used in an experiment.

Direct Calibration Method

Sensitivity can be measured by two independent methods. The most direct is to move a bead affixed to a coverslip through the laser focus with a nanopositioning piezocontrolled stage, or conversely by scanning the laser over the immobilized bead while measuring the detector signal. An example calibration curve is shown in Fig. 8 for a system using BFPI. The linear region within $\sim$150 nm on either side of zero is fit to a line to arrive at the sensitivity, β, equal to 6.8×10^{-4} V·nm^{-1}. The second method is discussed below under calibration with thermally driven motion.

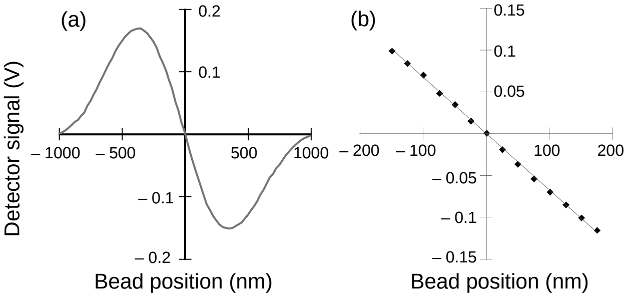


Figure 8. Optical tweezers sensitivity curve acquired with the direct method of moving the laser past an immobilized bead. The full curve is shown in a. The central linear region shown in b is fit to give a sensitivity of 6.8×10^{-4} V·nm^{-1}.

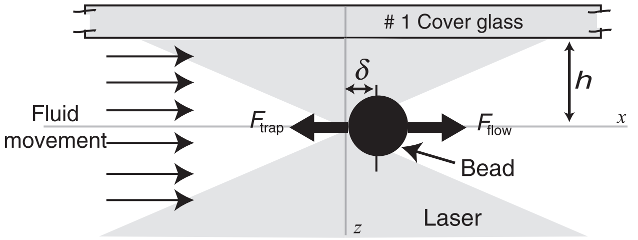


Figure 9. Schematic of drag force calibration. Uniform fluid flow from the left produces a force on the bead that can be calculated with the equations in the text. The bead is displaced to a distance, where the force due to fluid flow is balanced by the force due to the trap.

The direct calibration for sensitivity has the disadvantage of requiring the bead to be immobilized, typically to the coverslip, whereas experiments are typically conducted several microns from the surface. Additionally, with the bead fixed to a surface, its z-axis position is no longer controlled by the trapping forces, resulting in the possibility that the measured sensitivity does not correspond to the actual z axis position of a trapped particle. Estimates of sensitivity may also be rendered inaccurate if attachements to the surface slip or fail. Consequently, thermal force based calibration discussed in the next section are often more reliable.

The trap stiffness, κ, can be measured by applying drag force from fluid flow to the trapped particle. Figure 9 shows a schematic representation of the trap, particle, and fluid movement. Typically, a flow is induced around the bead by moving the entire experimental chamber on a motorized or piezodriven microscope stage. In the low Reynolds number regime where optical traps are typically applied, the drag coefficient, γ, on a sphere in the vicinity of a surface can be accurately calculated:

$$\gamma = 6\pi r c \eta \tag{2}$$

where r is the radius of the microsphere, η is the viscosity of the medium, and c is the correction for the distance from a surface given by

$$c = 1 + \frac{9}{16}\frac{r}{h} \tag{3}$$

where h is the distance from the center of the sphere to the surface. Equation 3 is an approximation and can be carried to higher order for greater accuracy (31).

The force due to fluid flow is readily calculated

$$F = \gamma V \tag{4}$$

where V is the velocity of the particle relative to the fluid medium.

When a flow force is applied to the trapped microsphere, the bead moves away from the center of the trap to an equilibrium position where the flow force and trapping force balance. Applying several drag forces (using a range of fluid flow velocities) and measuring the deflection in each case allows one to plot force of the trap as a function of bead position. The slope of this line is the trap stiffness, κ.

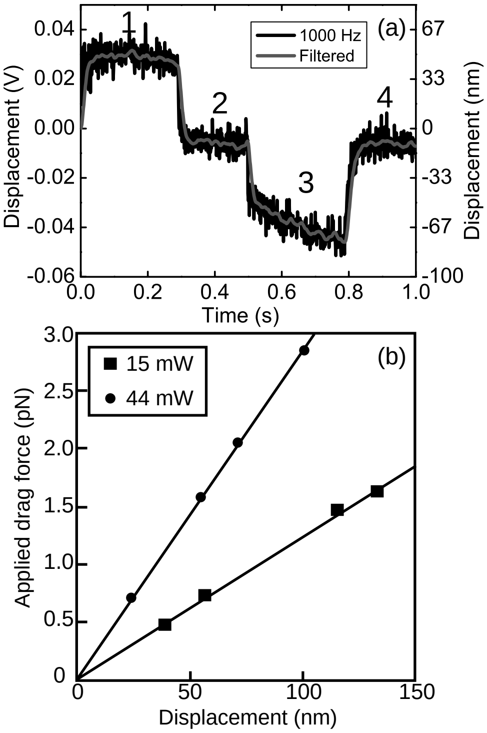


Figure 10. Results of a typical drag force calibration. (a) Typical signal of five averaged applications of uniform drag force on a trapped microsphere. In the region labeled 1, the bead quickly achieves a stable displacement due to the constant velocity. In region 2, the bead briefly returns to the center of the trap as the stage stops. In region 3, the bead is again displaced as the stage moves back to its home position and in region 4 the stage is at rest in its home position. The displaced position () in region 1 is averaged to determine the displacement for a given flow force. (b) Example of data from several different forces applied to the trapped microsphere reveal a linear relation to displacement.

Figure 10a shows a trace of a bead being displaced by fluid drag force provided by a piezoactuated stage moving at a known velocity. In the first region the bead quickly achieves and holds a stable position as long as the stage moves at a constant velocity, making measurement of the displacement relatively simple. When the stage stops the bead returns to the center of the trap as is seen in the second section of the record. The subsequent negative deflection in the third section is the result of the stage returning to its original position at a non-uniform speed. Figure 10b shows force as a function of position for two laser powers; the linear fitting coefficient gives the trap stiffness in each case.

In practice, calibration by viscous drag can be quite labor intensive, and the requirement for the sample to be moved repeatedly and rapidly is too disruptive to be performed "on the fly" during many types of experiments. A more efficient and less disruptive alternative relies on the fact that the thermal motion of the fluid molecules exerts significant, statistically predictable forces on the trapped microsphere; these result in fluctuations in bead position measurable with the optical tweezers.

Thermal Force Based Calibration Methods

At low Reynold's (Re) number the power spectrum of the position of a particle trapped in a harmonic potential well subject only to thermal forces takes the form of a Lorentzian (32):

$$S_x(f) = \frac{B}{f_c^2 + f^2} \tag{5}$$

with

$$B = \frac{k_B T}{\gamma \pi^2} \tag{6}$$

where k_B is the Boltzmann constant, f is the frequency, f_c is a constant called the corner frequency, and

$$f_c = \frac{\kappa}{2\pi\gamma} \tag{7}$$

There are several important practical considerations for applying the power spectrum to calibrate optical tweezers. The mathematical form of the power spectrum does not account for the electrical noise in the signal, so a large signal to noise ratio is required for accurate calibration. In addition, records of bead position must be treated carefully, especially with respect to instrument bandwidth and vibrations, to yield the correct shape and magnitude of the power spectrum.

A typical calibration begins by setting the position of the bead relative to any nearby surface, typically a microscope sample coverglass. With the bead positioned, a record of thermal motion of the bead is taken. For the example system in Figs. 3 and 4 a typical data collection for a calibration is 45 s at maximum bandwidth of the QPD and associated electronics. The data is then low pass (antialias) filtered using a high order Butterworth filter with a cut-off frequency set to one-half of the bandwidth.

Following acquisition and filtering, the power spectrum of the record of bead position is calculated. In practice, many power spectra should be averaged to reduce the inherently large variance of an individual power spectrum. For example, the 45 s of data is broken into 45, 1 s records of bead position, the power spectrum for each is calculated and the spectra are averaged together. To accurately compute the power spectrum the data must be treated as continuous; this is achieved by wrapping the end of the record back to meet the beginning. To avoid a discontinuity where the beginning and end are joined, each record is windowed with a triangle or similar function that forces the ends to meet but maintains the statistical variance in the original record. Figure 11a shows an example of an averaged power spectrum. Note the units on the y axis: often, power spectra are presented in different units according to their intended use. Accurate calibration of optical tweezers requires that the power spectrum be in $\text{nm}^2{\cdot}\text{Hz}^{-1}$ or, provided volts are proportional to displacement, $\text{V}^2{\cdot}\text{Hz}^{-1}$.

The average power spectrum is expected to fit the form of equation 5. To avoid aliased high frequency data corrupting the fit to the power spectrum, the spectrum is cropped well below the cut-off frequency of the antialiasing filter. The average, cropped power spectrum can be visually inspected for evidence of noise (e.g., spikes or other deviations from the expected Lorentzian form) and fit to equation 5, with f_c and B as fitting parameters. The drag coefficient is calculated using equations 2 and 3 and the stiffness, κ, is then calculated with a simple rearrangement of equation 7. The sensitivity, β, can also be calculated by recognizing that the value of B produced by the fit is proportional to the signal power. The value of B can be calculated from first principles by applying equation 6. In practice, the value of B is determined by fitting equation 5 with B as a parameter. Equation 6 gives B in units of $\text{nm}^2{\cdot}\text{Hz}^{-1}$ while the fitted value of B will typically be $\text{V}^2{\cdot}\text{Hz}^{-1}$. The parameter β is then determined by dividing the fitted value by the theoretical value and taking the square root. Figure 11b shows an example power spectrum used to calculate the trap stiffness and the sensitivity of the detection system with a 0.6 μm bead.

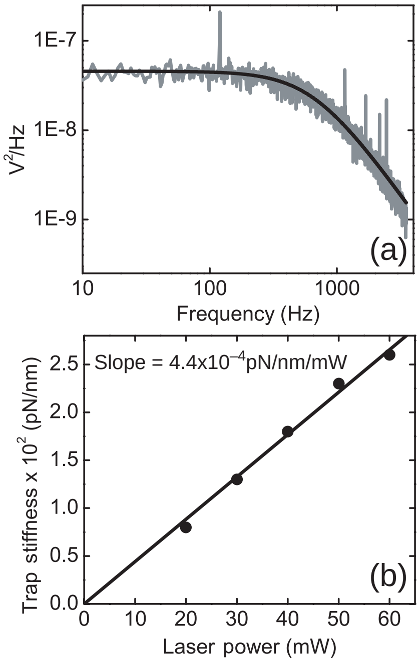


Figure 11. Example power spectrum calibration. (a) An example power spectrum (gray) and Lorentzian fit (black). (b) Power spectrum calibrations for five separate laser powers showing the linear dependence of stiffness on laser power.

Inspection of the power spectrum in Fig. 11a shows two distinct regions: the low frequency plateau and the high frequency roll off. These two sections can be qualitatively understood as representing two aspects of the motion of the trapped particle. The high frequency roll-off is representative of small, rapid oscillations of the bead in the trap. Since the distance the bead moves over these short time scales is generally small, the bead does not "feel" large changes in the force of the trap pushing it back to the center, and hence this motion has the character of free diffusion. At lower frequency, the spectrum is flat, because large-scale oscillations are attenuated by the trap pushing the bead back to the center. As a result low frequency oscillations do not exhibit the character of free diffusion; the bead feels the trap when it makes a large excursion.

The methods for calculating and then fitting the power spectrum above are adequate in most cases and for most experiments. However, several additional considerations can be added to this relatively simple approach to improve the accuracy in some situations. These include, but are not limited to, accounting for the frequency dependence of the drag coefficient, and theoretically accounting for aliasing that may be present in the signal (54).

Two other calibration methods also take advantage of thermal forces acting on trapped beads. These methods are not conceptually distinct from the power spectrum method, but treat the data differently, are susceptible to different errors, and thus provide a good cross-check with the power spectrum and direct methods above.

As one would expect, the range of bead position should decrease with increasing trap stiffness. For an overdamped harmonic potential, such as a spherical bead in an optical trap, this relation has a specific mathematical form:

$$\kappa = \frac{k_{\mathrm{B}}T}{\mathrm{var}(x)} \tag{8}$$

where the var(x) is the variance of the bead position in the trap, k_{B} is Boltzmann's constant and T is absolute temperature. This equation results from the equipartition theorem: the average thermal energy for a single degree of freedom is $1/2k_{\mathrm{B}}T$, and that the average potential energy stored in a Hookean trap is $1/2\kappa\langle x^2\rangle$. Setting these equal, rearranging and assuming that $\langle x\rangle = 0 =$ the center of the trap, results in equation 8. Thus by simply measuring the variance of the bead position and the temperature, one can determine the trap stiffness.

Alternatively the autocorrelation function of the bead position is used to determine the trap stiffness (33). This method is little different in principal than using the power spectrum, as the spectrum is the Fourier transform of the autocorrelation. The autocorrelation function is expected to exhibit an exponential decay with time constant τ, where

$$\tau = \frac{\gamma}{\kappa} \tag{9}$$

It is often easier to achieve a reliable fit to the autocorrelation than to the power spectrum, making the autocorrelation an attractive alternative to the more common power spectrum methods.

Each calibration method has its distinct advantages over the others. The methods involving direct manipulation are most often used in the course of building an optical tweezers device to verify that the thermal motion calibrations work properly. Once verified the thermal motion calibrations are much less labor intensive, and can be performed "on the fly". This allows sensitivity and stiffness for each bead to be calibrated as it is used in an experiment.

Each method relies on slightly different parameters and thus provides separate means to verify calibration accuracy. For example, the corner frequency of the power spectrum, and hence the stiffness can be determined with no knowledge of the detector sensitivity. However, this does depend on accurately calculating the drag coefficient. By calculating the stiffness from the variance one can avoid considering the drag coefficient altogether, but this method relies on accurate estimation of the sensitivity squared (β^2). Furthermore, the variance method can lead to inaccurate stiffness measurements because system drift will inflate the variance, leading to underestimated trap stiffness.

A safe course is to (1) inspect the power spectrum for evidence of external noise, (2) calculate the stiffness and sensitivity by fitting the power spectrum, (3) recalculate the stiffness with the variance method using the sensitivity determined from the power spectrum, and (4) compare the stiffness results from each method. If the stiffness agrees between the two it indicates that the drag coefficient and sensitivity are both correct. The only possibility for error would be that they are both in error in a manner that is exactly offsetting, which is quite unlikely. Comparison with the direct manipulation methods can alleviate this concern.

Optical Tweezers Compared with Other Approaches to Nanomanipulation

There are several alternatives to optical tweezers for working at the nanometer to micron scales and exerting piconewton forces. Most similar in application are magnetic tweezers, which use paramagnetic beads and an electromagnet to produce forces. The force profile of magnetic tweezers is constant on the size scale of microscopic experiments; the force felt by the bead is not dependent on displacement from a given point as with optical tweezers. This can be convenient in some circumstances, but is less desirable for holding form objects in specific locations.

A second option is the use of glass microneedles. These fine glass whiskers can be biochemically linked to a molecule of interest and their deflection provides a measure of forces and displacements. Stiffness must be measured for each needle with either fluid flow or methods relying on thermal forces similar to those described above. Glass needles can exert a broad range of forces but they only allow for force measurements along one axis, and are difficult for complex manipulations compared to optical tweezers devices.

Atomic force microscopy (AFM) is a third option for measuring small forces and displacements. Originally, this technique was used to image surface roughness by dragging a fine cantilever over a sample. A laser reflecting off the cantilever onto a photodiode detects cantilever deflection. To measure force, a cantilever of known stiffness can be biochemically linked to a structure of interest and deflections and forces measured with similar accuracy to optical tweezers. The AFM has the advantage that reliable AFMs can be purchased, while optical tweezers must still be custom built for most purposes. However, AFMs are unable to perform the complex manipulations that are simple with optical tweezers and typically are limited to forces >20 pN.

OPTICAL TWEEZERS RESEARCH

Within 4 years of the publication of the first demonstration of a single beam, three-dimensional (3D) trap, optical tweezers were being applied to biological measurements (2,7). In the years that followed, optical tweezers became increasingly sophisticated, and their contributions to biology and biophysics in particular grew rapidly. Some important experimental considerations, and contributions to basic science made possible by optical tweezers, highlighted representative assays, and results are discussed in this section.

Practical Experimental Concerns

Beyond the design and calibration challenges, there are additional concerns that come into play preparing an experimental assay for use with optical tweezers. The most general of these attached is a trappable object to the system being studied, and accounting for series compliances when interpreting displacements of this object.

Probably the most popular attachment scheme currently in use is the biotin-streptavidin linkage. Microbeads are coated with streptavidin, and the structure to be studied, if a protein, can be easily functionalized with biotin via a succinimidyl ester or other chemical crosslinker. When mixed, the streptavidin on the bead tightly binds to biotin on the structure to be studied. Beads coated in this manner tend to stick to the surfaces in the experimental chamber, which is inconvenient for setting up an experiment, and recently developed neutravidin is an attractive alternative to streptavidin with decreased nonspecific binding at close to neutral pH. Alternative attachment schemes usually involve coating the bead with a protein that specifically interacts with the structure to be studied. For example, a microtubule-associated protein might be attached to the bead; subsequently that bead sticks to microtubules, which can then be manipulated with the optical tweezers. Recombinant DNA technology can also be used; in this case the amino acid sequence of a protein is modified to add a particular residue (e.g., reactive cystein), which serves as a target for specific crosslinking to the bead. This approach provides additional control over the binding orientation of the protein, but runs the risk that the recombinant protein many not behave as the wild type.

Interpreting displacements measured with optical tweezers is complicated by compliances in the system under study, or the linkage to the trapped bead; these must be accounted for to determine the actual displacement of specific elements. Figure 12 diagrams the compliances in a sample system. A filament to which the trapped bead is linked is pulled to the right by the force generating process under study (not shown). This causes the bead to be displaced to the right within the trap by an amount, Δ Bead, which is directly measured. However, the other compliance in the system, κ_{linkage}, is also stretched and takes up some amount of the filament displacement, which can only be determined with knowledge of the link stiffness; thus the measured displacement is less than the actual displacement of the filament. In some situations it may be possible to measure the link stiffness directly, but generally this is not possible during an experiment. Furthermore, the link stiffness is usually nonlinear, and may be complicated by the bead rocking about the link under forces applied by the trap. When precision displacement measurements are desired these difficulties can be circumvented with a feedback system that maintains a constant force on the particle (force clamp). By maintaining constant force, the link is held at a fixed strain and the movements of the feedback controlled laser directly follow the positional changes of the filament.

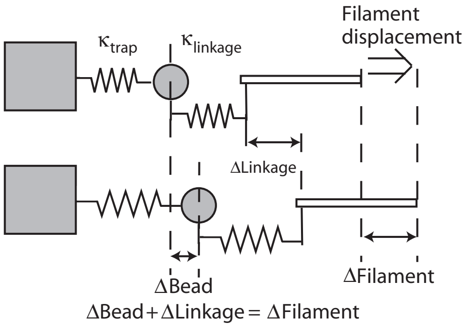


Figure 12. Schematic diagram of series compliances in optical tweezers experiments. Both the optical trap and the linkage between the bead and hypothetical filament are shown as springs. The optical trap is a well characterized spring, having been carefully calibrated prior to beginning experiments. The stiffness of the linkage, however, is unknown. When the filament is pulled to the right by a hypothetical motor or other biological system, both springs are stretched. Therefore the displacement of the trapped bead is not necessarily equal to the displacement of the filament by the motor. Examination of the system shows that the force on the trapped bead is the same as the force on the filament once all elements have reached their equilibrium. Knowledge of κ_{linkage} would allow for displacements at the bead to be used to calculate filament displacements.

Motor Proteins: Kinesin And Myosin

The kinesins and myosins are two large families of motor proteins that convert chemical energy released by adenosine triphosphate (ATP) hydrolysis into mechanical work. Kinesins move cargo along microtubules while myosins exert forces against actin filaments, most notably in muscle. Many of the mechanical and kinetic aspects of these molecules behaviors have been determined from measurements made with optical tweezers. These include the length of displacements during individual chemomechanical steps, single molecule force generation capabilities, and kinetic information about the enzymatic cycle that converts chemical into mechanical energy.

Due to the differences in the substrate along which myosin and kinesin motors move, and the nature of their movements, assays for studying them have significantly different geometries. In the case of kinesin (Fig. 13), generally a single bead coated with the motor is held in an optical trap (17,26). The motors are sparsely coated on the beads, so that only one motor interacts with the microtubule track, which has been immobilized onto a glass surface. The optical tweezers manipulate the coated bead onto the microtubule; bead movements ensue when a kinesin motor engages the microtubule lattice. If the

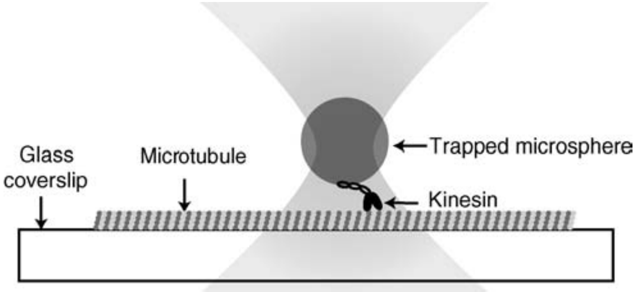


Figure 13. Schematic of an experiment to study kinesin. The motor protein kinesin is processive, meaning that the motor spends a large portion of its force generating cycle attached to the microtubule and is able to maintain movement when only a single motor is present. As a result, a single, sparsely coated bead held in an optical trap can be used to measure the force generating properties of kinesin.

optical tweezers are used in stationary mode, the records will indicate a fraction of the actual motor displacement, the remainder being taken up in the compliant link between the bead and the motor as described above. The actual displacements of the motor must then be inferred using estimates of the bead-motor link compliance, which is estimated in independent experiments (26).

Alternatively, a force clamp is applied to adjust the position of the trap to maintain a constant force on the bead as the kinesin motor travels along the microtubule. The movement of the motor is then inferred from the adjustments to the laser position, which directly reveal 8 nm steps, demonstrating that the kinesin molecule moves along the microtubule with the same periodicity as the microtubule lattice. Additional analysis of this data showed that kinesin stalls under loads of 5–8 pN, dependent on ATP concentration, and exhibits tight coupling between ATP hydrolysis and force generation (17,34).

Assays for studying myosin (Fig. 14) are somewhat more complicated, relying on two traps, each holding a microsphere, and a third large bead sparsely coated with myosin affixed to the surface of the microscope slide (35). An actin filament strung between the two smaller microspheres, each held by separate traps, is lowered into a position where myosin on the large fixed bead can pull against it. Myosin is not processive; it spends only a small fraction of the force generating cycle attached to the actin filament, and an unrestrained filament will diffuse away from the motor during the detached portion of the cycle. Thus the filament is held in close contact with the motor to allow repeated force generating interactions to be observed.

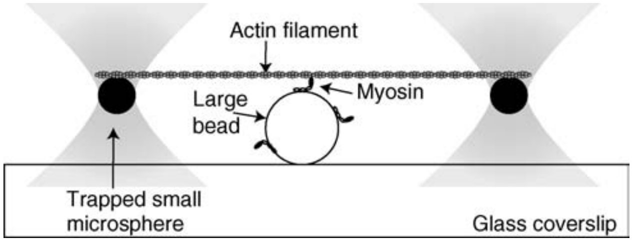


Figure 14. Schematic of an experiment to study muscle myosin. The motor protein myosin is not processive, meaning that a single copy of the motor cannot sustain movement along the actin filament. As a result an experimental arrangement that can keep the myosin in close proximity to the filament is necessary.

The actin filament and attached beads will move back and forth due to Brownian motion, which is limited by the drag on the particle and the force provided by the trap. This has two important consequences: the myosin molecule on the large bead is exposed to a number of possible binding sites along the actin filament, and attachment of the myosin cross bridge elevates the stiffness of the system sufficiently to greatly reduce the extent of thermally induced bead motion. With bandwidth sufficient to detect the full extent of microsphere Brownian motion, myosin binding events are identified by the reduction in the amplitude of the thermal motion. The distribution of positions of the trapped microspheres and filament at the onset of binding events is expected to be Gaussian of the same width as if the myosin-coated bead was absent. However, shortly after binding, the myosin motor displaces the filament and this shifts the center of the Gaussian by the distance of a single myosin step. Analysis of high resolution, high bandwidth traces of bead position with the above understanding lead to determination of the step length for single myosin subfragment-1 and heavy meromyosin: 3.5 and 5 nm, respectively (28).

Other Motor Proteins

In addition to the classic motor proteins that generate forces against cytoskeletal filaments, proteins may exhibit motor activity, not as their *raison d'etre*, but in order to achieve other enzymatic tasks. One such protein is ribonucleic acid (RNA) polymerase (RNAP), the enzyme responsible for transcription of genetic information from deoxyribonucleic acid (DNA) to RNA. The RNAP uses the energy of ATP hydrolysis to move along the DNA substrate, copying the genetic information at a rate of 10 base pairs per second. The movement is directed, and thus constitutes motor activity. The assay used to study RNAP powered movement along the DNA substrate is shown in Fig. 15. A piece of single-stranded DNA attached to the trapped glass microsphere is allowed to interact with an RNAP molecule affixed to the surface of the slide. Optical tweezers hold the bead so that the progress and force developed by RNAP can be monitored. As the RNAP molecule moves along the

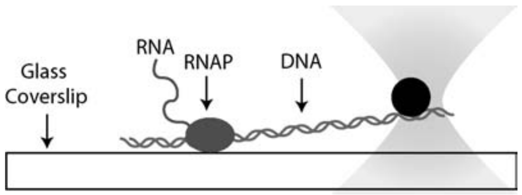


Figure 15. Schematic of an experiment to study RNA polymerase.

DNA, the load opposing the motor movement increases until the molecule stalls (36,37). Details of the kinetics of RNAP can be inferred from the relation between the force and speed of movement.

Nonstandard Trapping

A number of research efforts have studied the trapping of nonspherical objects and/or applied them to study biological processes. For example, the stiffness of a microtubule was measured using optical tweezers to directly trap and bend the rod-like polymer. Image data of the induced microtubule shape were then used in conjunction with mechanical elastic theory to determine the stiffness of the microtubule (38). Recently, there has been significant interest in using optical traps to exert torques. Generally, these techniques rely on nonspherical particles, non-Gaussian beams, or birefringent particles to create the necessary asymmetry to develop torque. One approach relies on using an annular laser profile (donut mode) with an interfering reference beam to create a trapping beam that has rotating arms (39). With such an arrangement torque generation is not dependent on the particle shape, but the implementation is relatively complicated. Two closely located traps can also be used to rotate objects by trapping them in multiple locations and moving the traps relative to one another. Using a spatial light modulator, which splits the beam into an arbitrary number of individual beams, to create and move the two traps allows rotation about an axis of choice (40).

For experimental applications, it is generally desirable to calibrate the torque exerted on the particles. The above techniques are not easily calibrated. A more readily calibrated altenative uses birefringent particles such as quartz microspheres (41). The anisotropic polarizability of the particle causes it to align to the beam polarity vector. The primary advantage of this technique is that the drag coefficient of the spherical particle is easily calculated, allowing torques can be calibrated by following rotational thermal motion of the particle, similarly to the techniques described for thermal motion calibrations above.

Modulation of trap position along with laser power or beam profile can create a laser line trap (42–44). The scheme is similar to time sharing a single laser between several positions to create several traps. However, to form a line trap the laser focus is rapidly scanned through positions along a single line. With simultaneous power modulation a linear trapping region capable of applying a single constant force to a trapped particle along the entire length of the trap is created. This technique can be used to replace the conventional force clamp relying on electronic feedback.

Some Other Optical Tweezers Assays

Optical tweezers have made numerous contributions to other subfields. A number of groups have utilized optical tweezers to study DNA molecules. Stretching assays have produced force extension relations for purified DNA, and stretching of individual nucleosomes revealed sudden drops in force indicative of the opening of the coiled DNA structure (45). Optical trapping has also been used to directly study the forces involved in packaging DNA into a viral capsid. Packaging was able to proceed against forces > 40 pN, and the force necessary to stall packaging was dependent on the how much of the DNA was already packaged into the capsid (46). Double-stranded DNA has also been melted (unzipped) by pulling the strands apart with optical tweezers: this established that lamba phage (a bacterial virus) DNA unzips and rezips in the range of 10–15 pN, dependent on the nucleotide sequence (47). Stretching RNA molecules to unfold loops and other secondary structures with a similar assay has been used to test a general statistical mechanics result known as the Jarzinsky Inequality (48).

Similar stretching experiments have been performed on proteins. A good example is the large muscle protein titin, which mediates muscle elasticity. Optical tweezers were applied to repeatedly stretch the molecule, providing evidence of mechanical fatigue of the titin molecule that could be the source of mechanical fatigue in repeatedly stimulated muscles (49).

Experiments that study interactions in larger, more complex systems are increasing common. Microtubules associated to mitotic chromosome kinetochores have been studied with optical tweezers. Forces of 15 pN were generally found to be insufficient to detach kinetochore bound microtubules and kinetochore attachment was found to modify microtubule growth and shortening (21). A number of studies have also applied optical tweezers to study structures inside intact cells and the force generating ability of highly motile cells such as sperm (6,50).

Beyond biological measurements, optical tweezers have utility in assembling micron scale objects in desired positions. Weakly focused lasers operating similarly to optical tweezers have been used to directly pattern multiple cell types on a surface for tissue engineering (55). Optical trapping has also been used for assembly and organization of nonbiological devices, such as groups of particles (51) and 3D structures, such as a crystal lattice (52).

Additionally, optical tweezers have been applied to great advantage in material and physical sciences. A substantial body of work has applied optical tweezers to study colloidal solutions and microrheology (53).

BIBLIOGRAPHY

1. Ashkin A. Acceleration and trapping of particles by radiation pressure. Phys Rev Lett 1970;24:156–159.
2. Ashkin A, Dziedzic JM, Bjorkholm JE, Chu S. Observation of a single-beam gradient force optical trap for dielectric particles. Opt Lett 1986;11:288–290.
3. Ashkin A, Dziedzic JM, Yamane T. Optical trapping and manipulation of single cells using infrared laser beams. Nature (London) 1987;330:769–771.
4. Ashkin A, Dziedzic JM. Optical trapping and manipulation of viruses and bacteria. Science 1987;235:1517–1520.
5. Block SM, Blair DF, Berg HC. Compliance of bacterial flagella measured with optical tweezers. Nature (London) 1989; 338:514–518.
6. Tadir Y, et al. Micromanipulation of sperm by a laser generated optical trap. Fertility Sterility 1989;52:870–873.
7. Block SM, Goldstein LS, Schnapp BJ. Bead movement by single kinesin molecules studied with optical tweezers. Nature (London) 1990;348:348–352.

8. Grover SC, Skirtach AG, Gauthier RC, Grover CP. Automated single-cell sorting system based on optical trapping. J Biomed Opt 2001;6:14–22.
9. Leach J, et al. 3D manipulation of particles into crystal structures using holographic optical tweezers. Opt Express 2004;12:220–226.
10. Ashkin A. Forces of a single-beam gradient laser trap on a dielectric sphere in the ray optics regime. Biophys J 1992;61: 569–582.
11. Visscher K, Brakenhoff G, Lindmo T, Brevik I. Theroretical study of optically induced forces on spherical particles in a single beam trap I: Rayleigh scatters. Optik 1992;89:174–180.
12. Nahmias YK, Odde DJ. A dimensionless parameter for escape force calculation and general design of radiation force-based systems such as laser trapping and laser guidance. Biophys J 2002;82:166A.
13. Nahmias YK, Gao BZ, Odde DJ. Dimensionless parameters for the design of optical traps and laser guidance systems. Appl Opt 2004;43:3999–4006.
14. Allersma MW, et al. Two-dimensional tracking of ncd motility by back focal plane interferometry. Biophys J 1998;74:1074–1085.
15. Brouhard GJ, Schek HT, Hunt AJ. Advanced optical tweezers for the study of cellular and molecular biomechanics. IEEE Trans Biomed Eng 2003;50:121–125.
16. Gittes F, Schmidt C. Interference model for back-focal-plane displacement detection in optical tweezers. Opt Lett 1998a;23:7–9.
17. Visscher K, Schnitzer MJ, Block SM. Single kinesin molecules studied with a molecular force clamp. Nature (London) 1999;400:184–189.
18. Ruff C, et al. Single-molecule tracking of myosins with genetically engineered amplifier domains. Nature Struct Bio 2001;8: 226–229.
19. Guck J, et al. The optical stretcher: A novel laser tool to micromanipulate cells. Biophys J 2001;81:767–784.
20. Smith SB, Cui YJ, Bustamante C. Overstretching B-DNA: The elastic response of individual double-stranded and single-stranded DNA molecules. Science 1996;271:795–799.
21. Hunt AJ, McIntosh JR. The dynamic behavior of individual microtubules associated with chromosomes in vitro. Mol Biol Cell 1998;9:2857–2871.
22. Kuo SC, Sheetz MP. Force of single kinesin molecules measured with optical tweezers. Science 1993;260:232–234.
23. Florin EL, Pralle A, Horber JKH, Stelzer EHK. Photonic force microscope based on optical tweezers and two-photon excitation for biological applications. J Struct Biol 1997;119:202–211.
24. Friese M, Rubinsztein-Dunlop H, Heckenberg N, Dearden E. Determination of the force constant of a single-beam gradient trap by measurement of backscattered light. Appl Opt 1996;35:7112–7116.
25. Pralle A, et al. Three-dimensional high-resolution particle tracking for optical tweezers by forward scattered light. Microsc Res 1999;44:378–386.
26. Svoboda K, Schmidt CF, Schnapp BJ, Block SM. Direct observation of kinesin stepping by optical trapping interferometry. Nature 1993;365:721–727.
27. Lang MJ, Asbury CL, Shaevitz JW, Block SM. An automated two-dimensional optical force clamp for single molecule studies. Biophys J 2002;83:491–501.
28. Molloy JE, et al. Movement and force produced by a single myosin head. Nature (London) 1995;378:209–212.
29. Visscher K, Brakenhoff GJ, Krol JJ. Micromanipulation by multiple optical traps created by a single fast scanning trap integrated with the bilateral confocal scanning laser microscope. Cytometry 1993;14:105–114.
30. Gensch T, et al. Transmission and confocal fluorescence microscopy and time-resolved fluorescence spectroscopy combined with a laser trap: Investigation of optically trapped block copolymer micelles. J Phys Chem *B* 1998;102:8440–8451.
31. Happel J, Brenner H. Low Reynolds number hydrodynamics. With special applications to particulate media. Leiden: Noordhoff International Publishing; 1973.
32. Gittes F, Schmidt C. Thermal noise limitations on micromechanical experiments. Eur Biophys J with Biophys Lett 1998b;27:75–81.
33. Meiners JC, Quake SR. Direct measurement of hydrodynamic cross correlations between two particles in an external potential. Phy Rev Lett 1999;82:2211–2214.
34. Schnitzer MJ, Block SM. Kinesin hydrolyses one ATP per 8-nm step. Nature (London) 1997;388:386–390.
35. Finer JT, Simmons RM, Spudich JA. Single myosin molecule mechanics - piconewton forces and nanometer steps. Nature (London) 1994;368:113–119.
36. Wang MD, et al. Force and velocity measured for single molecules of RNA polymerase. Science 1998;282:902–907.
37. Yin H, et al. Transcription against an applied force. Science 1995;270:1653–1657.
38. Felgner H, Frank R, Schliwa M. Flexural rigidity of microtubules measured with the use of optical tweezers. J Cell Sci 1996;109(Pt 2):509–516.
39. Paterson L, et al. Controlled rotation of optically trapped microscopic particles. Science 2001;292:912–914.
40. Bingelyte V, Leach J, Courtial J, Padgett MJ. Optically controlled three-dimensional rotation of microscopic objects. Appl Phys Lett 2003;82:829–831.
41. La Porta A, Wang MD. Optical torque wrench: Angular trapping, rotation, and torque detection of quartz microparticles. Phy Rev Lett 2004; 92.
42. Liesfeld B, Nambiar R, Meiners JC. Particle transport in asymmetric scanning-line optical tweezers. Phy Rev 2003; 68.
43. Nambiar R, Meiners JC. Fast position measurements with scanning line optical tweezers. Opt Lett 2002;27:836–838.
44. Nambiar R, Gajraj A, Meiners JC. All-optical constant-force laser tweezers. Bio J 2004;87:1972–1980.
45. Bennink ML, et al. Unfolding individual nucleosomes by stretching single chromatin fibers with optical tweezers. Nature Struct Biol 2001;8:606–610.
46. Smith DE, et al. The bacteriophage phi 29 portal motor can package DNA against a large internal force. Nature (London) 2001;413:748–752.
47. Bockelmann UP, et al. Unzipping DNA with optical tweezers: high sequence sensitivity and force flips. Biophys J 2000; 82:1537–1553.
48. Liphardt J, et al. Equilibrium information from nonequilibrium measurements in an experimental test of Jarzynski's equality. Science 2002;296:1832–1835.
49. Kellermayer MS, Smith SB, Granzier HL, Bustamante C. Folding-unfolding transitions in single titin molecules characterized with laser tweezers. Science 1997;276:1112–1116.
50. Aufderheide KJ, Du Q, Fry ES. Directed positioning of micronuclei in paramecium-tetraurelia with laser tweezers—absence of detectable damage after manipulation. J Eukaryotic Microbiol 1993;40:793–796.
51. Misawa H, et al. Multibeam laser manipulation and fixation of microparticles. Appl Phys Lett 1992;60:310–312.
52. Holmlin RE, et al. Light-driven microfabrication: Assembly of multicomponent, three-dimensional structures by using optical tweezers. Angew Chem Int Ed Engl 2000;39:3503–3506.
53. Lang MJ, Block SM. Resource letter: LBOT-1: Laser-based optical tweezers. Am J Phy 2003;71:201–215.

54. Berg-Sorensen, K & Flyvbjerg, H. (2004) Power spectrum analysis for optical tweezers. Review of Scientific Instruments 75, 594–612.
55. Odde, DJ & Renn, M.J. (2000) Laser-guided direct writing of living cells. Biotechnology and Bioengineering 67, 312–318.

See also FIBER OPTICS IN MEDICINE; MICROSURGERY; NANOPARTICLES.

ORAL CONTRACEPTIVES. See CONTRACEPTIVE DEVICES.

ORTHOPEDIC DEVICES, MATERIALS AND DESIGN. See MATERIALS AND DESIGN FOR ORTHOPEDIC DEVICES.

ORTHOPEDIC DEVICES MATERIALS AND DESIGN OF

AMIT BANDYOPADHYAY
SUSMITA BOSE
Washington State University
Pullman, Washington

INTRODUCTION

Musculoskeletal disorders are recognized as among the most significant human health problems that exist today, costing society an estimated $254 billion every year, and afflicting one out of seven Americans. Musculoskeletal disorders account for nearly 70 million physician office visits in the United States annually and an estimated 130 million total healthcare encounters including outpatient, hospital, and emergency room visits. In 1999, nearly 1 million people took time away from work to treat and recover from work-related musculoskeletal pain or impairment of function in the low back or upper extremities (1). There is still an ongoing debate on cause, nature and degrees of musculoskeletal disorders particularly related to work and how to reduce it. However, it is agreed unanimously that the number of individuals with musculoskeletal disorders will only increase over the coming years, as our population ages. According to the World Health Organization (WHO), these factors are called "work-related conditions", which may or may not be due to work exposures (1). Some of these factors include: (*1*) physical, organizational, and social aspects of work and the workplace, (*2*) physical and social aspects of life outside the workplace, including physical activities (e.g., household work, sports, exercise programs), economic incentives, and cultural values, and (*3*) the physical and psychological characteristics of the individual. The most important of the latter include age, gender, body mass index, personal habits including smoking, comorbidities, and probably some aspects of genetically determined predispositions (1). Among the various options to treat musculoskeletal disorders, use of orthopedic devices is becoming a routine, with the number of annual procedures approaching five million in the United States alone (2). Some of the common orthopedic devices include joint replacement devices for hip and knee and bone fixation devices such as pins, plates and screws for restoring lost structure and function.

Materials and design issues of orthopedic devices are ongoing challenges for scientists and engineers. Total hip replacement (THR) is a good example to understand some of these challenges. Total hip replacements are being used for almost past 60 years with a basic design concept that was first proposed by Charnley et al. (3). A typical lifetime for a hip replacement orthopedic device is between 10 and 15 years, which remained constant for the last five decades. From design point of view, a total hip prosthesis is composed of two components: the femoral component and the cup component. The femoral component is a metal stem, which is placed into the marrow cavity of the femoral bone, ending up with a neck section to be connected to the ball or head. The neck is attached to the head, a ball component that replaces the damaged femoral head. The implant can be in one piece where the ball and the stem are prefabricated and joined at the manufacturing facility, this is called a monobloc construction. It can also be in multiple pieces, called modular construction, which the surgeon put together during the time of the surgery based on patient needs, such as the size of the ball in the cavity. An acetabular component, a cup, is also implanted into the acetabulum, which is the natural hip socket, in the pelvic bone. The femoral component is typically made of metallic materials such as Ti or its alloys. The balls of the total prostheses are made either from metallic alloys or ceramic materials. The hip cups are typically made from UHMWPE (ultrahigh molecular weight polyethylene). Some part of the stem can be coated with porous metals or ceramics, which is called cementless implant, or used as uncoated in presence of bone cement, which is called cemented implants. Bone cements, which is primarily poly (methyl methacrylate) (PMMA) based, stabilizes the metallic stem in the femoral bone for cemented implants. However, for cementless implants, the porous coating helps in tissue bonding with the implants surface and this biological bonding helps implants to stabilize (Fig. 1).

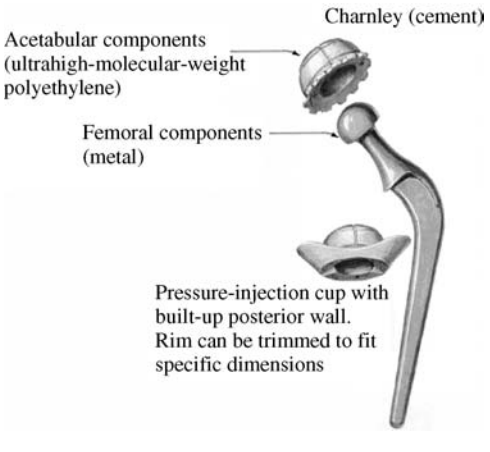

Charnley (cement)
Acetabular components (ultrahigh-molecular-weight polyethylene)
Femoral components (metal)
Pressure-injection cup with built-up posterior wall. Rim can be trimmed to fit specific dimensions

Figure 1. An example of a modern cemented hip prosthesis design and various components.

As it can be seen that THR by itself is a complex device that incorporates multiple materials and designs. However, all of these artificial materials mimic neither the composition nor the properties of natural bone. The inorganic part of natural bone consists of $\sim$ 70 wt%, consists of calcium phosphate. Moreover, patient pool is different in terms of their age, bone density, physiological environment, and postoperative activities. To complicate matters further, surgical procedure and total surgery time can also be different for various patients. Because of these complications it is difficult, if not impossible, to point out the exact reason(s) for the low *in vivo* lifetime for these implants, which remained constant for the past 50 years. However, it is commonly believed that stress shielding is one of the key factors for limiting the lifetime of these implants. Because a metal stem is introduced into the bone, which has a complex architecture including an outer dense surface or cortical bone and an inner porous surface or cancellous bone, during the total hip surgery, the load distribution within the body shifts and the load transfer between tissue and implant does not match with normal physiological system. Typically, more load will be carried by the metal implants due to their high stiffness which will cause excess tissue growth in the neck regions. At the same time, upper part of the femoral bone will carry significantly less load and become weaker, which will make it prone to premature fracture. Both of these factors contribute to the loosening of hip implant that reduces the implant lifetime. This is also called stress-shielding effect, in general, which means the loss of bone that occurs adjacent to a prosthesis when stress is diverted from the area. To reduce stress-shielding, an ideal hip implant needs to be designed in a way that it has similar stiffness as natural femoral bone. However, current biomedical industries mostly use materials for load bearing implants that are typically designed and developed for aerospace or automotive applications, instead of developing new materials tailored specifically for orthopedic devices needs. But the time has come when materials need to be designed for specific biomedical applications to solve long-standing problems like stress-shielding in THR. Though THR is used as an example to show the complexity of materials and design issues in orthopedic devices but THR is not alone. Most orthopedic devices suffer from complex materials and design challenges to satisfy their performance needs.

FACTORS INFLUENCING ORTHOPEDIC DEVICES

There are several factors that need to be considered to design an orthopedic device. From the materials point of view, usually mechanical property requirement, such as strength, toughness, fatigue degradation becomes the most important issues as long as the materials are nontoxic and biocompatible. However, as the body tissue interacts with the surface of the device during *in vivo* lifetime, surface chemistry becomes one of the most important aspects for orthopedic devices. Most complex device functions cannot be accomplished using only one material, and require applications of structures made of multimaterials. As a result, compatibility of multimaterials in design and manufacture becomes another issue. Figure 2 summarizes various parameters that are important toward design and development of an orthopedic implants. Four main areas of orthopedic devices include materials issues, design issues, biocompatibility issues and patient specific information. All four of them are complex in nature and their interactions are even more difficult to appreciate. The following section offers some basic understanding of all of those issues.

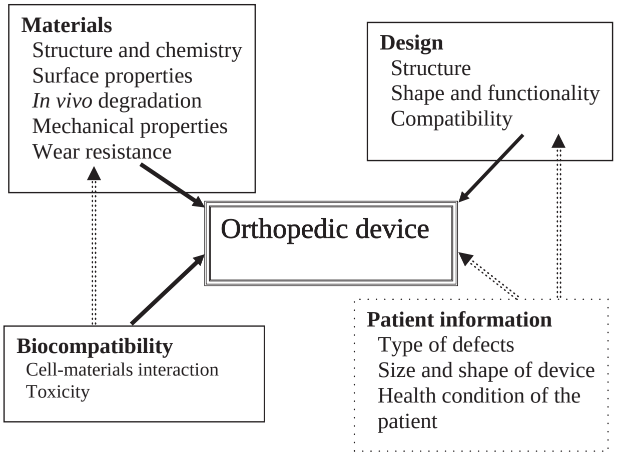


Figure 2. Materials and design parameters for orthopedic devices.

MATERIALS ISSUES IN ORTHOPEDIC DEVICES

Selection of appropriate material is probably the most important issue in successful design and development of orthopedic devices. Among various materials related issues, (*1*) physical and chemical properties, (*2*) mechanical properties, (*3*) surface properties, and (*4*) *in vivo* degradation or corrosion behavior are some of the most important ones. In general, it is most widely accepted to use metallic materials for load bearing, and polymers and ceramic–polymer composites for nonload bearing applications. Some ceramic compositions and glasses are also used for nonload bearing coatings and defect filling applications. Development of biomaterials and materials processing for different orthopedic device applications is currently a very active research area (4–6) and new materials are constantly being developed to meet the current and future needs.

Physical and Chemical Properties

Physical properties include density, porosity, particle size, and surface area type of information. Composition or chemistry is probably the most important chemical property. It is important to realize that orthopedic devices cannot be built with materials that are carcinogenic. Nontoxic materials that do not leach harmful metal ions *in vivo* are ideal. Apart from dense structures, partially or completely porous materials are also used for orthopedic devices. If a porous material is used, then some of the properties, such as pore size, pore volume, and pore–pore interconnectivity become important. Typically, for most porous materials, an optimum pore size between 100 and 500 μm are used in which cells can grow and stay healthy. Higher pore volume usually adds more space for cells to grow and naturally anchor the device. However, this also exposes higher surface area of the device material that can cause faster degradation or corrosion, which sometimes can be a concern depending on the materials used. For polymeric materials, chemistry, and structure are important because materials with the same chemistry, but different structure can show different *in vivo* response. This is particularly important for biodegradable polymers, such as poly lactic acids (PLA) and polyglycolic acids (PGA) and their copolymers. Trace element is another important factor in materials selection. Sometimes even a small amount of impurities can cause harmful effects *in vivo* (7). However, in calcium phosphate based ceramics, small addition of impurity elements actually proved to be beneficial for mechanical and biological response (8).

Mechanical Properties

Mechanical properties are important in selecting materials for orthopedic devices. Among various mechanical properties, uniaxial and multiaxial strength, elastic modulus, toughness, bending strength, wear resistance, fatigue resistance are some of the most important ones. Mechanical property requirements are tied to specific applications. For example, for the stem in THR, it should have high strength, low modulus, and very high fatigue resistance. As a result, due their low modulus, Ti and its alloys are usually preferred over high modulus metal alloys for the stem part of THR. However, for the acetabular component, high wear resistance requirement is the most important one and high-density polymers are preferred for the acetabular component. Mechanical properties are also linked on how they are processed. For example, casting devices in their near final shape can be a relatively inexpensive way to make complex shapes. However, material selection is critical in casting. The use of cast stainless steel for femoral hip stems is one experience that led to a high failure rate. This result generated significant debate in 1970s regarding processing of load bearing implants using casting and the options were considered by the ASTM Committee F04 on Medical and Surgical Materials and Devices to ban cast load bearing implants (8). For devices made of degradable polymers, strength loss due to degradation is an important factor. For example, materials compositions and structures in resorbable sutures are designed for different degradation times to achieve desired strength loss characteristics.

In vivo Degradation

Some orthopedic devices require materials to be bioresorbale or biodegradable, which will dissolve in body fluid as natural tissue repairs the site. Except for a few polymer and ceramic compositions, most materials are nondegradable in Nature. The degradation behavior is controlled by three basic mechanisms and they are (*1*) physiologic dissolution, which depends on pH and composition of calcium phosphate; (*2*) physical disintegration, which may be due to biochemical attack at the grain boundaries or due to high porosity; and (*3*) biological factors, such as phagocytosis.

In most materials, not just one mechanism but a combination of all three mechanisms control biodegradation behavior. Among them, biological factors are probably the most interesting ones. Though the actual process is quite complex (9), a simplistic action sequence can be viewed as osteoclastic cells slowly eat away the top surface of the foreign material and stimulate osteoblast cells. Osteoblast cells then come and deposit new bone to repair the site. Such dynamic bone remodeling is a continuous process within every human body. The rate at which osteoblastic deposition and osteoclastic resorption are taking place changes with the age of the person. This process controls the overall bone density. In terms of materials, *in vivo* degradation of polymeric materials is probably the most well-characterized field. Numerous products are available in which degradation kinetics has been tailored for specific applications. However, the same is not true for ceramics. Controlled degradation ceramics are not commercially available though degradation behavior of some calcium phosphate based ceramics is well documented (10). For metallic implants, the most serious concern regarding *in vivo* degradation is metal ion leaching or corrosion of the implants, which can cause adverse biological reactions. Corrosion products of nickel, cobalt, and chromium can form metal–protein complexes and lead to allergic reactions (7,11). Early reports of allergic reactions were reported with metal on metal (MOM) total hips (12). The inflammatory response to metallic wear debris from these devices may have been enhanced due to the high corrosion rate of the small wear particles. However, there is a lack of a predictable relationship between corrosion and allergies except in a few cases, such as vitallium implants (13,14). The number of patients with allergic reactions is not large, and it remains to be proven whether corrosion of devices causes the allergy, or the reactions are only manifest in patients with preexisting allergies. Most materials that are currently used in load bearing dental and orthopedic devices are plates and screws and they show minimum long-term degradation and health related concerns such as allergies.

Surface Properties

Surface property of materials is another important parameter for orthopedic device design. Once implanted, it is the surface that the body tissue will see first and interact. As a result, surface chemistry and roughness both are important parameters for device design. Devices that are designed for different joints, where wear is a critical issue, smooth surface is preferred there. For example, in knee joints UHMWPE is used to reduce wear debris. But most other places, where tissue bonding is necessary, rough surface or surface with internal porosity is preferred primarily to enhance physical attachment. However, biomechanical and biochemical bonding to device surfaces are still subject of active scientific investigation (15). Tailoring internal porosity and chemistry of metallic implants is still an active research area. Either metal on metal or ceramic on metal coatings are used to achieve this goal. Different

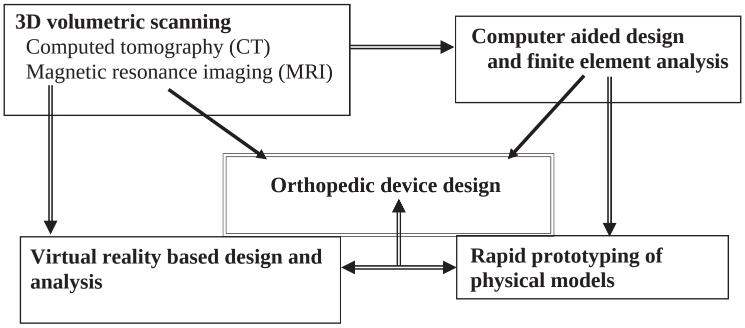

3D volumetric scanning
Computed tomography (CT)
Magnetic resonance imaging (MRI)
Computer aided design and finite element analysis
Orthopedic device design
Virtual reality based design and analysis
Rapid prototyping of physical models

Figure 3. Recent trends in orthopedic device design.

manufacturing techniques are used to modify surfaces of metal implants. Among them, partial sintering of metal powders on metallic implants is one approach (16). For example, Ti powders are sintered on cp-Ti or Ti6Al4V devices, where after sintering the sintered layers leave some porosity in the range of 100–500 μm for osteointegration by bone tissues. Similar techniques are also used for metal fibers to create a mesh with varying porosity for improved osteointegration (17). For ceramic coatings, such as calcium phosphate on Ti or Co–Cr based devices, plasma-spray technique is used. Typically, a direct current (DC) plasma gun is used under a controlled environment to coat metallic device with several hundred microns of calcium phosphate-based ceramics. Because of significantly better bioactivity of calcium phosphates, these coated devices show improved tissue–materials interaction and better long-term tissue bonding (18). In case of THRs, ceramics coated cementless implants are placed without any bone cement during surgery. During healing, body tissue forms strong bonds with the coating and anchors the device. This biological fixation is believed to be equal or better than cemented implants in which bone cement is used during surgery to anchor the device. Though the idea of cementless implants is great, but lack of interfacial strength at the implant metal and ceramic coating interface is still a concern and subject of active research. Formation of amorphous calcium phosphate during processing of plasma-sprayed ceramic coatings increase potential biodegradation rate for the coating material, which is another major concern for these devices. In general, though coated implants are promising, a significant number of uncoated implants are still used in surgery every day that has worked for a long time. In fact there are more research data available today on uncoated implants than on the coated ones.

DESIGN ISSUES IN ORTHOPEDIC DEVICES

Design of orthopedic devices is focused on the needs for that particular problem. As a result, for the same device (spinal grafting cage or THR), different designs can be found from various device makers. These devices can be in single piece or multiple-piece, made from the same or different material(s). As a result challenges are significantly different for multiple piece multimaterial devices like THR than single piece ones like bone screws or plates. For multiple piece multimaterial devices, compatibility among different materials/pieces and overall functionality becomes a more serious design issue.

Current practice in device design usually starts from biomechanical analysis of stress distribution and functionality of a particular device. If it is a joint related device, it is important that the patient can actually move the joint along multiple directions and planes to properly restore and recover functionality of that joint. Figure 3 shows some of the recent trends in orthopedic device designs. During the past 10 years, computer aided design (CAD) and rapid prototyping (RP) based technologies have played a significant role in orthopedic device design. Using this approach, real information from patients can be gathered using a computed tomography (CT) or magnetic resonance imaging (MRI) scans, which then can be visualized in three dimensions (3D). This 3D data can be transformed to a CAD file using different commercially available software.

The CAD file can be used to redesign or modify orthopedic devices that will be suitable to perform patient's need. If necessary, the device can also be tested in a virtual world using finite element analysis (FEA) to optimize its functionality. Optimized device can then be fabricated using mass manufacturing technologies such as machining and casting. If small production volume is needed, then RP technologies can be used. In RP, physical objects can be directly built from a CAD file without using any part specific tooling or dies. Rapid Prototyping is an additive or layer by layer manufacturing process in which each layer will have a small thickness, but the X and Y dimensions will be based on the part geometry. Because no tooling is required, batches as small as 1 or 2 parts can be economically manufactured. Most RP processes are capable of manufacturing polymer parts with thermoset or thermoplastic polymers. Some of the RP techniques can also be used to manufacture metal parts. Figure 4 shows a life-sized human femur made of Ti6Al4V alloy using laser engineered net shaping (LENS) process. The LENS technology uses metal powders to create functional parts that can be used in many demanding applications. The process uses up to 2 kW of Nd:YAG laser power focused onto a metal substrate to create a molten puddle on the substrate surface. Metal powder is then injected into the molten puddle to increase the material volume. The substrate is then scanned relative to the deposition apparatus to write lines of the metal with a finite width and thickness. Laser engineered net shaping is an exciting technology for orthopedic devices because it can directly build functional parts that can be used for different applications instead of

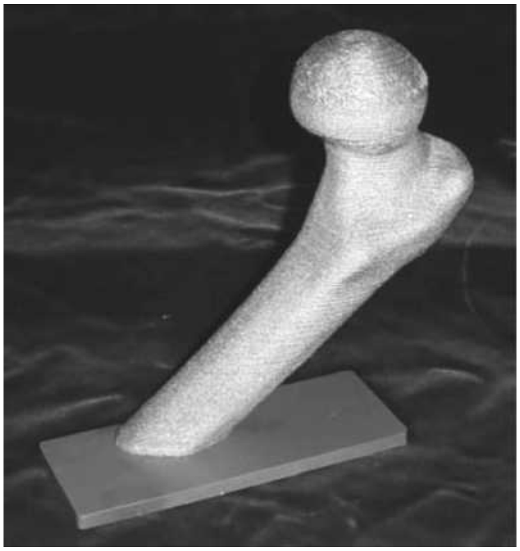


Figure 4. A life-sized human femur made of Ti6Al4V alloy using LENS.

other RP parts that are typically used for "touch and feel" applications. Commercial RP processes have also been modified to make ceramic parts for orthopedic devices. Figure 5 shows one such example in which reverse engineering of horse's-knuckle was used to show how fused deposition modeling (FDM), a commercial RP process, can be used to create tailored porosity bone implants in which pore size and pore volume can be varied simultaneously keeping the outside geometry constant (19,20). The FDM process was used to make polymer molds of to cast porous ceramic structures. The mold was designed from the CAD file of the horse's knuckle. The CAD file was created from the 3D volumetric data received from the CT scan of the bone. Such examples demonstrate the feasibility of patient specific implants through novel design and manufacturing tools.

BIOCOMPATIBILITY ISSUES IN ORTHOPEDIC DEVICES

Biocompatibility issue is an important issue in orthopedic device design and development, but it is usually considered during materials selection and surface modification. From cell materials interaction point of view, materials can be divided into three broad categories: (*1*) toxic; (*2*) nontoxic and bioinert, and (*3*) nontoxic and bioactive. For any application in the physiological environment, a material must be nontoxic. A bioinert material is nontoxic, but biologically inactive such as Ti metal. A bioactive material is the one that elicits a specific biological response at the interface of the biological tissue and the material, which results in formation of bonding between tissue and material. An

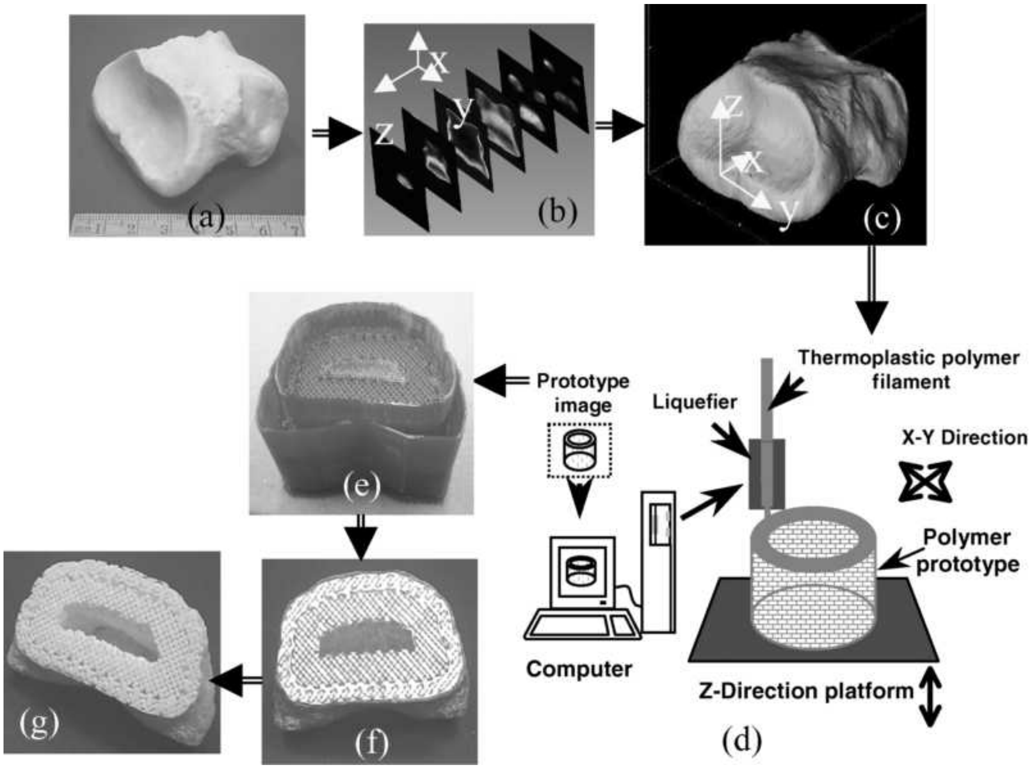


Figure 5. (a) A real bone; (b) CT scans of the bone; (c) a CAD file from the CT scans; (d) FDM process. (e) a polymer mold processed via FDM of the desired bone; (f) alumina ceramic slurry infiltrated polymer mold; (g) controlled porosity alumina ceramic bone graft.

example of bioactive material will be hydroxyapatite, $Ca_{10}(PO_4)_6(OH)_2$, which is similar in composition to the inorganic part of natural bone. The terms biodegradable, bioresorbable, and bioabsorbable are often used interchangeably. For most orthopedic devices, bioactive surfaces are ideal for better cell–materials attachment.

In vivo cell–materials interaction is a fairly complex process. In simple terms, when orthopedic devices are placed inside, our body will try to isolate the device by forming a fibrous tissue around it, which is particularly true for devices made with bioinert materials. If the material is bioactive, bone cells will first attach to the implant surface and then grow or proliferate. A material shows good biocompatibility when cells will attach and grow quickly on the device surface. Growth factors, such as bone morphogenic proteins (BMP), are sometimes used to stimulate cell attachment and growth behavior *in vivo*. After proliferation, cells will differentiate or produce new bones, which will repair the site and anchor the device. All three stages, attachment, proliferation, and differntiation of bone cells are important for repair and reconstruction of musculoskeletal disorders.

PATIENT SPECIFIC ORTHOPEDIC DEVICE: A FUTURE TREND

Patient specific device is the future trend for repair and reconstruction of musculoskeletal disorders. Due to the advancement of CAD and RP based small volume manufacturing technologies, orthopedic devices for people with special needs will be designed to meet the specific requirements. Such activities are currently pursued in academic research and hope to translate into standard industrial practice in the next one or two decades. Three-dimensions (3D) visualization of disorders using computer images and physical models, and follow-up discussion between patients and physicians will help to better educate the patient population about their problems and possible options. These options can then be transformed to physical models for trials before placing them *in vivo*. The goal is to reduce revision surgeries and increase device lifetime while maintaining the quality of life for the patient population. Various innovative scientific and engineering advancements toward novel materials and design options are helping us to make significant developments to achieve this goal.

BIBLIOGRAPHY

Cited References

1. Musculoskeletal Disorders and the Workplace: Low Back and Upper Extremities. New York: National Academic Press; 2001.
2. Materials science. Encyclopædia Britannica. 2005. Encyclopædia Britannica Premium Service.
3. Charnley J, Kamangar A, Longfield MD. The optimum size of prosthetic heads in relation to the wear of plastic sockets in total replacement of the hip. Med & Biol Eng 1969;1:31–39.
4. Hench LL. Bioceramics: From concept to clinic. J Amer Ceram Soc 1991;74(7):1487–510.
5. Burg KJL, Porter S, Kellam JF. Biomaterials development for bone tissue engineering. Biomaterials 2000;21:2347–2359.
6. Berndt CC, Haddad GN, Farmer AJD, Gross KA. Thermal spraying for bioceramic applications. Mater Forum 1990 14:161–173.
7. Mayor MB, Merritt K, Brown SA. Metal allergy and the surgical patient. Am J Surg 1980;139:447–479.
8. Chao EYS, Coventry MB. Fracture of the femoral component after total hip replacement. J Rone Joint Surg 1981;63A:1078–1094.
9. Roodman GD. Mechanisms of bone metastasis. N Eng J Med 2004;350(16):1655–1664.
10. Ravaglioli A, Krajewski A. Bioceramics: Materials, Properties, Application. London: Chapman and Hall, 1992. pp. 156–197.
11. Fontanna MG, Greene ND. Corrision Engineering. New York: McGraw-Hill; 1978.
12. Evans EM, Freeman MAR, Miller AJ, Vemon-Roberts B. Metal sensitivity as a cause of bone necrosis and loosening of prostheses in total joint replacements. J Bone Joint Surg 1974;56B:626–642.
13. Halpin DS. An unusual reaction in muscle in association with Vitallium plate: A report of possible metal hypersensitivity. J Bone Joint Sur Br 1975;57(4):451–453.
14. Garcia DA. Biocompatibility of dental implant materials measured in an animal model. J Dental Res. 1981;60(1):44–49.
15. Roberts WE. Osseous adaptation to continuous loading of rigid endosseous implants. Am J Orthod 1984;86(2):95–111.
16. Pilliar RM. Powder metal-made orthopaedic implants with porous surfaces for fixation by tissue ingrowth. Clin Orthop 1983;176:42–51.
17. Ducheyne P, Martens M. Orderly oriented wire meshes as porous coatings on orthopaedic implants II: The pore size, interfacial bonding and microstructure after pressure sintering of titanium OOWM. Clin Mats 1986;1:91–98
18. Ducheyne P, Qiu Q. Bioactive ceramics: the effect of surface reactivity on bone formation and bone cell function. Biomaterials 1999;20:2287–2303.
19. Darsell J, Bose S, Hosick H, Bandyopadhyay A. From CT Scans to Ceramic Bone Grafts. J Am Ceramic Soc 2003; 86(7):1076–1080.
20. Bose S, et al. Pore Size and Pore Volume Effects on Calcium Phosphate Based Ceramics. Mat Sci Eng 2003; C 23:479–486.

ORTHOPEDICS PROSTHESIS FIXATION FOR

Patrick J. Prendergast
Trinity Centre for
Bioengineering
Dublin, Ireland

INTRODUCTION

The fixation of an orthopedic implant should secure it rigidly to the underlying bone. The ideal fixation will sustain high forces, pain free, for the remaining lifetime of the patient. Difficulties in achieving this objective arise because (1)

1. The loads are often several times body weight in the lower extremity.The loads are fluctuating, or cyclic, and furthermore extremely high loads can occur occasionally (2).

2. The presence of the implant alters the stress transfer to the underlying bone leading to bone remodelling or fibrous tissue formation at the bone/implant interfaces. This can threaten the long-term mechanical integrity of the prosthetic replacement.
3. The range of materials that can be placed in contact with bone is limited by biocompatibility issues.

The fixation of an orthopedic implant may be catagorized as either cemented fixation or biological fixation.

Cemented fixation involves securing the implant into the bone with a "bone cement." By far the most common bone cement is based on the (polymer polymethylmethacrylate (PMMA)). PMMA bone cement is polymerized *in situ* during the surgery. It contains radiopacificiers in the form of particles of barium sulphate ($BaSO_4$) or zirconia (ZrO_2), which make it visible in radiographs (3). It also contains an inhibitor (hydroquinone) to prevent spontaneous polymerization and an initiator (benzoyl peroxide) to allow polymerization at room temperature. Antibiotics to prevent infection (e.g., gentimacin) may also be added. Table 1 lists typical components of bone cement and their roles. Polymerization begins when a powder of the PMMA polymer is mixed with the MMA monomer liquid. The mixing can either be done by hand in a mixing bowl just before to its use in the surgery or a mechanical mixing system may be used; these have the advantage of reducing the porosity of the bone cement and increasing its fatigue life. The cement is applied in a doughy state to the bone before placement of the implant.

In biological fixation, the implant is secured to the bone by a process known as "osseointegration." Osseointegration occurs by bone ingrowth onto the surface of the implant. The surface of the implant must have a structure so that, when the bone grows in, sufficient tensile and shear strength is created. Bone ingrowth requires a mechanically stable environment and an osteoconductive surface. An osteoconductive surface can be achieved by various treatments, e.g., plasma spraying with hydroxyapatite. Ingrowth occurs over approximately 12 weeks, and during this period, implant stability is required: Initial stability can be achieved by press-fitting the implant into the bone, or by using screws.

Hybrid fixation refers to the use of both cemented and biological techniques for the fixation of a prosthesis. For example, a hip replacement femoral component may be fixated using cement, whereas the acetabular cup may be fixated into the pelvic bone by osseointegration.

Failure of prosthesis fixation is observed as loosening and pain for the patient. If loosening occurs without infection it is called *aseptic* loosening. Loosening is a multifactorial process and does not have just one cause. Loosening of cemented fixation often occurs by fatigue failure of the bone cement, but loosening can have several root causes: fatigue from pores in the cement and stress concentrations at the implant/cement interface, debonding at the prosthesis/cement interface or cement/bone interface, or bone resorption causing stresses to rise in the cement. Loosening of biological fixation occurs if the relative micromotion between the bone and the implant is too high to allow osseointegration, i.e., if the initial stability of the implant is insufficient. Huiskes (4) proposed the concept of *failure scenarios* as a method for better understanding the multifactorial nature of aseptic loosening. The failure scenarios are

1. Damage accumulation failure scenario: the gradual cracking of bone cement, perhaps triggered by interface debonding, pores in the cement, or increased stresses due to peripheral bone loss.
2. Particulate reaction failure scenario: wear particles emanating from the articulating surfaces or from metal/metal interfaces in modular prostheses (fretting wear) can migrate into the interfaces causing bone death (osteolysis).
3. Failed ingrowth failure scenario: High micromotion of the implant relative to the bone can prevent bone ingrowth, as can large gaps (> 3 mm). If the area of ingrowth is insufficient, then the strength of the fixation will not be high enough to sustain loading when weight-bearing commences.
4. Stress shielding failure scenario: Parts of the bone can be "shielded" from the stresses they would normally experience because of the rigidity of the implant. This can lead to resorption of the bone and degeneration of the fixation.
5. Stress bypass failure scenario: In biological fixation, ingrowth can be patchy leading to stress transfer over localized areas. When this happens, some bone tissue is "bypassed," and in these regions, bone atrophy can occur because the stress is low.

Table 1. Components of Bone Cement and Their Roles

Components	Role	Amount
Liquid		20 mL
Methyl methacrylate (monomer)	Wetting PMMA particles	97.4 v/o
N,N,-dimethyl-p-toluidine	Polymerization accelerator	2.6 v/o
Hydroquinone	Polymerization inhibitor	75 + 15 ppm
Solid powder		40 g
Polymethyl methacrylate	Matrix material	15.0 w/o
Methyl methacrylate-styrene-copolymer	Matrix material	75.0 w/o
Barium sulphate ($BaSO_4$), USP	Radiopacifying agent	10.0 w/o
Dibenzoyl peroxide	Polymerization initiator	0.75 w/o

From Park (3).

Note: v/o: % by volume; w/o: % by weight.

6. Destructive wear failure scenario: In some joint replacement prostheses, e.g., hip and knee, wear can occur to such a degree that the component eventually disintegrates.

CEMENTED FIXATION

It is common to classify cementing techniques according to "generation": The first generation involved hand-mixing and finger packing of the cement, and the second generation improved the procedure by using a cement gun and retrograde filling of the canal, with a bone-plug to contain the cement within the medullary canal. This allows pressurization of the cement and therefore better interdigitation of the cement into the bone. Third generation (called modern cementing) uses, in addition, mechanical mixing techniques for the cement to remove pores and pulsative lavage to clean the bone surface of debris. The most common mechanical mixing technique is "vacuum mixing," where the powder and monomer are placed together in a mixing tube and the air is removed under pressure; often the tube can then be placed into an injection gun from which it can be extruded into the bone cavity. Another mechanical mixing technique is centrifugation (i.e., spinning the polymerizing bone cement in a machine), which is found to remove pores and increase the fracture strength (3). Precoating the implant with a PMMA layer or addition of a roughened surface strengthens the implant/bone cement interface.

Fixation strength using bone cement relies on an interdigitation of the bone cement with the bone; i.e., it is a mechanical interlock between the bone and the solidified cement that maintains the strength and not a chemical bond. Good interdigitation requires that the bone bed be rough. Creating a rough surface is done by appropriate broaching during preparation of the bony bed; it also requires lavage to clean the bed of loose debris and marrow tissue. Mann et al. (5) found the strength of the bone cement/bone interface to be positively correlated with the degree of interdigitation (Fig. 1). To achieve superior interdigitation, it was thought useful to develop "low viscosity" cements, and although higher penetration was achieved, the clinical outcomes using low viscosity cements in hips were not superior (6).

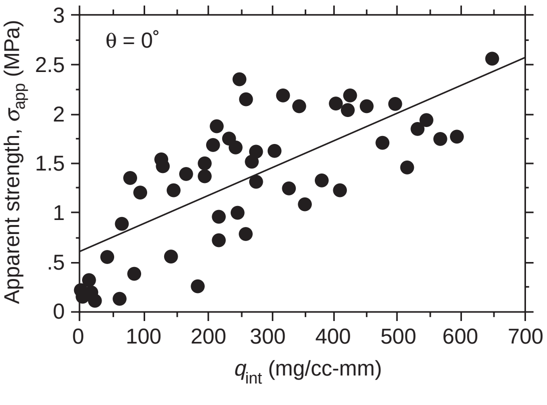


Figure 1. Interdigitation of the bone cement into the bone increases the strength of the bone cement interface. q_{int} is the product of the average value of the thickness of the interdigitated region and the density of the interface region measured using a CT scan. See Mann et al. (5) for details.

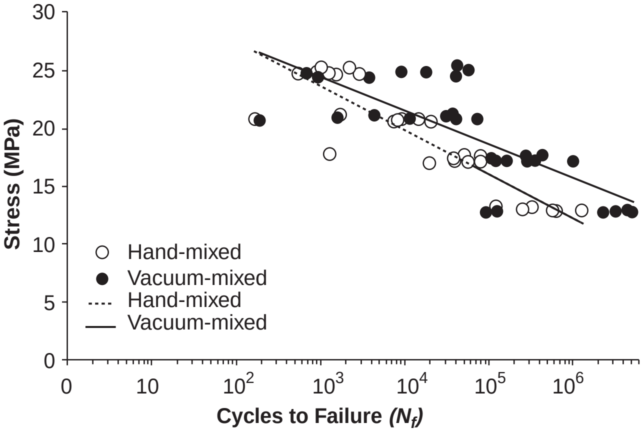


Figure 2. A comparison of the fatigue strength of hand-mixed and vacuum-mixed bone cement. After Murphy and Prendergast (10).

PMMA bone cement undergoes an exothermic polymerization reaction. This means that heat is produced on polymerization and this can cause necrosis of the surrounding bone tissue. Another consequence of heating is that the cement expands and contracts on cooling. As solidification occurs before to full cooling, residual stresses are generated in the cement (7). This is one reason to minimize the thickness of the cement layer. Also, metallic stems, because they conduct heat, can minimize the peak temperature transmitted to the bone, cooling the metallic implant before implantation has also been suggested. Bioactive cements have also been proposed; see the review by Harper (8). These cements have filler particles added to create a bioactive surface on the cement; fillers can be hydroxyapatite powder or fibers, bone particles, or human growth hormone. Alternatives to PMMA are bisphenol-*a*-glycidyl methacrylate (BIS-GMA) or poly(ethylmethacrylate) (PEMA)/*n*-butylmethacrylate (*n*BMA) cement. However, these cements are not yet widely used.

The mechanical strength depends on the brand of cement used and on the mixing technique (9). To prevent the damage accumulation failure scenario (see above), sufficient fatigue strength is required. This has been measured as a function of mixing technique (Fig. 2) (10). Being a polymer operating close to its melting temperature, bone cement is also subject to creep, i.e., viscoplasticity, and the creep strain as a function of stress has been measured under dynamic loading (11). However, it is clear that the *in vitro* testing conditions may not account for many of the extremely complex *in vivo* conditions, so these results should be interpreted with caution (12).

OSSEOINTEGRATION (CEMENTLESS FIXATION)

There is no simple definition of osseointegration, although Albrektsson (13) advocates the following: Osseointegration means *a relatively soft-tissue-free contact between implant*

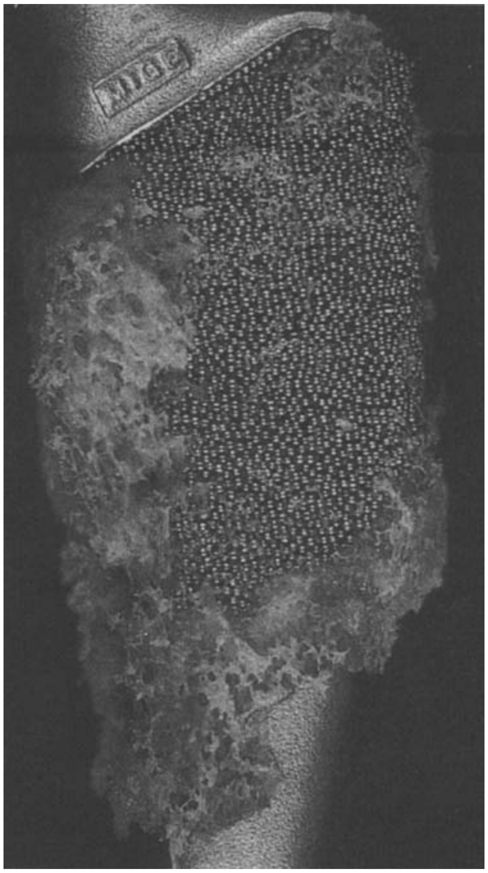


Figure 3. Bone ingrowth into a multilayer of a proximal part of a femoral hip prosthesis. After Eldridge and Learmonth (15).

and bone, leading to a clinically stable implant. Early in the study of the osseointegration concept, Skalak (14) found osseointegration was promoted by a micro-rough surface more so than a smooth one. Since then, many animal experiments investigating the effect of plasma spraying the surface and various methods of creating a porous surface have been reported. For orthopedic fixation, porous surfaces with beads in one or more layers have been used, as have wire meshes attached to surfaces, and plasma spraying the surface with hydroxyapatite.

Figure 3 shows bone ingrowth into a multilayer of a proximal part of a femoral hip prosthesis (15). It can be observed that ingrowth is patchy; this is what is commonly found, even with successful implants retrieved at autopsy (16); it is evident, therefore, that ingrowth is not required everywhere on the prosthesis for a successful fixation. Ingrowth is controlled by a combination of the mechanical environment and the size of the pores; the spacing between the pores should not be greater than the degree of micromotion or else the new bone ingrowth path will be continuously sheared as the tissue attempts to grow in. In experiments in dogs, Søballe (17) studied the relationship between implant coating and micromotion and found that hydroxyapatite coating increased the rate of bone ingrowth, and that a relative motion between implant and bone of 150 μm allowed osseointegration, whereas a relative motion of 500 μm inhibited it. The mechanobiological consequences of these different shearing magnitudes was analyzed by Prendergast et al. (18). The depth of the porosity will also affect the strength, with multilayer beaded surfaces having the potential for greater tensile strength (1).

FIXATION OF PROSTHESIS DESIGNS

Each implant design has specialized fixation features. In the following sections, examples are provided of the fixation approaches used in the main orthopedic implant categories.

Hip Prostheses

Although hip arthroplasty may involve replacement of the femoral side only, *total* hip arthroplasty (THA) involves replacement of the proximal femur and the acetabular socket. Both cemented and cementless fixation is used for both the femoral component (the "stem") and the acetabular component (the "cup"). Selection is a matter of surgeon choice, although there is some agreement that the cementless fixation is preferable in younger patients because cementless implants are easier to revise than cemented where complete removal of the cement mantle may be problematic.

Considering the femoral side first, cemented fixation takes two categories: stem designs in which a bond is encouraged between the stem and the cement (referred here as bonded stems) and designs that discourage a bond (referred here as unbonded stems). Stem bonding can be achieved through roughening of the stem surface to create a mechanical interlock between the metallic stem or cement or through use of a PMMA precoat to create a chemical bond between the precoat/cement interface. Bonded stems usually contain a collar that rests on the bone surface preventing subsidence and often containing ridges, dimples, and undercoats to provide additional interlock with the cement. As long as the bonded stems remain bonded, they have the theoretical benefit of reducing the stress levels in the cement. However, if the bonded stems fail, the roughened surface could generate debris particles and initiate a loosening process. In contrast to the bonded stems, unbonded stems discourage a bond between the stem and the cement through use of a smooth, polished stem surface in combination with a stem design that typically has no collar or macrofeatures to lock into the cement. With the lack of a bond, the polished stems facilitate some stem subsidence within the cement mantle and thereby allow wedging of the implant within the medullary canal. Lennon et al. (19) compared the damage accumulation around polished with matt stems and did not find a difference in the damage accumulated in their cement mantles. Another point of comparison between cemented and cementless fixation is that cementless stems will have a larger cross-sectional area than cemented stems because they must fill the medullary canal; this will make cementless hip prostheses stiffer and predispose them to the stress shielding failure scenario. Recognizing this, it is usual for the osseointegration surface to be on the proximal part of cementless stems to ensure proximal load transfer;

furthermore, patches of osseointegrative surface may be limited to the posterior and anterior faces of the stem.

Considering the acetabular side, the cup is either made from ultra-high-molecular-weight polyethylene (UHMWPE), ceramic, or metal. UHMWPE cups may be metal-backed. As the head can be either ceramic (a modular head can be connected to a metal femoral component using a Morse taper) or metal (modular or monobloc), this means that several combinations of bearing materials are possible. Polyethylene cups and metal heads are the most common, but the others, such as metal-on-metal, are advocated as well. The selection of bearing materials is important for the fixation because a high frictional torque predisposes to loosening of the cup or stem and because the wear particles produced can provoke the particulate reaction failure scenario. Polyethylene cups are cemented into the acetabulum using bone cement. Metal-backing of the cup is designed to decrease stresses in the polyethylene "liner," which should lead to lower wear rates although it is also predicted to increase stress concentrations in the fixation at the periphery of the cup (20). Metal, ceramic, and metal-backed UHMWPE cups may be threaded on the outside so that they can be screwed into the acetabelum, or they may be fixated by osseointegration.

The interrelationship between design factors and fixation of hip implants is complicated and involves maximizing strength of the cement/metal interface, the cement itself, and the bone/cement interface. According to the design philosophy of polished stems, it is better to safeguard the vital bone/cement interface by allowing the cement/metal interface to fail first and facilitating subsidence (21). Not only should the interfaces have the required strength, but the stresses should be minimized to ensure the most durable fixation (22)— the measures to achieve this are listed in Table 2.

Knee Prostheses

Total knee replacement involves femoral and tibial components, and a component for patellar resurfacing (a patellar "button") is also often used. Both cemented and cementless fixation is used in knee arthroplasty. The femoral component may be fixated with an intramedullary stem that may be cemented, or it may have a porous surface for osseointegration with medial and lateral "posts" to aid initial stability. The tibial component consists of a metal "tray" and a polyethylene insert; the tray may also be fixated with an intramedullary stem cemented into the tibia, perhaps accompanied by medical and lateral posts/pegs for rotational stability. Figure 4 shows a design fixated by osseointegration (23). Walker (24) gives a thorough description of the options available for knee prostheses.

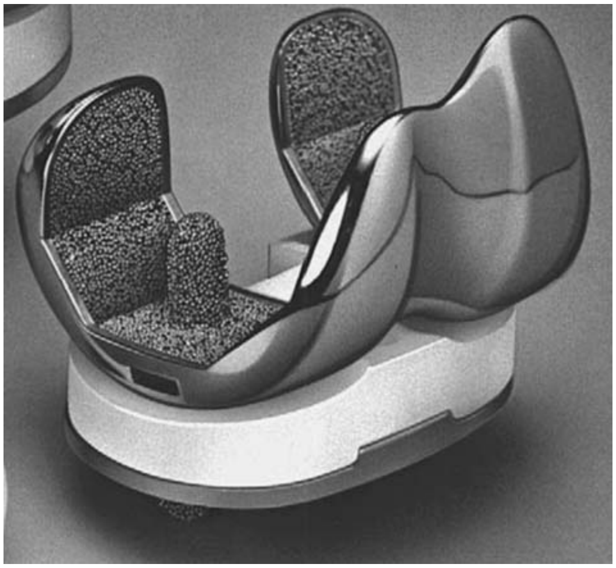


Figure 4. A knee replacement prosthesis showing porous-coating for osseointegration and posts for fixation. From Robinson (23).

Upper Extremity and IVD Prostheses

Upper extremity prostheses include the shoulder, elbow, and wrist (1). Total shoulder arthroplasty (TSA) consists of a humeral component with an intermedullary stem and a

Table 2. Measures that Maximize Strength and Minimize Stress in Total Hip Replacement Structures

	Cement/Metal Interface	Cement	Cement/Bone Interface
Maximize strength	Grit-blasted metal PMMA-coated metal	Optimal preparation Pressurization Cement restrictor	Careful reaming Pressurization Minimal polymerization heat Minimal monomer Bone lavage Minimal wear debris
Minimize stress	Reduce patient weight Reduce patient activity Anatomical reconstruction of the femoral head Minimal friction No impingement or subluxation Bonded cement/metal interface Optimal implant and cement mantle design Optimal implant material Optimal (reproducible) placement		

Adapted from Huiskes (22).

glenoid component inserted into the glenoid cavity of the scapula. The glenoid component is either all-polyethylene; in which case, it is cemented; or metal-backed; in which case, it may be fixated by osseointegration. Glenoid components may have several pegs, or they may have one central “keel” for fixation (25). Elbow prostheses consist of humeral, ulnar, and radial components, all which may be fixated with or without cement. Wrist prostheses replaces the radial head and the schapoid and lunate bones of the wrist and may be cemented and uncemented (1). Intervertebral disk (IVD) prostheses replace the degenerated disk with a polymer; there are several strategies for fixation: The endplates may be porous coated and plasma sprayed for osseointegration to the cancellous bone with vertical fins to increase stability. IVD prostheses may also be fixed to adjacent vertebral bodies with screws (26).

EVALUATION OF FIXATION AND FUTURE STRATEGIES

One of the key issues in orthopedic implant fixation is whether to use cemented fixation or biological fixation, with surgeons on both sides of the debate (16,27). Cemented fixation has the advantage of immediate postoperative, stability whereas concerns may be raised about the reliability of bone cement’s fatigue strength; furthermore, there is a school of thought that the exothermic polymerization reaction should be avoided if at all possible. Biological fixation by osseointegration has the advantage of avoiding the use of the PMMA cement but runs the risk of the failed ingrowth failure scenario; furthermore, immediate postoperative weight-bearing is not possible. Finally cementless implants are easier to revise if they fail.

Another key issue in orthopedic implant fixation is that of preclinical testing and regulatory approval of new fixation technologies. Considerable challenges exist in achieving consensus around regulatory tests that safeguard patients against ineffective devices while still allowing innovation (4). Preclinical tests can use either (1) finite element models of the direct postoperative situation, e.g., for the hip (28), knee (29), or shoulder (25), or computer simulations of a failure scenario, e.g., damage accumulation (30); (2) physical model “bench” testing with simulators (24,31); or (3) animal testing. Animal testing is not ideal for testing the biomechanical efficacy of orthopedic implant fixation because the implant geometry must be modified to fit the animal skeleton. Furthermore, an important emerging concept is that of patient-specific implants based on computational analysis of a patient’s medical images (32).

One useful clinical method to assess implant fixation is through the use of radiostereometric analysis (RSA). With this approach, the migration of the implant relative to the bone can be determined and is used to determine designs that may be at risk of early loosening. Retrospective and prospective clinical studies are also very useful to determine designs or materials that have promising or poor clinical results. On a larger scale, implant registries performed in many countries in Western Europe can provide information on how designs, materials, and surgical techniques rank in terms of risk of failure. All of these clinical tools can aid in understanding the role of implant fixation in success of joint replacements.

A final issue is the degree to which broader technological innovations in surgery and medicine will affect orthopedics. For example, minimally invasive therapy (33) requires special implants and associated instrumentation. Tissue-engineering and regenerative medicine also has the potential to change the nature of orthopedics, not only by reducing the need for joint arthroplasty implants but by integrating tissue engineering concepts with conventional implant technologies, for example, cell seeding into implant surfaces to promote biological fixation.

ACKNOWLEDGMENTS

Research funded by the Programme for Research in Third-Level Institutions, administered by the Higher Education Authority. Dr. A. B. Lennon and Ms. S. Brown are thanked for their comments.

BIBLIOGRAPHY

1. Prendergast PJ. Bone prostheses and implants. In: Cowin SC, editor. Bone Mechanics Handbook. Boca Raton: CRC Press; 2001; 35(1)–35(29).
2. Prendergast PJ, van der Helm FCT, Duda G. Analysis of joint and muscle loading. In: Mow VC, Huiskes R, editors. Basic Orthopaedic Biomechanics and Mechanobiology. Philadelphia: Lippincott Williams & Williams; 2005;29–89.
3. Park JB. Orthopaedic prosthesis fixation. In: Bronzino JD, editor. The Biomedical Engineering Handbook. Boca Raton: CRC Press; 1995. pp. 704–723.
4. Huiskes R. Failed innovation in total hip replacement. Diagnosis and proposals for a cure. Acta Orthopaedica Scandinavica 1993;64:699–715.
5. Mann KA, Mocarski R, Damron LA, Allen MJ, Ayers DC. Mixed-mode failure response of the cement-bone interface. J Orthop Res 2001;19:1153–1161.
6. Balderston RA, Rothman RH, Booth RE, Hozack WJ. The Hip. New York: Lea & Febiger; 1992.
7. Lennon AB, Prendergast PJ. Residual stress due to curing can initiate damage in porous bone cement: experimental and theoretical evidence. J Biomechan 2002;35:311–321.
8. Harper EJ. Bioactive bone cements. Proc Inst Mech Eng Part H J Eng Med 1998;212:113–120.
9. Lewis G. The properties of acrylic bone cement: A state-of-the-art review. J Biomed Mater Res 1997;38:155–182.
10. Murphy BP, Prendergast PJ. On the magnitude and variability of fatigue strength in acrylic bone cement. Int J Fatigue 2000;22:855–864.
11. Verdonschot N, Huiskes R. The dynamic creep behaviour of acrylic bone cement. J Biomed Mater Res 1995;29:575–581.
12. Prendergast PJ, Murphy BP, Taylor D. Discarding specimens for fatigue testing of orthopaedic bone cement: A comment on Cristofolini et al. (2000). Fatigue Fracture Eng Mater Structures 2002;25:315–316.
13. Albrektsson T. Biological factors of importance for bone integration of implanted devices. In: Older J, editor. Implant Bone Interface. New York: Springer; 1990;7–19.
14. Skalak R. Biomechanical considerations in osseointegrated prostheses. J Prosthetic Dentistry 1983;49:843–860.

15. Eldridge JDI, Learmonth ID. Component bone interface in cementless hip arthroplasty. In: Learmonth ID, editor. Interfaces in Total Hip Arthroplasty. London: Springer; 1999: 71–80.
16. Sychterz CJ, Claus AM, Eng CA. What we have learned about cementless fixation from long-term autopsy retrievals. Clin Orthop Related Res 2002;405:79–91.
17. Søballe K. Hydroxyapatite ceramic coating for bone-implant fixation. Mechanical and histological studies in dogs. Acta Orthop Scand 65: (Suppl. 255).
18. Prendergast PJ, Huiskes R, Søballe K. Biophysical stimuli on cells during tissue differentiation at implant interfaces. J Biomechan 1997;30:539–548.
19. Lennon AB, McCormack BAO, Prendergast PJ. The relationship between cement fatigue damage and implant surface finish in proximal femoral prostheses. Med Eng Phys 2003; 25:833–841.
20. Dalstra M Biomechanical aspects of the pelvic bone and design criteria for acetablar prostheses. Ph. D. Thesis, University of Nijmegen, 1993.
21. Lee AJC. Rough or polished surface on femoral anchorage stems. In: Buchhorn GH, Willert HG, editors. Technical Principles, Design and Safety of Joint Implants. Seattle: Hogrefe & Huber Publishers; 1994:209–211.
22. Huiskes R. New approaches to cemented hip-prosthetic design. In: Buchhorn GH, Willert HG, editors. Technical Principles, Design and Safety of Joint Implants. Seattle: Hogrefe & Huber Publishers; 1994:227–236.
23. Robinson RP. The early innovators of today's resurfacing condylar knees. J Arthroplasty 2005;20 (suppl. 1):2–26.
24. Walker PS. Biomechanics of total knee replacement designs. In: Mow VC, Huiskes R, editors. Basic Orthopaedic Biomechanics and Mechanobiology. Philadelphia: Lippincott Williams & Wilkins; 2005:657–702.
25. Lacroix D, Murphy LA, Prendergast PJ. Three-dimensional finite element analysis of glenoid replacement prostheses: A comparison of keeled and pegged anchorage systems. J Biomech Eng 2000;123:430–436.
26. Szpalski M, Gunzburg R, Mayer M. Spine arthroplasty: A historical review. European Spine J 2002;11 (suppl. 2): S65–S84.
27. Harris WH. Options for the primary femoral fixation in total hip arthroplasty—cemented stems for all. Clin Orthop Related Res 1997;344:118–123.
28. Chang PB, Mann KA, Bartel DL. Cemented femoral stem performance—effects of proximal bonding, geometry, and neck length. Clin Orthop Related Res 1998;355:57–69.
29. Taylor M, Barrett DS. Explicit finite element simulation of eccentric loading in total knee replacement. Clin Orthop Related Res 2003;414:162–171.
30. Stolk J, Maher SA, Verdonschot N, Prendergast PJ, Huiskes R. Can finite element models detect clinically inferior cemented hip implants? Clin Orthop Related Res 2003;409:138–160.
31. Britton JR, Prendergast PJ. Pre-clinical testing of femoral hip components: an experimental investigation with four prostheses. J Biomechan Eng. In press.
32. Viceconti M, Davinelli M, Taddei F, Capello A. Automatic generation of accurate subject-specific bone finite element models to be used in clinical studies. J Biomechanics 2004; 37:1597–1605.
33. DiGioia AM, Blendea S, Jaramaz B. Computer-assisted orthopaedic surgery: minimally invasive hip and knee reconstruction. Orthop Clin North Am 2004;35:183–190.

See also Biocompatibility of materials; bone and teeth, properties of; bone cement, acrylic; hip joints, artificial; materials and design for orthopedic devices.

ORTHOTICS. See Rehabilitation, orthotics in.

OSTEOPOROSIS. See Bone density measurement.

OVULATION, DETECTION OF. See Contraceptive devices.

OXYGEN ANALYZERS

Susan McGrath
Suzanne Wendelken
Dartmouth College
Hanover, New Hampshire

INTRODUCTION

Oxygen is essential for all aerobic life on Earth. It is the most abundant element as it comprises a little more than one-fifth of the weight of air, nine-tenths of the weight of water, and almost one-half of the weight of the earth's crust (1).

Because of its role in supporting and sustaining life, it is often important to monitor the level of oxygen in the atmosphere. Too much oxygen can lead to a toxic atmosphere where as too little oxygen causes asphyxia and eventually death. A relatively constant level of oxygen is required for most aerobic processes.

Oxygen gas monitoring is used for a number of purposes: (*1*) **Medical:** anesthesia (drug delivery, airway monitoring), respiratory oxygen content monitoring (inhaled and exhaled), controlled environments, incubators. (*2*) **Physiological:** exercise (rate of oxygen consumption), aircraft, spacecraft, scuba diving, fire fighting, mountain climbing, spelunking. (*3*) **Biological:** metabolism (oxygen uptake and consumption), fermentation, beverage and food packing. (*4*) **Industrial:** combustion control, fuel and pollution management, safe operation of chemical plants, monitoring gas purity.

This article gives an overview of the analyzers used to measure gaseous oxygen in medicine, physiology, and biology. Measurement of dissolved or bound oxygen is also important in medicine and is discussed in detail elsewhere in this Encyclopedia.

History and Relevance

Oxygen was not known to exist until the 1770s when it was discovered by French scientist Antoine Lavoisier and English clergyman and scientist Joseph Priestly through experiments on combustion. Previously, air was considered to be an element composed of a single substance. Combustible materials were thought to have a substance called phlogiston, from the Greek word meaning to be set on fire, which escaped as a material was burned. Lavoisier, however, believed that combustion resulted from a combination of fuel and air. He conducted experiments where he burned a candle in a sealed jar and observed that only one-fifth of the air was consumed. He named this unconsumed portion of the air oxygen from the Greek word meaning acid producing. Although his thoughts about oxygen being the corrosive agent in acidic compounds was wrong, the name stuck and the study of oxygen was born (2).

Oxygen is essential for most life on Earth as it plays a key role in aerobic metabolism as a final electron acceptor due to its high electron affinity. Metabolic rate can be indirectly measured by monitoring oxygen consumption as >95% of energy is produced by reactions of oxygen with other food (3). This method is called indirect calorimetry and is a much more cost effective and timely method for measuring metabolic rate as compared to direct calorimetry (the direct measure of heat energy produced).

Oxygen availability is a function of its partial pressure and the total pressure of the gas mixture in which it resides. At sea level, the partial pressure of oxygen is roughly 21%. With decreasing atmospheric pressure, as accompanies increasing altitude, the total amount of available oxygen decreases (Table 1). For example, at 18,000 ft. (5.48 km) above sea level, although the partial pressure of oxygen is still 21%, there is roughly half the amount of available oxygen. At ~29,000 ft. (8.33 km) above sea level on the top of Mt. Everest, there is less than a third the amount of total available oxygen compared to sea level. At such altitudes, most humans require the use of supplemental oxygen. In addition, the body will compensate for the reduced oxygen availability by increasing the heart and respiration rate to keep up with the metabolic demands (3). A climber's resting heart rate at this altitude is double to triple their normal resting heart rate. Long-term exposure to high altitude prompts the body to produce more red blood cells per unit blood volume thus increasing the number of oxygen carriers and making respiration easier. If the body does not properly adapt to such conditions, altitude sickness, pulmonary and cerebral edema, and potentially death may result (3).

Table 1. Atmospheric Pressure, the Fraction of Available Oxygen Compared to Sea Level, and Temperature All Decrease with Increasing Altitude[a]**.**

Altitude, ft.	Barometric Pressure, mmHg	Fraction Available Oxygen	Temperature °C
0	760	1.00	15
1,000	733	0.96	13
5,000	632	0.83	5.1
10,000	523	0.69	−5.4
14,000	447	0.59	−12.7
16,000	412	0.54	−16.7
18,000	380	0.50	−20.7
20,000	349	0.46	−24.6
22,000	321	0.42	−28.6
24,000	295	0.39	−32.6
26,000	270	0.36	−36.5
28,000	247	0.33	−40.5
30,000	228	0.30	−44.4
32,000	206	0.27	−48.4
34,000	188	0.25	−52.4
36,000	171	0.23	−56.3

[a] See Ref. 4.

The partial pressure of oxygen remains a fairly constant 21% until very high altitudes [i.e., >50,000 ft. (15.24 km) (5)]. At these altitudes it is necessary to maintain a pressurized, enclosed environment, such as aircraft, spacecraft, or space suite, in order to sustain human life.

Oxygen availability can be decreased by displacement by other gases, such as nitrogen, carbon dioxide, methane, and anesthetics. Oxygen availability is also easily decreased by combustion and oxidation processes. Thus it is necessary to monitor the atmospheric oxygen level in situations where these gases or combustion is present such as in enclosed environments, closed breathing circuits, and fire fighting.

Each year ~20 deaths occur as a result of asphyxiation due to displacement of oxygen by another gas in air (6). Accidental asphyxia usually occurs in industry as a result of oxygen depletion by carbon dioxide, CO_2, methane (CH_4), or a hydrocarbon gas in a confined space, such as a tunnel, laboratory, sewer, mine, grain silo, storage tank, or well (7). For example, in 1992, a barge operator in Alaska died from asphyxiation and a rescuer lost consciousness during rescue efforts due to a low level of oxygen (6%) in a confined space (8). In anesthesia, accidental asphyxia has resulted from incorrectly connected gas delivery tubes (9).

In choosing an oxygen analyzer for a particular need, it is important to be acquainted with the properties of operation, characteristics, and limitations of these devices. The primary methods for oxygen detection are based on the paramagnetic susceptibility, electrochemical properties, and light absorption properties of oxygen.

PARAMAGNETIC OXYGEN ANALYZERS

All mater exhibits some form of magnetism when placed in a magnetic field. Magnetic susceptibility is the measure of the strength of a material's magnetic field when placed in an external magnetic field. Diamagnetic substances, such as gold and water, align perpendicularly to an external magnetic field causing them to be repelled slightly. This property arises from the orbital motion of electrons that produces a small magnetic moment. In substances with paired valence electrons, these moments cancel out. However, when an external magnetic field is applied it interferes with the motion of the electrons causing the atoms to internally oppose the field and be slightly repelled by it. Diamagnetism is a property of all materials, but is very weak and disappears as soon as the external magnetic field is withdrawn. In materials with unpaired valence electrons (e.g., nickel and iron) an external magnetic field aligns the small magnetic moments in the direction of the field, which increases the magnetic flux density. Materials with this behavior are weakly attracted to magnetic fields and are classified as paramagnetic. Ferromagnetism is a special case of paramagnetism where materials (e.g., iron and cobalt) are strongly attracted to magnetic fields. Paramagnetic materials have a high susceptibility (10).

Oxygen has a relatively high susceptibility when compared to other gases (see Table 2). This property is the key principle behind paramagnetic oxygen analyzers.

Table 2. Relative magnetic susceptibility values on a scale of Oxygen = 100 and Nitrogen ~= 0 at 20° C (11–13).

Gas	Relative Magnetic Susceptibility
Argon	−0.58
Acetylene	−0.38
Air (dry air)	21.00
Ammonia	−0.58
Carbon dioxide	−0.61
Carbon monoxide	0.06
Chlorine	−0.13
Ethane	−0.83
Helium	0.29
Hydrogen	−0.12
Methane	−0.37
Nitrogen	**−0.42**
Nitrous oxide	−0.58
Nitrogen monoxide	43.80
Nitrogen dioxide	28.00
Oxygen	**100.00**

The three main types of paramagnetic oxygen analyzers are (*1*) thermomagnetic (magnetic wind); (*2*) magnetodynamic (dumbbell or autobalance); (*3*) magnetopneumatic (differential pressure).

Paramagnetic analyzers are typically used for monitoring the quality of breathing air in open and enclosed environment, biological laboratory measurements, and in industrial combustion analysis (2,14).

Limitations of Paramagnetic Analyzers. Because the magnetic susceptibility of oxygen depends on temperature, it is necessary to operate at a constant temperature or to have some temperature compensation ability (1,15). The output of the sensor is also proportional to the absolute atmospheric pressure and thus pressure compensation is sometimes necessary (1,15).

Paramagnetic devices are typically delicate instruments with moving parts and are thus adversely influenced by vibrations. They generally are not used as portable devices (13).

Paramagnetic sensors work well for percent oxygen measurement, but are not recommended for trace oxygen measurements. In addition, these sensors should not be used when interference effects cannot be compensated for (i.e., sample gas containing other paramagnetic or diamagnetic gases, or varying background gas composition) (14). The effects of background gases used in anesthesia are small but not always negligible. These effects are summarized in Table 3.

Table 3. Errors in Paramagnetic Analyzer measurements due to anesthesia gases[a].

Gas	Error in Instrument Reading in %O_2 Due to 1% Vapor
Diethyl ether	−0.0068
Halothane	−0.0157
Nitrous oxide	−0.0018
Methoxyflurane	0.0000
Trichloroethylene	−0.0033

[a] See Ref. 15.

Thermomagnetic (Magnetic Wind)

Thermomagnetic analyzers are based on the fact that magnetic susceptibility decreases inversely with the square of temperature.

Principles of Operation. A schematic diagram of a thermomagnetic analyzer can be seen in Fig. 1. A gas sample is admitted into the inlet that branches into equal segments and converges again at the outlet. These tubes are connected by another tube halfway between the inlet and outlet. This cross-tube is heated by a platinum coil that is separated in the center by a thermal resistance bridge. These two heater coil segments form two arms of a Wheatstone bridge with the third arm being the output of the sensor. A magnetic field is applied to one-half of the coil. Any oxygen in the gas is attracted to the magnetic field in the cross-section. These oxygen molecules are subsequently heated by the heater coil and immediately begin to lose their magnetic susceptibility. They are then displaced by cooler oxygen molecules with higher magnetic susceptibility. This flow of gas through the cross-tube, referred to as the magnetic wind, cools the magnetized heating coil and heats the unmagnetized coil causing an imbalance in the bridge resulting from the difference in resistance between the two coils. The bridge output is then calibrated by passing a gas with known oxygen concentration through the chamber (1,11,13).

Limitations of Thermomagnetic Analyzers. Measurement by themomagnetic oxygen analyzers is affected by the magnetic susceptibility and thermal conductivity of the

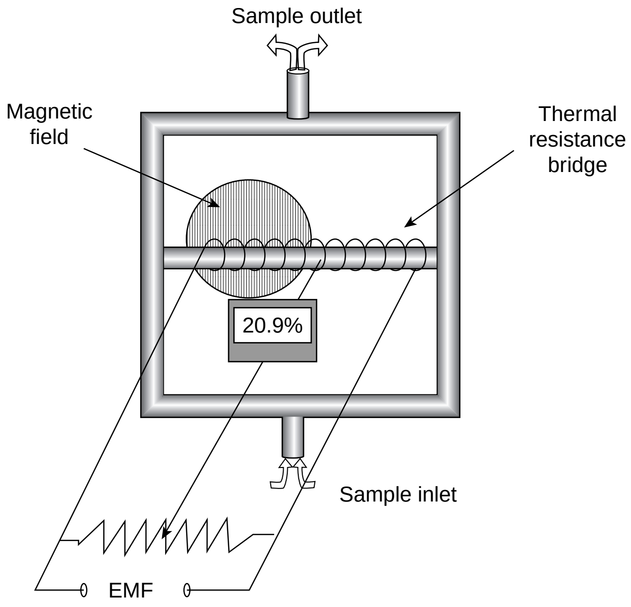


Figure 1. Schematic diagram of a thermomagnetic oxygen analyzer (1).

carrier gas, the sample gas temperature, ambient temperature, tilt, sample flow, and pressure (1,16).

Magnetodynamic (Dumbbell or Autobalance)

Developed by Faraday in 1884, this is the most popular method of paramagnetic oxygen analyzers and the earliest developed oxygen analyzer (13). Magnetodynamic oxygen analyzers are based on property that oxygen will be drawn into a magnetic field because it is paramagnetic. These analyzers essentially measure the magnetic susceptibility of sample gas.

Principles of Operation. A simple form of this device consists of small, dumbbell shaped body made of quartz and filled with nitrogen or some gas with small or negative magnetic susceptibility, an optical lever system, and a nonuniform magnetic field (Fig. 2). The dumbbell is suspended in a closed chamber by a quartz or platinum wire between two permanent magnets that are specially shaped to have a nonuniform magnetic field. The dumbbell is free to rotate. Since the dumbbell is slightly diamagnetic, it will naturally rotate away from the highest magnetic field intensity. Oxygen in a sample gas will be attracted to the region of maximum field intensity and displace the dumbbell even further. This deflection is measured by an optical lever system in which a light source outside the test chamber shines a beam onto a mirror which is mounted in the center of the dumbbell. The beam is then reflected onto a scale outside the chamber. The amount of defection is directly proportional to the partial pressure of oxygen in the sample (1,17).

Modern designs of this sensor are self-nulling and have temperature compensation capabilities. A single turn coil is wound around the dumbbell. The coil produces a magnetic field when current flows through it which will in turn cause the dumbbell to rotate in the external magnetic field. The deflection of the dumbbell is measured by an optical lever system that uses photocells to detect the light reflected from the mirror.

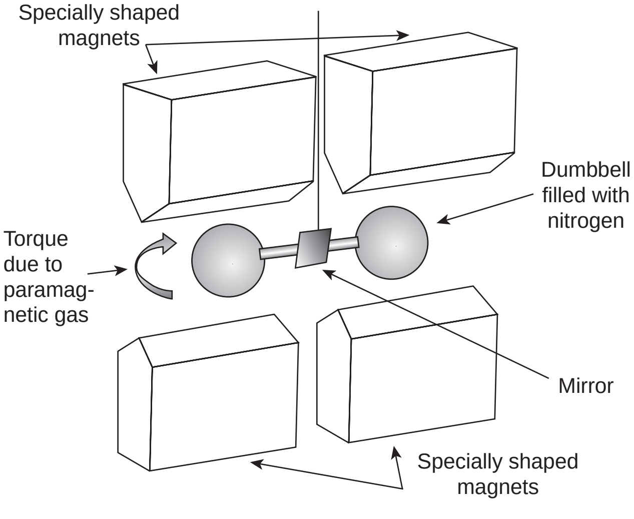


Figure 2. Schematic diagram of paramagnetic “dumbbell” sensor. Adapted from (1).

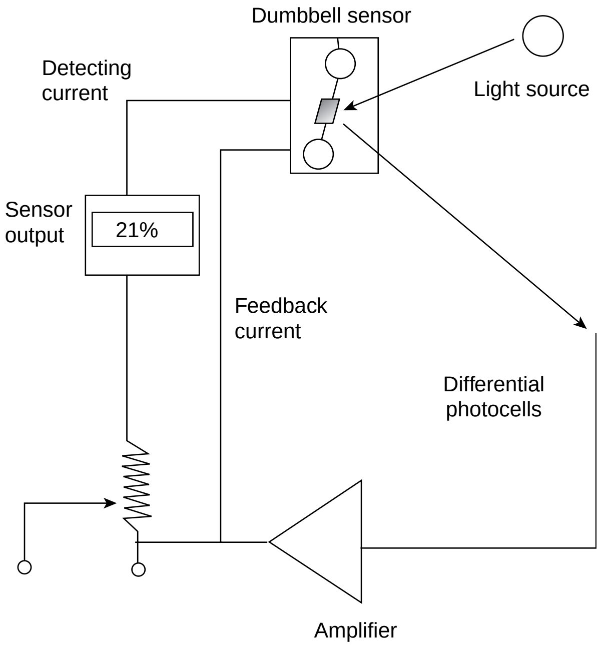


Figure 3. Self-nulling paramagnetic analyzer. Adapted from Refs. 1 and 17.

The photocells are connected to a feedback loop that controls the amount of current through the coil, keeping the dumbbell centered with respect to the photocell detectors. As the paramagnetic components of a gas sample move into the strongest part of the magnetic field, the dumbbell is displaced and begins to rotate. The photocells detect this motion and drive an amplifier to produce the necessary current in the coil to keep the dumbbell in the original zero position. The system is zeroed using a sample of pure nitrogen. In this case, the dumbbell is at an equilibrium position and there is no current flowing through the coil. The current is directly proportional to the magnetic susceptibility of the sample. The system is calibrated using a sample of known oxygen content. See Fig. 3 for a diagram of this design.

Limitations of Dumbbell Analyzers. The main limitation of the dumbbell design is its slow response time ($\sim$10 s) (15). Thus, the dumbbell analyzer is not recommended for uses where real-time oxygen analysis is needed. These analyzers also have moving parts and are extremely sensitive to tilt and vibrations.

Magnetopneumatic (Differential Pressure)

This sensor operates on the principle that a differential pressure will be generated when a sample containing oxygen is drawn into a nonuniform magnetic field with a reference gas of different oxygen content. Differential pressure sensors directly measure the magnetic susceptibility of sample gas and are thus not influenced by thermal properties of background gas.

Principles of Operation. A reference gas is admitted at a constant rate into a chamber like the one seen in Fig. 4. The

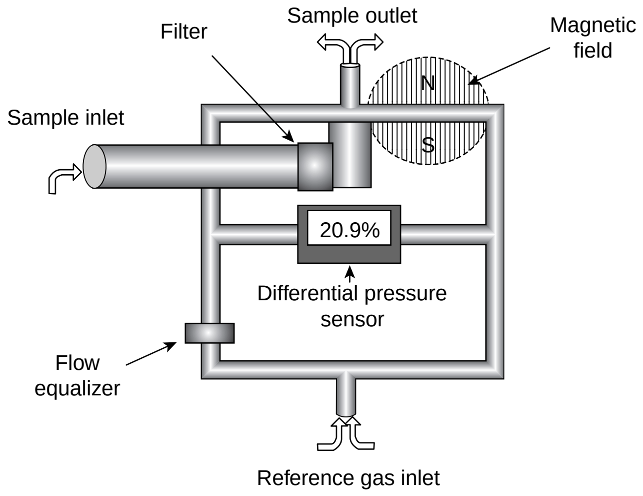


Figure 4. Schematic diagram of a differential pressure sensor. (Adapted from Ref. 1.)

reference gas is split into two paths with a flow equalizer to ensure equal flow in each path. Each path is joined at the midpoint by a connector pipe containing a differential pressure sensor (e.g., a capacitive differential pressure sensor or a microflow sensor). The two paths reconnect at an outlet where the sample gas is admitted. There is a strong nonuniform magnetic field placed over one of the reference gas outlets. The reference gas and the sample gas combine in the outlet. Oxygen or other paramagnetic gases in the sample gas will be drawn into the nonuniform magnetic field and cause a pressure build up on that side of the reference gas path.

The differential pressure is proportional to the magnetic susceptibility of the sample gas only. This imbalance is sensed by the differential pressure sensor in the cross-tube. The output of this sensor is calibrated in terms of oxygen concentration by using a reference and sample gas of known oxygen content. The output is zeroed by using a sample gas that is the same as the reference gas. In this case, the output of the differential pressure sensor will be zero (1,15).

Limitations of Magnetopneumatic Analyzers. Differential pressure sensors are sensitive to tilt and vibrations. An alternating magnetic field reduces the effects of background flow and tilt on the sensor (1).

ELECTROCHEMICAL OXYGEN ANALYZERS

There are two main types of electrochemical oxygen analyzers: those with aqueous electrolytes, and those with solid electrolytes. These sensors use the chemical reactivity of oxygen to produce a measurable current that is proportional to the partial pressure of oxygen in a sample gas.

Aqueous Electrolyte Sensors

Galvanic Oxygen Analyzer. Galvanic oxygen analyzers are commonly called a Hersch cell after the inventor. They are essentially a battery that produces energy when it is exposed to oxygen. Fig. 5. Galvanic sensors are typically insensitive to vibration and tilt. They are usually packaged small and made out of inexpensive and sometimes disposable materials (14). Disposable capsules containing galvanic cells are fairly inexpensive (~$85) and typically last 1–5 years (18). Recently, small, portable galvanic sensors have been manufactured and approved for medical breath analysis purposes (Fig. 6) (19).

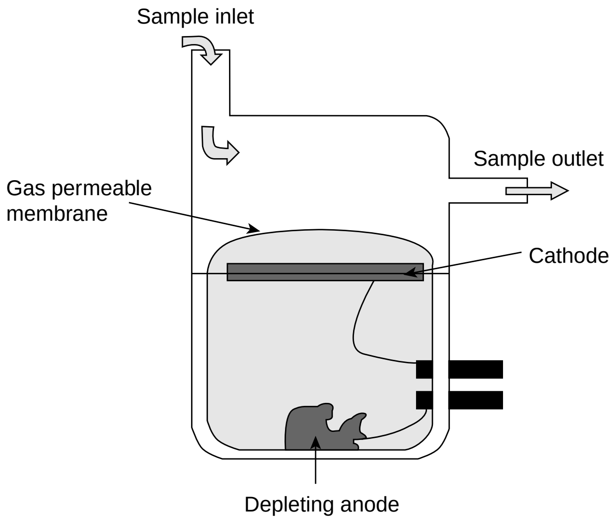


Figure 5. Schematic diagram of a galvanic sensor. (Adapted from Ref. (14).)

Figure 6. A small, handheld galvanic oxygen sensor (model AII 2000, Analytical industries Inc.) (19).

Galvanic sensors are typically used for industrial purposes, such as validating the quality of semiconductor grade gases and for environmental monitoring (e.g., monitoring the quality of breathing air or monitoring the oxygen content in potentially hazardous or explosive environments) (14).

Principles of Operation. Basic cells consist of a cathode made of a precious metal (platinum, gold, silver, etc.) and an anode made of a base metal (lead, cadmium, antimony). These electrodes are in contact with a liquid or semisolid electrolyte, such as potassium hydroxide. A gas sample is admitted into the cell and diffuses through a membrane made of a thin material, such as Teflon or silicone, which is permeable to oxygen but not to the electrolyte. The oxygen in the solution is chemically reduced at the cathode to form hydroxyl ions that flow to the anode where an oxidation reaction occurs. This oxidation–reduction reaction results in an electromotive force (EMF), which is proportional to the oxygen concentration in the solution and the partial pressure of oxygen in the sample gas. The electron flow is measured by an external galvanometer connected to the electrodes.

Reactions at the cathode and anode are as follows 1, 20–22):

1. Cathode reaction:

$$O_2 + 2\,H_2O + 4\,e^- \rightarrow 4\,OH^-$$

2. Anode reaction:

$$2\,Pb + 6\,OH^- \rightarrow 2\,PbO_2H^- + 2\,H_2O + 4\,e^-$$

3. Overall reaction:

$$O_2 + 2\,Pb + 2\,OH^- \rightarrow 2\,PbO_2H^-$$

Designs that allow for every oxygen molecule passing through the cell to react are called coulometers and are suitable for trace (parts per million, ppm) measurements (1,20–22).

Limitations of Galvanic Analyzers. Because the anode is consumed by oxidation, the cell has a limited life. These devices tend to lose sensitivity as they age resulting in falsely low readings (14).

There are some designs that lessen the rate of anode consumption by using a third inert electrode made of platinum that is kept at a constant positive potential. This results in the majority of the current to be conducted through this electrode instead of the consumable anode (1,4).

Acidic gases in the sample (i.e., SO_2 CO_2, Cl_2) react with the electrolyte and must be scrubbed out. There are some coulometric cell designs that overcome this problem by the use of additional electrodes (1,23).

Exposure to very high oxygen concentration can lead to oxygen shock where the sensor is saturated and does not recover for hours (14).

Polarographic Oxygen Analyzers. This sensor responds to changes in the partial pressure of oxygen in a sample gas. Polarographic sensors can be used for measuring oxygen in dissolved liquid (14) or in a gas sample. They are often used in anesthesia gas delivery systems. Polarographic sensors are insensitive to shock, vibration, and tilt. The effects of flow are minimal because diffusion is controlled by a membrane (14).

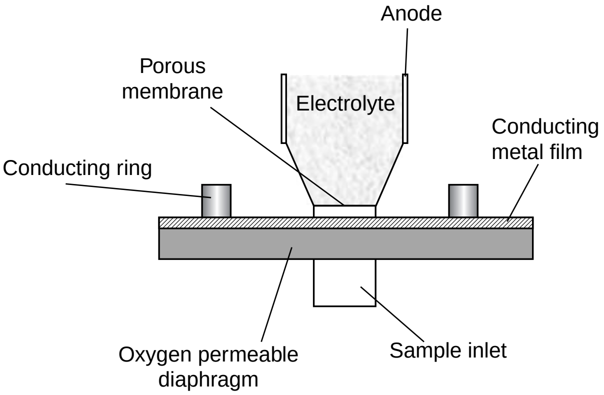


Figure 7. Schematic diagram of a polarographic sensor. (Adapted from Refs. 1,24,25.)

Principles of Operation. A polarographic cell, as seen in Fig. 7, consists of two electrodes, usually a silver anode and a gold cathode, immersed in an electrolyte, such as potassium chloride (1). An EMF is applied across the electrodes inducing oxidation–reduction reactions when a sample containing oxygen is admitted into the cell. Like the galvanic cell, oxygen diffuses through a thin membrane that is preferentially permeable to oxygen and not to the electrolyte. This membrane is usually made from poly(tetrafluoroethylene) (PTFE) and controls the rate of oxygen flux to the cathode. The current flow in the cell is proportional to the applied EMF and the partial pressure of oxygen in the sample. As seen in Fig. 8, there are four main regions in the EMF–current curve of importance (1):

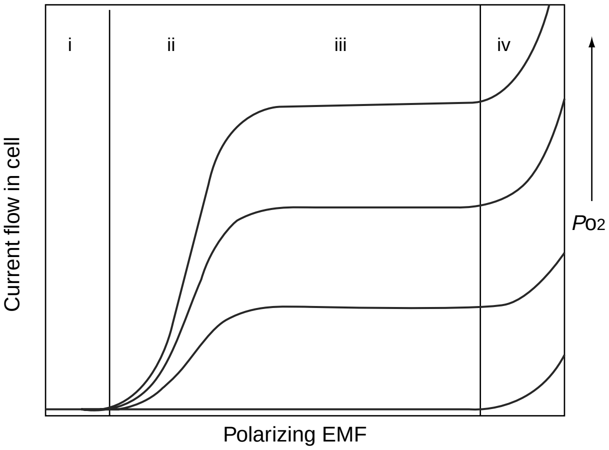


Figure 8. Diagram showing the operating regions of a polarographic sensor. (Adapted from Ref. 10.)

Region i: If the applied EMF is very low, then the presence of oxygen hardly has any effect on the current. There are very little reactions occurring at the electrodes.

Region ii: Oxygen molecules begin to react at the electrodes causing a measurable increase in current. For a given level of oxygen partial pressure, an increase in the EMF produces a sharp increase in the current.

Region iii: This region is the polarized or working region for the polarographic sensor. Here, the current plateaus and an increase in the EMF does not alter the current. In this region, all the oxygen molecules are reduced immediately at the cathode. A calibration curve is used in this region relating oxygen concentration to sensor current.

Region iv: In this region, an increase in EMF leads to a sharp, nonlinear increase in current as a result of the breakdown of the electrolyte solution.

The reactions at the cathode and anode are as follows (1,15):

1. Cathode reaction:

$$O_2 + 2\,H_2O + 4\,e^- \rightarrow 4\,OH$$

2. Anode reaction:

$$4\,Ag + 4\,Cl^- \rightarrow 4\,AgCl + 4\,e^-$$

Limitations of Polarographic Analyzers. Polarographic sensors generally have slow response because oxygen must diffuse through membrane. They are also sensitive to pressure and temperature and compensation for these factors is sometimes required. In addition, these sensors lose sensitivity over time due to degradation of anode and electrolyte solution giving falsely low readings. Polarographic sensors that consume all the injected oxygen change the content of the sample gas and are not good for closed-loop systems, such as closed-circuit anesthesia systems (26).

Capacitive Coulometry. Capacitive coulometry (also referred to as coulometric microrespiometry) is another aqueous electrolytic method for oxygen analysis. This method is based on the replacement of oxygen consumed by an organism in a closed system with electrolytic oxygen produced by discharging a capacitor through a solution of $CuSO_4$ (27). Such analyzers can be used to monitor oxygen consumption and metabolism rate of tissues or microorganisms. However, this type of sensor is not used as frequently as other more common types of aqueous electrolyte sensors.

Solid Electrolyte Cells

Some ceramics conduct electricity at very high temperatures. This conductivity is largely a result of the oxygen mobility in this solid solution. Oxygen mobility is the principle behind solid electrolyte or concentration cells.

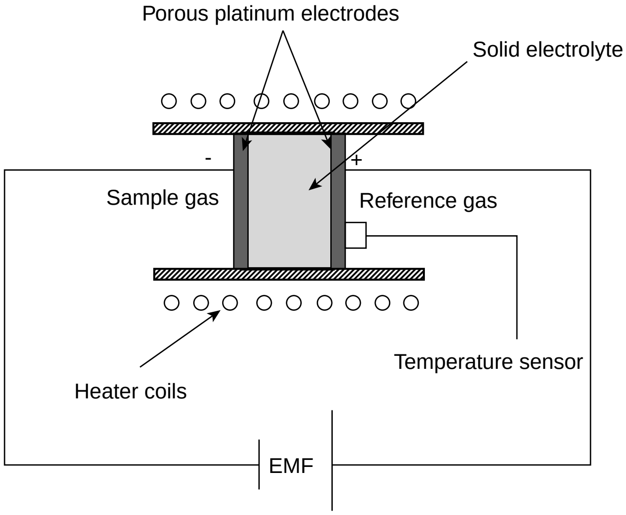


Figure 9. Schematic diagram of a solid electrolyte concentration cell. (Adapted from Refs. 1,15.)

There exists a family of solid electrolytes, such as MgO or CaO in ZrO_2 (ceramic), whose electrical conductivity is mostly due to the mobility of $O_2{}^-$ as opposed to electrons in the solid. At room temperature, the conductivity is low, but at high temperatures (>600° C) the conductivity is comparable to that of an aqueous electrolyte and electron mobility can be neglected (1).

A solid electrolyte sensor is commonly referred to as a concentration cell. Solid electrolyte oxygen sensors are commonly used in anesthesia and patient monitoring for breath to breath analysis (15). Solid electrolyte cells typically have a fast response (<150 ms) (15) and are good for real-time oxygen analysis.

Principles of Operation. A concentration cell is made by separating a test chamber and a reference chamber by a solid, oxygen conducting electrolyte, such as ZrO_2 or Y_2O_3 with a porous electrode on either side (Fig. 9). When the temperature is increased by an external heater, the solid electrolyte begins to conduct O_2 and an EMF is established between the electrodes. The EMF is related to the partial pressure of oxygen in the test chamber by the Nernst equation:

$$\text{outputEMF} = \frac{RT}{4F}\ln\left(\frac{P'_{O_2}}{P''_{O_2}}\right)$$

where R is the universal gas constant (8.314 $J \cdot K^{-1} \cdot mol^{-1}$), T is the operating temperature in kelvin, F is Faraday's constant (9.6485×10^4 $C \cdot mol^{-1}$), P'_{O_2} is the reference partial pressure of oxygen, and P''_{O_2} is the sample partial pressure of oxygen.

Fuel cell: An alternative configuration of a concentration cell is a fuel cell. If a fuel gas such as hydrogen is admitted into one of the chambers, the cell converts

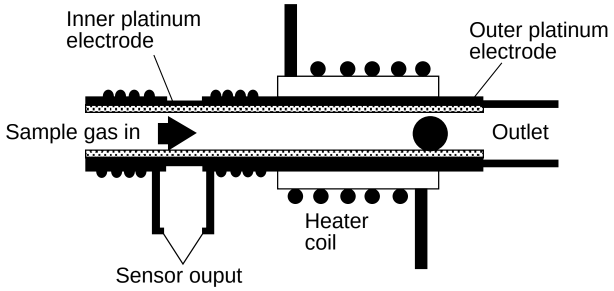


Figure 10. Schematic diagram of a flow through tube sensor. (Adapted from Refs. (1,28,29).)

the chemical energy of the fuel into electrical energy which may be delivered to a load placed across the electrodes (1,15).

Oxygen pump: A concentration cell may also be used as an oxygen pump. An EMF applied across the electrodes will pump oxygen from one chamber to another, the rate and direction depending on the strength and polarity of the applied EMF (1,15).

Sensor design: There are a number of different solid electrolyte sensor designs typically used: a flow through tube sensor, a test tube sensor, and a disk sensor.

Flow-through tube and test tube sensor: The flow-through tube sensor is made from a solid electrolyte tube with a porous platinum electrode on the outside and inside of the tube (Fig. 10). An external heater is used to raise the temperature of the solid electrolyte to its operating temperature where it conducts oxygen. Ambient air outside the tube is used as the reference gas. The sample gas is admitted into the central part of the tube. This simple design was one of the first forms used (15,30,31).

A common variation on this design is the miniature test tube sensor (Fig. 11), which uses a sealed tube containing a reference gas with known oxygen partial pressure (1,32).

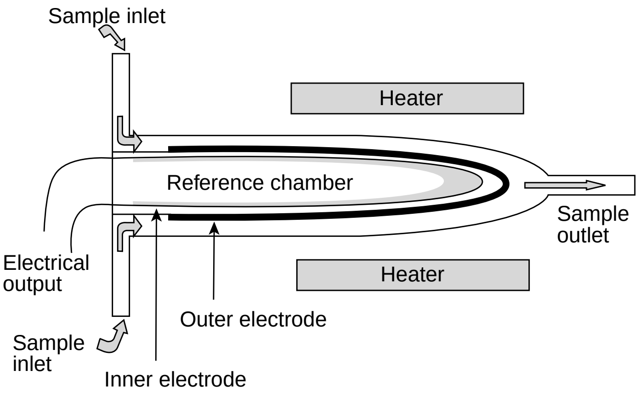


Figure 11. Schematic diagram of a test tube sensor. (Adapted from Refs. 1,33.)

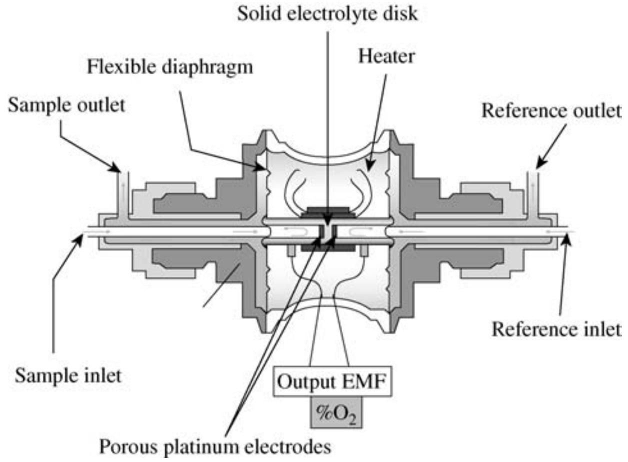


Figure 12. Schematic diagram of a miniature disk sensor. (Adapted from Refs. 1,34,35.)

Like the flow-through sensor, the test tube sensor is made of a solid electrolyte tube with porous electrodes on the inside and outside of the tube, but it is hermetically sealed at one end. The closed end of the tube is place in an external heater. The inside of the tube is used as the reference chamber and ambient air is used for the reference gas. The sample gas is admitted into a chamber surrounding the tube and flows around the outer electrodes. Because of its simple design, it is the most common solid electrolyte cell (1).

Disk sensor: The disk sensor is made with a solid electrolyte disk with porous electrodes on each side. The disk is attached to a metal tube of equal thermal expansion coefficient. Sample gas is admitted into a chamber on one side of the disk and a reference gas on the other. A heater is placed inside the tube on the reference chamber side (1,15,34).

A miniature design of the disk sensor (Fig. 12) has symmetrically placed porous platinum electrodes on either side of a thin solid electrolyte disk. The disk divides a small ceramic cylinder into two hermetically sealed chambers: one for the reference gas and one for the sample gas. An electrical heater brings the cell to the operating temperature. A temperature sensor and feedback loop attached to the heater ensures that the electrodes are also heated to the same temperature as each other and to the disk. If there is a difference in oxygen concentration between the reference and sample chambers, a voltage is generated between the two electrodes. A thin metallic holder is used to suspend the ceramic cylinder and ensures that the ceramic will not crack due to sudden thermal expansion.

Limitations of Ceramic Analyzers. Because the electrode is catalytic, any combustible gas will react with oxygen on the electrode causing it to age and the sensor to give a falsely low reading.

The high operating temperature of these sensors precludes its use around combustible gases.

Ceramics fracture due to thermal shock. Miniaturization helps considerably. These sensors usually require a warm up time (1,15,34).

Pressure on both sides of the disk must be the same or the reading will be inaccurate. This can be accomplished by venting both sides to the ambient atmosphere (15).

OPTICAL SENSORS

Fluorescence Quenching

One of the more recent oxygen analyzer designs uses a technique known as florescence quenching. Fluorescence dyes, such as perylene dibutyrate fluoresce for a certain amount of time in an atmosphere without oxygen. The presence of oxygen quenches this fluorescence. The fluorescence time is thus inversely proportional to the partial pressure of oxygen. In addition to oxygen sensing, fluorescence sensing can be used to detect glucose, lactate, and pH in the laboratory setting (36).

Fiber optic sensors: Fiber optic sensors use florescence quenching to measure the partial pressure of oxygen. A fiber optic strand delivers an optical pulse (usually blue light) to the fluorescent dye. The dye molecules are held in place by small beads of clear plastic. The beads are enclosed by a porous polypropylene membrane that is gas permeable and hydrophobic. The fluorescence is sensed by a photodetector at the end of a second fiber optic strand. The configuration of this sensor can be seen in Fig. 13. The time of fluorescence is calibrated for oxygen concentration. Because the fiber optic is only used to deliver the optical pulse and the fluorescence quenching is used to sense oxygen, the term fiber optic sensor is somewhat of a misnomer.

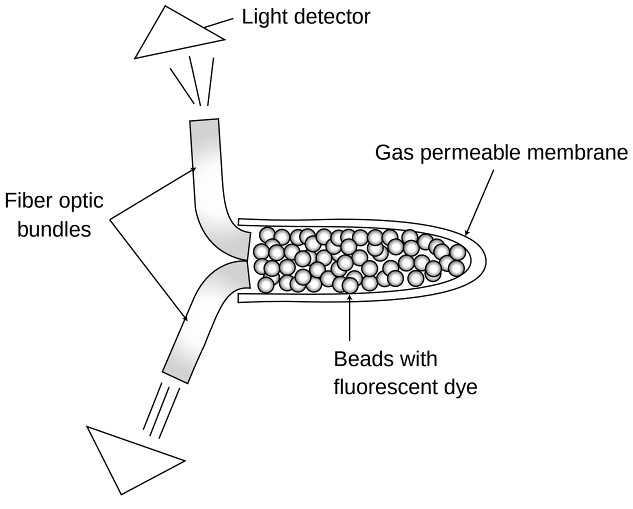


Figure 13. Schematic diagram of a fiber optic sensor. (Adapted from Refs. 1,37.)

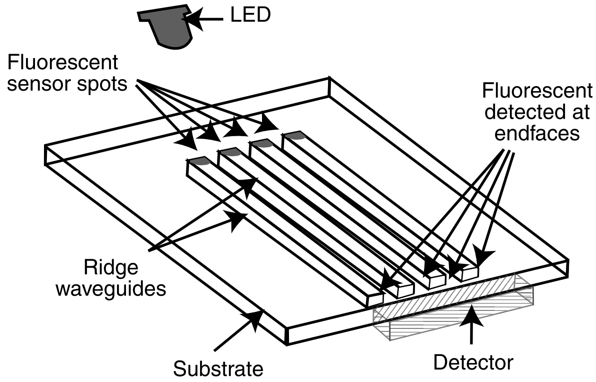


Figure 14. Schematic diagram of an integrated optic oxygen sensor chip. (From Ref. (38).)

Integrated optic oxygen sensor chip: Some sensors that can detect and identify multiple gases in a sample are called multianalyte sensors. One such sensor is an integrated optic multianalyte sensor. This recent development is a miniature optical sensor that has both biomedical and commercial applications. It is based on fluorescence quenching by oxygen.

This sensor, shown in Fig. 14, consists of a multimode ridge waveguide deposited on a dielectric substrate of a higher refractive index than the ridge. Spots of solgel, doped with an oxygen sensitive fluorescent dichlororuthenium dye complex, are deposited at the end of the waveguide and directly excited by a blue LED. Optical detectors at the other end of the waveguide detect emissions from the fluorescent spots. The fluorescence is efficiently coupled to the waveguide as the fluorescent spots are oriented to preferentially emits photons at an angle exceeding the critical angle defined by the two mediums. The theory of fluorescence emission at a dielectric interface is discussed further in (39).

The main limitation of this type of device is the response time. Typically, a 10 s integration period is used for each partial pressure oxygen measurement.

The main advantage of this device is its size and relative ease of fabrication (38). These chips have a very small foot print (<1 cm^2). They can be quickly manufactured using soft lithography.

Polarization-based oxygen sensor: Another sensor based on fluorescence quenching is the polarization based oxygen sensor. This sensor uses an oxygen-sensitive film ($\text{Ru(dpp)}_3\text{Cl}_2$) and an oxygen-insensitive film (Styrl7). A diagonally polarized source illuminates these films that fluoresce in different ways. The oxygen-insensitive film is stretched so that the molecules preferentially emit vertically polarized photons. The $\text{Ru(dpp)}_3\text{Cl}_2$ film emits mostly unpolarized photons. Orthogonally oriented polarizers select for the vertical and horizontal components of the combined emitted light. The overall polarization of

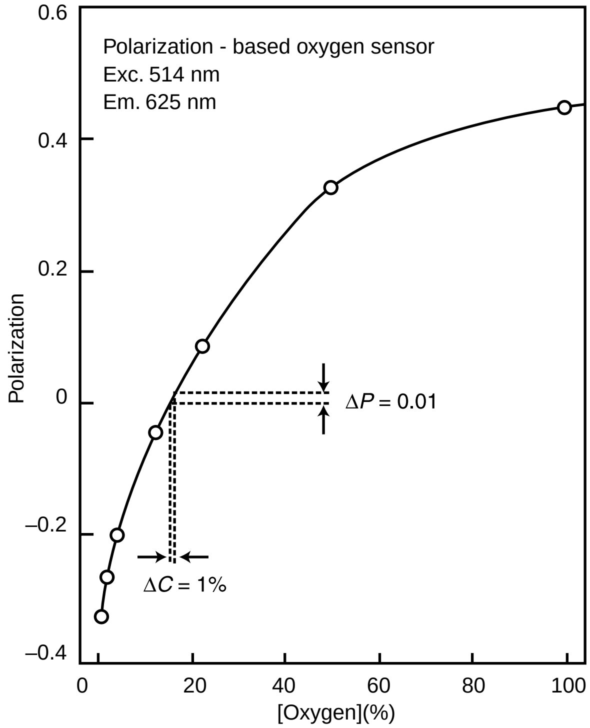


Figure 15. The polarization of the combined film photon emission is proportional to the partial pressure of oxygen. (From Ref. (40).)

the combined emission is sensitive to the partial pressure of oxygen in a sample (Fig. 15). The theory of polarization sensing is discussed further in (40).

UV Absorption

Spectroscopy can also be used to detect oxygen. Oxygen has a maximum absorption coefficient around a wavelength of 0.147 μm, which is in the ultraviolet (UV) range. Most other gases have a much smaller absorption coefficient at this wavelength (1).

A simple device uses an ultraviolet light source, such as a discharge lamp or UV laser as the light source. The beam is split into two paths by a vibrating mirror and directed into either a reference cell filled with nitrogen or a sample cell filled with the gas in question. A photomultiplier tube is used for detection at the end of each path. The ratio of energy received through the sample and reference cell is related to the partial pressure of oxygen in the sample.

Raman Spectroscopy

Raman spectroscopy can be used to monitor multiple gases in a sample. Raman spectroscopy is commonly used in medicine for real-time breath-to-breath analysis or for monitoring respiratory gas mixtures during anesthesia (41–44).

This technique uses the inelastic or Raman scattering, of monochromatic light. The frequencies of the returned light give information about the vibrational, rotational, or other low frequency modes of atoms or molecules in a sample. This information is specific to given elements and compounds and can thus be used to identify and distinguish different gases in a mixture.

In Raman spectroscopy, a sample is illuminated with a laser beam, usually in the visible, near-IR, or near-UV range. Most of the light is scattered elastically or by Rayleigh scattering and is of the same frequency as the incident light. However, a small portion of the light is scattered in-elastically. Phonons, which are quanta of vibrational energy, are absorbed or emitted by the atom or molecule causing the energy of the incident photons to be shifted up or down. This shift in energy corresponds to a shift of frequency. Frequencies close to the laser line are filtered out and frequencies in a certain spectral window are dispersed on to a photomultiplier tube or CCD (charged couple device) camera.

OTHER GASEOUS OXYGEN SENSORS

Gas Chromatography–Mass Spectrometer

A gas chromatography–mass spectrometer (GCMS) combines GC and MS to identify substances in a gas sample. It can be used to detect a variety of compounds in a mixture or it can simply be used to detect the presence of oxygen. The GCMS analyzers are commonly used to measure gas composition in respiratory circuits (42).

Principles of Operation. The gas chromatograph separates compounds into the molecular constituents by retaining some molecules longer than others. These molecules are broken up into ionized fragments that are identified by the mass spectrometer based on the molecules' mass/charge ratio (m/z).

The GC consists of an injector port, an oven, a carrier gas supply, a separation column, and a detector. The injected sample is vaporized in a heated chamber and pushed through the separation column by an inert carrier gas. The separation column is typically a capillary column made of a long, small diameter (usually 1–10 m in length and 0.5 mm in diameter) tube of fused silica (high quality drawn glass) or stainless steel formed into a coil. The components of the sample are separated by two different mediums inside the column that control the speed of travel. These mediums are either coated on the inner surface of the column or packed in the column. Part of the media, known as the stationary phase, absorbs molecules for a certain amount of time before releasing them. The amount of time depends on the chemical properties of the molecule and thus certain molecules are detained longer than others. This, in effect, separates the molecules in the mixture in time. The molecules then travel to the detector. The output of the detector is processed by an integrator. The response of the detector over time is the chromatograph (Fig. 16).

The mass spectrometer separates ions from the gas chromatograph by their charge to mass ratio (m/z) by using an electric or magnetic field. Mass spectrometers usually consist of an ion source, a mass analyzer and a detector.

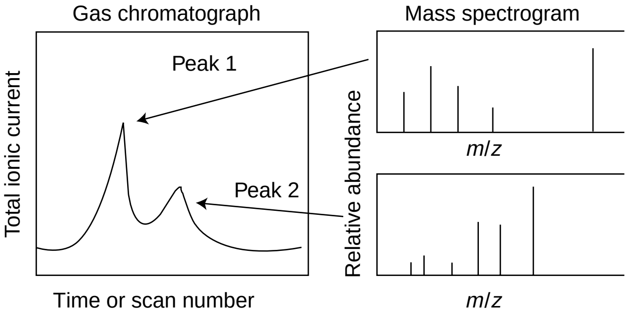


Figure 16. An example gas chromatograph and corresponding mass spectrograms.

The ion source ionizes the sample gas usually with an electron gun. The charged particles are accelerated with an electric field and are steered by the mass analyzer on to the detector by means of a varying electric or magnetic field. The speed and deflection of the particle depends on its mass an charge. When a charged particle comes near or strikes the surface of the detector, a current is induced and recorded. This current is typically amplified by an electron multiplier or Faraday cup. The resulting ionic current plot is the mass spectrum of the ions.

Limitations of Gas Chromatography–Mass Spectrometry. The main limitation of the GCMS is its extremely high price. Mass spectrometers alone run ~$40 k. However, a GCMS may be used for any time of gas measurement.

The Warburg Apparatus

A much older method of oxygen analysis was pioneered by German biochemist Otto Heinrich Warburg (45). The Warburg apparatus was used for measuring cellular respiration and fermentation. This method is based on Boyle's law, which relates the pressure and volume of a gas, and Charle's law, which relates the pressure and temperature of a gas. Combining these laws yields the ideal gas law ($PV = nRT$, where P = pressure, V = volume, n = number moles, R = universal gas constant, T = temperature). At a constant temperature and volume, any change in the amount of gas can be measured as a change in pressure. A typical Warburg apparatus consists of a detachable flask, a waterbath, and a barometer. The sample is placed in a flask and immersed in a bath of water held at a constant temperature. Pressure is measured periodically to determine the amount of gas produced or absorbed by the sample. A variation of this device, used to measure gas production in plants directly from the stem, is the Scholander–Hemmel pressure bomb (46).

CONCLUDING REMARKS

There are three primary types of oxygen sensors available: paramagnetic, electrochemical and spectrographic. Each type of oxygen sensor has limitations and uses dependent on their design, cost and operational environment. Other available oxygen sensors include semiconductor sensors, fluidic sensors, and electron capture oxygen sensors (1). These devices are typically not used in medicine or biology.

BIBLIOGRAPHY

Cited References

1. Kocache R. The measurement of oxygen in gas-mixtures. J Phys E Sci Instrum 1986;19:401–412.
2. O_2 Guide. Electronic Source, Delta-f Corporation. Available at http://www.delta-f.com/O2Guide/o2guide.htm; Accessed 2005.
3. Guyton A, Hall J. Medical Physiology. New York: W.B. Saunders C. 2000.
4. Allsopp PJ. Electrical Cell. UK patent, 969,608. 1964.
5. Johnson T, Rock P. Acute Mountain Sickness. N Eng J Med 1988;319:841–845.
6. Safety_Advisory_Group. Campaign Against Asphyxia. European Industrial Gases Association (EIGA) Safety Newsletter (Special ed.). Brussels.
7. Watanabe T, Morita M. Asphyxia due to oxygen deficiency by gaseous substances. Forensic Sci Int 1998;96:47–59.
8. Operator Dies in Oxygen Defficient Compartment of Ice-Making Barge. Alaska Fatality Assessment & Control Evaluation.
9. Bonsu AK, Stead AL. Accidental cross-connection of oxygen and nitrous oxide in an anesthetic machine. Anaesthesia 1983; 38:767–769.
10. Cheng DK. Field & Wave Electromagnetics. New York: Addison-Wesley; 1984.
11. Instruction Manual: Paramagnetic Oxygen Analyzer: Fuji Electric Co. Ltd.
12. Havens G. The magnetic suceptibilities of some common gases. Phys Rev 1933;43:992–1000.
13. Servomex. Paramagnetic oxygen analysis. Electronic Source, Servomex Corporation. Available at http://www.servomex.com/Servomex.nsf/GB/technology_paramagnetic.html. Accessed 2005.
14. Corporation D-f. O_2 Guide. Electronic Source, Delta-f Corporation. Available at http://www.delta-f.com/O2Guide/o2guide.htm; Accessed 2005.
15. Kocache R. Oxygen analyzers. Encyclopedia of Medical Devices and Instrumentation, Webster JG. editor. New York: Wiley & Sons; 1988. p. 2154–2161.
16. Medlock RS. Oxygen Analysis. Inst Eng 1952;1:3–10.
17. Pauling L. Apparatus for determining the partial pressure of oxygen in a mixture of gases. US patent, 2,416,344, 1947.
18. Meyer RM. Oxygen analyzers: failure rates and life spans of galvanic cells. J Clin Monitoring 1990;6:196–202.
19. Analytical_Industries_Inc. Oxygen analysis pocketed and ABS protected. MD Med. Design, 2001.
20. Hersch P. Electrode assembly for oxygen analysis. UK patent, 913,473. 1962.
21. Hersch P. Improvements relating to the analysis of gases. UK patent, 707,323. 1954.
22. Hersch P. Improvements relating to the analysis of gases. UK patent, 750254. 1956.
23. Gallagher JP. Apparatus and method for maintaining a stable electrode in oxygen analysis. US patent, 3,929,587. 1975.
24. Bergman I. Improvements in or relating to electrical cells. UK patent, 1,344,616. 1974.

25. Bergman I. Improvements in or relating to membrane electrodes and cells. UK patent, 1,200,595. 1970.
26. Li S, Wang Z, Zeng B, Liu J. Multiple respiratory gas monitoring causes changes of inspired oxygen concentration in closed anesthesia system. J Tong Med Univ 1997;17:54–56.
27. Heusner AA, Hurley JP, Arbogast R. Coulometric microrespirometry. Am J Physiol 1982;243:R 185–192.
28. Hickam WM. Device for monitoring oxygen content of gases. US patent, 3,347,767. 1967.
29. Hickam WM. Oxygen control and measuring apparatus. US patent, 3,650,934. 1972.
30. Sodal IE, Micco AJ, Weil JV. An improved fast response oxygen analyzer with high accuracy for respiratory gas analysis. Biomed Sci Instrumentat 1975;11:21–24.
31. Weil JV, Sodal IE, Speck RP. A modified fuel cell for the analysis of oxygen concentration. J Appl Physiol 1967;23: 419–422.
32. Deportes CH, Henault MPS, Tasset F, Vitter GRR. Electrochemical gauge for measuring partial pressure of oxygen. US patent, 4,045,319. 1977.
33. Sayles DA. Method and apparatus for continuously seeing the condition of a gas stream. US patent, 3,869,370. 1975.
34. Kocache RMA, Swan J, Holman DF. A miniature rugged and accurate solid electrolyte oxygen sensor. J Phys E Sci Instrum 1984;17:477–482.
35. Servomex. Zirconia oxygen analysis. Electronic Source, Servomex Corporation. Available at http://www.servomex.com/Servomex.nsf/GB/technology_zirconia.html. Accessed 2005.
36. Gryczynski Z, Gryczynski I, Lakowicz JR. Fluorescence-sensing methods. Methods Enzymol 2003;360:44–75.
37. Peterson JI, Fitzgerald RV, Buckhold DK. Fibre-optic probe for *in vivo* measurement of oxygen partial pressure. Anal Chem 1984;56:62–67.
38. Burke CS, et al. Development of an integrated optic oxygen sensor using a novel, generic platform. The Analyst 2005;130: 41–45.
39. Polerecki L, Hamrle J, MacCraith BD. Theory of the radiation of dipoles placed within a multilayer system. Appl Op 2000; 39:3968–3977.
40. Gryczynski I, Gryczynski Z, Rao G, Lakowicz JR. Polarization-based oxygen sensor. Analyst 1999;124:1041–1044.
41. VanWagenen RA, et al. Dedicated monitoring of anesthetic and respiratory gases by Raman scattering. J Clin Monitoring 1986;2:215–222.
42. Westenskow DR, Coleman DL. Can the Raman scattering analyzer compete with mass spectrometers: an affirmative reply. J Clin Monitoring 1989;5:34–36.
43. Westenskow DR, Coleman DL. Raman scattering for respiratory gas monitoring in the operating room: advantages, specifications, and future advances. Biomed Inst Technol 1989;23:485–489.
44. Westenskow DR, et al. Clinical evaluation of a Raman scattering multiple gas analyzer for the operating room. Anesthesiology 1989;70:350–355.
45. Definition of Warburg Apparatus. Electronic Source, MediciNet, Inc. Available at http://www.medterms.com/script/main/art.asp?articlekey=7150; Accessed 2005.
46. Scholand Pf, Hammel HT, Bradstre E, Hemmings E. Sap pressure in vascular plants - negative hydrostatic pressure can be measured in plants. Science 1965;148: 339.

See also FIBER OPTICS IN MEDICINE; GAS AND VACUUM SYSTEMS, CENTRALLY PIPED MEDICAL; OXYGEN MONITORING.

OXYGEN SENSORS

STEVEN J. BARKER
University of Arizona
Tucson, Arizona

INTRODUCTION

This article reviews recent advances in the monitoring of patient oxygenation. We summarize the transport of oxygen from the atmosphere to the cell, and describe monitors that function at four stages of the O_2 transport process. These four stages include respired gas, arterial blood, tissue, and venous blood. History and recent developments in pulse oximetry will be discussed. Continuous intraarterial blood-gas sensors will be described and contrasted with other oxygen monitors. Finally, tissue oxygen monitoring and mixed-venous oximetry are discussed.

OXYGEN TRANSPORT IN THE HUMAN BODY

At rest, we consume $\sim 10^{23}$ molecules of oxygen per second. Our complex cardiopulmonary system has developed to rapidly transport this large amount of oxygen from the atmosphere to every cell in the body (Fig. 1).

The equation for arterial blood oxygen content ($C\mathrm{aO}_2$) shows that $\sim$ 99% of the oxygen in arterial blood is in the form of hemoglobin-bound oxygen:

$$C\mathrm{aO}_2 = 1.38(S\mathrm{aO}_2/100)(\mathrm{Hb}) + 0.003\,P\mathrm{aO}_2 \qquad (1)$$

where $C\mathrm{aO}_2$ is in units of milliliters per deciliter of blood (also called vols%); $S\mathrm{aO}_2$ is the arterial hemoglobin saturation in percent; Hb is the total hemoglobin concentration in grams per deciliter; and $P\mathrm{aO}_2$ is the arterial oxygen tension (partial pressure) in millimeters of mercury. Upon inserting typical arterial values [$S\mathrm{aO}_2 = 100\%$, $\mathrm{Hb} = 15\ \mathrm{g{\cdot}dL^{-1}}$, $P\mathrm{aO}_2 = 100$ mmHg (13.33 kPa)], we find that normal $C\mathrm{aO}_2$ is $\sim 21\ \mathrm{mL{\cdot}dL^{-1}}$. The amount of oxygen delivered to the tissues in the arterial blood is then given by the cardiac output (CO) times the $C\mathrm{aO}_2$ (neglecting the small dissolved oxygen term):

$$D\mathrm{O}_2 = 13.8(\mathrm{CO})(\mathrm{Hb})(S\mathrm{aO}_2/100) \qquad (2)$$

(The factor 1.38 becomes 13.8 because Hb is normally measured in grams per deciliter, while CO is measured in liters per minute. There are 10 dL in 1 L.)

Finally, the oxygen consumption by the tissues ($V\mathrm{O}_2$) is determined by the difference between arterial oxygen delivery and venous oxygen return:

$$V\mathrm{O}_2 = 13.8(\mathrm{Hb})(\mathrm{CO})(S\mathrm{aO}_2 - S\mathrm{vO}_2)/100 \qquad (3)$$

This Fick equation can be solved for any of the four oxygen variables involved.

If we substitute normal resting values into the equation, we predict a resting $V\mathrm{O}_2$ of

$$\begin{aligned} V\mathrm{O}_2 &= 13.8\,(15\,\mathrm{g \cdot dL^{-1}})(5\,\mathrm{L \cdot min^{-1}})(99 - 75)/100 \\ &= 248\ \mathrm{mL \cdot min^{-1}} \end{aligned}$$

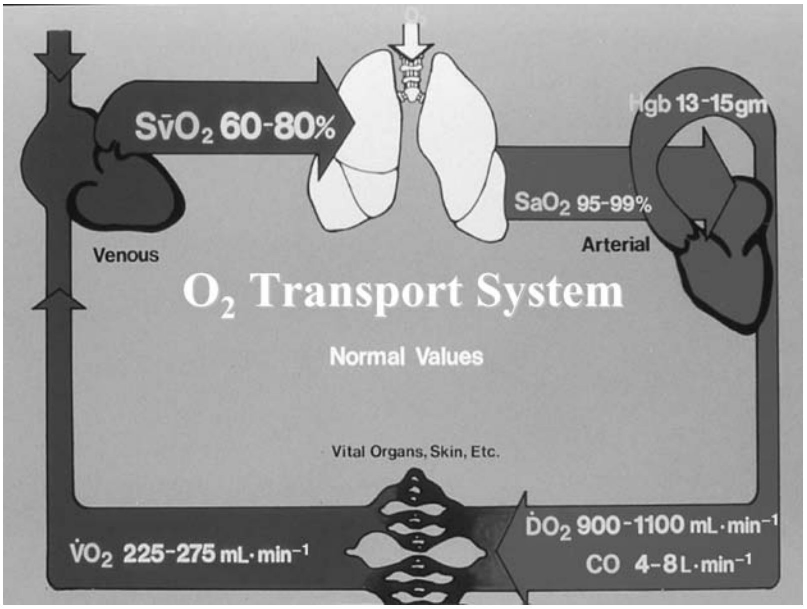


Figure 1. Schematic of the oxygen transport system. Arterial blood leaving the left ventricle (right side of figure) has and oxygen content of 21 mL·dL^{-1}, and the total oxygen delivery (DO$_2$) is roughly 1000 mL·min^{-1}. At rest, 0.25 of this oxygen delivery is consumed (VO$_2$), leaving a mixed venous saturation of 75%.

During exercise or stress we can rapidly increase cardiac output to at least 20 L·min^{-1}, and decrease venous saturation to $\sim$ 40%, yielding a VO$_2$ of

$$\begin{aligned} VO_2 &= 13.8\,(15\,\mathrm{g}\cdot\mathrm{dL}^{-1})(20\,\mathrm{L}\cdot\mathrm{min}^{-1})(99-40)/100 \\ &= 2029\ \mathrm{mL}\cdot\mathrm{min}^{-1} \end{aligned}$$

The human ability to rapidly adjust cardiac output (CO) and mixed-venous oxygen saturation (SvO$_2$) can be used to compensate for disease processes that affect other transport variables, such as anemia (hemoglobin) or hypoxemia (SaO$_2$). For example, consider a severely anemic patient with Hb value of 2.5 g·dL^{-1}, who compensates by increasing cardiac output to 15 L·min^{-1} and decreasing venous saturation to 50%:

$$\begin{aligned} VO_2 &= 13.8\,(2.5\,\mathrm{g}\cdot\mathrm{dL}^{-1})(15\,\mathrm{L}\cdot\mathrm{min}^{-1})(99-50)/100 \\ &= 254\ \mathrm{mL}\cdot\mathrm{min}^{-1} \end{aligned}$$

This extremely anemic patient can thus maintain a normal oxygen consumption by adaptations in CO and SvO$_2$ that are milder than those we use during normal exercise.

OXYGEN IN THE ARTERIAL BLOOD: PULSE OXIMETRY

Physiologic Considerations

The normal relationship between SaO$_2$ and PaO$_2$ is the familiar oxyhemoglobin dissociation curve, shown in Fig. 2. Three convenient points on this curve to remember are $P\mathrm{aO}_2 = 27$ mmHg (3.60 kPa), $S\mathrm{aO}_2 = 50\%$; $P\mathrm{aO}_2 = 40$ mmHg (5.33 kPa), $S\mathrm{aO}_2 = 75\%$; and $P\mathrm{aO}_2 = 60$ mmHg (8.00 kPa), $S\mathrm{aO}_2 = 90\%$. The curve will be shifted toward the right by acidosis, hypercarbia, or increasing 2,3-DPG. At PaO$_2$, values greater than 80 mmHg (10.66 kPa), SaO$_2$ is almost 100% and thus becomes virtually independent of PaO$_2$. It is important to remember this fact during SaO$_2$ monitoring in the operating room, where elevated inspired oxygen fraction (FIO$_2$) values will yield PaO$_2$ values much $>$ 80 mmHg (10.66 kPa) most of the time.

A knowledge of the relationship between SaO$_2$ and PaO$_2$ allows us to predict the physiologic limitations of saturation monitoring by pulse oximetry. Specifically, the pulse oximeter will give no indication of downward trends in PaO$_2$ during anesthesia at elevated FIO$_2$ until PaO$_2$ values $<$ 80–90 mmHg (10.66–12.00 kPa) are reached. In an animal study, intentional endobronchial intubations at $\mathrm{F_IO_2}$ values $>$ 30% were not detected by the pulse oximeter (1). This results from the fact that the PaO$_2$ after endobronchial intubation did not decrease below $\sim$ 80 mmHg (10.66 kPa) when $F_\mathrm{I}\mathrm{O}_2$ was elevated.

Technology

Oximetry, a subset of spectrophotometry, determines the concentrations of various hemoglobin species by measuring the absorbances of light at multiple wavelengths. The absorbance spectra of the four most common hemoglobin species are shown in Fig. 3. The number of oximeter light wavelengths used must be equal to or greater than the number of hemoglobin species present in the sample. A laboratory CO-oximeter, which uses four or more wavelengths, can measure the concentrations of reduced hemoglobin, oxyhemoglobin, methemoglobin, and carboxyhemoglobin. If all four hemoglobin species are present in significant concentrations, then an oximeter must have at least four light wavelengths to determine the concentration of *any* of the four species.

The conventional pulse oximeter is a two-wavelength oximeter that functions *in vivo*. Conventional pulse oximetry first determines the fluctuating or alternating current (ac) component of the light absorbance signal. At each of its two wavelengths the oximeter divides the ac signal by the

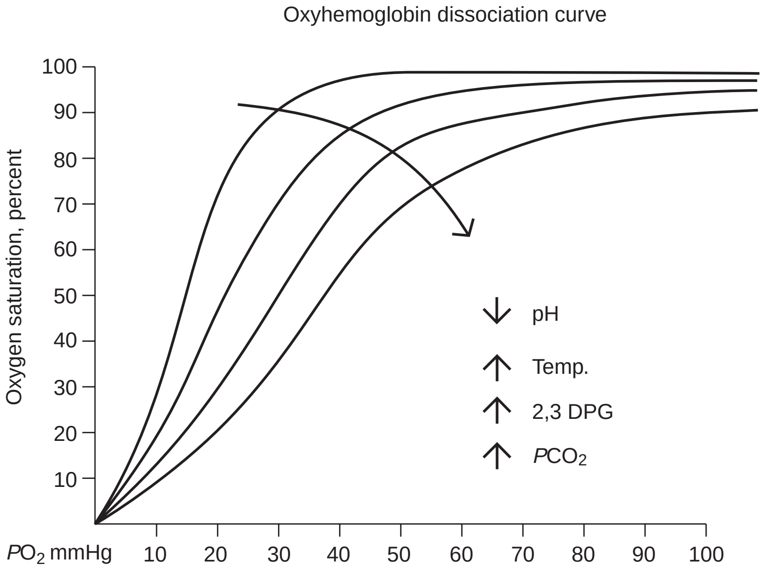


Figure 2. The oxyhemoglobin dissociation curve: a plot of hemoglobin oxygen saturation as a function of oxygen tension ($P\mathrm{O}_2$). The curve is shifted to the right with decreasing pH or increasing temperature, $P\mathrm{CO}_2$, or 2,3-DPG.

corresponding fixed or direct current (dc) absorbance component, to obtain a pulse-added absorbance. It then calculates the ratio of the two pulse-added absorbances (one for each wavelength), and this ratio R is related to arterial hemoglobin saturation by a built-in calibration algorithm. The resulting pulse oximeter saturation estimate is called $S\mathrm{pO}_2$. The calibration curve of the pulse oximeter is empirical; that is, it is based on human volunteer experimental data.

Sources of Error

Dyshemoglobins. As previously noted, the pulse oximeter uses two wavelengths and can distinguish only two hemoglobin species: reduced hemoglobin and oxyhemoglobin. If either carboxyhemoglobin (COHb) or methemoglobin (MetHb) is present, the pulse oximeter effectively has fewer equations than unknowns, and it cannot find any of the hemoglobin concentrations. It is thus unclear *a priori* how the pulse oximeter will behave in the presence of these dyshemoglobins.

Two animal experiments have characterized pulse oximeter behavior during methemoglobinemia and carboxyhemoglobinemia. In one of these, dogs were exposed to carbon monoxide (220 ppm) over a 3–4 h period (2). At a COHb level of 70% (meaning that 70% of the animal's hemoglobin was in the COHb form), the $S\mathrm{pO}_2$ values were $\sim 90\%$, whereas the actual $S\mathrm{aO}_2$ was 30% (Fig. 4). The pulse oximeter thus "sees"

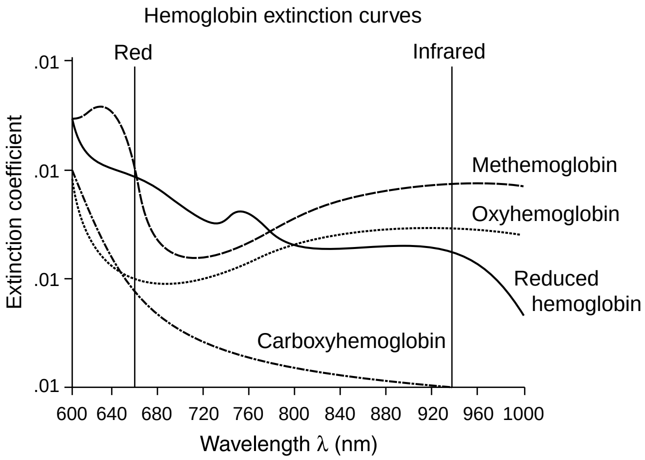


Figure 3. Extinction coefficient (or light absorbance) versus wavelength of light for four different hemoglobins. reduced Hb, $\mathrm{O_2Hb}$, COHb, and MetHb. The two wavelengths used by most pulse oximeters (660 nm, 930 nm) are indicated by vertical lines.

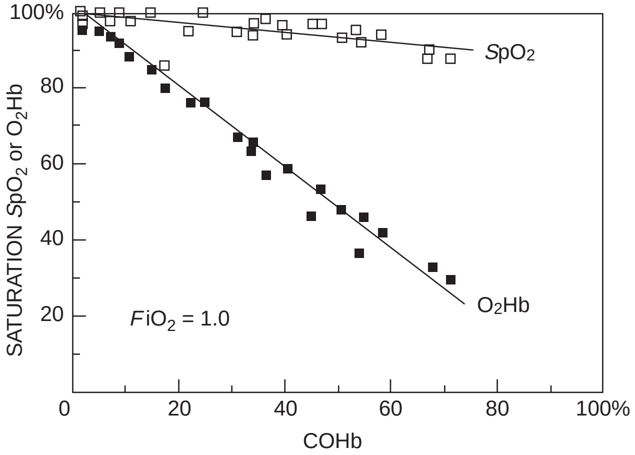


Figure 4. The effect of carbon monoxide on pulse oximetry. Plots of Hb saturation measured by ■ laboratory CO-oximeter, and □ conventional pulse oximeter, as functions of COHb level. As COHb increases, CO-oximeter shows linear decline in saturation, while pulse oximeter remains > 90% saturation (2).

carboxyhemoglobin as if it were composed mostly of oxyhemoglobin.

In a similar animal experiment, increasing methemoglobin concentrations (up to 60%) produced SpO_2 readings that gradually decreased to $\sim 85\%$ (3). As these animals were further desaturated by lowering the FIO_2 during methemoglobinemia, the pulse oximeter SpO_2 reading failed to track either functional or fractional saturation. On the other hand, the presence of fetal hemoglobin has little effect on pulse oximeter accuracy, which is fortunate in the treatment of premature neonates. There are conflicting anecdotal reports on the influence of sickle hemoglobin, and it is impossible to perform volunteer hypoxia studies on patients with sickle-cell disease.

As of this date, no commercially available pulse oximeter can either measure dyshemoglobins or function accurately in the presence of substantial levels of COHb or MetHb. Masimo Inc. has very recently (March, 2005) announced the development of a new Rainbow Technology pulse oximeter that uses multiple light wavelengths to measure COHb and SaO_2 simultaneously. There are as yet no published data regarding the success of this approach, but it is a potentially important new advancement.

Intravenous Dyes. As abnormal hemoglobin species can adversely affect the accuracy of pulse oximetry, so can intravenous dyes injected during surgery or critical care. Two studies found that intravenous methylene blue causes large, rapid decreases in displayed SpO_2 without changes in actual saturation, and that indocyanine green causes smaller false decreases in SpO_2 (4,5). Intravenous fluorescein or indigo carmine appeared to have little effect.

Reductions in Peripheral Pulsation; Ambient Light. Several studies have examined the effects of low perfusion upon SpO_2 (6,7). In a clinical study of critically ill patients during a wide range of hemodynamic conditions, extremes in systemic vascular resistance were associated with loss of pulse oximeter signal or decreased accuracy. During reduced pulse amplitude, pulse oximeters may become more sensitive to external light sources, such as fluorescent room lights (8). Most modern pulse oximeters effectively measure and correct for ambient light intensity.

Motion Artifact. Patient motion, which causes a large fluctuating absorbance signal, is a very challenging artifact for pulse oximetry. Motion artifact rarely causes great difficulty in the operating room, but in the recovery room and intensive care unit it can make the pulse oximeter virtually useless. Design engineers have tried several approaches to this problem, beginning with increasing the signal averaging time. Most current pulse oximeters allow the user to select one of several time averaging modes. However, improving motion performance by simply increasing averaging time is potentially dangerous—it can cause the instrument to miss significant, but short-lived hypoxemic events, which are very common in neonates.

Masimo, Inc. has developed a completely different approach to the analysis of the oximeter light absorbance signals, using adaptive digital filtering. This has led to improved accuracy and reliability during motion artifact, both in laboratory studies (9,10) and in the neonatal intensive care unit (11). The new technology has spurred other manufacturers (e.g., Nellcor, Philips, Datex-Ohmeda) to improve their signal analysis methods, so that today's generation of pulse oximeters has much improved performance during motion.

Venous Pulsations. Conventional pulse oximeter design is predicated on the assumption that the pulsatile component of tissue light absorbance is entirely caused by arterial blood. However, the light absorbance of venous blood can also have a pulsatile component, and this may affect SpO_2 values under some conditions (12). Conventional pulse oximeters may read falsely low values or may fail to give any reading in circumstances leading to venous congestion. This can occur, for example, when using an earlobe sensor on a patient who is undergoing mechanical ventilation, or who is in the Trendelenberg position.

Penumbra Effect. When a pulse oximeter sensor is not properly positioned on the finger or earlobe, the light traveling from the source to the detector may pass through the tissues at only grazing incidence. This penumbra effect reduces the signal/noise ratio, and may result in SpO_2 values in the low 1990s in a normoxemic subject. More importantly, a volunteer study has shown that in hypoxemic subjects, the penumbra effect can cause SpO_2 to either overestimate or underestimate actual SaO_2 values, depending on the instrument used (13). A pulse oximeter with a malpositioned sensor may therefore indicate that a patient is only mildly hypoxemic when in fact he or she is profoundly so.

OXYGEN IN THE ARTERIAL BLOOD: CONTINUOUS INTRAARTERIAL PO_2 MEASUREMENT

There have been a number of efforts to monitor intraarterial oxygen tension directly and continuously, using miniaturized sensors passed through arterial cannulas. The first practical approach to this problem employed the Clark electrode, the same oxygen electrode used in the conventional laboratory blood-gas analyzer. Although miniaturized Clark electrodes have been used in several clinical studies, the technique never achieved popularity because of problems with calibration drift and thrombogenicity (14). More recently, the principle of fluorescence quenching was used to develop fiberoptic "optodes" that can continuously monitor pH, $PaCO_2$, and PaO_2 through a 20 gauge radial artery cannula (Fig. 5).

Fluorescence quenching is a result of the ability of oxygen (or other substances to be measured) to absorb energy from the excited states of a fluorescent dye, thus preventing this energy from being radiated as light.

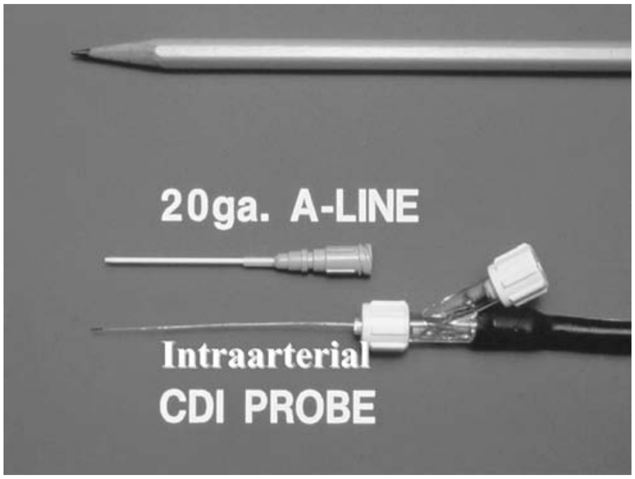


Figure 5. An intraarterial fiber-optic optode sensor. The optode is 0.6 mm in diameter and fits easily through a 20 gauge radial artery cannula, also shown.

Lubbers and Opitz (15) developed the first fluorescence quenching optode that simultaneously measured oxygen and carbon dioxide tensions in gases or liquids. In the 1980s, optodes were successfully miniaturized for intraarterial use, and several studies were reported in both animal and humans (16,17).

Clinical Studies

Several clinical studies suggested the usefulness of intraarterial optodes in the operating room (18). The scatter (random error) of optode oxygen tension values is lowest at low oxygen tensions, a characteristic of these sensors. The accuracy of the optode appeared to be within the clinically acceptable range when 18-gauge arterial cannulas were used. The optode can display complete blood-gas data continuously at the patient's bedside, with a time response measured in seconds. Nevertheless, the high costs of the disposable sensors ($\sim$ \$300 each) and their inconsistent reliability have caused the intraarterial optodes to disappear from the clinical market. These devices have other potential applications in tissues and organs, which may be realized in the future. One manufacturer today is marketing an optode sensor for assessment of the viability of tissue grafts.

OXYGEN IN TISSUE: TRANSCUTANEOUS OXYGEN

Physiologic Considerations

The transcutaneous oxygen ($P_{tc}O_2$) sensor is a Clark electrode that measures oxygen diffusing through the surface of the skin from dermal capillaries (Fig. 6). The sensor must be heated to at least 43 °C (in adults) to facilitate diffusion through the stratum corneum. Surface heating also produces local hyperemia of the dermal capillaries, which tends to "arterialize" the blood and cause a rightward shift in the oxyhemoglobin dissociation curve. The effects above tend to increase $P_{tc}O_2$, and these are counterbalanced by other effects that decrease it, namely diffusion gradients and metabolic oxygen consumption by the skin. In neonates, these competing effects nearly cancel and $P_{tc}O_2$ is approximately equal to PaO_2. In adults, the stratum corneum is thicker and hence the $P_{tc}O_2$ is usually lower than PaO_2. The transcutaneous index, $P_{tc}O_2/PaO_2$, has average values of 1.0 in neonates, 0.9 in pediatric patients, 0.8 in adults, and 0.6–0.7 in the elderly.

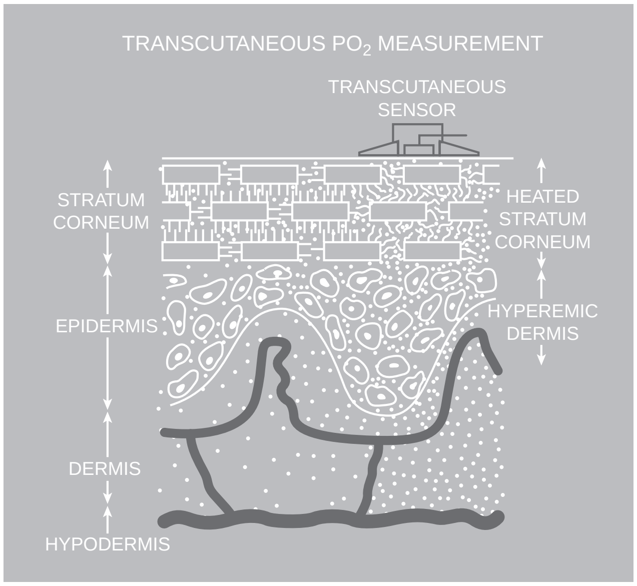


Figure 6. Schematic of transcutaneous PO_2 sensor on skin surface. Heat from sensor "melts" the diffusion barrier of the stratum corneum layer, and "arterializes" the blood in the dermal capillaries beneath.

The most serious challenges with the interpretation of $P_{tc}O_2$ values are their dependence upon cardiac output and skin perfusion. Several studies have shown that the transcutaneous index falls when the cardiac index decreases below its normal range (19). Animal shock studies have shown that $P_{tc}O_2$ decreases when either PaO_2 or cardiac index decreases, and that it closely follows trends in oxygen delivery (i.e., the product of CO and CaO_2). In other words, $P_{tc}O_2$ monitors oxygen delivery to the tissues rather than oxygen content of arterial blood.

Technical Problems

There are several practical problems associated with the use of $P_{tc}O_2$ sensors. The transcutaneous electrode must be gas calibrated before each application to the skin, and then the sensor requires a 10–15 min warm-up period. In children, the warm-up period is usually shorter. The sensor membrane and electrolyte must be replaced at least once a week. The heated $P_{tc}O_2$ electrode can cause small skin burns, particularly at temperatures of 44 °C or greater. Lower probe temperatures (43 or 43.5 °C) should be used on premature infants and neonates, and the sensor location should be changed every 2–3 h. In adults with a sensor temperature of 44 °C, we have used the same location for 6–8 h with no incidence of burns.

Summary

Transcutaneous oxygen sensors provide continuous, noninvasive monitoring of oxygen delivery to tissues. By contrast, pulse oximetry provides continuous monitoring of arterial hemoglobin saturation. The dependence of $P_{\mathrm{tc}}\mathrm{O}_2$ on blood flow as well as $P\mathrm{aO}_2$ sometimes makes it difficult to interpret changing values. If $P_{\mathrm{tc}}\mathrm{O}_2$ is normal or high, we know that the tissues are well oxygenated. When $P_{\mathrm{tc}}\mathrm{O}_2$ is low, we must determine whether this is the result of low $P\mathrm{aO}_2$ or a decrease in skin perfusion.

OXYGEN IN VENOUS BLOOD: PULMONARY ARTERY OXIMETRY

Physiology of Mixed-Venous Saturation

Oxygen saturation in venous blood is related to venous oxygen content $C\mathrm{vO}_2$ by an equation similar to equation 1:

$$C\mathrm{vO}_2 = 1.38(\mathrm{Hb})(S\mathrm{vO}_2)/100 + 0.003(P\mathrm{vO}_2) \qquad (4)$$

The normal $C\mathrm{vO}_2$ value [with $S\mathrm{vO}_2 = 75\%$, $P\mathrm{vO}_2 = 40$ mmHg (5.33 kPa)] is 15.6 $\mathrm{mL{\cdot}dL^{-1}}$. If we solve equation 3 (the Fick equation) for the venous saturation ($S\mathrm{vO}_2$), we obtain:

$$S\mathrm{vO}_2 = S\mathrm{aO}_2 - V\mathrm{O}_2/[(13.8)(\mathrm{Hb})(\mathrm{CO})] \qquad (5)$$

Equation 8 shows how $S\mathrm{vO}_2$ depends on the four oxygen transport variables: $S\mathrm{aO}_2$, $V\mathrm{O}_2$, Hb, and CO.

When $V\mathrm{O}_2$ falls behind oxygen demand, lactic acidosis will result, eventually leading to death if the problem is not corrected. When this begins to occur in disease (e.g., anemia), the patient's body will try to maintain normal $V\mathrm{O}_2$ using the same two compensatory mechanisms described above during exercise: increasing CO and/or decreasing $S\mathrm{vO}_2$. In the case of anemia, we saw that such compensation can maintain normal $V\mathrm{O}_2$ values even at hemoglobin levels < 3 $\mathrm{g{\cdot}dL^{-1}}$. Thus, a decrease in $S\mathrm{vO}_2$ indicates that a patient is using oxygen reserves to compensate for a supply–demand imbalance. Decreasing oxygen supply may result from low CO (shock), low hemoglobin (anemia), abnormal hemoglobin (carboxyhemoglobinemia), or low $P\mathrm{aO}_2$ (hypoxemia). On the other hand, increasing oxygen demand can result from fever, malignant hyperthermia, thyrotoxicosis, or shivering.

The aforementioned are possible clinical causes of a decrease in $S\mathrm{vO}_2$. There are also conditions that can increase $S\mathrm{vO}_2$ above its normal range of 68–77%. High $S\mathrm{vO}_2$ values can result from decreased tissue uptake of oxygen, peripheral arteriovenous shunting, and inappropriate increases in CO. Clinical conditions that produce elevated $S\mathrm{vO}_2$ values include sepsis, Paget's disease of bone, excessive use of inotropes, cyanide poisoning, and hypothermia. A wedged pulmonary artery catheter will also cause a high $S\mathrm{vO}_2$ reading, but this is a measurement artifact. This can actually be a useful artifact, since it warns the clinician of an inadvertently wedged catheter.

Technical Considerations

Pulmonary artery $S\mathrm{vO}_2$ catheters use the technology of reflectance spectrophotometry; that is, they measure the color of the blood in a manner similar to pulse oximetry. The $S\mathrm{vO}_2$ catheters use fiberoptic bundles to transmit and receive light from the catheter tip. Light-emitting diodes provide monochromatic light sources at two or three wave-

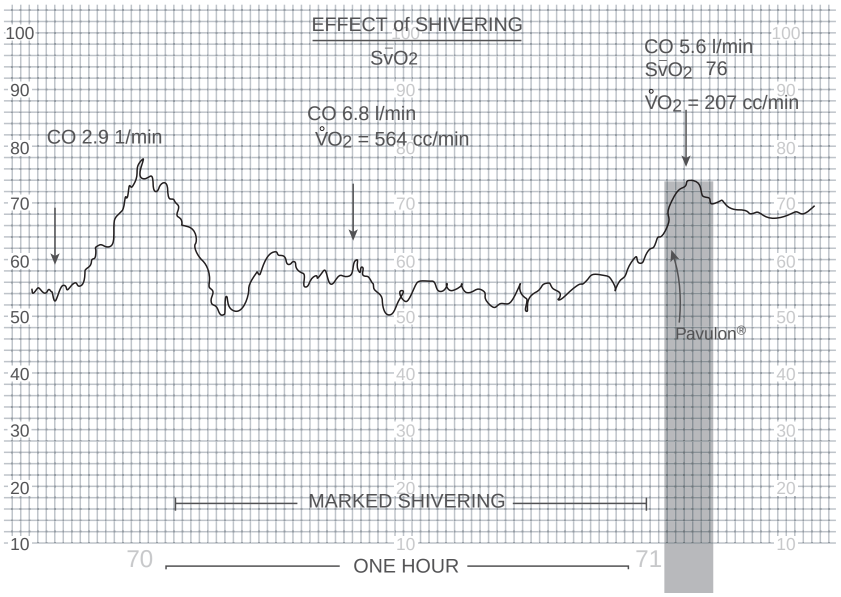


Figure 7. Effect of oxygen consumption ($V\mathrm{O}_2$) upon mixed venous saturation ($S\mathrm{vO}_2$). After weaning from cardiopulmonary bypass, the $S\mathrm{vO}_2$ reaches a normal value of 75%, but then falls to 55%. Measured $V\mathrm{O}_2$ at this time is 564 $\mathrm{mL{\cdot}min^{-1}}$, or twice normal resting value. Patient noted to be shivering. Administration of muscle relaxant stops shivering, restores $V\mathrm{O}_2$ to normal (207 $\mathrm{mL{\cdot}min^{-1}}$), and $S\mathrm{vO}_2$ also returns to normal values.

lengths. (Currently, the Edwards system uses two wavelengths, while the Abbott instrument employs three.) A theoretical advantage of a three wavelength system is that its measurements should not depend on the total hemoglobin level (20). Another problem common to all SvO_2 catheters is the so-called wall artifact, whereby reflection from a vessel wall can produce a signal that is interpreted as an SvO_2 of 85–90%. This problem has been reduced by the addition of digital filtering to the processor, which effectively edits out sudden step increases in SvO_2. However, a persistently high value of SvO_2 may alert the user that the catheter is in the wedged condition, as noted above.

Applications and Limitations

When interpreting continuous SvO_2 versus time tracings in the operating room and intensive care unit, we must always consider equation 5, the Fick equation solved for SvO_2. When SvO_2 changes, we should ask which term(s) in equation 5 are responsible. In the operating room, the terms most likely to change significantly are cardiac output (CO) and hemoglobin (Hb). During general anesthesia with mechanical ventilation, SaO_2 and VO_2 are usually constant, with the exception that VO_2 will decrease during hypothermia. On the other hand, this is not the case in the intensive care unit. Patients in respiratory failure will have varying degrees of arterial desaturation (low SaO_2). Note that SvO_2 is directly related to SaO_2; if SaO_2 decreases by 20% and nothing else changes, then SvO_2 will decrease by 20%. Critical care unit patients may also have frequent changes in VO_2, which can be increased by agitation, shivering, coughing, fever, pain, seizures, defecation, or eating, to name just a few (Fig. 7).

Continuous SvO_2 is a valuable adjunct in the treatment of ventilator-dependent patients. As positive end-expiratory pressure (PEEP) is slowly increased to improve oxygenation, SaO_2 will usually increase, but eventually the cardiac output will begin to decrease as venous return is compromised. At this point, oxygen delivery to tissue may begin to decrease (and SvO_2 begins to decrease) even though SaO_2 is still increasing. SvO_2 is a reflection of oxygen delivery in this situation, and can thus provide a means to optimize positive end-expiratory pressure without the need of serial blood gases and CO measurements.

In summary, continuous SvO_2 monitoring is a valuable technology for the operating room and the critical care unit. It reflects the overall health and functional state of the oxygen transport system. To realize the most benefit from this monitor, it is essential to thoroughly understand the physiology of SvO_2 and how it relates to the other oxygen transport variables.

CONCLUSIONS

Monitoring of oxygen in the respired gases and arterial blood is the standard of care during all anesthetics today. None of us would consider administering general anesthesia without both an FiO_2 monitor and a pulse oximeter. New advances in pulse oximetry will make these instruments more reliable in moving or poorly perfused patients, but they will still be subject to the fundamental limitations of saturation monitoring. Further developments will include pulse oximeters that can function in the presence of COHb and MetHb. In the near future, noninvasive monitors of oxygenation in specific organs and tissues (heart, brain) will become available. Finally, mixed venous oxygen saturation indicates how much is “left over” at the end of the oxygen transport process, which gives an indication of the status of the transport system and the degree to which reserves are being used.

BIBLIOGRAPHY

1. Barker SJ, Tremper KK, Hyatt J, Heitzmann H. Comparison of three oxygen monitors in detecting endobronchial intubation. J Clin Monitoring 1988;4:240–243.
2. Barker SJ, Tremper KK. The effect of carbon monoxide inhalation on pulse oximetry and transcutaneous PO_2. Anesthesiology 1987;66:677–679.
3. Barker SJ, Tremper KK, Hyatt J. Effects of methemoglobinemia on pulse oximetry and mixed venous oximetry. Anesthesiology 1989;70:112–117.
4. Sidi A, et al. Effect of fluorescein, indocyanine green, and methylene blue on the measurement of oxygen saturation by pulse oximetry (abstract). Anesthesiology 1986;65(3A): A132.
5. Scheller MS, Unger RJ, Kelner MJ. Effects of intravenously administered dyes on pulse oximetry readings. Anesthesiology 1986;65:550–552.
6. Lawson D, et al. Blood flow limits and pulse oximeter signal detection. Anesthesiology 1987;67:599–603.
7. Narang VPS. Utility of the pulse oximeter during cardio--pulmonary resuscitation. Anesthesiology 1986;65:239–240.
8. Eisele JH, Downs D. Ambient light affects pulse oximeters. Anesthesiology 1987;67:864–865.
9. Barker SJ, Shah NK. The effects of motion on the performance of pulse oximeters in volunteers. Anesthesiology 1997;86:101–108.
10. Barker SJ. Motion Resistant pulse oximetry. A comparison of new and old models. Anesth Analg 2002;95:967–972.
11. Bohnhorst B, Peter C, Poets CF. Pulse oximeters' reliability in detecting hypoxia and bradycardia: Comparison between a conventional and two new generation oximeters. Crit Care Med 2000;28:1565–1568.
12. Kim JM, et al. Pulse oximetry and circulatory kinetics associated with pulse volume amplitude measured by photoelectric plethysmography. Anesth Analg 1986;65:133–139.
13. Barker SJ, et al. The effect of sensor malpositioning on pulse oximeter accuracy during hypoxemia. Anesthesiology 1993;79: 248–254.
14. Rithalia SVS, Bennett PJ, Tinker J. The performance characteristics of an intraarterial oxygen electrode. Intensive Care Med 1981;7:305–307.
15. Lubbers DW, Opitz N. Die pCO_2/pO_2-optode: eine neue pCO_2 bzw. pO_2-Messonde zur Messung des pCO_2 oder pO_2

von Gasen and Flussigkeiten. Z Naturforsch 1975;30:532–533.

16. Barker SJ, et al. Continuous fiberoptic arterial oxygen tension measurements in dogs. J Clin Monitoring 1987;39: 48–52.
17. Barker SJ, et al. A clinical study of fiberoptic arterial oxygen tension. Crit Care Med 1987;15:403.
18. Barker SJ, Hyatt J. Continuous measurement of intraarterial pHa, $PaCO_2$, and PaO_2 in the operating room. Anesth Analg 1991;73:43–48.
19. Tremper KK, Waxman K, Shoemaker WC. Effects of hypoxia and shock on transcutaneous PO_2 values in dogs. Crit Care Med 1979;7:52.
20. Gettinger A, DeTraglia MC, Glass DD. *In vivo* comparison of two mixed venous saturation catheters. Anesthesiology 1987;66:373–375.

See also BLOOD GAS MEASUREMENTS; SAFETY PROGRAM, HOSPITAL.

OXYGEN TOXICITY. See HYPERBARIC OXYGENATION.

PACEMAKERS

Alan Murray
Newcastle University
Medical Physics
Newcastle upon Tyne,
United Kingdom

INTRODUCTION

The primary function of a cardiac pacemaker is for the treatment of bradyarrhythmias, when the heart beat stops or responds too slowly. The clinical condition can be intermittent or permanent. If permanent, the pacemaker will control the heart continuously. If temporary, the pacemaker will respond only when necessary, avoiding competition with the heart's own natural response. As these devices are battery-powered, allowing the pacemaker to pace only when necessary also conserves pacemaker energy, extending its lifetime and reducing the frequency of replacement surgery.

Since their clinical introduction in the late 1950s and early 1960s, pacemakers have significantly improved the ability of many patients to lead normal lives. They also save lives by preventing the heart from suddenly stopping. The small size and long life of pacemakers allow patients to forget that they have one implanted in their chest. The first pacemakers were simple devices designed primarily to keep patients alive. Modern pacemakers respond to patients' needs and can regulate pacing function to enable the heart to optimize cardiac output and blood flow.

CLINICAL USE OF PACEMAKERS

The clinical problem with bradyarrhythmias is often associated with sick sinus syndrome. The heart's own natural pacing function originates from the sinus node in the right atrium. The rate of impulse formation at the sinus node is controlled by nerves feeding the node. Impulses arriving via the vagal nerve act to slow the heart down, as part of the body's parasympathetic response. When a higher heart rate is needed, the vagal nerve impulse rate to the sinus node is slowed down and the sinus node impulse rate increases.

Impulses propagating through the heart tissue are created via a series of action potential changes. Action potential changes can be triggered either naturally from one of the heart's pacing cells or by contact with a neighboring cell, causing the outside of the cell to produce a negative voltage with respect to the inside of the cell. These voltage changes are in the order of only 90 mV, but as we will see, an external voltage from a pacemaker has often to be several volts before depolarization is initiated.

Without pacemaker control, patients with bradyarrhythmias suffer from dizziness and can collapse without warning and, hence, risk injuring themselves. Heart pauses of the order of 10 s will cause unconsciousness. Most patients who collapse will recover their normal heart rhythm. They can then subsequently be examined clinically, and, if necessary, a pacemaker can be implanted to prevent recurrence of a further collapse. Sometimes the heart will stop and not recovers its normal pumping function and the patient will die, but usually there will have been preceding warning events allowing a pacemaker to be fitted to prevent death.

Many good texts exist that explain the clinical background to cardiac pacing, and these texts should be consulted (1–3). This text is primarily a description of the medical device itself.

Practical cardiac pacing started in the 1950s with the first clinical device, which was external to the body and required connection to a main power supply. This device was followed by an implantable pacemaker developed by Elmquist and surgically implanted by Senning in Sweden (4). The device only lasted a short time before failing, but it did show the potential for implanted pacemakers. This work was followed by Greatbatch and Chardack in the United States (5,6), first in an animal and then in a patient two years later. An interesting early review of this period has been given by Elmqvist (7). These first pacemakers were very simple devices and paced only at a fixed rate, taking no account of the heart's natural rhythm. Although this approach was less than ideal, it did provide the necessary spur for both clinical expectations and technical and scientific developments by research bioengineers and industry.

The next major technical development allowed pacemakers to pace on demand, rather than only at a fixed rate. Other pacing functions developed, including pacemakers that could pace more than one heart chamber, and pacemakers that could change their response rate as a function of patient physiological requirements. Three- or four- chamber pacing was an extension of basic pacing. Pacing functions have also been included in implantable defibrillators. More complicated pacing algorithms have been developed for controlling tachyarrhythmias, including ventricular tachycardia and rhythms that can deteriorate to ventricular fibrillation.

With the evolution of smaller devices and leads, their use in pediatrics has grown, including for children with congenital heart problems. Devices as thin as 6 mm are available. Reduction in size has also aided the move from epicardial to endocardial fixation of the lead. When pacemakers are implanted in children, special consideration has to be given to the type of device as children are usually active, the lead length as children continue to grow, and lead fixation as future lead replacement must be considered.

No doubt exists that, with continuing experience, pacing techniques and pacemaker devices will continue to evolve.

PHYSIOLOGICAL FUNCTION

Understanding the physiological function of a pacemaker is more important than knowing the technical details of the pacemaker. The main functional characteristics are the ones that are important for the physician or cardiologist who will want to know how the device will operate when implanted in a patient. A series of clinical questions exist. The first question asks where you want the device to sense the heart's rhythm, in the ventricle, the most common location, in the atrium, or in both. The second question asks in which chamber or chambers you would like the pacemaker to pace. This location is usually the ventricle, but can be the atrium or both. The third question relates to how you want the device to work when it encounters natural heartbeats. It can either be inhibited, which is by far the most common approach, or it can be triggered to enhance the natural beat. It is also possible to switch off the device's ability to sense the heart's natural rhythm, but is rarely done as there could then be competition between the pacemaker output and the heart's natural rhythm.

The answers to these three questions provide the first three codes given to any pacemaker. This code is an international code developed by the Inter-Society Commission on Heart Disease (ICHD) (8). It was subsequently expanded to a five-code system by the North American Society of Pacing and Electrophysiology (NASPE) and the British Pacing and Electrophysiology Group (BPEG) (9,10). The first version of the NASPE/BPEG codes allowed for programmability and communication, but as they became universal functions, the latest version of the codes simplified the codes in the fourth and fifth letter positions for use with rate modulation and multisite pacing only. These codes are used throughout the world. A summary of the coding is given in Table 1.

It is useful to give a few examples to illustrate how the codes are used. VVI pacemakers, which are in common use, allow the pacemaker to sense natural heartbeats in the ventricle (V), and, if they are absent, to pace in the ventricle (V), ensuring that the pacemaker inhibits (I) its output if a natural beat is detected. DDD pacemakers can sense in both the atrium and the ventricle (D), and, if required, pace in the atrium, ventricle, or both (D), with inhibiting and triggering (D). With programming techniques, the pacemaker's mode can be changed after the device is implanted, and so a manufacturer may list a very large number of modes for some pacemakers.

The codes in Table 1 also show the fourth and fifth letters. The fourth tells the user if the device has an internal function for modulating its pacing rate, known as a rate responsive (R) mode. If no code is quoted in the fourth position, it can be assumed that the device is not rate responsive. The fifth letter is for multisite pacing and is used if at least two atrial pacing sites or two ventricular pacing sites exist.

TEMPORARY EXTERNAL PACEMAKERS

This review primarily concerns implantable pacemakers, but the role of temporary external pacemakers must not be forgotten. These devices provide essential support after some cardiac surgery and after some myocardial infarctions, allowing time for the recovery of the heart's own pacing function. The way these devices function is very similar to implantable pacemakers, but they are generally simpler and provide the clinician with access to controls such as for pacing rate and pacing voltage. The pacing leads do not have the tip features required for permanent fixation, and the connector to the temporary pacemaker is simpler. Also, the leads are bipolar with two electrode contacts.

CLINICAL IMPLANTATION

Briefly, pacemakers are implanted most commonly at one of three sites (Fig. 1). Implantation is undertaken by a surgeon or cardiologist using an aseptic technique. The pacemaker pulse generator and leads are delivered in sterile packages with clear use-by dates. Venous insertion of the lead allows it to be pushed through the right atrium and tricuspid valve, and into the right ventricle, where the electrode can be positioned in the apex where it is less likely to move or displace in comparison with other possible positions. If an atrial lead exists, it is positioned in the right atrium. Good contact with the atrial wall is harder to achieve, and active fixation such as with a screw contact can be used, in comparison with ventricular apex passive fixation.

Patients must be followed up at regular intervals to ensure that the device is working correctly, that its output pulse characteristics are appropriate, and that the end-of-life of the internal battery is estimated. This follow-up interval may be over a period of months initially, and then annually, with more frequent follow-up visits toward the end of the device's life.

Most countries have national registration schemes, which enables information on specific patients to be obtained if, for example, a patient develops a problem while away from home. If, however, this information is not available, the device type can be recognized by a unique

Table 1. International Pacemaker Codes

1	2	3	4	5
Chamber sensed	Chamber paced	Response	Rate modulation	Multisite pacing
O = none	O = none	O = none	O = none	O = none
A = atrium	A = atrium	I = inhibited	R = rate modulation	A = artium
V = ventricle	V = ventricle	T = triggered		V = ventricle
D = A+V	D = A+V			D = A+V

The following letters are used sequentially in 5 positions.

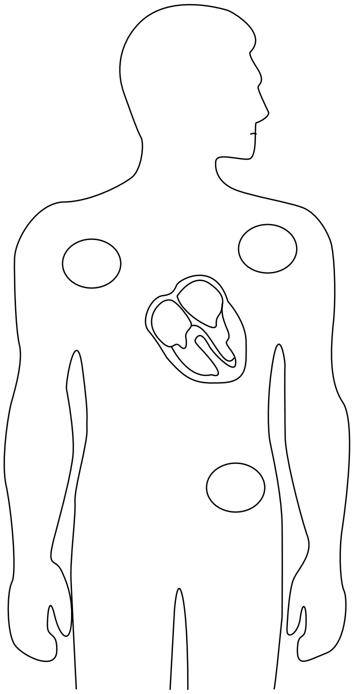


Figure 1. The location of normal pulse generator implantation sites.

radiopaque code that can be obtained by X ray. There has been discussion on whether such codes could be retrieved by a standard external interrogating device without the need for an X ray, but no such device is as yet universally available for all pacemakers.

Another goal of pacemaker registration is to provide useful data on the range of device lifetimes for each pacemaker type and information on sudden pacemaker failures, which enables clinical staff to plan any necessary replacement, and manufactures to act when it appears that failures are not random and may relate to their manufacturing process. This advance has enabled manufacturers to withdraw faulty or potentially faulty devices from the marketplace and correct production faults.

MARKET

Pacemakers have made a remarkable impact on clinical medicine. Over half a million new patients worldwide receive a pacemaker each year (11). In addition, approximately 100,000 patients worldwide receive a replacement pacemaker (11). Most implants are in the United States. When implants are related to the population size, the countries with the greatest new implant rates are Germany, the United States, and Belgium, with between approximately 700 and 800 implants per million population. Many countries with poor economies have very low implant rates. Within Europe, the implant rates are generally high, with, for example, the United Kingdom falling towards the bottom end of the implant rate, at approximately 300 per million population (12), where there are approximately 25,000 implants per year, and, of these, 75% are for new implants and 25% for replacements (12).

Table 2. Example Ranges of Pulse Generator Features

Volume	6–20 ml
Length/Width	30–60 mm
Depth	6–14 mm
Mass	13–50 g
Battery	0.8–2 Ah
Life	5–14 years
Sense threshold	0.1–15 mV
Refractory period	100–800 ms
Lower pulse rate	approximately 20/min
Upper pulse rate	approximately 185/min
Pulse amplitude	0–10 V
Pulse width	0.1–2 ms

FEATURES

Clinically, the most important pacemaker features relate to the device code, discussed above. Next in importance for both the clinician and patient is likely to be pacemaker size and lifetime. As a guide, example ranges of pacemaker features are included in Table 2. With continuous developments, these should be taken only as a guide. The shape and size of some pulse generators are shown in Fig. 2.

For a health-care system, the cost of devices is important. Costs vary significantly in different countries and also relate to the numbers purchased, so no specific figures can be given. It is, however, interesting to note the current relative costs of different types of devices. For guidance, approximate costs relative to a standard ventricular demand pacemaker type VVI are shown in Fig. 3, where the cost of the VVI pulse generator is given as unity. As the proportion of different types used will change, so will the relative costs.

The use of the unique radiopaque code for each pulse generator type is useful when a patient is referred to a medical center away from home, allowing that center to determine the pacemaker pulse generator being used.

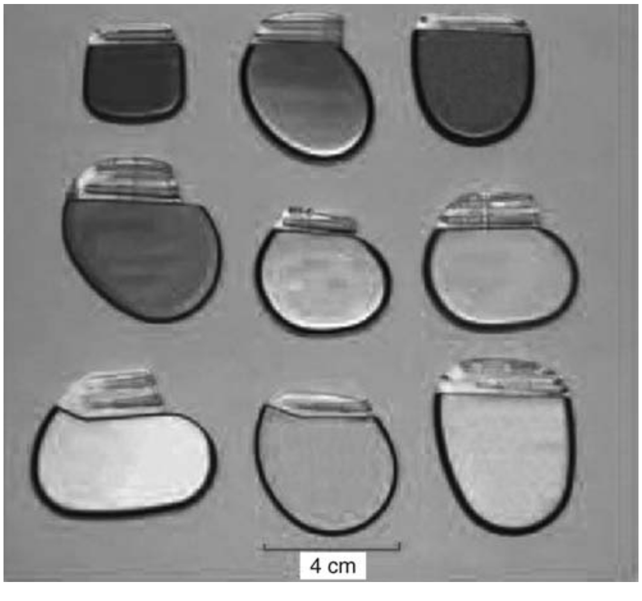


Figure 2. Illustration of some pacemaker shapes and sizes.

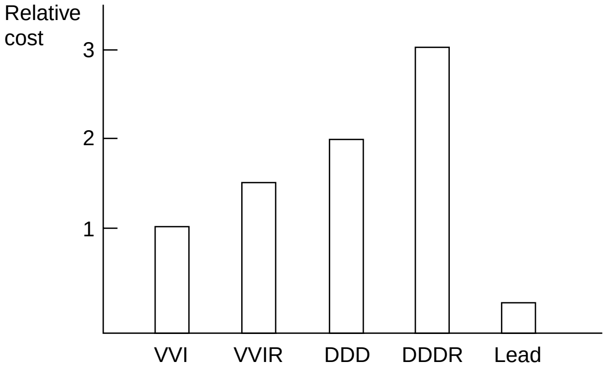


Figure 3. Relative approximate costs of pulse generators and pacing leads. The demand pacemaker (VVI) is taken as the reference.

A unipolar device has only one electrode directly in contact with the heart. In this case, the electrode is at the distal end of the lead. To complete the pacing circuit, another electrode contact is required, and this is on the pacemaker case with current flowing via the muscle in contact with the case electrode to the heart. Bipolar electrodes are also common. Here, both electrode contacts are on the lead, one at the tip and another several centimeters away. The second electrode makes contact with the ventricular wall simply by lying against the wall with the tip firmly located at the apex of the ventricle.

PACEMAKER COMPONENTS

A pacemaker refers to all components necessary for a complete clinical pacing device. At least two components will always exist, the pulse generator and the lead (Fig. 4). More than one lead any exist, such as for dual-chamber pacing, in both the atrium and ventricle. Unusual pulse generator and lead combinations may require an adaptor, but extra components should be avoided whenever possible. Extra components add to the areas where failure might occur.

When a pacemaker has been implanted or is, subsequently, to be checked, external devices will be required. An ECG recorder will confirm correct pacing, and some pacemakers generate impulses that can be visualized on an ECG recorder for relaying pacing information. In addition, a magnet or other technique may be used to put the pacemaker into various test modes, which is now commonly achieved with a programmer that communicates with the pacemaker via an electromagnetic wand using communication technology with coded sequences to prevent external interference such as from a radio telephone accidentally reprogramming the pacemaker. As well as controlling the pacing functions, the programmer can interrogate the pacemaker about the frequency of pacing, provided, of course, these features are available. Telemetry may also be available, where intracardiac waveforms can be relayed as they occur and selected pacing episodes can be recovered from the pacemaker memory.

Further useful technical information can be obtained from the books by Schaldach (13) and Webster (14).

PULSE GENERATOR

When pulse generators were first used in the early 1960s, they were simple devices with a battery power supply and a circuit to produce a regular pulse rate with a defined pulse voltage and pulse width output. Modern pulse generators are much more complex, with sensing and output control. Special electronic circuitry has been developed, often with sophisticated microprocessor control. Battery technology has also developed significantly. A block diagram of a complete pulse generator is shown in Fig. 5. Each major part is now described.

Power Supply

The first battery power supplies were made up from separate zinc-mercury cells. They could often be seen through the casing before implantation, after removal, or on the X ray. The voltage of these cells quickly fell to approximately 1.35 V, which was held until the cell reached the end of its life. Unfortunately, these cells could power the early pacemakers for only about two years.

These batteries encouraged the search for other power sources, including, for a short time, nuclear power, but safety concerns discouraged these developments. Rechargeable batteries, where the recharging energy was transmitted to the pulse generator via an external coil, were also employed, and had in fact been used in the first clinical pacemaker. However, reliability and frequency of charging inhibited their use.

Fortunately, a solution was found in the form of lithium-iodide cells. Their use in pacemakers was pioneered by Greatbatch (15) and introduced into clinical use in 1971. The cell has an initial open circuit voltage of 2.8 V, which falls slowly with use until its end of life is approached, when the voltage fall is more rapid. Although other types of lithium cells have been researched, they have not replaced the lithium-iodide cell for pacemakers. Some pacemakers

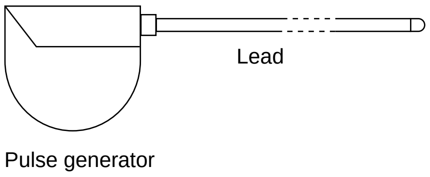


Figure 4. Pulse generator and lead are the two main components of a pacemaker.

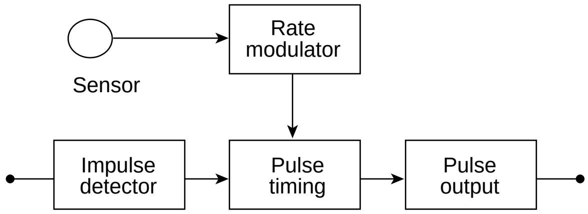


Figure 5. Block diagram of a complete pulse generator.

can now function for ten years, and, for low demand situations, 14 years can be achieved.

Ventricular and Atrial Sensing

All pacemakers sense natural cardiac impulses, which is achieved very reliably. However, the sensing circuitry needs to be able to differentiate between impulses coming from the heart and those due to external interference. The possibility of interference cannot be neglected, especially as the cardiac impulse is, at most, only a few tens of millivolts. In addition, in unipolar pacing, the sensing circuitry cannot differentiate between signals coming from the electrode at the end of the lead or the electrode on the pulse generator case, which is described more fully under leads below. It is not always possible to detect heart impulses reliably during periods of excessive interference, and in these cases, the device needs to be programmed to respond in a clearly defined way, such as by initiating fixed-rate pacing.

Pulse Interval Generator

At the heart of the pacemaker is the interval generator. As with other pulse generator functions, this is achieved with electronic circuitry or microprocessor control that generates the pacing interval measured in seconds or pacing rate measured in beats per minute. Without any means to detect patient activity or the need for a higher heart rate, this interval is fixed, other than for programmed changes that can be made at the clinic. The interval generator is connected to the sensing circuitry so that output from the interval generator can be synchronized or inhibited. Interactions for a VVI pulse generator are illustrated in Fig. 6. If a natural heartbeat is sensed, the pacemaker will time its pulse interval from that beat. If, however, the beat occurs in the refractory period set in the pacemaker, it will not respond to the beat, and the pulse interval timing will not be altered. For devices that can pace in both the atria and ventricle, such as a DDD device, the interval between atrial and ventricular pacing, known as the AV interval, will be set to optimize heart pumping function (Fig. 7).

Interval Modulator

The interval modulator, if it is available, simply attempts to change the response of the interval generator to that required by the patient. Electronic modulation of the interval is easy to achieve, and this function will usually be seen as part of the interval generator. However, determining how to modulate is difficult, and it requires a sensor to detect the patient's need for a higher or lower heart rate, usually in response to patient activity, whether determined directly or indirectly via physiological functions.

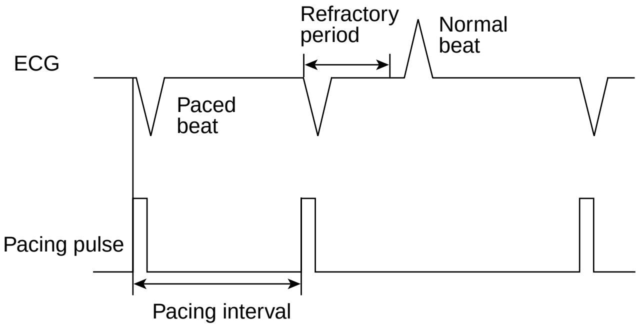


Figure 6. Illustrative example of some pacemaker-heart interactions.

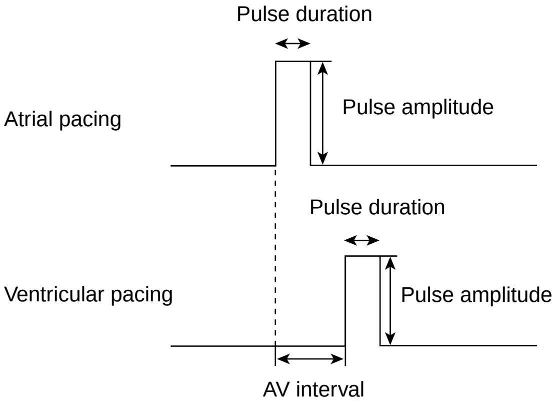


Figure 7. Atrioventricular pacing with the AV interval set within the pacemaker.

Activity Sensor

Many different sensors have been used. Some have been more successful than others. Changes in blood oxygen levels, through oxygen partial pressure or saturation levels, suggest the need for a higher heart rate when these levels fall, but this technique requires a sensor with continuous blood contact that will continue to work over many years, a difficult specification to achieve. Increased respiration rates also suggest the need for a higher rate. To detect respiration, some devices used a special lead to detect respiratory movement or electrical impedance changes but were not always successful. Special leads with inbuilt sensors add to the complexity of the pacemaker. Respiration can now be monitored using electrical impedance changes from a standard lead.

Other successful techniques used the intracardiac electrogram obtained from the sensing electrode or activity sensing from an inbuilt accelerometer. With careful analysis of the electrogram, it is possible to obtain a measure of the repolarization interval, which is known to change systematically with heart rate changes and, for this application, has been shown to shorten even without any increase in paced heart rate when the need for a higher rate is physiologically required. The use of a motion sensor using a piezoelectric sensor has been successful and is most widely applied because the technology is simple. The sensor can be built into the pulse generator housing with better reliability than for techniques requiring additional patient contacts. Movement of the sensor produces an output voltage proportional to acceleration. Algorithms have then been developed to relate changes in sensed activity to changes in pacing interval. They can also, to some extent, be programmed to individual patients. One drawback is that the sensor senses movement of the pacemaker that can be in a patient, for example, driving a car where an increase in heart rate may not be appropriate. Improvements to deal with such situations are under continuous development, as is research into different sensor techniques.

Without doubt, rate-responsive pacemakers have made a great contribution, and patients welcome the ability of the pacemaker to adapt to their needs, even if the pacemaker response is not physiologically perfect (12).

Lead Connector

Leads and pulse generators are provided separately, which gives greater flexibility in their use and enables different lead lengths and lead types to be selected. However, a connector is then needed. Ideally, this connector should be able to connect any appropriate lead to any pulse generator, and international standards have gone a long way to achieving this. For single-chamber pacemakers, the connector takes one lead that can be either for a unipolar or bipolar lead with one or two electrode contacts.

It is important for the connector to make a good electrical contact, while preventing any body fluids penetrating into the pulse generator, which in the past, have caused pacemakers to fail. Connectors allow the electrical contact to be made easily and provide seals between leads and the pulse generator.

Case

The primary function of the case is to protect the inner electronics from mechanical damage and from penetration of blood or other fluids. It is essential for the case to be biocompatible so that the patient does not attempt to reject the pulse generator as a foreign body. Titanium has been a successful material.

For unipolar devices, the case must contain an electrode, which acts as the reference for the lead electrode. The sensed voltage is that between the case and the lead electrode. Also, when pacing, the stimulation voltage appears between the case and lead electrode. To minimize the possibility of muscle stimulation at the case electrode, this electrode has a large surface area, so reducing the current density in comparison with that at the lead electrode. Also, the output pulse is positive at the case electrode with respect to the negative voltage at the pacing electrode, which confers preference to stimulation at the negative site in the heart.

Telemetry Function and Programming

For programming and telemetry functions, the pacemaker needs to be able to communicate with an external device, usually with coded signals using electromagnetic transmission between the pacemaker and a wand of the programming unit. These devices use standard techniques, with only the coding being specific to pacemakers.

Computer Algorithms

Pulse generators usually can be seen as small microprocessor devices. As such they contain computer code or computer algorithms to control their function, delivering advantages to the patient and clinician, as the pacemaker's mode of operation can be changed, and also to the manufacturer, as it is easier to develop new and improved devices. However, reliable software is notoriously difficult to develop and test. Manufacturers have discovered, to their cost-unusual errors in their software only after devices have been implanted, necessitating a recall of devices not yet implanted and careful follow-up of patients with devices already implanted. High quality software development is taken seriously by the manufacturers and cannot be stressed enough.

Table 3. Example Ranges of Lead Features

Length	20–120 cm
Diameter	1.2–3.5 mm
Tip diameter	0.7–3.3 mm

LEAD

The lead has four main features. It needs a connector to connect it to the pulse generator, a long flexible wire, a biocompatible sheath over the wire, and at least one electrode to make contact with the heart. Table 3 provides illustrative ranges of lead features.

Connector

The connector needs to be compatible with that on the pulse generator. As with the pulse generator, there should be no ingress of fluid, this time into the wire. Also, as the wire can move with each heartbeat, the construction needs to ensure that no extra stress exists on the lead wire near the connector.

Lead Wire

The most important characteristic of the lead wire is that is has to be flexible. Normal wire easily fractures when bent repeatedly. A pacing lead wire has to move with each heartbeat, which averages approximately 100,000 movements each day. Good flexibility is achieved by using a spiral construction. All wires have some impedance, which is taken into account by the pulse generator output.

Insulated Lead Sheath

The sheath covering the lead wire also needs to be flexible and must not become brittle with age. The sheath material must be biocompatible so as not to be rejected by the body. Materials used are silicone rubber or polyurethane.

Electrode

The design of the electrode is very important. In particular, the fixation, contact area, and contact material are essential features. Illustrative examples of basic features of lead-tip electrodes are shown in Fig. 8. Unipolar electrode leads have a single-electrode contact at the tip. Bipolar electrode leads have two contacts, one at the tip and the other a few centimeters distant from the tip.

When the lead is implanted and a suitable electrode site found, the electrode needs to stay in position, which is

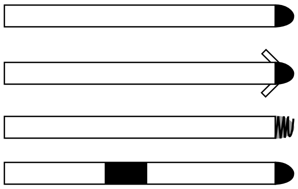


Figure 8. Illustrative example of pacing electrodes.

achieved by mechanical features at the tip of the lead, which can be, for example, tines or a helical electrode construction. With time, tissue will grow over the tip holding it in place. One problem is that this fixation can become so good that the lead can be difficult to remove if a problem occurs and it needs to be replaced. Electrodes positioned in the ventricular apex are easy to locate and also tend to stay in position easily, and hence require only passive fixation, such as with tines at the end of the lead. Other tip locations may require active fixation such as with a screw tip.

The electrode area needs to be high enough to ensure good electrical contact. The greater the contact area, the lower the contact impedance, which in turn reduces the electrode-tissue interface impedance and ensures that most of the pulse generator voltage appears at the cardiac tissue.

As with any electrode, the electrode contact material is important. The aim in selecting the material is to reduce polarization effects. Many electrode-coating materials have been studied, including steroid-eluting electrodes to reduce inflammation. Changes in electrode polarization are the cause of the increase in stimulation voltage in the days and early months after implantation, to be subsequently followed by a lowering of the effect and also of the required stimulation voltage.

STIMULATION THRESHOLDS

Stimulation success is a function of both pulse amplitude and pulse width. A minimum voltage and energy is required. The voltage has to be greater than that required to initiate the approximate 90 mV change in action potential. However, because of polarization and other effects, the voltage required is usually in the order of several volts and can reach 10 V soon after implantation. After a few months, this voltage will have reduced to the level of a few volts.

The initial research on stimulation pulse energy was with stimulating nerves, but the results obtained have been shown generally to hold when stimulating or pacing cardiac tissue. The energy used should be the minimum possible to induce stimulation reliably, which is controlled by varying the pulse width. Early work on nerve stimulation showed that no matter how wide the pulse width was, a minimum pulse voltage existed, called the **rheobase** voltage. At about twice this voltage, with a lower pulse width, the minimum energy required is found. If the pulse width is reduced further, the greater voltage required results in increased energy requirements to induce pacing. The pulse width for lowest energy is called the chronaxie time, shown in Fig. 9.

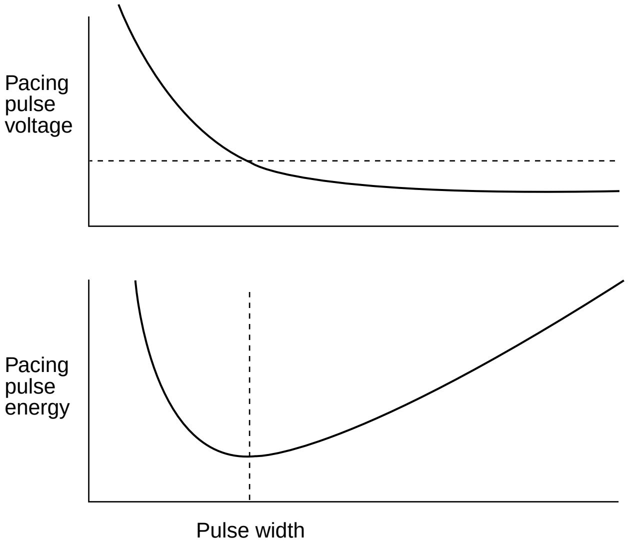


Figure 9. Pulse energy as a function of pulse width.

As the pulse energy is only to initiate cardiac depolarization, and not for providing pumping energy, the energy levels required are low and on the order of only a few microjoules (16).

PROBLEMS IN USE

Interference

Interference is a well-understood problem. Muscle interference at the case electrode in a unipolar system can be a problem in active people, especially when the pulse generator is implanted beside the pectoral muscles. If this particular problem is anticipated, a bipolar system can be used.

Threshold Voltage Changes

Threshold voltages do change. If the pacing voltage is set too high, energy will be wasted, reducing the lifetime of the device. If set too low, changes in threshold voltage may result in the pacing voltage being below the pacing threshold. These factors have to be balanced. Some devices automate the selection of an appropriate voltage output.

Early Failure

Pacemakers are complex devices, and like all devices they can fail. Failure is not a common problem, but because pacemakers are implanted and are life supporting, failure can have fatal consequences. Reporting of individual problems is essential, allowing manufacturers and national health bodies to identify a common problem early and, if necessary, withdraw stocks before they are used in new patients and take action to review patients who already have the device implanted.

SPECIAL DEVICES

This review has concentrated on the main use of pacemakers for treatment of bradyarrhythmias. Other options are used, but these options are based on the standard pacing approaches.

Implantable defibrillators can have an additional pacing function so that if the heart stops rather than developing ventricular fibrillation, pacing can be initiated. The pacing technology is exactly the same as for pacemakers described above, except that the electrode system will be different.

Devices are used for control of tachyarrhythmias. These devices, rather than using the regular pacing interval, usually use a series of pacing intervals at different rates to terminate the arrhythmia.

Some patients in heart failure were, in the past, often assumed to be untreatable unless by heart transplantation. Much can now be done for these patients, including pacing in all four cardiac chambers, which maximizes the pumping function of the heart by pacing the left as well as the right heart chambers, and pacing the atria and ventricles with an appropriate atrio-ventricular delay. This solution requires complex and multiple leads, and as these leads are used in sick patients, success is not always assured. Many of these patients may also require a defibrillation function.

FUTURE

Cardiac pacing had a small beginning but has grown at a steady rate each decade. With an aging population, the need for pacing will continue to grow. The development and production of pacemakers will remain a major medical device industry.

Of those devices currently available, increased use of physiological or rate responsive devices is likely as clinical studies prove their clinical value to patients, especially those who are active.

Technical advances will, to some extent, be dependent on the production of improved batteries, and then the decision will be either to make them smaller, last longer, or power more microprocessor technology. Improved electrode design to reduce energy requirements could also make a significant impact in reducing pacing pulse energy, and hence overall energy requirements. Improvements in setting the optimum AV delay will help many patients and, in particular, children who are active. Increased ability to store intracardiac data for review will ensure more research into effective use of pacing.

Pacing will continue as an essential therapeutic technique, saving lives and bringing some normality to patients with abnormal physiological heart rate control.

BIBLIOGRAPHY

1. Trohman RG, Kim MH, Pinski SL. Cardiac pacing: The state of the art. Lancet 2004;364:1701–1719.
2. ACC/AHA/NASPE 2002 guideline update for implementation of cardiac pacemakers and antiarrhythmia devices. American College of Cardiology and the American Heart Association; 2002.
3. Gold MR. Permanent pacing: New indications. Heart 2001;86: 355–360.
4. Senning A. Problems in the use of pacemakers. J Cardiovasc Surg 1964;5:651–656.
5. Greatbatch W, Chardack W. A transistorized implantable pacemaker for the long-term correction of complete heart block. Trans Northeast Electron Res Eng Meet Conf 1959; 1:8.
6. Chardack WM, Gage AA, Greatbatch W. A transistorized, self-contained, implantable pacemaker for the long-term correction of complete heart block. Surgery 1960;48:643–654.
7. Elmqvist R. Review of early pacemaker development. PACE 1978;1:535–536.
8. Parsonnet V, Furman S, Smyth NP. Implantable cardiac pacemakers: Status report and resource guideline (ICHD). Circulation 1974;50:A21–35.
9. Bernstein AD, Camm AJ, Fletcher RD, Gold RD, Rickards AF, Smyth NPD, Spielman SR, Sutton R. The NASPE/BPEG generic pacemaker code for antibradyarrhythmia and adaptive-rate pacing and antiachyarrhythmia devices. PACE 1987;10:794–799.
10. Bernstein AD, Daubert J-C, Fletcher RD, Hayes DL, Luderitz B, Reynolds DW, Schoenfeld MH, Sutton R. The revised NASPE/BPEG generic code for antibradycardia, adaptive-rate, and multisite pacing. PACE 2002;25:260–264.
11. Mond HG, Irwin M, Morillo C, Ector H. The world survey of cardiac pacing and cardiovertor defibrillators: Calendar year 2001. PACE 2004;27:955–964.
12. National Institute of Clinical Excellence. Technology Appraisal 88, Dual-chamber pacemakers for symptomatic bradycardia due to sick sinus syndrome and/or artrioventricular block. London, UK: National Institute of Clinical Excellence; 2005.
13. Schaldach M. Electrophysiology of the Heart: Technical Aspects of Cardiac Pacing. Berlin: Springer-Verlag; 1992.
14. Webster JG, ed. Design of Cardiac Pacemakers. Piscataway, NJ: IEEE Press; 1995.
15. Greatbatch W, Lee J, Mathias W, Eldridge M, Moser J, Schneider A. The solid state lithium battery. IEEE Trans Biomed Eng 1971;18:317–323.
16. Hill WE, Murray A, Bourke JP, Howell L, Gold R-G. Minimum energy for cardiac pacing. Clin Phys Physiol Meas 1988;9: 41–46.

See also AMBULATORY MONITORING; BIOELECTRODES; BIOTELEMETRY; DEFIBRILLATORS; MICROPOWER FOR MEDICAL APPLICATIONS.

PAIN CONTROL BY ELECTROSTIMULATION. See TRANSCUTANEOUS ELECTRICAL NERVE STIMULATION (TENS).

PAIN SYNDROMES. See BIOFEEDBACK.

PANCREAS, ARTIFICIAL

ROMAN HOVORKA
University of Cambridge
Cambridge, United Kingdom

INTRODUCTION

In 2000, some 171 million people worldwide had diabetes. By 2030, a conservative forecast suggests that this number will increase to 366 million attaining epidemic proportions as the prevalence increases from 2.8 to 4.4% in all age groups (1) due to, primarily, a relative increase in developing countries (2).

Diabetes is a group of heterogeneous chronic disorders characterized by hyperglycemia due to relative or absolute insulin deficiency. Two major categories of diabetes are recognized according to aetiology and clinical presentation, type 1 diabetes and type 2 diabetes. More than 90% cases are accounted for by type 2 diabetes. Regional and ethnic differences in diabetes incidence and prevalence exist.

Type 1 diabetes is one of the most common chronic childhood disease in developed nations (3), but occurs at all ages. Type 1 diabetes is caused by autoimmune destruction of pancreatic islet beta-cells resulting in the absolute loss of insulin production. Treatment demands the administration of exogenous insulin. Type 1 diabetes is associated with a high rate of complications normally occurring at young ages placing a considerable burden on the individual and the society.

Type 2 diabetes is caused by insulin resistance and relative insulin deficiency, both of which are normally present at the diagnosis of the disease. Environmental and polygenic factors contribute to these abnormalities (4), but specific reasons for their development are not known. A considerable number of subjects with type 2 diabetes progresses to insulin dependency.

The persistent hyperglycemia in diabetes is associated with long-term complications and dysfunction of various organs, especially the eyes, kidneys, nerves, heart, and blood vessels. The Diabetes Control and Complications Trial (DCCT) (5) and the United Kingdom Prospective Diabetes Study (UKPDS) (6) demonstrated that tight glycaemic control reduces the risk of long-term complications of type 1 and type 2 diabetes reducing the cost to the healthcare system (7). There is no threshold for the relationship between blood glucose, that is, glycosylated hemoglobin (HbA_{1C}) and reduced risk. This indicates that glucose levels in subjects with type 1 or 2 diabetes should be as close as possible to those observed in healthy subjects. However, tight glucose control is associated with an increased risk of hypoglycemia (8), which acts as a limiting factor to the effective diabetes management.

In health, insulin is secreted by the pancreas in a highly controlled fashion to maintain the plasma glucose concentration within a narrow physiological range. In type 1 diabetes, insulin is delivered exogenously to mimic the basal and postprandial insulin needs. The standard therapy is based on multiple insulin injections using a combination of short- and long-acting insulin analogs supported by blood glucose self-monitoring (9). Treatment by the continuous subcutaneous insulin infusion (CSII), that is, using insulin pumps, is on the rise (10).

The present review is limited to the artificial electromechanical endocrine pancreas, which has the potential to revolutionize diabetes management. The artificial pancreas replaces the glucose sensing and insulin delivery by beta-cells and is therefore sometimes called an "artificial beta-cell". It consists of three components, a glucose monitor to measure continuously glucose concentration, a titrating algorithm to compute the amount of insulin to be delivered, and an insulin pump to deliver the insulin. Only few prototypes have been developed and tested in a controlled clinical environment. Further progress depends on advancements of continuous glucose monitoring (11).

HISTORICAL BACKGROUND

The hormone insulin was discovered by Banting and Best in 1921. The first patient was treated a year later in 1922.

The first reported continuous *ex vivo* glucose measurement in humans was by Weller *et al.* in 1960 (12). In 1964, Kadish (13) was first to use continuous real-time glucose monitoring in a subject with diabetes to close the loop. The system employed an intravenous (iv) infusion of insulin and glucose, which could be switched on or off, denoted as an "on–off system". At that time, no suitable computational means were available.

In 1974, two groups developed a true "artificial endocrine pancreas". Albisser *et al.* (14,15) in Toronto and Pfeiffer *et al.* (16) in Ulm combined continuous glucose monitors with algorithms implemented on a microcomputer to automate iv delivery of insulin and glucose. The first commercial device, the Biostator (17) (Life Science Instruments, Miles, Elkhart, IN) was put into production in 1977 thanks to the determination by Clemens. The golden age of the Biostator was between late 1970s and early 1980s. It is still used for research purposes.

The last two decades have witnessed a considerable technological progress. Between 1999 and 2005, five continuous or semicontinuous monitors have received regulatory approval in the United States or Europe and further are under development (18). Since the introduction of continuous subcutaneous insulin infusion (CSII) (19), insulin pumps have been miniaturized and their reliability improved (20). Advanced titrating algorithms have been developed.

PHYSIOLOGICAL CONSIDERATIONS

Pancreas

The pancreas has digestive and hormonal functions. It is divided into the endocrine tissues secreting hormones insulin, glucagons, and somatostatin, and the exocrine tissues secreting digestive enzymes.

The endocrine tissues consist of many small clusters of cells called islets of Langerhans. Humans have roughly 1 million islets. Three major cell types are located in islets. Alpha-cells secrete the hormone glucagon. Beta-cells produce insulin and are the most abundant of the islet cells. Delta-cells secrete the hormone somatostatin.

Endogenous Insulin Secretion

Pancreatic beta-cells secrete insulin by sensing the levels of nutrients, such as glucose and certain amino acids. The beta-cells therefore integrate the sensing and secreting functions and are efficient in maintaining glucose homeostasis.

Four different phases of insulin secretion can be identified (21). Basal insulin secretion represents insulin released in the postabsorptive state. The cephalic phase of insulin secretion is evoked by the sight, smell, and taste of food before its digestion or absorption and is mediated by pancreatic innervations. The early-phase relates to the first 0–30 min postmeal insulin secretion, whereas the late-phase relates to the secretion between 60 and 120 min. During all phases, insulin is secreted in a pulsatile fashion with rapid pulses occurring every 8–15 min superimposed on slower, ultradian oscillations occurring every 80–120 min.

Insulin secretion is induced by other energetic substrates besides glucose, such as amino acids and drugs. Incretin hormones, such as glucagon-like peptide-1 (GLP–1) and to a lesser extent, glucose-dependent insulinotropic polypeptide (GIP), are responsible, in part, for the higher insulin secretory response after oral compared to the intravenous glucose administration.

COMPONENTS

The artificial pancreas consists of three components, a glucose monitor to measure glucose concentration, an algorithm to decide the amount of insulin to be delivered, and a device delivering insulin. This is a minimum setup. Some argue that a safe system should include a device for the delivery of glucose but all existing prototypes, with the exception of the Biostator, avoid the delivery of glucose.

The glucose monitor could be an implantable or extracorporeal device and based on a minimally or noninvasive technology (22). Generally, the implantable sensors are projected to have several months to years lifetime whereas the nonimplantable devices have, at present, lifetime of one-half of a day to several days.

Similarly, the insulin pump can be implanted or extracorporeal. The implantable pump normally delivers insulin intraperitoneally whereas the extracorporeal insulin pump delivers insulin subcutaneously.

The control algorithm can be implemented on a separate device or on the same platform as the insulin pump. The communication between the devices can be achieved using wire or wireless technologies. The latter are becoming prevalent for the transfer of data from insulin pumps onto diabetes management systems. Integrated systems exist which allow wireless transfer of data between glucose meters and insulin pumps such as the “all-in-one” CozMore Insulin Technology System (Smiths Medical MD, Inc. MN).

TYPES OF ARTIFICIAL PANCREAS

Meal Time Insulin Delivery

Artificial pancreas can handle meal delivery in different ways. In a “fully closed-loop” setting, the artificial pancreas delivers insulin without information about the time or size of the meal. Insulin is administered purely by evaluating the glucose excursions and the system works autonomously.

Alternatively, the artificial pancreas is provided with information about the time and size of the meal. The controller generates an advice, in an open-loop manner, on prandial insulin bolus. This can be termed “closed–loop with meal announcement” or “semiclosed-loop” control.

Other ways exist to handle the meal-related insulin delivery, but most systems adopt a fully closed-loop or semiclosed-loop setting.

Body Interface

Depending on body interface, three major types of artificial pancreas are recognized, (i) the subcutaneous (sc) sensing and sc delivery (sc–sc) system, (ii) the iv sensing and intraperitoneal (ip) delivery (iv–ip) system, and (iii) the iv glucose sensing and iv insulin delivery (iv–iv) system. The approaches differ in their invasiveness and associated kinetic delays (11).

Subcutaneous: Subcutaneous Body Interface

As a minimally invasive solution, the sc–sc approach has the potential to achieve a widespread application. However, it is unlikely to be compatible with a fully closed-loop system due to considerable delays disallowing effective compensation of large disturbances, such as meals.

The overall delay from the time of insulin delivery to the peak of its detectable glucose lowering effect is 100 min (11). This consists of a 50 min delay due to insulin absorption with short-acting insulin analogs (23), 30 min and more due to insulin action (24), 10 min due to interstitial glucose kinetics (25), and 10–30 min due to the transport time for *ex vivo* based monitoring system, such as those based on the microdialysis technique (26).

It is likely that users of the sc–sc approach will have to enter nutritional information to assist in the delivery of the prandial insulin dose. Most present prototypes adopt the sc–sc approach.

Intravenous: Intraperitoneal Body Interface

The iv–ip can benefit from existing intraperitoneal insulin pumps. The delays in the system are about 70 min, which comprises a 40 min time-to-peak of plasma insulin following intraperitoneal administration and a 30 min delay due to insulin action (11). Additionally, a delay due to kinetic properties of the glucose sensor applies, such as a 16 min kinetic and transport delay introduced by the long-term sensor system (27). It is unclear whether a fully closed-loop system can be developed under such circumstances.

The drawback of the iv–ip route is considerable invasiveness and relative inexperience with intraperitoneal compared to subcutaneous insulin pumps. Only > 1000 intraperitoneal pumps have been implanted so far (28) compared to $> 200{,}000$ subcutaneous pump users (29). Intraperitoneal insulin can be delivered by an implantable insulin pump Minimed 2007 (28) or via an indwelling intraperitoneal catheter such as DiaPort by Disetronic.

Intravenous: Intravenous Body Interface

The iv–iv approach was the first to have been investigated. It is embodied by the Biostator device. At present, the iv–iv approach is usable at special situations, such as in critically ill patients, surgical operations, or for research investigations. The drawback of the approach is its invasiveness requiring vascular access for both glucose monitoring and insulin delivery and is associated with a high risk of complications arising from, for example, biocompatibility issues.

The benefit of the approach is that the kinetic delays, ~ 30 min due to the delay in insulin action, are minimized enabling the development of a fully closed-loop system.

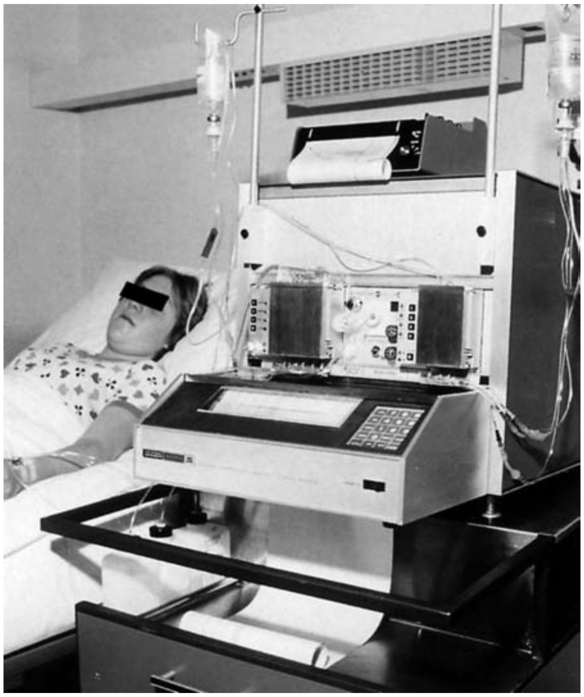


Figure 1. Biostator is the first commercial artificial endocrine pancreas. (Courtesy of Dr. Freckmann, Institute for Diabetes Technology, Ulm, Germany.)

PROTOTYPES

Biostator

Introduced in 1977, the Glucose-Controlled Insulin Infusion System (GCIIS), trademark name Biostator, is a modular, computerized, feedback control system for control of blood glucose concentrations (17), see Fig. 1. The Biostator is an example of an iv–iv system working in the fully closed-loop mode.

The Biostator was developed to normalise glucose in acute metabolic disturbances such as during diabetic ketoacidosis. However, its primary use has been in research investigating insulin sensitivity by the method of the glucose clamp and assessing insulin requirements and associated inter and intrasubject variability in subjects with type 1 diabetes and other conditions.

The rapid on-line glucose analyzer uses whole blood utilising a glucose oxidase sensor in the measurement process. The analyzer demonstrated both short- and long-range stability based on a two-point calibration.

The nonlinear proportional-derivative controller uses a five-point moving average smoothing and titrates insulin or dextrose intravenous infusion using a multichannel peristaltic infusion system to achieve user-defined glucose concentration. A printer records, on a minute-by-minute basis, the glucose value measured, the insulin and/or dextrose infusion rates, and the cumulative total of the insulin infused. A serial RS232 link allows these data to be downloaded to an external computer. The system response is < 90 s including transport of blood from the patient.

Although a pioneering device, the Biostator suffers from serious limitations. It needs constant technician's supervision. It discards continuously venous blood at a rate of 50 mL per 24 h. The control algorithm is oversimplistic. The original insulin titrating algorithm was linked to the rate of glucose change by Albisser *et al.* (14) with modifications, for example, by Botz (30), Marliss *et al.* (31), and Kraegen *et al.* (32) to reduce postprandial hyperglycemia and hyperinsulinemia. The algorithms require individualization by assigning values to constants. No formal adaptive approach was used to support the assignment, which is based on heuristics. These and similar algorithms were reviewed by Broekhuyse *et al.* (33), who concluded that none of the algorithms was superior and that further work was required to achieved normalization of the glucose concentration.

Over 200 devices have been sold worldwide. The Biostator contributed to the development and acceptance of the present gold standard in the diabetes management by multiple daily injections. At present, it is used for research purposes to evaluate diabetes drugs and technologies. The number of functioning prototypes counts in tens as spare parts run out. The Glucostator (mtb GmbH, Lonsee, Germany) is a CE-marked device recently marketed to replace the aging Biostator devices.

Shichiri's Group

Professor Shichiri and co-workers, Kumamoto, Japan, has developed as early as in 1975, a prototype of an iv–iv artificial endocrine pancreas (34) made later into a compact bedside version, STG-22 (Nikkiso Co. Ltd., Japan) (35) with a similar properties to the Biostator. The device is still marketed. STG-22 uses a glucose sensor for continuous glucose monitoring by combining the immobilised glucose oxidase membrane glucose enzyme sensor measuring hydrogen peroxide.

Following on, the group developed a prototype wearable artificial pancreas using the sc–sc route with the regular (36,37) and short acting insulin (38), and the sc–ip route (39). The latest versions use a microdialysis-type (40) or a ferrocene-mediated needle-type (39) glucose sensor working over a period of 7 days without any *in vivo* calibration (i.e., without using blood glucose measurement to calibrate the glucose sensor) followed by 14 days with one point calibration (41).

The results of the performance of their closed-loop system are even more impressive. With a fully closed sc–sc route using short acting insulin Lispro, the group claimed to have achieved "perfect" normalization of blood glucose over 24 h (38,42).

These results are surprising given that the control algorithm was a simple, nonadaptive PD controller in the form

$$\mathrm{IIR}(t) = K_{\mathrm{P}}G(t) + K_{\mathrm{D}}\frac{dG(t)}{dt} + K_{\mathrm{C}}$$

where $\mathrm{IIR}(t)$ is insulin infusion rate, $G(t)$ is the monitored glucose concentration, and K_{P}, K_{D}, and K_{C} are constants, which are dependent on the type of insulin delivery, subcutaneous versus intravenous, and also on the type of insulin, regular versus short-acting insulin lispro (38,43).

These enviable groundbreaking results, however, failed to be confirmed by other groups. The achievements of the group are summarized in an edited monograph (34).

Minimed: Medtronic

The Continuous Glucose Monitoring System (CGMS; Medtronic MiniMed, Northridge CA) (44) is the first commercial continuous glucose monitor. Approved in 1999, CGMS adopts a Holter-style monitoring to store up to 3-day data for retrospective analysis.

The CGMS employs an electrochemical sensor inserted into the subcutaneous tissue adopting the hydrogen peroxide-based enzyme electrode (45), which provides signal every 10 s. Calibration is achieved using self-monitoring of blood glucose. The new "gold" sensor introduced in November 2002 is more accurate than the original sensor [the mean absolute deviation 0.83 vs 1.11 $\text{mmol}\cdot\text{L}^{-1}$ (46)].

Employing the CGMS sensor, an external physiologic insulin delivery (ePID) has been developed by Minimed–Medronic. The system uses a PID controller (47), which was designed to reproduce the first phase insulin secretion by linking insulin administration to the rate of change in glucose concentration (the proportional component of the controller) and the second phase by linking insulin administration to the difference between the ambient and target glucose (the integrative component of the controller).

First studies with a fully closed loop were executed in dogs (48). The example presented in (49) shows peak postmeal glucose of 15 $\text{mmol}\cdot\text{L}^{-1}$ with the set point reached in 11 h indicating a suboptimal performance of a fully closed loop with the sc–sc approach. The adaptation of the PID controller was achieved by assigning the proportional gain K_P a value resulting in a normal daily insulin dose of the dog at euglycemia (48).

An evaluation of the ePID system in six subjects with type 1 diabetes > 27.5 h resulted in preprandial and postprandial (2 h) glucose levels at 5.8 ± 1.2 and 9.8 ± 1.6 $\text{mmol}\cdot\text{L}^{-1}$ (mean ± SD) (50). Morning glucose after overnight control was 6.8 ± 1.0 $\text{mmol}\cdot\text{L}^{-1}$.

Roche Diagnostics

The sc–sc closed-loop prototype with meal announcement (51,52) developed by Roche adopted the subcutaneous continuous glucose monitor (SCGM1; Roche Diagnostics GmbH, Manheim, Germany), which has been designed to monitor glucose in the subcutaneous interstitial fluid for up to 4–5 days (53).

SCGM1 is based on the microdialysis technique with an *ex vivo* glucose measurement. The sensor produces a signal every second. This is reduced to one glucose measurement every 5 min. Calibration is required once every 24 h (26,53). SCGM1 has a low flow rate ($0.3\ \mu\text{L}\cdot\text{min}^{-1}$), achieves a 100% recovery of the subcutaneous glucose in the dialysate, but has a 30 min technical lag. *In vitro* performance is excellent with a mean absolute difference of 0.2–3.8% in 10 sensor units (53).

An "empirical algorithm" (51) was develop to titrate sc insulin. A set of rules, derived from clinical observations, determine the insulin bolus administered every 10 min.

The closed-loop system with meal announcement was tested in 12 well-controlled ($\text{HbA}_{1\text{C}} < 8.5\%$) subjects with type 1 diabetes (51). Control lasted over 32 h and included the digestion of breakfast, lunch, dinner, and a snack. The target glucose concentration for the algorithm was 6.7 $\text{mmol}\cdot\text{L}^{-1}$. Prandial bolus was calculated from the carbohydrate content of the meal.

The algorithm achieved a near-target monitored glucose concentration (6.9 vs. 6.2 $\text{mmol}\cdot\text{L}^{-1}$; mean, algorithm vs. self-directed therapy) and reduced the number of hypoglycemia interventions from 3.2 to 1.1 per day per subject. During the algorithm therapy, 60% of SCGM1 values were within the 5–8.3 $\text{mmol}\cdot\text{L}^{-1}$ range compared to 45% with the self-directed therapy.

Adicol Project

The project Advanced Insulin Infusion using a Control Loop (Adicol) (54) was an EC funded project that completed at the end of 2002. The Adicol's sc–sc closed loop with meal announcement consisted of a minimally invasive subcutaneous glucose system, a handheld PocketPC computer, and an insulin pump (Disetronic D-Tron) delivering subcutaneously insulin lispro, see Fig. 2.

As continuous sensor was developed in parallel with the control algorithm and was not sufficiently stable, throughout the Adicol project, the intravenous glucose measurement was used, delayed by 30 min to simulate the lag associated with sc glucose sampling.

Adicol adopted an adaptive nonlinear model predictive controller (MPC) (55), which included a model based on a two compartment representation of glucose kinetics (24)

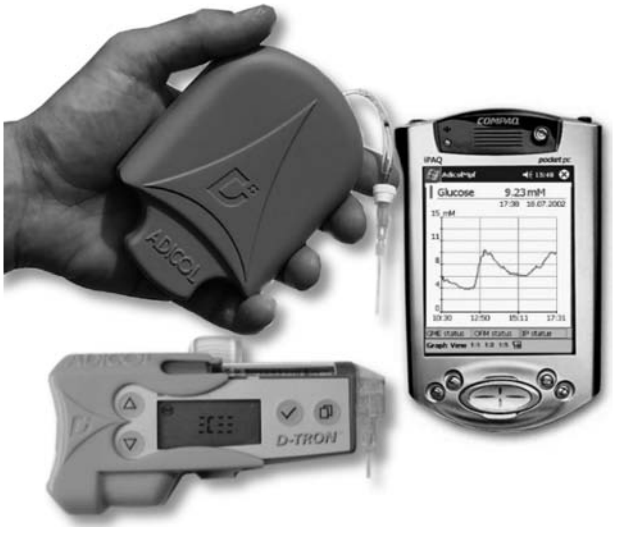


Figure 2. Components used by the Adicol's biomechanical artificial pancreas. Top left corner shows the microperfusion probe connected to the glucose monitor, which includes microfluidics components, Bluetooth communication, and the sensor. The handheld iPAQ PocketPC maintains wireless communication with the other two components, runs the MPC controller. Disetronic D-Tron insulin pump is equipment with a special sleeve visible on the left hand side of the pump which converts the Bluetooth radiofrequency signal to an infrared signal accepted by the pump (reprinted with permission from (54)).

extended by submodels representing the absorption of short acting insulin lispro, the insulin kinetics, the renal clearance of glucose, and the gut absorption. The MPC approach was combined with an adaptive Bayesian technique to individualize the glucoregulatory model to represent the inter- and intrasubject variability. The individualization was integrated within the control algorithm and was executed at each 15 min control cycle.

The largest clinical study performed in the Adicol project assessed the efficacy of the MPC controller with 30 min delayed glucose sampling > 26 h in 11 subjects with type 1 diabetes. Glucose was normalized from 1400 to 1800. Dinner followed with an individually determined prandial bolus at 1800, and control by the MPC from 1930 to 2200 the following day.

One hypoglycemia event (touchdown at 3.3 $\text{mmol}\cdot\text{L}^{-1}$) due to the MPC control was recorded. The highest glucose concentration was 13.3 $\text{mmol}\cdot\text{L}^{-1}$ following breakfast; 84% of glucose measurements were between 3.5 and 9.5 $\text{mmol}\cdot\text{L}^{-1}$ (56).

Following the completion of the Adicol project, a viscometric sensor (57) was tested with the MPC algorithm. Five subjects with type 1 diabetes treated by CSII were studied for 24 h (58). No hypoglycemia event (< 3.3 $\text{mmol}\cdot\text{L}^{-1}$) due to the MPC control was observed. Overall, 87% sensor values were between 3.5 and 9.5 $\text{mmol}\cdot\text{L}^{-1}$. Outside the 3 h postmeal periods, 74% of sensor measurements were in the range 3.5–7.5 $\text{mmol}\cdot\text{L}^{-1}$.

Institute for Diabetes Technology, Ulm

Building on foundations laid by Professor Pfeiffer in the early 1970s, the work in Ulm continues (59).

The group used the amperometric–enzymatic approach in combination with the microdialysis technique. The continuous flow method uses a slow continuous flow through the tubing achieving nearly a 100% recovery with a 30 min lag (59). The comparative method does not require calibration (60). Saline with glucose (5.5 $\text{mmol}\cdot\text{L}^{-1}$) is pumped through the probe in a stop-flow mode. During the stop mode, a nearly 100% equilibrium between the interstitial plasma glucose and the perfusate is achieved. In the flow mode, the dialysate is pumped rapidly to the sensor chamber. The technique facilitates sensor internal calibration for each measuring cycle and yields five glucose measurements per hour.

The group developed and tested an sc–sc closed-loop approach with meal announcement (61,62), see Fig. 3.

The algorithm uses the basal insulin need, determined from an individual insulin need, and a postprandial insulin need, expressed as an insulin/carbohydrate ratio. A model exploits these values and predicts future glucose excursions. The algorithm was tested in eight subjects with type 1 diabetes over a period of 24 h. The average glucose value was 7.8 ± 0.7 $\text{mmol}\cdot\text{L}^{-1}$ (mean ± SD). The postprandial increases were at 2.9 ± 1.3 $\text{mmol}\cdot\text{L}^{-1}$ with largest excursions recorded after breakfast. One hypoglycemia (< 3.3 $\text{mmol}\cdot\text{L}^{-1}$) was observed (62).

EVADIAC Group

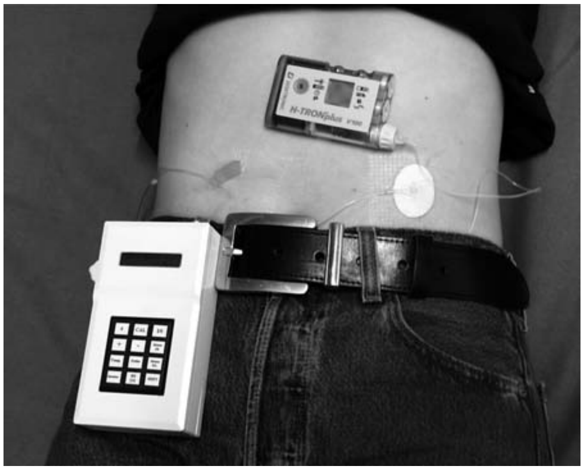


Figure 3. The system V4-IDT from the ULM group. The system uses a glucose monitor based on microdialysis integrated with a portable computer and an H-Tron pump, Disetronic. (Courtesy of Dr. Freckmann, Institute for Diabetes Technology, Ulm, Germany.)

Exploiting the progress made by the French group on implantable pumps "Evaluation dans le Diabete du Traitement par Implants Actifs" (EVADIAC), the work by Renard *et al.* is at the forefront of the fully closed-loop iv–ip approach. The group has developed the implantable physiologic insulin delivery (iPID) system, which uses a long-term sensor system (LTSS) (63,64).

Long-term sensor system, an intravenous enzymatic oxygen-based sensor developed by Medtronic MiniMed (Northridge CA), is implanted by direct jugular access in the superior vena cava. It is connected by a subcutaneous lead to an insulin pump delivering insulin intraperitoneally and implanted in the abdominal wall, see Fig. 4. The pump implements a PD controller similar to that used by the ePID system.

The system has been investigated in subjects with type 1 diabetes with collected data per sensor of ~ 280 days (65). Most investigations with LTSS have adopted the open-loop approach. The fully closed-loop system was tested > 48 h reducing % time spent at < 3.9 $\text{mmol}\cdot\text{L}^{-1}$ from 18 to 6%, and % time spent at > 13.3 $\text{mmol}\cdot\text{L}^{-1}$ from 17 to 2%. The addition of insulin bolus at meal time, all glucose values were inside the range 3.9–13.3 $\text{mmol}\cdot\text{L}^{-1}$.

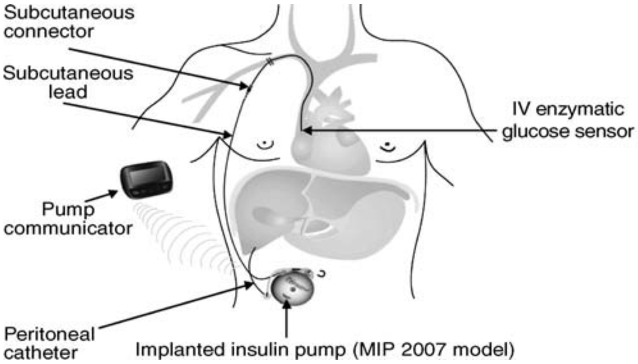


Figure 4. Scheme of human implantation of the Long-Term Sensor System (LTSS, Medtronic-MiniMed), a prototype of artificial pancreas. (Courtesy of Dr. Renard, Lapeyronie Hospital, Montpellier, France.)

Recently, the iPID system was evaluated in four elderly lean subjects with type 1 diabetes over 48 h (66). During the second 24 h control period following empirical tuning of the algorithm, 4 and 7% of time was spent $< 4.4\ \text{mmol}\cdot\text{L}^{-1}$ in the postprandial (0–2 h) and outside meal conditions, respectively, 12 and 32% was spent in the region 4.4–$6.7\ \text{mmol}\cdot\text{L}^{-1}$, 63 and 60% was spent in the region 6.7–$13.3\ \text{mmol}\cdot\text{L}^{-1}$, and 20 and 2% was spent $> 13.3\ \text{mmol}\cdot\text{L}^{-1}$.

CLINICAL STUDIES

Clinical studies performed with prototypes of an artificial pancreas in subjects with type 1 diabetes are summarized in Table 1. All experiments were performed in hospital environment. No prototype has yet been studied in home settings.

Table 1 excludes numerous experiments carried out with the Biostator as the invasiveness and the setup adopted by the device does not permit development into a routinely used system.

Numerous experiments have been carried out in pancreatectomized dogs especially in the early phases of prototype development. This includes the ePID system (48), but some approaches such as that adopted by the Adicol project, did not use testing on animal models, but adopted testing on a simulation environment (82).

INDICATIONS

Artificial pancreas has the potential to be used in various disease conditions. For subjects with type 1 diabetes, the system offers “cure” especially if implemented as a fully closed-loop iv–ip system. Realistically, first prototypes for home setting are expected to adopt the sc-sc route with meal announcement and thus participation of the subjects with type 1 diabetes in the disease management process is required. Fail-safe procedures are needed for such a setup.

An increasing proportion of subjects with type 2 diabetes is treated by insulin. It has been reported that nearly 50% of subjects with type 2 diabetes require insulin treatment at some stage of their disease. The artificial pancreas could provide a solution for a subgroup of subjects with type 2 diabetes, but this has to be justified by a cost-benefit analysis.

The sc–sc route may require a continuing close subject involvement in the disease management. This would restrict the treatment group to those well motivated, who generally have good glucose control. The greatest treatment benefit would be for those with poor glucose control, but other aspects, such as psychological factors might impair or even prevent the system deployment.

A recent study in adult critically ill subjects revealed that glucose control $< 6.1\ \text{mmol}\cdot\text{L}^{-1}$ is associated with reduced mortality by 43%, overall inhospital mortality by 34%, newly developed kidney failure requiring dialysis by 41%, bacteremia by 46%, the number of red blood cell transfusions by 50%, and critical illness polyneuropathy by 44% (83). Although still awaiting confirmation by another prospective study, the results indicate that artificial pancreas for critically ill is likely to bring about major improvements in therapeutic outcomes. Whether these results apply to a broader category of inpatients be it a general ward or pediatric population is yet to be determined. It is also unclear what is the most appropriate setup of a “hospital-based” artificial pancreas.

COST

The cost of the artificial pancreas can only be inferred from the cost of existing technology. The cost of the Biostator was ~ $70,000 prohibiting its wider use.

Insulin pumps cost from $5,000 to 6,000. The monthly cost of pump treatment is ~ $100. The CGMS monitor costs ~ $4,000 and the single-use 3 day sensors cost $50 each. It is likely that the sc-sc artificial pancreas will be at least as expensive as the combined cost of the insulin pump and the glucose monitor.

The cost needs to be set against the total cost of diabetes which, in developed countries such as in the United States or the United Kingdom, is 10% of the total health care budget (84). Most of the direct expenditure is on treating diabetes complications, which could be delayed or prevented with tight glucose control.

OUTLOOK

Present technology has made considerable advances toward a truly personal wearable treatment system. The lack of availability of a glucose monitor with adequate properties appears to hinder further progress and the development of a commercially viable system. The algorithms need to be improved and subjected to rigorous clinical testing.

With regard to the existing glucose monitors, it is possible that their potential has not been fully exploited. The regulatory-driven research and development to achieve an equivalence between finger-prick glucose measurements and values provided by continuous glucose monitors in the interstitial fluid hinders the engagement of the industry and the academia in the development and testing of closed-loop systems.

In the first instance, the artificial pancreas is most likely to find its wider use in the supervised hospital environment, such as at the intensive care units. The application in home settings will most likely be gradual starting with a supervised system (by the treated subjects) and increasing its autonomous function following on from wider experience.

The regulatory bodies will play an important role in the introduction of the artificial pancreas into the clinical practice. Until recently, the perception of closed-loop systems was not overly positive by the regulatory authorities. Artificial pancreas with its potential to “cure” diabetes, but also to lead to life-threatening complications, if malfunctioned, will have to pave the way to a new generation of closed-loop home-based biomedical devices if it is ever to succeed.

Table 1. Clinical Studies with sc–sc or iv–ip Closed-Loop Control in Subjects with Type 1 Diabetes[a]

N	Duration of Control, h	Sensing–Infusion	Sensor	Insulin	Algorithm	Control Interval, min	Type of Control	Performance	References
6	72	sc–sc	needle-type Ref. 37	regular	PD[b]	1[c]	f–cl[d]	M-value (67) 15 ± 4; mean glucose 6.1 ± 0.5 mmol·L^{-1}; Mage (68) 73 ± 14 mg·dL^{-1}	37
5	5	sc–sc	needle-type Ref. 69	regular	PD[b]	1[c]	f–cl	postprandial glucose (1.5 h) 10.6 ± 0.9 mmol·L^{-1}; late postprandial glucose (5 h) 2.8 ± 0.4 mmol·L^{-1};	38,70
5	5	sc–sc	needle-type Ref. 71	lispro	PD[b]	1[c]	f–cl	no hypoglycemia (< 2.8 mmol·L^{-1}); postprandial glucose (1 h) 8.5 ± 0.5 mmol·L^{-1};	38,72
5	24	sc–sc	needle-type Ref. 73	regular	PD[b]	1[c]	f–cl	hypoglycemia observed; glucose between peak (b'fast postprandial) 12.5 ± 1.0 mmol·L^{-1} and nadir (before dinner) 2.7 ± 0.3 mmol·L^{-1}	38,74
5	24	sc–sc	needle-type Ref. 75	lispro	PD[b]	1[c]	f–cl	no hypoglycemia; Near normal control	38,76
9	8	iv–sc	offline[d]	lispro	MPC	15	cl–ma[e]	[f]no hypoglycemia (< 3.3 mmol·L^{-1}); [h]6.1 ±0.6 mmol·L^{-1}	77
6	8	simulated[g] sc–sc	offline[d]	lispro	MPC	15	cl–ma	[f]no hypoglycemia (< 3.3 mmol·L^{-1}); [h]6.6 ± 0.8 mmol·L^{-1}	78,79
6	14	simulated[g] sc–sc	offline[d]	lispro	MPC	15	cl–ma	[f]no hypoglycemia (< 3.0 mmol·L^{-1}); preprandial 7.0 ± 1.1 mmol·L^{-1}; [i]6.3 ± 1.6 mmol·L^{-1}	80
11	26.5	simulated[g] sc–sc	offline[d]	lispro	MPC	15	cl–ma	1 hypoglycemia (< 3.3 mmol·L^{-1}); 84% glucose values between 3.5 and 9.5 mmol·L^{-1}	56
5	24	sc–sc	viscometric Ref. 57	lispro	MPC	15	cl–ma	no hypoglycemia (< 3.3. mmol·L^{-1}); 87% sensor values between 3.5 and 9.5 mmol·L^{-1}	58
6	27.5	sc–sc	CGMS/Guardi an Refs. 44,81	lispro	PID	1–5[c]	f–cl	hypoglycemia not reported; preprandial glucose 5.8 ± 1.2 mmol·L^{-1}; postprandial glucose (2 h) 9.8 ± 1.6 mmol·L^{-1}	50
4	48	iv–ip	LTSS Refs. 63,64	U400	PID	1–5[c]	f–cl	hypoglycemia not reported; 84% glucose values between 4.4 and 13.3 mmol·L^{-1}	66
12	32	sc–sc	SCGM1 Ref. 53	lispro	empirical	10	cl–ma	1.1 hypoglycemia per day per subject; 56% glucose values between 5.0 and 8.3 mmol·L^{-1}	51
12	7	sc–sc	comparative microdialysis Ref. 60	lispro	MPC	12	cl–ma	mean glucose 8.0 ± 2.3 mmol·L^{-1}; postprandial glucose increase 3.0 ± 1.6 mmol·L^{-1}	61
8	24	sc–sc	comparative microdialysis Ref. 60	lispro	MPC	12 (day) 36 (night)	cl–ma	one hypoglycemia (< 3.3 mmol·L^{-1}) mean glucose 7.8 ± 0.7 mmol·L^{-1}; postprandial glucose increase 2.9 ± 1.3 mmol·L^{-1}	59,62

[a] Adapted from (11).

[b] Proportional-derivative controller.

[c] Not reported, an estimate from plot(s).

[d] Beckman Glucose Analyzer 2.

[e] f–cl: fully closed–loop; cl–ma: closed-loop with meal announcement.

[f] Number of hypoglycemias due to the controller.

[g] IV glucose measurements delayed by 30 min.

[h] Mean ± SD in the last two hours of control.

[i] Over 4 hours following meal.

ACKNOWLEDGMENT

Support by Disetronic Medical Systems AG, Burgdorf, Switzerland, and the EU Clincip Project (IST-2002-506965) is acknowledged. Dr Malgorzata E Wilinska helped with the preparation of the background material.

BIBLIOGRAPHY

1. Wild S, et al. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. Diabetes Care 2004; 27:1047–1053.
2. Diabetes Atlas, International Diabetes Federation, 2003.
3. LaPorte R, Matsushima M, Chang Y. Prevalence and incidence of insulin-dependent diabetes. In NDDG, edition. Diabetes in America, NIH; 1995. p 37–46.
4. Bell GI, Polonsky KS. Diabetes mellitus and genetically programmed defects in beta-cell function. Nature (London) 2001; 414:788–791.
5. Diabetic Control and Complications Trial Research Group. The effect of intensive treatment of diabetes on the development and progression of long term complications in insulin-dependent diabetes mellitus. N Engl J Med 1993;329:977–986.
6. Turner RC, et al. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). Lancet 1998;352:837–853.
7. Gilmer TP, O'Connor PJ, Manning WG, Rush WA. The cost to health plans of poor glycemic control. Diabetes Care 1997;20: 1847–1853.
8. Cryer PE, Davis SN, Shamoon H. Hypoglycemia in diabetes. Diabetes Care 2003;26:1902–1912.
9. Standards of Medical Care in Diabetes. Diabetes Care 2005; 28:S4–S36.
10. Pickup J, Keen H. Continuous subcutaneous insulin infusion at 25 years: evidence base for the expanding use of insulin pump therapy in type 1 diabetes. Diabetes Care 2002;25:593–598.
11. Hovorka R. Continuous glucose monitoring and closed-loop systems. Diabetic Med 2005; (in press).
12. Weller C, et al. Continuous *In vivo* determination of blood glucose in human subjects. Ann N Y Acad Sci 1960;87:658–668.
13. Kadish AH. Automation control of blood sugar. I. A. servomechanism for glucose monitoring and control. Am J Med Electron 1964;39:82–86.
14. Albisser AM, et al. An artificial endocrine pancreas. Diabetes 1974;23:389–404.
15. Albisser AM, et al. Clinical control of diabetes by the artificial pancreas. Diabetes 1974;23:397–404.
16. Pfeiffer EF, Thum C, Clemens AH. The artificial beta cell - A continuous control of blood sugar by external regulation of insulin infusion (glucose controlled insulin infusion system). Horm Metab Res 1974;6:339–342.
17. Clemens AH, Chang PH, Myers RW. The development of Biostator, a Glucose Controlled Insulin Infusion System (GCIIS). Horm Metab Res 1977; (Suppl. 7):23–33.
18. Klonoff DC. Continuous glucose monitoring:roadmap for 21st century diabetes therapy. Diabetes Care 2005;28: 1231–1239.
19. Pickup JC, Keen H, Parsons JA, Alberti KG. Continuous subcutaneous insulin infusion: an approach to achieving normoglycaemia. Br Med J 1978;1:204–207.
20. Pickup J, Keen H. Continuous subcutaneous insulin infusion at 25 years: evidence base for the expanding use of insulin pump therapy in type 1 diabetes. Diabetes Care 2002;25:593–598.
21. Caumo A, Luzi L. First-phase insulin secretion: does it exist in real life? Considerations on shape and function. Am J Physiol Endocrinol Metab 2004;287:E371–E385.
22. Klonoff DC. Continuous glucose monitoring: roadmap for 21st century diabetes therapy. Diabetes Care 2005;28: 1231–1239.
23. Plank J, et al. A direct comparison of insulin aspart and insulin lispro in patients with type 1 diabetes. Diabetes Care 2002;25:2053–2057.
24. Hovorka R, et al. Partitioning glucose distribution/transport, disposal, and endogenous production during IVGTT. Am J Physiol 2002;282:E992–E1007.
25. Rebrin K, Steil GM, Van Antwerp WP, Mastrototaro JJ. Subcutaneous glucose predicts plasma glucose independent of insulin: implications for continuous monitoring. Am J Physiol 1999;277:E561–E571.
26. Heinemann L. Continuous glucose monitoring by means of the microdialysis technique: underlying fundamental aspects. Diabetes Technol Ther 2003;5:545–561.
27. Steil GM, Panteleon AE, Rebrin K. Closed-loop insulin delivery-the path to physiological glucose control. Adv Drug Deliv Rev 2004;56:125–144.
28. Selam JL. External and implantable insulin pumps: current place in the treatment of diabetes. Exp Clin Endocr Diab 2001;109:S333–S340.
29. Pickup J, Keen H. Continuous subcutaneous insulin infusion at 25 years: evidence base for the expanding use of insulin pump therapy in type 1 diabetes. Diabetes Care 2002;25:593–598.
30. Botz CK. An improved control algorithm for an artificial beta-cell. IEEE Trans Biomed Eng 1974;23:252–255.
31. Marliss EB, et al. Normalization of glycemia in diabetics during meals with insulin and glucagon delivery by the artificial pancreas. Diabetes 1977;26:663–672.
32. Kraegen EW, et al. Control of blood glucose in diabetics using an artificial pancreas. Aust N Z J Med 1977;7:280–286.
33. Broekhuyse HM, Nelson JD, Zinman B, Albisser AM. Comparison of algorithms for the closed-loop control of blood glucose using the artificial beta cell. IEEE Trans Biomed Eng 1981;28:678–687.
34. Shichiri M. Artificial Endocrine Pancreas: Development and Clinical Applications. Kumamoto: Kamome Press; 2000.
35. Goriya Y, Kawamori R, Shichiri M, Abe H. The development of an artificial beta cell system and its validation in depancreatized dogs: the physiological restoration of blood glucose homeostasis. Med Prog Technol 1979;6:99–108.
36. Shichiri M, et al. Wearable artificial endocrine pancrease with needle-type glucose sensor. Lancet 1982;2:1129–1131.
37. Kawamori R, Shichiri M. Wearable artificial endocrine pancreas with needle-type glucose sensor. In Nose Y, Kjellstrand C, Ivanovich P. editors. Progress in Artificial Organs, Cleveland: ISAO Press; 1986. pp. 647–652.
38. Shimoda S, et al. Closed-loop subcutaneous insulin infusion algorithm with a short-acting insulin analog for long-term clinical application of a wearable artificial endocrine pancreas. Front Med Biol Eng 1997;8:197–211.
39. Shimoda S, et al. Development of closed-loop intraperitoneal insulin infusion algorithm for a implantable artificial endocrine pancreas. Diabetes 2001;50(Suppl. 2):A3.
40. Hashiguchi Y, et al. Development of a miniaturized glucose monitoring system by combining a needle-type glucose sensor with microdialysis sampling method. Long-term subcutaneous tissue glucose monitoring in ambulatory diabetic patients. Diabetes Care 1994;17:387–396.

41. Shichiri M, Sakakida M, Nishida K, Shimoda S. Enhanced, simplified glucose sensors: long-term clinical application of wearable artificial endocrine pancreas. Artif Organs 1998;22:32–42.
42. Shichiri M, Sakakida M, Nishida K, Shimoda S. Enhanced, simplified glucose sensors: long-term clinical application of wearable artificial endocrine pancreas. Artif Organs 1998; 22:32–42.
43. Shichiri M, Sakakida M, Nishida K, Shimoda S. Enhanced, simplified glucose sensors: long-term clinical application of wearable artificial endocrine pancreas. Artif Organs 1998; 22:32–42.
44. Mastrototaro J. The MiniMed continuous glucose monitoring system (CGMS). J Pediatr Endocr Met 1999;12:751–758.
45. Johnson KW, et al. *In vivo* evaluation of an electroenzymatic glucose sensor implanted in subcutaneous tissue. Biosens Bioelectron 1992;7:709–714.
46. Accuracy of the GlucoWatch G2 Biographer and the continuous glucose monitoring system during hypoglycemia: experience of the Diabetes Research in Children Network. Diabetes Care 2004;27:722–726.
47. Steil GM, et al. Modeling beta-cell insulin secretion–implications for closed-loop glucose homeostasis. Diabetes Technol Ther 2003;5:953–964.
48. Steil GM, et al. Tuning closed-loop insulin delivery based on known daily insulin requirements. Diabetes 2002;51 (Suppl. 2):510.
49. Steil GM, Panteleon AE, Rebrin K. Closed-loop insulin delivery-the path to physiological glucose control. Adv Drug Deliv Rev 2004;56:125–144.
50. Steil GM, et al. Continuous automated insulin delivery based on subcutaneous glucose sensing and an external insulin pump. Diabetes 2004;53(Suppl. 2):A2.
51. Galley P, et al. Use of subcutaneous glucose measurements to drive real-time algorithm-directed insulin infusion recommendations. Diabetes Technol Ther 2004;6:245–246.
52. Galley PJ, Thukral A, Chittajallu SK, Weinert S. Diabetes management system. US Pat. 2003. 6,544,212:1–17.
53. Schoemaker M, et al. The SCGM1 System:subcutaneous continuous glucose monitoring based on microdialysis technique. Diabetes Technol Ther 2003;5:599–608.
54. Hovorka R, et al. Closing the loop: The Adicol experience. Diabetes Technol Ther 2004;6:307–318.
55. Hovorka R, et al. Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes. Physiol Meas 2004;25:905–920.
56. Schaller HC, et al. Avoidance of hypo- and hyperglycaemia with a control loop system in patients with Type 1 DM under daily life conditions. Diabetes Metab 2003;29:A2225.
57. Beyer U, et al. Recording of subcutaneous glucose dynamics by a viscometric affinity sensor. Diabetologia 2001;44:416–423.
58. Vering T. Minimally invasive control loop system for sc-sc control on patients with type 1 diabetes. Diabetes Technol Ther 2004;6:278.
59. Freckmann G, et al. Recent advances in continuous glucose monitoring. Exp Clin Endocr Diab 2001;109:S347–S357.
60. Hoss U, et al. A novel method for continuous online glucose monitoring in humans: the comparative microdialysis technique. Diabetes Technol Ther 2001;3:237–243.
61. Kalatz B, et al. Development of algorithms for feedback-controlled subcutaneous insulin infusion with insulin lispro. Acta Diabetol 1999;36:215.
62. Kalatz B. Algorithmen zur Glucosegesteuerten Insulininfusion bei Diabetes Mellitus-Entwicklung und Experimentelle Untersuchung, Ulm: Medical Dissertation. 1999.
63. Renard E, Costalat G, Bringer J. From external to implantable insulin pump, can we close the loop? Diabetes Metab 2002;28:S19–S25.
64. Renard E. Implantable closed-loop glucose-sensing and insulin delivery: the future for insulin pump therapy. Curr Opin Pharmacol 2002;2:708–716.
65. Renard E, et al. Sustained safety and accuracy of central IV glucose sensors connected to implanted insulin pumps and short-term closed-loop trials in diabetic patients. Diabetes 2003;52(Suppl. 2):A36.
66. Renard E, et al. Efficacy of closed loop control of blood glucose based on an implantable iv sensor and intraperitoneal pump. Diabetes 2004;53(Suppl. 2):A114.
67. Schlichtkrull J, Munck O, Jersild M. The M-value, an index of blood sugar control in diabetics. Acta Med Scand 1965;177:95–102.
68. Service FJ, et al. Mean amplitude of glycemic excursions, a measure of diabetic instability. Diabetes 1970;19:644–655.
69. Hashiguchi Y, et al. Development of a miniaturized glucose monitoring system by combining a needle-type glucose sensor with microdialysis sampling method. Long-term subcutaneous tissue glucose monitoring in ambulatory diabetic patients. Diabetes Care 1994;17:387–396.
70. Shichiri M, Sakakida M, Nishida K, Shimoda S. Enhanced, simplified glucose sensors: long-term clinical application of wearable artificial endocrine pancreas. Artif Organs 1998;22:32–42.
71. Hashiguchi Y, et al. Development of a miniaturized glucose monitoring system by combining a needle-type glucose sensor with microdialysis sampling method. Long-term subcutaneous tissue glucose monitoring in ambulatory diabetic patients. Diabetes Care 1994;17:387–396.
72. Shichiri M, Sakakida M, Nishida K, Shimoda S. Enhanced, simplified glucose sensors: long-term clinical application of wearable artificial endocrine pancreas. Artif Organs 1998; 22:32–42.
73. Hashiguchi Y, et al. Development of a miniaturized glucose monitoring system by combining a needle-type glucose sensor with microdialysis sampling method. Long-term subcutaneous tissue glucose monitoring in ambulatory diabetic patients. Diabetes Care 1994;17:387–396.
74. Shichiri M, Sakakida M, Nishida K, Shimoda S. Enhanced, simplified glucose sensors: long-term clinical application of wearable artificial endocrine pancreas. Artif Organs 1998; 22:32–42.
75. Hashiguchi Y, et al. Development of a miniaturized glucose monitoring system by combining a needle-type glucose sensor with microdialysis sampling method. Long-term subcutaneous tissue glucose monitoring in ambulatory diabetic patients. Diabetes Care 1994;17:387–396.
76. Shichiri M, Sakakida M, Nishida K, Shimoda S. Enhanced, simplified glucose sensors: long-term clinical application of wearable artificial endocrine pancreas. Artif Organs 1998;22: 32–42.
77. Schaller HC, et al. MPC algorithm controls blood glucose in patients with type 1 diabetes mellitus under fasting conditions using the IV-SC route. Diabetes Technol Ther 2002;4:234.
78. Schaller HC, et al. Feasibility of the SC-SC route for an extracorporeal artificial pancreas. Diabetes 2002;51(Suppl. 2): 462.
79. Schaller HC, Schaupp L, Bodenlenz M, Wilinska ME, Chassin LJ, Wach P, Vering T, Hovorka R, Pieber TR. On-line adaptive algorithm with glucose prediction capacity for subcutaneous closed loop control of glucose: Evaluation under fasting conditions in patients with type 1 diabetes. Diabetic Med 2005; (in press).

80. Canonico V, et al. Evaluation of a feedback model based on simulated interstitial glucose for continuous insulin infusion. Diabetologia 2002;45(Suppl. 2):995.
81. Bode B, et al. Alarms based on real-time sensor glucose values alert patients to hypo- and hyperglycemia: the guardian continuous monitoring system. Diabetes Technol Ther 2004; 6:105–113.
82. Chassin LJ, Wilinska ME, Hovorka R. Evaluation of glucose controllers in virtual environment: Methodology and sample application. Artif Intell Med 2004;32:171–181.
83. Van den Berghe G, et al. Intensive insulin therapy in the surgical intensive care unit. N Engl J Med 2001;345:1359–1367.
84. Hogan P, Dall T, Nikolov P. Economic costs of diabetes in the US in 2002. Diabetes Care 2003;26:917–932.

See also GLUCOSE SENSOR; HEART, ARTIFICIAL.

PARENTERAL NUTRITION. See NUTRITION, PARENTERAL.

PCR. See POLYMERASE CHAIN REACTION.

PERCUTANEOUS TRANSLUMINAL CORONARY ANGIOPLASTY. See CORONARY ANGIOPLASTY AND GUIDEWIRE DIAGNOSTICS.

PERINATAL MONITORING. See FETAL MONITORING.

PERIPHERAL VASCULAR NONINVASIVE MEASUREMENTS

CHRISTOPH H. SCHMITZ
HARRY L. GRABER
RANDALL L. BARBOUR
State University of New York
Brooklyn, New York

INTRODUCTION

The primary task of the peripheral vasculature (PV) is to supply the organs and extremities with blood, which delivers oxygen and nutrients, and to remove metabolic waste products. In addition, peripheral perfusion provides the basis of local immune response, such as wound healing and inflammation, and furthermore plays an important role in the regulation of body temperature. To adequately serve its many purposes, blood flow in the PV needs to be under constant tight regulation, both on a systemic level through nervous and hormonal control, as well as by local factors, such as metabolic tissue demand and hydrodynamic parameters. As a matter of fact, the body does not retain sufficient blood volume to fill the entire vascular space, and only $\sim$ 25% of the capillary bed is in use during resting state. The importance of microvascular control is clearly illustrated by the disastrous effects of uncontrolled blood pooling in the extremities, such as occurring during certain types of shock.

Peripheral vascular disease (PVD) is the general name for a host of pathologic conditions of disturbed PV function. Peripheral vascular disease includes occlusive diseases of the arteries and the veins. An example is peripheral arterial occlusive disease (PAOD), which is the result of a buildup of plaque on the inside of the arterial walls, inhibiting proper blood supply to the organs. Symptoms include pain and cramping in extremities, as well as fatigue; ultimately, PAOD threatens limb vitality. The PAOD is often indicative of atherosclerosis of the heart and brain, and is therefore associated with an increased risk of myocardial infarction or cerebrovascular accident (stroke).

Venous occlusive disease is the forming of blood clots in the veins, usually in the legs. Clots pose a risk of breaking free and traveling toward the lungs, where they can cause pulmonary embolism. In the legs, thromboses interfere with the functioning of the venous valves, causing blood pooling in the leg (postthrombotic syndrome) that leads to swelling and pain.

Other causes of disturbances in peripheral perfusion include pathologies of the autoregulation of the microvasculature, such as in Reynaud's disease or as a result of diabetes.

To monitor vascular function, and to diagnose and monitor PVD, it is important to be able to measure and evaluate basic vascular parameters, such as arterial and venous blood flow, arterial blood pressure, and vascular compliance.

Many peripheral vascular parameters can be assessed with invasive or minimally invasive procedures. Examples are the use of arterial catheters for blood pressure monitoring and the use of contrast agents in vascular X ray imaging for the detection of blood clots. Although they are sensitive and accurate, invasive methods tend to be more cumbersome to use, and they generally bear a greater risk of adverse effects compared to noninvasive techniques. These factors, in combination with their usually higher cost, limit the use of invasive techniques as screening tools. Another drawback is their restricted use in clinical research because of ethical considerations. Although many of the drawbacks of invasive techniques are overcome by noninvasive methods, the latter typically are more challenging because they are indirect measures, that is, they rely on external measurements to deduce internal physiologic parameters. Noninvasive techniques often make use of physical and physiologic models, and one has to be mindful of imperfections in the measurements and the models, and their impact on the accuracy of results. Noninvasive methods therefore require careful validation and comparison to accepted, direct measures, which is the reason why these methods typically undergo long development cycles.

Even though the genesis of many noninvasive techniques reaches back as far as the late nineteenth century, it was the technological advances of the second half of the twentieth century in such fields as micromechanics, microelectronics, and computing technology that led to the development of practical implementations. The field of noninvasive vascular measurements has undergone a developmental explosion over the last two decades, and it is still very much a field of ongoing research and development.

This article describes the most important and most frequently used methods for noninvasive assessment of the PV; with the exception of ultrasound techniques, these are not imaging-based modalities. The first part of this article, gives a background and introduction for each of these measuring techniques, followed by a technical description of the underlying measuring principles and technical implementation. Each section closes with examples of clinical applications and commercially available systems. The second part of the article briefly discusses applications of modern imaging methods in cardiovascular evaluation. Even though some of these methods are not strictly noninvasive because they require use of radioactive markers or contrast agents, the description is meant to provide the reader with a perspective of methods that are currently available or under development.

NONIMAGING METHODS

Arterial Blood Pressure Measurement

Arterial blood pressure (BP) is one of the most important cardiovascular parameters. Long-term monitoring of BP is used for the detection and management of chronic hypertension, which is a known major risk factor for heart disease. In this case, it is appropriate to obtain the instantaneous BP at certain intervals, such as days, weeks, or months, because of the slow progression of the disease.

In an acute care situation, such as during surgery or in intensive care, continuous BP measurements are desired to monitor heart function of the patients. The following sections describe the most important techniques.

Instantaneous BP Measurements

The most widely used approach is the auscultatory method, or method of Korotkoff, a Russian military physician, who developed the measurement in 1905. A pressure cuff is inflated to $\sim$ 30 mmHg (3.99 k Pa) above systolic pressure on the upper extremity. While subsequently deflating the cuff at a rate of $\sim$ 2 (0.26)–3 mmHg (0.39 kPa) (1), the operator uses a stethoscope to listen to arterial sounds that indicate the points at which cuff pressure equals the systolic and diastolic pressure. The first is indicated by appearance of a “tapping” sound, while the latter is identified by the change from a muffled to vanishing sound.

A second widespread BP measurement technique is the oscillatory method. Here, the cuff contains a pressure sensor that is capable of measuring cuff pressure oscillations induced by the arterial pulse. The cuff is first inflated to achieve arterial occlusion, and then deflated at rate similar to that for the auscultatory method. During deflation, the sensor registers the onset of oscillations followed by a steady amplitude increase, which reaches maximum when the cuff pressure equals the mean ABP. Beyond that, oscillations subside and eventually vanish. Systolic and diastolic pressure are given by the cuff pressure values at which the oscillatory signal amplitude is 55 and 85% of the maximum amplitude, respectively. These objective criteria, based on population studies, make this method superior to the auscultatory method, which relies on the subjective judgment of changes in sounds. Oscillatory measurements are typically used in automated BP monitors.

Continuous BP Monitoring

Currently, the standard of care for obtaining continuous central blood pressure is the insertion of a Swan–Ganz catheter into the pulmonary artery. The device has to be placed by a trained surgeon, and its use is restricted to the intensive care unit. In addition, besides bearing the risk of serious complications, the procedure is costly. There is clearly a need for noninvasive continuous blood pressure monitoring methods, which could be more widely applied, and which would reduce the patient risk. In the following, we describe two such techniques, the vascular unloading method of Peňáz, and arterial tonometry, both of which have been developed into commercial products.

Vascular Unloading. Many noninvasive BP measurements rely on vascular unloading (i.e., the application of distributed external pressure to the exterior of a limb to counter the internal pressure of the blood vessels). Typically, this is achieved with an inflatable pressure cuff under manual or automated control. Because tissue can be assumed essentially incompressible, the applied pressure is transmitted onto the underlying vessels, where it results in altered transmural (i.e., external minus internal) pressure. If the external pressure P_{ext} exceeds the internal pressure P_{int}, the vessel collapses. For the case $P_{\text{ext}} = P_{\text{int}}$ the vessel is said to be unloaded (1).

In 1973, Czech physiologist Jan Peňáz proposed a noninvasive continuous BP monitoring method based on the vascular unloading principle (2). The approach, which was first realized by Wesseling in 1985, employs a servo-controlled finger pressure cuff with integrated photoplethysmography (see below) to measure digital arterial volume changes (3). The device uses a feedback mechanism to counter volume changes in the digital arteries through constant adjustment of cuff pressure, hence establishing a pressure balance that keeps the arteries in a permanently unloaded state. The applied cuff pressure serves as a measure of the internal arterial pressure. The cuff pressure is controlled with a bandwidth of at least 40 Hz to allow adequate reaction to the pulse wave (4). The method was commercialized in the late 1980s under the name Finapres. The instrument has a portable front end, which is worn on the wrist and contains an electropneumatic pressure valve, the cuff, and the PPG sensor. This part connects to a desktop unit containing the control, air pressure system, and data display–output. Two successor products are now available, one of which is a completely portable system.

One problem of this method is that the digital BP can significantly differ from brachial artery pressure (BAP) in shape, because of distortions due to pulse wave reflections, as well as in amplitude because of flow resistance in the small arteries. The former effect is corrected by introducing a digital filter that equalizes pressure wave distortions. The second problem is addressed by introducing a correction factor and calibrating the pressure with an independent return-to-flow BP measurement. It has been demonstrated that the achievable BP accuracy lies well within the American Association for Medical Instrumentation (AAMI) standards (5).

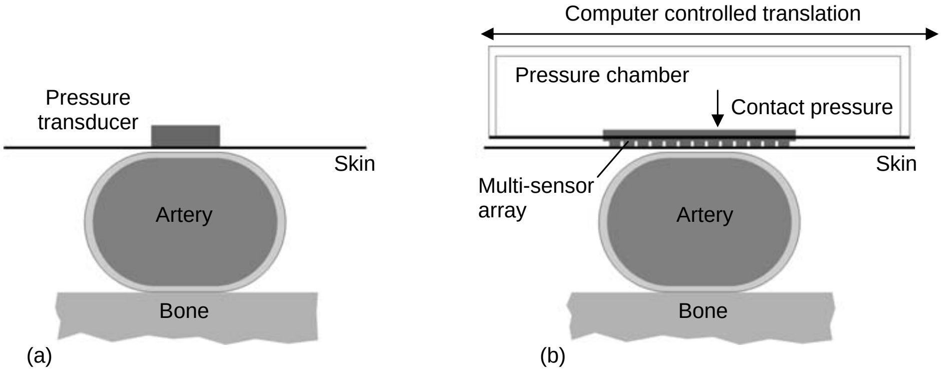


Figure 1. Applanation tonometry principle. (a) Single-element transducer. (b) Sensor array with pneumatic contact pressure control.

Applanation Tonometry. First conceived and implemented by Pressman and Newgard in the early 1960s, applanation tonometry (AT) measures the pulse pressure wave of a superficial artery with an externally applied transducer (6). The method requires the artery to be supported by an underlying rigid (i.e., bone) structure. Therefore, the method has been applied mainly to the temporal and the radial arteries; the last one being by far the most frequently employed measurement site. Figure 1a shows the general principle of the method. A pressure transducer is placed over the artery, and appropriate pressure is applied so as to partially flatten, or applanate, the artery. This ensures that the vessel wall does not exert any elastic forces perpendicular to the sensor face; therefore the sensor receives only internal arterial pressure changes caused by the arterial pulse. To obtain an accurate measurement it is crucial that the transducer is precisely centered over the artery, and that it is has stable support with respect to the surrounding tissue.

The original design used a single transducer that consisted of a rod of 2.5 mm^2 cross-sectional area, which was pressed against the artery, and which transmitted arterial pressure to a strain gauge above it. This early design suffered from practical difficulties in establishing and maintaining adequate sensor position. In addition, Drzewiecki has shown that for accurate pressure readings, the transducer area needs to be small compared to artery diameter (ideally, < 1 mm wide), a requirement that early designs did not meet (1).

The development of miniaturized pressure sensor arrays in the late 1970s has alleviated these difficulties, leading to the development of commercial AT instruments by Colin Medical Instruments Corp., San Antonio, TX. These sensor arrays use piezoresistive elements, which essentially are membranes of doped silicon (Si) that show a change in electrical resistance when subjected to mechanical stress. Piezoresistivity is a quantum mechanical effect rooted in the dependence of charge carrier motility on mechanical changes in the crystal structure. Pressure-induced resistance changes in a monocrytalline semiconductor are substantially greater than in other available strain gauges. This sensitivity together with the possibility of using semiconductor fabrication techniques to create miniaturized structures makes piezoresistive elements ideal candidates for AT applications. The change in resistance is measured with a Wheatstone bridge (e.g., pictured in Fig. 2), which, together with suitable amplification, can be integrated on the same chip as the sensing element. While piezoresistance shows linear change with strain, the devices are strongly influenced by ambient temperature. Therefore, appropriate compensation measures have to be taken.

Figure 1b shows the schematic of a modern AT pressure sensor. Thirty-one piezoeresistive elements form the 4.1×10 mm large sensor array, which is created from a monolithic Si substrate. After placing the sensor roughly over the artery, the signal from each element is measured to determine whether the transducer is appropriately centered with respect to the vessel. If necessary, its lateral position is automatically adjusted with a micromotor drive. The sensor contact pressure is pneumatically adjusted to achieve appropriate artery applanation. To provide probe stability suitable for long-term studies, the device is strapped to the wrist together with a brace that immobilizes the hand in a slightly overextended position so as to achieve better artery exposure to the sensor field.

Because AT can accurately measure the pulse wave shape, not its absolute amplitude, the AT signal is calibrated with a separate oscillatory BP measurement on the ipsilateral arm. Calibration is performed automatically at predetermined intervals.

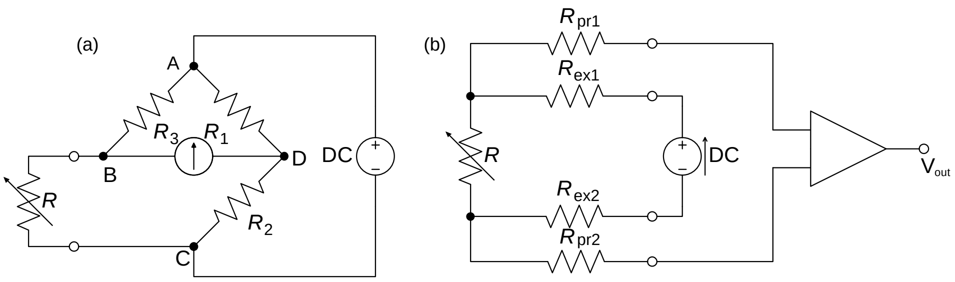


Figure 2. (a) Wheatstone bridge for sensitive detection of resistor changes; gauge lead resistance disturbs measurement. (b) Four-wire strain gauge measurement; influence of lead resistance is eliminated.

In addition to fully automated long-term monitoring devices, simpler single-element transducers are offered for clinical and research applications (PulsePen by DiaTecne, Milan, Italy and SPT-301 by Millar Instruments, Inc., Houston, TX).

Plethysmography

Plethysmography (derived from the Greek words for "inflating" and "recording") is the general name of an investigative method that measures volume change of the body, or a part of it, in response to physiologic stimulation or activity. Specifically, it allows measuring dynamics of blood flow to/from the extremities for the diagnosis of peripheral vascular diseases.

Different approaches have been conceived to measure volume change of a limb, the earliest of which, reaching back to the end of the nineteenth century, were based on measuring the volume displacement in a water filled vessel in which the limb was sealed in (7,8). Even though accurate in principle, the method suffers from practical limitations associated with the need to create a satisfactory watertight seal; it has therefore largely been supplanted by other approaches. The most important methods currently used are strain gauge PG, impedance PG, and photo PG. The working principles and applications of each of these methods will be described in the following sections.

Strain Gauge Plethysmography. Introduced by Whitney in 1953 (9), strain gauge plethysmography (SPG) measures changes in limb circumference to deduce volume variations caused by blood flow activity. The strain gauge is a stretchable device whose electrical resistance depends on its length; for SPG, it is fitted under tension in a loop around the limb of interest. Pulsatile and venous blood volume changes induce stretching/relaxing of the gauge, which is translated into a measurable resistance variation. Measurement interpretation is based on the premise that the local circumference variation is representative of the overall change in limb volume.

Whitney introduced a strain gauge consisting of flexible tubing of length l made from silastic, a silicone-based elastomer, filled with mercury. Stretching of the tubing increases the length and decreases the cross-sectional area a of the mercury-filled space, leading to an increase in its resistance R according to

$$R = \rho_m \frac{l^2}{v} \tag{1}$$

where ρ_m denotes mercury's resistivity (96 $\mu\Omega$ cm), and v is the mercury volume, $v = l\ a$. Differentiation of Eq. 1. shows that the gauge's change in resistance is proportional to its length variation (10):

$$\frac{\Delta R}{R} = 2\frac{\Delta\lambda}{l} \tag{2}$$

Whitney's original strain gauge design is still the most commonly used in SPG. Because of typical gauge dimensions and mercury's resistivity, the devices have rather small resistance values, in the range of 0.5–5 Ω.

The measured limb is often approximated as a cylinder of radius r, circumference C, length L, and volume V. By expressing the cylinder volume in terms of its circumference, $V = C^2L/(4\pi)$, and then differentiating V with respect to C, it is shown that the fractional volume change is proportional to relative variations in circumference:

$$\frac{dV}{V} = 2\frac{dC}{C} \tag{3}$$

Because changes in C are measured by the strain gauge according to Eq. 1, the arm circumference is proportional to the tubing length, and so the following relationship holds:

$$\frac{\Delta V}{V} = \frac{\Delta R}{R} \tag{4}$$

Whitney used a Wheatstone bridge, a measurement circuit of inherently great sensitivity, to detect changes in gauge resistance. In this configuration, shown in Fig. 2a, three resistors of known value together with the strain gauge form a network, which is connected to a voltage source and a sensitive voltmeter in such a manner that zero voltage is detected when $R_1/R_2 = R_3/R$. In this case, the bridge is said to be balanced. Changes in strain, and therefore R, cause the circuit to become unbalanced, and a nonzero voltage develops between points B and D according to

$$\frac{V_{BD}}{V_{AC}} = \frac{R_2}{R_1 + R_2} - \frac{R}{R_3 + R} \tag{5}$$

One disadvantage of the circuit is its nonlinear voltage response with respect to variations in R. For small changes, however, these nonlinearities remain small, and negligible errors are introduced by assuming a linear relationship.

A more significant shortcoming of the Wheatstone setup is that the measurement is influenced by the voltage drop across the lead resistance; especially for small values of R, such as those encountered in SPG applications, this can be a significant source of error. Therefore, modern SPG instruments use a so-called four-wire configuration, which excludes influences of lead resistance entirely. Figure 2b shows the concept. An electronic source is connected to the strain gauge with two excitation leads of resistance R_{ex1}, R_{ex2}, sending a constant current I through the strain gauge. Two probing leads with resistances R_{pr1}, R_{pr2} connect the device to an instrumentation amplifier of high impedance $R_{\text{amp}} \gg R_{\text{pr1}}, R_{\text{pr2}}$, which measures the voltage drop $V_{\text{SG}} = I \times R$ across the strain gauge. Because V_{SG} is independent of R_{ex1} and R_{ex2}, and there is negligible voltage drop across R_{pr1} and R_{pr2}, lead resistances do not influence the measurement of R.

Recently, a new type of plethysmography strain gauge has been introduced, which measures circumference variations in a special band that is worn around the limb of interest. The band, which has a flexible zigzag structure to allow longitudinal stretching, supports a nonstretching nylon loop, whose ends are connected to an electromechanical length transducer. Changes in circumference are thus translated into translational motion, which the transducer measures on an inductive basis, with 5 μm accuracy (11). In evaluation studies, the new design performed comparable to traditional strain gauge designs (12).

Impedance Plethysmography. Electrical conductivity measurements on the human body for the evaluation of cardiac parameters were performed as early as the 1930s and 1940s. Nyboer is widely credited with the development of Impedance Plethysmography (IPG) for the measurement of blood flow to organs, and its introduction into clinical use (13–15).

The frequency-dependent, complex electrical impedance *Z(f)* of tissue is determined by the resistance of the inter- and intracellular spaces, as well as the capacitance across cell membranes and tissue boundaries. The IPG measurements are performed at a single frequency in the 50–100 kHz range, and only the impedance magnitude Z (not the phase information) is measured. Therefore, Z can be obtained by applying Ohms law

$$Z = \frac{\hat{V}}{\hat{I}} \tag{6}$$

where $\hat{V}$ and $\hat{I}$ denote voltage and current amplitude values, respectively.

In the mentioned frequency range, the resistivity of different tissues varies by about a factor of 100, from $\sim 1.6\ \Omega\cdot\text{m}$ for blood to $\sim 170\ \Omega\cdot\text{m}$ for bone. Tissue can be considered a linear, approximately isotropic, piecewise electrically homogeneous volume conductor. Some organs, however, notably the brain, heart and skeletal muscles, show highly anisotropic conductivity (16).

Figure 3 shows a schematic for a typical four-electrode IPG setup. Two electrodes are used to inject a defined current into the body part under investigation, and two separate electrodes between the injection points measure the voltage drop that develops across the section of interest. The impedance magnitude Z is obtained, via Eq. 6, from the known current amplitude and the measured voltage amplitude. The four-electrode arrangement is used to eliminate the influence of the high skin impedance ($Z_{s1} = Z_{s4}$), which is $\sim$ 2–10 times greater than that of the underlying body tissue (17). If the same two electrodes were used to inject the current as well as to pick up the voltage drop, the skin resistance would account for most of the signal and distort the information sought.

The current source generates a sinusoidal output in the described frequency range and maintains a constant amplitude of typically 1 mA. This provides sufficient signal noise ratio (SNR) to detect physiologic activity of interest but is $\sim$ 50 times below the pain threshold for the employed frequency range, and therefore well below potentially hazardous levels.

The voltage difference between the pick-up electrodes is measured with an instrumentation amplifier, whose input impedance is much greater than that of skin or underlying tissue. Therefore, the influence of skin impedance can be neglected, and the measurement yields the voltage drop caused by the tissue of interest. The output of the instrumentation amplifier is a sinusoidal voltage with amplitude proportional to the impedance of interest. The signal needs to be demodulated, that is, stripped of its carrier frequency, to determine the instantaneous amplitude value. This is done with a synchronous rectifier followed by a low pass filter, a technique also known as synchronous, lock-in, or homodyne detection. Figure 3 shows a possible analog circuit implementation; a discriminator, or zero-crossing detector, actuates a switch that, synchronously with the modulation frequency, alternately connects the buffered or inverted measured signal to a low pass filter. If phase delays over transmission lines can be neglected, this will generate a noninverted signal during one-half of a wave, say the positive half, and an inverted signal during the negative half wave. As a result, the carrier frequency is rectified. The low pass filter averages the signal to remove ripple and generates a signal proportional to the carrier frequency amplitude. Inspection shows that frequencies other than the carrier frequency (and its odd harmonics) will produce zero output.

Increasingly, the measured voltage is digitized by an analog-to-digital converter, and demodulation is achieved by a microprocessor through digital signal processing.

The measured impedance is composed of a large direct current (DC), component onto which a small (0.1–1%)

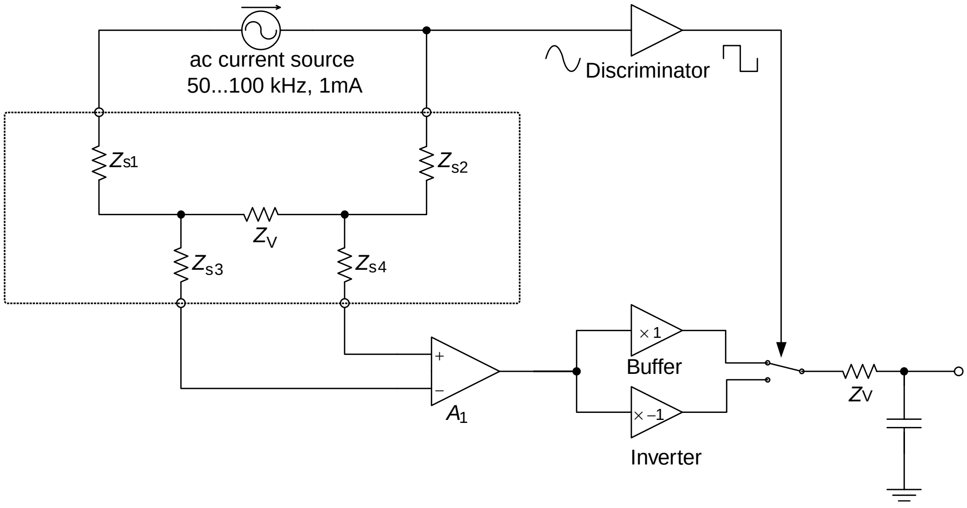


Figure 3. Four-lead tissue impedance measurement; the configuration mitigates influence of skin resistance (A_1, instrumentation amplifier). Synchronous detection removes carrier signal.

time-varying component is superimposed. The former is the constant impedance of the immobile tissue components, such as bone, fat, and muscle, and the latter represents impedance variations induced by volume changes in the fluid tissue components:most significantly, blood volume fluctuations in the vascular component. To obtain a quantitative relationship between measured impedance variations ΔZ and the change in blood volume, a simple electrical tissue equivalent may be considered, consisting of two separate, parallel-connected volume conductors of equal length L. One of these represents the immobile tissue components of impedance Z_0. The other is assigned resistivity ρ and a variable cross-sectional area A, thereby modeling impedance changes caused by variations in its volume V according to the relationship

$$Z_v = \rho\frac{L}{A} = \rho\frac{L^2}{V} \qquad (7)$$

Here Z_v denotes the variable compartment's impedance, which is generally much greater than that of the immobile constituents. Therefore, the parallel impedance of both conductors can be approximated $Z \approx Z_0 - Z_v$. To obtain a functional relationship between the measured changes in impedance ΔZ and variations in blood volume ΔV, Eq. 7 is solved for V and differentiated with respect to Z. Making use of $Z \approx Z_0$ and $dZ \approx -dZ_v$ yields

$$\Delta V = -\frac{\rho L^2}{Z_0^2}\Delta Z_v \qquad (8)$$

The IPG measurements do not allow independent determination of Z_0 and Z_v; however, the dc component of the IPG signal serves as a good approximation to Z_0, while the ac part closely reflects changes in Z_v. Low pass filtering of the IPG signal extracts the slowly varying dc components. Electronic subtraction of the dc part from the original signal leaves a residual that reflects physiologic impedance variations. The ac/dc separation can be implemented with analog circuitry. Alternatively, digital signal processing may be employed for this task. Digital methods help alleviate some of the shortcomings of analog circuits, especially the occurrence of slow signal drifts. In addition, software-based signal conditioning affords easy adjusting of processing characteristics, such as the frequency response, to specific applications.

Air Plethysmography. Air plethysmography (APG), also referred to as pneumoplethysmography, uses inflatable pressure cuffs with integrated pressure sensors to sense limb volume changes. The cuff is inflated to a preselected volume, at which it has snug fit, but at which interference with blood flow is minimal. Limb volume changes due to arterial pulsations, or in response to occlusion maneuvers with a separate cuff, cause changes in the measurement cuff internal pressure, which the transducer translates into electrical signals. The volume change is given by

$$\Delta V = V\frac{\Delta P}{P} \qquad (9)$$

The APG measurements can be calibrated with a bladder that is inserted between the limb and the cuff, which is filled with a defined amount of water. Because temperature changes influence air pressure inside the cuff, and it needs to be worn for a few minutes after inflation before starting the measurement, so the air volume can reach thermal equilibrium.

Photoplethysmography. Optical spectroscopic investigations of human tissue and blood reach back as far as the late nineteenth and early twentieth century, and Hertzman is widely credited with introducing photoplethysmography (PPG) in 1937 (18). The PPG estimates tissue volume changes based on variations in the light intensity transmitted or reflected by tissue.

Light transport in tissue is governed by two principle interaction processes; elastic scattering, that is, the random redirection of photons by the microscopic interfaces of the cellular and subcellular structures; second, photoelectric absorption by molecules. The scattering power of tissue is much greater (at least tenfold, depending on wavelength) than is the absorption, and the combination of both interactions cause strong dampening of the propagating light intensity, which decays exponentially with increasing distance from the illumination point. The greatest penetration depth is achieved for wavelengths in the 700–1000 nm range, where the combined absorption of hemoglobin (Hb) and other molecules show a broad minimum. Figure 4 shows the absorption spectra of Hb in its oxygenated (HbO_2) and reduced, or deoxygenated, forms.

The PPG measurements in transmission are only feasible only for tissue thickness of up to a few centimeters. For thicker structures, the detectable light signal is too faint to produce a satisfactory SNR. Transmission mode measurements are typically performed on digits and the earlobes.

Whenever tissue is illuminated, a large fraction of the light is backscattered and exits the tissue in the vicinity of the illumination point. Backreflected photons that are detected at a distance from the light source are likely to have descended into the tissue and probed deeper lying structures (Fig. 5). As a general rule, the probing depth

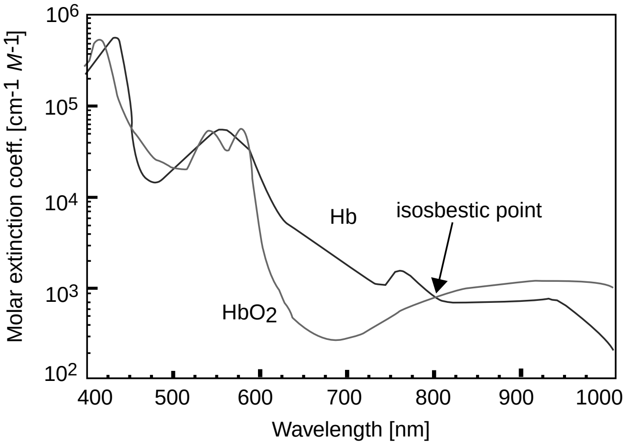


Figure 4. Hemoglobin spectra; typical PPG wavelength for oxygen saturation measurement are 660 and 940 nm.

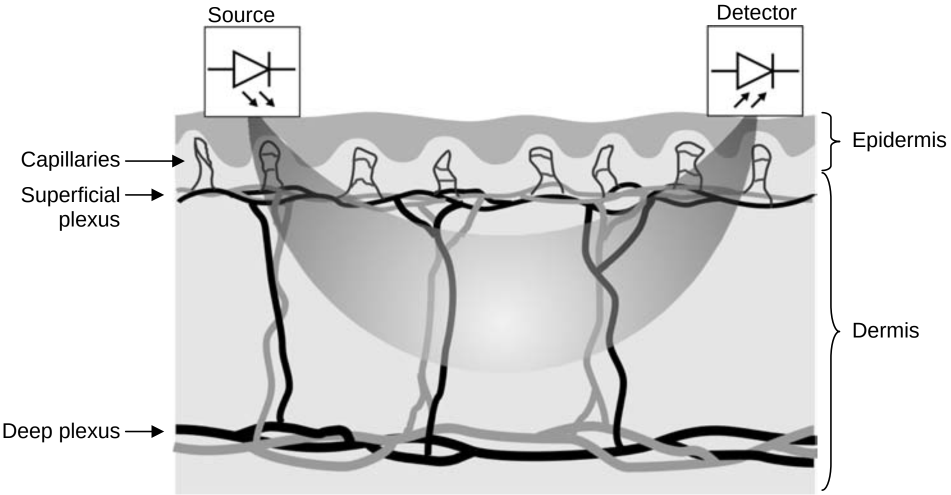


Figure 5. Schematic of the probed tissue volume in PPG reflection geometry.

equals about half the source-detector separation distance. The separation represents a tradeoff between probing volume and SNR; typical values are on the order of a few millimeters to a few centimeters. Reflection-geometry PPG can be applied to any site of the body.

Originally, incandescent light sources were used for PPG. These have been replaced with light emitting diodes (LED), owing to the latter devices' superior lifetime, more efficient operation (less heat produced), and desirable spectral properties. The LED technology has vastly evolved over the last 20 years, with a wide range of wavelengths—from the near-infrared (NIR) to the ultraviolet (UV) —and optical power of up to a tens of milliwatts currently available. These devices have a fairly narrow emission bandwidth, $\sim$ 20–30 nm, which allows spectroscopic evaluation of the tissue. The light-emitting diode (LEDs) are operated in forward biased mode, and the produced light intensity is proportional to the conducted current, which typically is on the order of tens of milliamps. Because in this configuration, the device essentially presents a short circuit to any voltage greater than its forward drop V_{LED} (typically 1–2 V), some form of current control or limiting circuitry is required. In the simplest case, this is a current-limiting resistor R_{lim} in series with the diode (Fig. 6a). The value of R_{lim} is chosen so that $I_{\mathrm{LED}} = (V_{\mathrm{cc}} - V_{\mathrm{LED}})/R_{\mathrm{lim}}$ is sufficient to drive the LED, typically on the order of tens of milliamps, but does not exceed the maximum permissible value. Depending on the voltage source, this results in adequate output stability. Often, there is a requirement to modulate or adjust the diode output intensity. In this case, an active current source is better suited to drive the LED. Figure 6b shows the example of a voltage-controlled current source with a boost field effect transistor. Here, the light output is linearly dependent on the input voltage from zero to maximum current; the latter is determined by LEDs thermal damage threshold.

Either phototransistors (PT) or photodiodes (PD) are used as light sensors for PPG. Both are semiconductor devices with an optically exposed *pn* junction. Photons are absorbed in the semiconductor material, where they create free charges through internal photoelectric effect. The charges are separated by the junction potential, giving rise to a photocurrent I_{p} proportional to the illumination intensity. While a PD has no internal gain mechanism and requires external circuitry to create a suitable signal, the PT (like any transistor) achieves internal current amplification h_{FE}, typically by a factor of $\sim$ 100. Figure 7a shows how a load resistor is used to convert the PT output current to voltage that is approximately proportional to

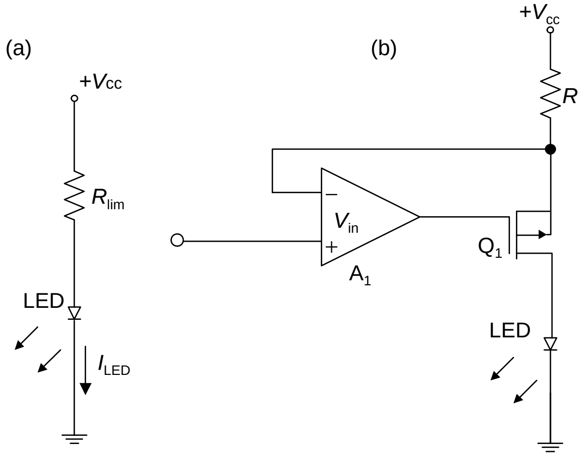


Figure 6. LED drive circuits. (a) Simple current limiting resistor. (b) Voltage controlled current source.

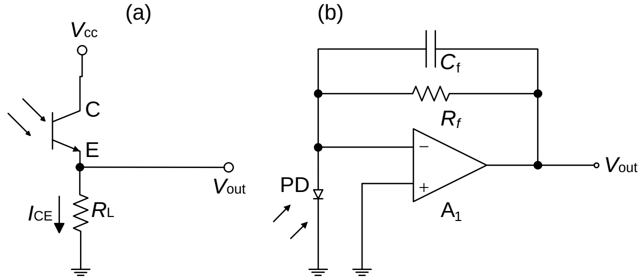


Figure 7. Light detection circuits. (a) Phototransistor operation (V_{CE} = collector-emitter voltage). (b) Use of photodiode with current-to-voltage converter (transimpedance amplifier).

the incident light according to

$$V_L = I_{CE}R_L = h_{FE}I_pR_L \tag{10}$$

h_{FE} unavoidably depends on a number of factors including I_p and V_{CE}, thus introducing a nonlinear circuit response. In addition, the PT response tends to be temperature sensitive. The light sensitivity of a PT is limited by its dark current, that is, the output at zero illumination, and the associated noise. The largest measurable amount of light is determined by PT saturation, that is, when V_L approaches V_{cc}. The device's dynamic range, that is, the ratio between the largest and the smallest detectable signal is on the order of three-to-four orders of magnitude.

Figure 7b shows the use of a PD with photoamplifier. The operational amplifier is configured as a current-to-voltage converter and produces an output voltage

$$V_o = I_pR_f \tag{11}$$

Capacitor C_f is required to reduce circuit instabilities. Even though compared to (a) this circuit is more costly and bulky because of the greater number of components involved, it has considerable advantages in terms of temperature stability, dynamic range, and linearity. The PD current varies linearly with illumination intensity over seven-to-nine orders of magnitude. The circuit's dynamic range is determined by the amplifier's electronic noise and its saturation, respectively, and is on the order of four decades. Proper choice of R determines the circuit's working range; values in the 10–100 MΩ range are not uncommon.

It is clear from considering the strong light scattering in tissue that the PPG signal contains volume-integrated information. Many models of light propagation in tissue based on photon transport theory have been developed that are capable of computing realistic light distributions even in inhomogeneous tissues, such as the finger (19). However, it is, difficult, if not impossible, to exactly identify the sampled PPG volume because this depends critically on the encountered optical properties as well as the boundary conditions given by the exact probe placement, digit size and geometry, and so. These factors vary between individuals, and even on the same subject are difficult to quantify or even to reproduce. Therefore, PPG methods do generally not employ a model-based approach, and the origin of the PPG signal as well as its physiologic interpretation has been an area of active research (20–23).

The PPG signals are analyzed based on their temporal signatures, and how those relate to known physiology and empirical observations. The detected light intensity has a static, or dc component, as well as a time-varying ac part. The former is the consequence of light interacting with static tissues, such as skin, bone, and muscle, while the latter is caused by vascular activity. Different signal components allow extraction of specific anatomical and physiological information. For example, the signal component showing cardiac pulsation, also called the digital volume pulse (DVP), can be assumed to primarily originate in the arterial bed, a premise on which pulse oximetry, by far the most widely used PPG application, is based. This method obtains two PPG signals simultaneously at two wavelengths that are spectrally located on either side of the isosbestic point. The ratio of the DVP peak amplitudes for each wavelength, normalized by their respective dc components, allows assessment of quantitative arterial oxygen saturation.

A number of other applications of the DVP signal have been proposed or are under investigation. For example, the DVP waveform is known to be related to the arterial blood pressure (ABP) pulse. Using empirically determined transfer functions, it is possible to derive the ABP pulse from PPG measurements (24).

Another potential use of arterial PPG is the assessment of arterial occlusive disease (AOD). It is known that the arterial pulse form carries information about mechanical properties of the peripheral arteries, and PPG has been investigated as a means to noninvasively assess arterial stiffness. To better discriminate features in the PPG signal shape, the second derivative of the signal is analyzed (second-derivative plethysmography, SDPTG) (23).

The PPG is furthermore used for noninvasive peripheral BP monitoring in the vascular unloading technique.

Continuous Wave Doppler Ultrasound

Doppler ultrasound (US) methods are capable of measuring blood flow velocity and direction by detecting the Doppler shift in the ultrasound frequency that is reflected by the moving red blood cells. The acoustic Doppler effect is the change in frequency of a sound wave that an observer perceives who is in relative motion with respect to the sound source. The amount of shift Δf in the US Doppler signal is given by (25)

$$\Delta f = \frac{2\, f_0 v_b \cos\Theta}{c} \tag{12}$$

where f_0 is the US frequency, v_b is the red blood cell velocity, Θ is the angle between the directions of US wave propagation and blood flow, and c is the speed of sound in tissue. Equation 12 demonstrates a linear relationship between blood flow and US Doppler shift. It is also seen that the shift vanishes if the transducer is perpendicular to the vessel because there is no blood velocity component in the direction of wave propagation. The algebraic sign of the shift depends on the flow direction (toward/away from) with respect to the transducer and is hence influenced by angle Θ. A factor of two appears in Eq. 12 because the Doppler effect takes place twice; the first occurs when the red blood cell perceives a shift in the incoming US wave, and the second shift taks place when this frequency is backreflected toward the transducer. Using typical values for the quantities in Eq. 12 ($f_0 = 5$ MHz, $v_b = 50$ cm·s^{-1}, $\Theta = 45^\circ$) yield frequency shifts of 2.3 kHz, which falls within the audible range (25).

Because the shift is added to the US frequency, electronic signal processing is used to remove the high frequency carrier wave. Analog demodulation has been used for this purpose; by mixing, that is, multiplying, the measured frequency with the original US frequency and low pass filtering the result, the carrier wave is removed, and audio-range shift frequencies are extracted. In so-called quadrature detection, this demodulation process yields two separate signals, one for flow components toward the detector, and one for flow away from it. Modern instruments typically employ digital signal processing, such as fast Fourier transformation (FFT), to accomplish this.

Continuous wave (CW) Doppler refers to the fact that the measurement is performed with a constant, nonmodulated US wave (i.e., infinite sine-wave). This technology does not create echo pulses, and hence does not allow any form of depth-profiling or imaging. Its advantages are its simplified technology, and that it is not restricted to a limited depth or blood velocity. Because the signal is generated continuously, CW transducers require two separate crystals, one for sound generation, and one for detection. The two elements are separated by a small distance and are inclined toward each other. The distance and angle between the elements determine the overlap of their respective characteristics, and thus determine the sensitive volume. All blood flow velocity components within this volume contribute to the Doppler signal.

In the simplest case, the measured frequency shift is made audible through amplification and a loudspeaker, and the operator judges blood velocity from the pitch of the tone (higher pitch = faster blood movement). These devices are used in cuff-based arterial pressure measurements, to detect the onset of pulsation. They also permit qualitative assessment of arterial stenosis and venous thrombosis. More sophisticated instruments display the changing frequency content of the signal as a waveform (velocity spectral display). Doppler spectra are a powerful tool for the assessment of local blood flow in the major vessels, especially when combined with anatomical US images (Duplex imaging, see section on imaging methods).

Peripheral Vascular Measurements

The following is a description of peripheral vascular measures that can be derived with the nonimaging techniques described in the preceding sections.

Assessment Of Peripheral Arterial Occlusive Disease. Peripheral arterial occlusive disease (PAOD) is among the most common peripheral arterial diseases (26), with a reported prevalence as high as 29% (27). The PAOD is usually caused by atherosclerosis, which is characterized by plaque buildup on the inner-vessel wall, leading to congestion, and hence increased blood flow resistance. Risk factors for PAOD are highly correlated with those for coronary artery disease, and include hypertension, smoking, diabetes, hyperlipidemia, age, and male sex (27–29). A symptom of PAOD is intermittent claudication, that is, the experience of fatigue, pain, and cramping of the legs during exertion, which subsides after rest. Ischemic pain at rest, gangrene, and ulceration frequently occur in advanced stages of the disease (29,30).

Ankle Brachial Index Test. The Ankle Brachial Index (ABI) is a simple and reliable indicator of PAOD (31), with a detection sensitivity of 90% and specificity of 98% for stenoses of >50% in a major leg artery (27,29). The ABI for each leg is obtained by dividing the higher of the posterior and anterior tibial artery systolic pressures for that leg by the greater of the left- and right-brachial artery systolic pressures. In normal individuals both systolic values should be nearly equal, and ABIs of 1.0–1.1 are considered normal. In PAOD, the increased peripheral arterial resistance of the occluded vessels causes a drop in blood pressure, thus diminishing the ABI. Because of measurement uncertainties, ABI values of < 0.9 generally are considered indicative of PAOD (27,29). Table 1 shows a scheme for staging of disease severity according to ABI values, as recommended by the Society of Interventional Radiology (SIR) (32).

Table 1. Ankle-Brachial Index (ABI) staging after ()

AB Ratio Range	Stage
1.0–1.1	Normal
< 1.0	Abnormal, possibly asymptomatic
0.5–0.9	Claudication
0.3–0.5	Claudication, rest pain, severe occlusive disease
< 0.2	Ischemia, gangrenes

The ABI test is performed with the patient in a supine position, and inflatable pressure cuffs are used to occlude the extremities in the aforementioned locations. The systolic pressure is determined by deflating the cuff and noting the pressure at which pulse sounds reappear. The audio signal from a CW Doppler instrument is used to determine the onset of arterial pulse sounds. Although the test can be performed manually with a sphygmomanometer, there are dedicated vascular lab instruments that automatically control the pressure cuff, register the pressure values, and calculate the ABI. Such instruments are commercially available, for example, from BioMedix,MN, Hokanson, WA, and Parks Medical Electronics, OR.

The primary limitation of the ABI test is that arterial calcifications, such as are common in diabetics, can resist cuff pressure thus causing elevated pressure readings. In those patients, the toe pressure as determined by PPG or SPG may be used because the smaller digital arteries do not develop calcifications (29). The normal toe systolic pressure (TSP) ranges from 90 mmHg to 100 mmHg, or 80% to 90% brachial pressure. The TSP < 30 mmHg (3.99 kPa) is indicative of critical ischemia (33). Also, pulse volume recordings (PVR, see below) are not affected by calcifications (26).

Segmental Systolic Pressure. Segmental Systolic Pressure (SSP) testing is an extension of the ABI method, wherein the arterial systolic pressure is measured at multiple points on the leg, and the respective ratios are formed with the higher of the brachial systolic readings. As is the case for ABI testing, a low ratio for a leg segment indicates occlusion proximal to the segment. By comparing pressures in adjacent segments, and with the contralateral side, SSP measurements offer a way to locate pressure drops and thus roughly localize occlusions. Typically, three to four cuffs are placed on each leg; one or two on the thigh, one below the knee, and one at the ankle.

The SSP technique suffers from the same limitation as does the ABI test, that is, it produces falsely elevated readings for calcified arteries. Again, TSP and PVR are advised for patients where calcifications might be expected because of known predisposition (e.g., diabetics), or unusual high ABI (>1.5) (26).

Pulse Volume Recording/Pulse Wave Analysis. Pulse volume recording (PVR), a plethysmographic measurement, is usually combined with the ABI or SSP test. Either the pressure cuffs used in these measurements are equipped with a pressure sensor so they are suitable for recordind volume changes due to arterial pulsation, or additional strain gauge sensors are used to obtain segmental volume pulsation. In addition to the leg positions, sensors may be located on the feet, and PPG sensors may be applied to toes. Multiple waveforms are recorded to obtain a representative measurement; for example, the Society for Vascular Ultrasound (SVU) recommends a minimum of three pulse recordings (34). The recorded pulse amplitudes reflect the local arterial pressure, and the mean amplitude, obtained by integrating the pulse, is a measure of the pulse volume. The pulsatility index, the ratio of pulse amplitude over mean pulse volume serves as an index for PAOD, with low values indicating disease (26).

Besides its amplitude and area, the pulse contour can be evaluated for indications of PAOD in a method referred to as pulse wave analysis (PWA). The normal pulse has a steep (anacrotic) rise toward the systolic peak, and a slower downward slope, representing the flow following the end of the heart's left ventricular contraction. The falling slope is interrupted by a short valley (the dicrotic notch), which is caused by the closing of the aortic valve. The details of the pulse shape are the result of the superposition of the incoming arterial pulse and backward-traveling reflections from arterial branchings and from the resistance caused by the small arterioles, as well as from atherosclerotic vessels (35). The pulse shape distal to occlusions tends to be damped and more rounded, and loses the dicrotic wave. Combined SSP and PVR recording has been reported over 90% accurate for predicting the level and extent of PAD (26).

Recently, PWA has increasingly been used to infer central, that is, aortic, arterial parameters from peripheral arterial measures, typically obtained from tonometric measures on the radial artery. These methods, which are commercialized for example by Atcor Medical, Australia and Hypertension Diagnostics, Inc, MN, rely on analytic modeling of transfer functions that describe the shaping of the pulse from the aorta to the periphery. These analytical functions are typically based on models mimicking the mechanical properties of the arterial tree by equivalent circuits (so-called Windkessel models) representing vascular resistances, capacitances, and inductances (36–38). From the derived aortic waveform, parameters are extracted that indicate pathologic states, such as hypertension and central arterial stiffness. One example is the augmentation index, the pressure difference in the first and second systolic peak in the aorta (39,40).

Pulse wave velocity (PWV) depends on arterial stiffness, which is known to correlate with increased cardiovascular risks. The PWV is established by measuring pulse waves on more than one site simultaneously, and dividing their respective separation difference from the heart by the time difference between the arrivals of comparable wave form features. Peripheral PWV measurements on toes and fingers have been conducted with PPG (41).

It was recently shown that arterial stiffness assessed from PWA and PVA on multisite peripheral tonometric recording correlates with endothelial function. The latter was assessed by US measurements of brachial arterial diameter changes in response to a reactive hyperemia maneuver (42). There is increasing clinical evidence that PWA can serve as a diagnostic tool for hypertension and as a cardiac disease predictor (43).

CW Doppler Assessment. Velocity spectral waveform CW Doppler is useful for assessing peripheral arterial stenoses. The Doppler waveform obtained from a normal artery shows a triphasic shape; there is a strong systolic peak, followed by a short period of low amplitude reversed flow, reversing again to a small anterograde value during mid-diastole. Measuring at a stenosis location shows a waveform of increased velocity and bi- or monophasic behavior. Measurements distal to stenosis appear monophasic and dampened (26,33).

Venous Congestion Plethysmography. The described PG methods are most valuable for the assessment of peripheral vascular blood flow parameters for the diagnosis of peripheral vascular diseases (PVD). Venous congestion PG (VCP), also called venous occlusion PG (VOP), allows the measurement of a variety of important vascular parameters including arterial blood flow (Q_a), venous outflow (Q_{vo}), venous pressure (P_v), venous capacitance (C_v), venous compliance (C), and microvascular filtration. In VCP, a pressure cuff is rapidly inflated on the limb under investigation to a pressure, typically between 10 (1.33 kPa) and 50 mmHg (6.66 kPa), sufficient to cause venous occlusion, but below the diastolic pressure so that arterial flow is not affected. Blocking the venous return in this fashion causes blood pooling in the tissue distal to the cuff, an effect that can be quantified by measuring volume swelling over time with PG. Figure 8 sketches a typical

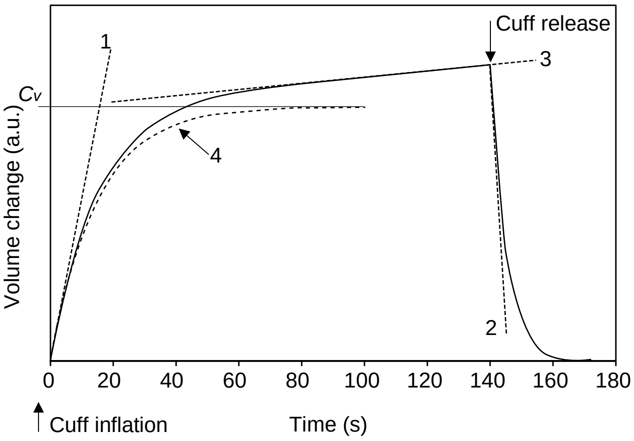


Figure 8. Schematic of typical VCP response curve. (*1*), asymptotic arterial blood flow; (*2*), asymptotic venous outflow; (*3*), swelling due to filtration; (*4*), effect of venous blood pooling only, obtained by subtracting the filtration component (*3*); C_v, venous capacitance.

volume response curve. Upon cuff inflation, an exponential increase in volume is observed, which is caused by the filling of postcapillary compliance vessels, and for which the time constant is $\sim$ 15 s in healthy individuals (44). The initial rate of swelling (indicated by asymptote 1), expressed in (mL·min^{-1}), is a measure of the arterial blood flow. When the venous pressure reaches the cuff pressure, blood can again flow out, and the relative volume increase reaches a plateau, which equals C_v. If C_v is measured for different occlusion pressure values, the slope of the C_v versus pressure curve yields C (in mL·mmHg^{-1} $\times$ 10^{-2}). Upon cuff release, the VCP curve shows fast exponential decay (on the orders of seconds), whose initial rate (indicated by asymptote 2) is a measure for Q_{vo} (45,46). With rising venous pressure an increase in fluid leakage, or filtration, takes place from the veins into the interstitial space, leading to additional tissue volume increase. Therefore, the PG response shows a second slow exponential component with a time constant on the order of 800 s (asymptote 3), from which can be deduced the capillary filtration capacity *CFC* (11).

Another application of VCP is the noninvasive diagnosis of deep vein thrombosis (DVT). The presence of venous blood clots impacts on venous capacitance and outflow characteristics. It has been shown that the combined measurement of C_v and Q_{vo} may serve as a diagnostic discriminator for the presence of DVT (45). A computerized SPG instrument is available (Venometer, Amtec Medical Ltd, Antrim, Northern Ireland), which performs a fully automated VCP measurement and data analysis for the detection of DVT (47).

Laser Doppler Flowmetry

Laser Doppler flowmetry (LDF) is a relatively young method of assessing the perfusion of the superficial microvasculature. It is based on the fact that when light enters tissue it is scattered by the moving red blood cells (RBC) in the superficial vessels (see PPG illustration in Fig. 5), and as a result the backscattered light experiences a frequency (Doppler) shift. For an individual scattering event, the shift magnitude can be described exactly; it depends on the photon scattering angle and the RBCs velocity, and it is furthermore related to the angular orientation of blood flow with respect the path directions of the photon before and after the scattering event. In actual tissue measurements, it is impossible to discern individual scattering events, and the obtained signals reflect stochastic distributions of scattering angles and RBC flow directions and velocities present in the interrogated volume.

In its simplest form, an LDF measurement is obtained by punctual illumination of the tissue of interest with a laser source, and by detecting the light that is backscattered at (or close to) the illumination site with a photodetector, typically a photodiode. The laser is a highly coherent light source, that is, its radiation has a very narrow spectral bandwidth. The spectrum of the backscattered light is broadened because it contains light at Doppler frequencies corresponding to all scattering angles and RBC motions occurring in the illuminated volume. Because of typical flow speeds encountered in vessels, the maximum Doppler shifts encountered are on the order of 20 kHz, which corresponds to a relative frequency variation of $\sim 10^{-10}$. Frequency changes this small can be detected because the shifted light components interfere with the unscattered portion of the light, causing a "beat signal" at the Doppler frequency. This frequency, which falls roughly in the audio range, is extracted from the photodetector signal and further processed. The small amount of frequency shift induced by cell motion is the reason that spectrally narrow, high quality laser sources (e.g., HeNe gas lasers or single-mode laser diodes) are required for LDF measurements.

The basic restriction of LDF is its limited penetration depth. The method relies on interference, an effect that requires photon coherence. Multiple scattering, however, such as experienced by photons in biological tissues, strongly disturbs coherence. For typical tissues, coherence is lost after a few millimeters of tissue. The sensitive volume of single-point LDF measurements is therefore on the order of 1 mm^3.

Single-point LDF measurements are usually performed with a fiberoptic probe, which consist of one transmitting fiber and one adjacent receiving fiber. Typical distances between these are from a few tenths of a millimeter to $>$ 2 mm. According to light propagation theory (see section on PPG), farther separations result in deeper probing depths, however, because of coherence loss, the SNR decreases exponentially with increasing distance, limiting the maximum usable separation.

The use of a probe makes a contact-based single-point measurement very convenient. Fiber-based probe have been developed that implement several receivers at different distances from the source (48). This allows a degree of depth discrimination of the measured signals, within the aforementioned limits.

The LDF signal is analyzed by calculation the frequency power spectrum of the measured detector signal, usually after band-pass filtering to exclude low frequency artifacts and high frequency noise. From this the so-called flux or perfusion value—a quantity proportional to the number of RBCs and their root-mean-squared velocity, stated in arbitrary units—is obtained. It has been shown that the flux is proportional to the width the measured Doppler power spectrum, normalized to optical power fluctuations in the setup (49).

Laser Doppler imaging (LDI) is an extension of the LDF technique that allows the instantaneous interrogation of extended tissue areas up to tens of centimeters on each side. Two approaches exist. In one implementation, the laser beam is scanned across the desired field of view, and a photodetector registers the signal that is backreflected at each scanning step. In another technology, the field of view is broadly illuminated with one expanded laser beam, and a fast CMOS camera is used to measure the intensity fluctuations in each pixel, thus creating an image. This second approach, while currently in a stage of relative infancy, shows the potential for faster frame rates than the scanning imagers, and it has the advantage of avoiding mechanical components, such as optical scanners (50).

The LDF/LDI applications include the assessment of burn wounds, skin flaps, and peripheral vascular perfusion problems, such as in Raynaud's disease or diabetes. The vascular response to heating (51) and the evaluation of carpal tunnel syndrome also have been studied (52).

IMAGING METHODS

This section provides an overview of existing medical imaging methods, and how they relate to vascular assessment. Included here are imaging modalities that involve the use of contrast agents or of radioactive markers, even though these methods are not considered noninvasive in the strictest sense of the word.

Ultrasound Imaging

Structural US imaging, especially when combined with Doppler US methods, is at present the most important imaging technique employed in the detection of PVD.

Improvements in ultrasound imaging technology and data analysis methods have brought about a state in which ultrasonic images can, in some circumstances, provide a level of anatomical detail comparable to that obtained in structural X-ray CT images or MRIs (53,54). At the same time, dynamic ultrasound imaging modalities can readily produce images of dynamic properties of macroscopic blood vessels (55,56). The physical phenomenon underlying all the blood-flow-sensitive types of ultrasound imaging is the Doppler effect, wherein the frequency of detected ultrasonic energy is different from that of the source, owing to the interactions of the energy with moving fluid and blood cells as it propagates through tissue. Several different varieties have become clinically important.

The earliest, most basic version of dynamic ultrasound imaging is referred to as color flow imaging (CFI) or color Doppler flow imaging (CDFI). Here, the false color value assigned to each image pixel is a function of the average frequency shift in that pixel, and the resulting image usually is superimposed upon a coregistered anatomic image to facilitate its interpretation. Interpretation of a CDFI image is complicated by, among other factors, the dependence of the frequency shift on the angle between the transducer and the blood vessels in its field of view. In addition, signal/noise level considerations make it difficult to measure low but clinically interesting flow rates. Many of these drawbacks were significantly ameliorated by a subsequently developed imaging modality known as either power flow imaging (PFI) or power Doppler imaging (PDI) (57). The key distinction between PDI and CDFI is that the former uses only the amplitude, or power, of ultrasonic energy reflected from erythrocytes as its image-forming information; in consequence, the entire contents of a vessel lumen have a nearly uniform appearance in the resulting displayed image (58).

The greatest value of Doppler imaging lies in the detection and assessment of stenoses of the larger arteries, as well as in the detection of DVT.

An increasingly employed approach to enhancing ultrasound images, at the cost of making the technique minimally invasive, is by introducing a contrast agent. The relevant contrast agents are microbubbles, which are microscopic, hollow, gas-filled polymer shells that remain confined in the vascular space and strongly scatter ultrasonic energy (59). It has been found that use of microbubbles permits a novel type of Doppler-shift-based ultrasound imaging, called contrast harmonic imaging (CHI). The incident ultrasound can, under the correct conditions, itself induce oscillatory motions by the microbubbles, which then return reflections to the transducer at not only the original ultrasonic frequency, but also at its second and higher harmonics. While these signals are not high in amplitude, bandpass filtering produces a high SNR, with the passed signal basically originating exclusively from the vascular compartment of tissue (60). The resulting images can have substantially lower levels of "clutter" arising from nonvascular tissue, in comparison to standard CDFI or PDI methods.

Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) is certainly the most versatile of the imaging modalities, and give the user the ability to study and quantify the vasculature at different levels. That is, depending on the details of the measurement, the resulting image may principally confer information about bulk blood flow (i.e., rates and volumes of blood transport in macroscopic arteries and veins), or about perfusion (i.e., amount of blood delivered to capillary beds within a defined time interval). The methods commonly employed to perform each of these functions are given the names magnetic resonance angiography (MRA) and perfusion MRI, respectively. As is the case for the other types of imaging treated here, each can be performed either with or without administration of exogenous contrast agents, that is, in either a noninvasive or a minimally invasive manner (61,62).

Magnetic Resonance Angiography. The MRA methods fall into two broad categories. The basic physical principle for one is flow-related enhancement (FRE), in which the pulse sequence employed has the effect of saturating the spins of ^{1}H nuclei in the nonmoving portion of the selected slice (63,64). Blood flow that has a large component of motion perpendicular to that slice remains significantly less saturated. Consequently, more intense MR signals arise from the blood than from the stationary tissues, and the contents of blood vessels are enhanced in the resulting image. Subtraction of a static image obtained near the start of the data collection sequence permits even greater enhancement of the blood vessel contents. A drawback of this approach is that vessels that lie within the selected slice, with the direction of blood flow basically parallel to it, do not experience the same enhancement.

A phase contrast mechanism is the basic physical principle for the second category of MRA techniques (65,66). Position-selective phase shifts are induced in ^{1}H nuclei of the selected slice via the sequential imposition of at least two transverse gradients that sum to a constant value across the slice. The effect is to induce zero net phase shifts among nuclei that were stationary during the gradient sequence, but a nonzero, velocity-dependent net phase shift on those that were in motion. The net phase shifts associated with the flowing blood are revealed in the resulting image, again, especially following subtraction of an image derived from data collected before the phase contrast procedure.

Perfusion MRI. Early developments in this field necessarily involved injection of a MR contrast agent that has the effect of inducing a sharp transient drop in signal detected as the contrast bolus passes through the selected slice (67). As with many dynamic MRI techniques, the bulk of the published work is geared toward studies of the brain, where, provided that the blood–brain barrier is intact, the method's implicit assumption that the contrast agent remains exclusively within the vascular space usually is justified. A sufficiently rapid pulse sequence allows the operator to generate curves of signal intensity vs. time, and from these one can deduce physiological parameters of interest, such as the cerebral blood flow (CBF), cerebral blood volume (CBV), and mean transit time (MTT) (67,68).

The minimally invasive approach described in the preceding paragraph remains the most common clinically applied type of perfusion MRI. A noninvasive alternative would have advantages in terms of permissible frequency of use, and would be of particular value in situations where pathology or injury has disrupted the blood–brain barrier. It also would make it possible to obtain perfusion images of other parts of the body than just the brain. Such an alternative exists, in the form of a set of methods known collectively as arterial spin labeling (ASL) (67) or arterial spin tagging (69). The common feature of all these techniques is that the water component of the blood is itself used as a contrast agent, by applying a field gradient that inverts the spins of ^{1}H nuclei of arterial blood before they enter the slice selected for imaging. The signal change induced by this process is smaller than that resulting from injection of a contrast agent, however, so that requirements for high SNRs are more exacting, and subtraction of a control image a more necessary step.

An increasingly popular and important functional MRI technique is blood-oxygen-level-dependent (BOLD) imaging (67,70). This type of imaging produces spatial maps determined by temporal fluctuations in the concentration of deoxygenated hemoglobin, which serves as an endogenous, intravascular, paramagnetic contrast agent. The physiological importance of BOLD images is that they reveal spatial patterns of tissue metabolism, and especially of neuronal activity of the brain. However, careful examination of the BOLD signal indicates that it depends, in a complex way, on many vascular and non-vascular tissue parameters (67,71). As such, it is not (yet) a readily interpretable method for specifically studying peripheral vasculature.

Fast X Ray Computed Tomography

As data acquisition speeds, and consequently repetition rates, for X ray computed tomography (CT) imaging have increased over the last couple of decades, previously unthinkable dynamic imaging applications have become a reality. The most direct approach taken along these lines is to rapidly acquire sequences of images of a slice or volume of interest and then examine and interpret, at any desired level of mathematical sophistication, temporal variations in the appearance of tissue structures in the images. Of course this approach is well suited to studying only organs whose functionality entails changes in shape or volume, such as the heart and lungs, and these have been the subject of many fast CT studies.

The functioning of many other organs, such as the kidneys (72), is related to the flow of blood and/or locally formed fluids, but does not involve gross changes in size or shape. In these cases it is necessary to introduce one or more boluses of X ray contrast agents, and to use fast CT to monitor their transit (73). While these methods violate the strict definition of noninvasive procedure, we include them in this synopsis out of consideration of the fact that the health risks associated with the contrast agents ordinarily are minor in relation to those imposed by the ionizing radiation that forms the CT images. For quantitation of regional blood flow and other dynamic vascular parameters, these techniques invariably employ some version of indicator dilution theory (73).

Indicator Dilution Approach

In one clinically important example, it has been found that the detected X ray CT signal changes in a quantifiable manner following inhalation of a gas mixture containing appreciable levels of nonradioactive isotopes of xenon (typically 33% Xe and 67% O_2, or 30% Xe, 60% O_2, 10% air) (75–77). A variety of techniques based on this phenomenon, and referred to as stable xenon-enhanced CT or ^{s}Xe–CT, were subsequently developed. The common feature is inhalation of a Xe-enriched gas mixture followed by repetitive scanning to monitor the wash-in and/or wash-out of the Xe-affected signal (75). However, while negative side effects are uncommon, they are known to occur (77). Consequently, the ^{s}Xe/CT technique is applied almost exclusively to cerebral vascular imaging studies, in cases where there is a diagnosed injury or pathology (75).

Qualitatively similar approaches, known collectively as perfusion CT, that are based on monitoring the time course of passage of an injected bolus of an iodinated contrast agent, also have been developed (74,78,79). Of course, these are even more invasive than the approaches based on inhalation an inert gas. Conflicting assertions (80) have been made regarding which of two approaches, injection or inhalation based, give superior results.

Positron Emission Tomography with ^{15}O

Strictly speaking, all forms of positron emission tomography (PET) imaging (as do all other nuclear medicine procedures) violate the formal definition of noninvasive measurements. The 2.1 min half-life of the isotope ^{15}O, which is brief relative to those of the other commonly used positron emitters, makes ^{15}O-labeled water a useful indicator of blood flow in dynamic PET measurements (81,82). Of course, as a practical matter radioisotope imaging methods cannot be used on a large scale for research or for clinical screening purposes. Medical literature on ^{15}O-PET-based vascular studies, understandably, also focuses almost exclusively on studies of circulation in the brain, in subjects with diagnosed vascular pathologies such as arteriovenous malformations (81).

Temporal-Spectral Imaging

A new area of investigation and technology development involving assessment of blood delivery to the periphery includes the use of optical array measurements combined with image formation and analysis capabilities. The basic concept, referred to as temporal-spectral imaging (TSI) (83), considers the functional interactions between peripheral tissues and their blood supply, and the added information that is attainable from a time series of volumetric images of tissue, where the latter are reconstructed from optical measurements of the hemoglobin signal. The information content of the measurement comprises three elements: first, expected differential response properties of tissues as a consequence of their varying blood supply and responses to internal and external stimuli; second, added information available from analysis of a time series, such as measures of rates, time delays, frequency content, and so on; third, further information available from a spatial deconvolution of diffusely scattered optical signals arising from deep tissue structures, to provide a volumetric image. In combination, the method provides for the assignment of multiple characteristics to each tissue type, and these can serve to distinguish one tissue from another, and healthy from diseased tissues. Because the supporting technology can be made compact, it is feasible to consider performing the optical array measurements on multiple sites of the body simultaneously in order to identify regional redistribution of blood, as can occur, for example, in response to shock.

The TSI approach builds upon an expanding investigative field known as diffuse optical tomography (DOT), which was first introduced late 1980s (84–88), and later extended to explore time-varying states (89,90). The DOT technology, unlike laser Doppler techniques, is suitable for exploring deep tissue structures and, similar to PPG, employs near-IR optical sources to assess the hemoglobin signal. In the backreflection mode, penetration depths of 3–4 cm are possible, depending on the tissue type. In the transmission mode, greater penetration is possible, including up to 12 cm of breast tissue and 6–8 cm in the forearm. For larger structures, for example, involving the lower extremities, full transmission measurements are not feasible, although off-axis measures partially inscribing the structure can be made.

The DOT measurements are made using an array of optical emitters and detectors that measure the light field reemitted from tissue over a relatively broad area. Measurements encompassing many tens of square centimeters of tissue surface, with interrogation of subsurface structures to the depths indicated above, are achievable. In practice, the DOT technique can be applied to explore the full volume of a breast, or partial volumes of the head, limbs or portions of the torso.

An example of the DOT approach extended to include TSI is presented below. First, however, it is useful to appreciate how DOT is accomplished. Figure 9 shows a schematic of the measurement strategy applied to time-series DOT. Being a tomographic technique, The DOT employs a multisource, multidetector measurement covering a relative wide area of tissue. Although light migrating in tissue is randomly scattered, predictions can be made as to the volume distribution of paths taken by photons propagating between any one source and detector. These take the shape of a fuzzy banana, and are conceptually similar to the linear trajectories taken by X rays as applied to CT imaging. Collectively, an array measurement can be processed to produce a cross-sectional or volumetric image which, when measured over time, produces a time series of images.

The TSI extends the time-series DOT technique in two ways. First, as noted, it recognizes that a wealth of differential information is available due to the significant

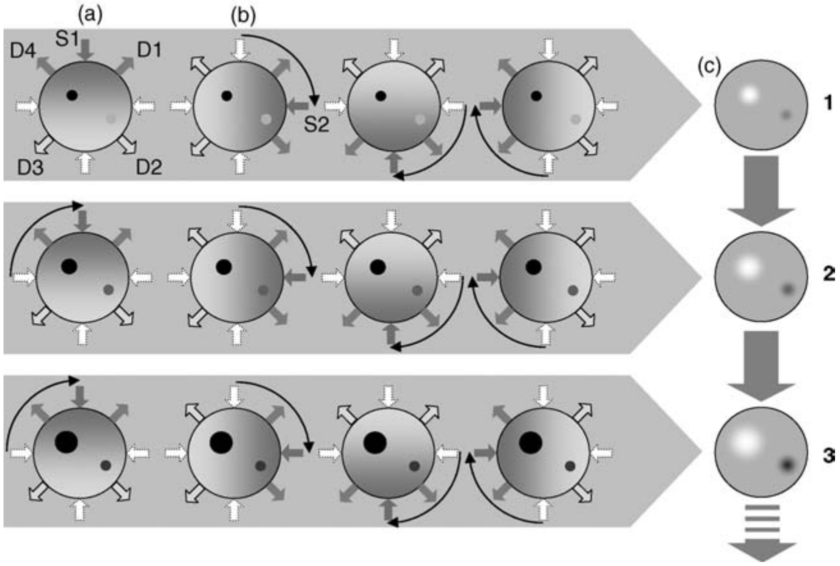


Figure 9. Temporal Spectral Imaging principle. (a) A NIR light source (S1) illuminates the target of interest in a specific location, and the reflected and transmitted light is measured at several locations (D1–D4). The source moves on to the next location (S2), and the measurement is repeated (b). After one full scan, a tomographic image can be reconstructed (c). Steps (a–c) are repeated to obtain successive image frames that show target dynamics, such as local changes in blood volume and oxygenation.

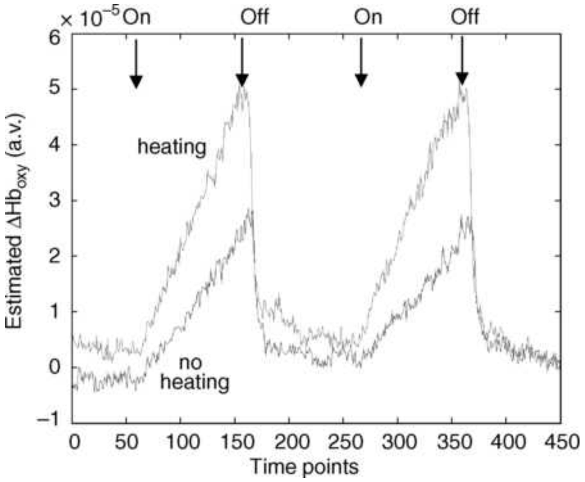


Figure 10. Time course of computed average oxyhemoglobin response to two cycles of mild venous occlusion in the forearm. Dark curve, no heat applied; light curve, with heat. [Reprinted with permission from R.L. Barbour et al., Functional imaging of vascular bed by dynamic optical tomography. Proceedings of SPIE, vol. 5369, in Physiology, Function, and Structure from Medical Images, 2004.]

variations in vascular density among the different tissue types, and in their responses to stimuli. Second, it also recognizes that for any given tissue, multiple response characteristics can be defined that serve to discriminate, as noted, one tissue type from another, or healthy from diseased. An example of this discriminatory capability is the following.

Shown in Fig. 10 is the response to partial cuff occlusion, applied to the arm in order to induce venous engorgement as recorded from a representative site about the forearm. Also shown is an enhanced response to a similar maneuver when the forearm was warmed by a thermal blanket prior to the measurement. We interpret the enhanced response as evidence of a decrease in local peripheral vascular resistance due to vasodilatation. Using the response curve shown in Fig. 10, it is possible to identify where similar behavior is seen in the corresponding cross-sectional image time series. This is shown in Fig. 11, together with an MR map of the forearm for comparison to anatomy. Regions of high correlation to the response curve are seen in many areas, with distinctively lower correlation occurring in regions that correspond to the ulna and radius, a finding consistent with known differences between soft tissue and interosseous hemodynamics (see overlay in e).

Additional discriminatory information about the considered stimulus can be gleaned from other analyses of the image time series. Motivating the approach taken is knowledge that any measure of bulk tissue properties is actually a composite of multiple overlapping features due to known differences, for example, in vascular compliance attributable to the principal elements of the vascular tree (i.e., arteries, veins and microvessels). Separation of the individual components comprising the whole is attainable using a class of methods known as blind source separation techniques.

Data in Figure 12 is an example of use of these methods, applied to four consecutive mild inflation–deflation cycles of the type illustrated in Fig. 10. The dark curve corresponds to the first principal component (PC) and accounts for ~ 80% of the signal variance. The light curve is the second PC, accounting for ~ 10% of total variance. Inspection reveals that the time courses of the two functions differ, and that they change from one application of mild occlusion to another. In the case of the first PC, it is seen that Hb_{oxy} levels increase promptly upon inflation. Also seen is that the magnitude of this response increases modestly following the second challenge. We interpret this to represent blood volume changes in the venous

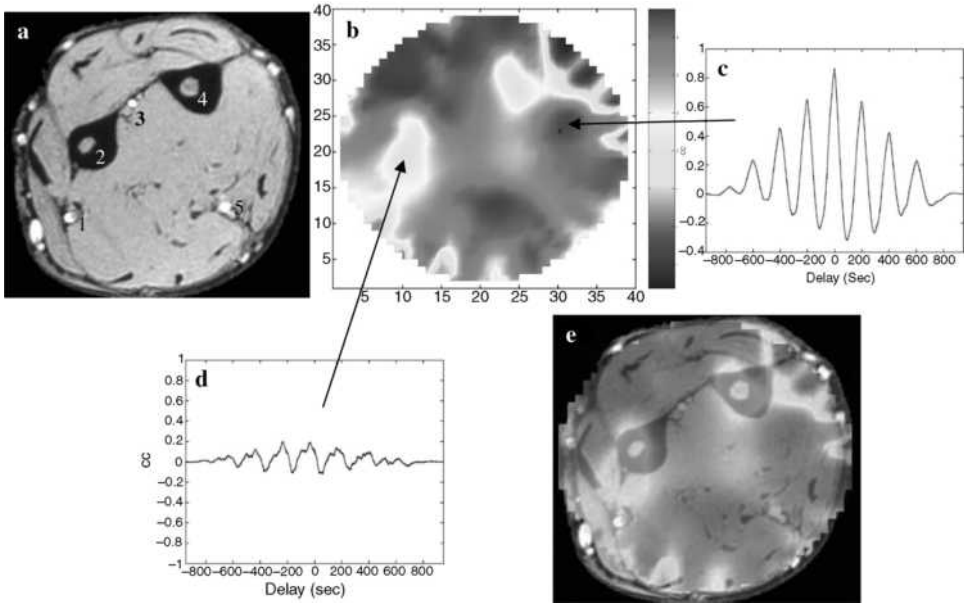


Figure 11. (a) MR cross-section of arm. (b) Cross-correlation (CC) map between (dark) model function in Fig. 10 with Hb image time series. (c) Overlay of a and b. (d and e), Time dependence of CC at indicated locations. (1) Radial artery, (2) radius, (3) interosseous artery, (4) ulna, (5) ulnar artery. [Reprinted with permission from R.L. Barbour et al., Functional imaging of vascular bed by dynamic optical tomography. Proceedings of SPIE, vol. 5369, in Physiology, Function, and Structure from Medical Images, 2004.]

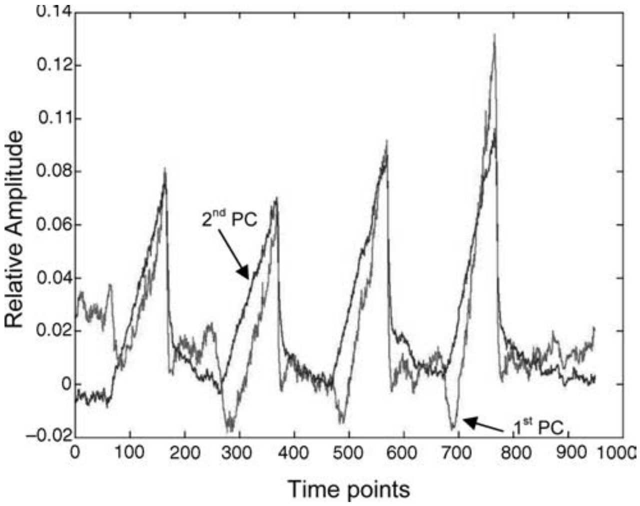


Figure 12. First (light curve) and second (dark curve) principal component (PC) of oxyhemoglobin signal computed for reconstructed image time series for four consecutive cuff inflation cycles. [Reprinted with permission from R.L. Barbour et al., Functional imaging of vascular bed by dynamic optical tomography. Proceedings of SPIE, vol. 5369, in Physiology, Function, and Structure from Medical Images, 2004.]

tree, as these distend most easily relative to the other elements of the vascular tree. The response seen in the second PC is more complex. Early on in the inflation cycle, a decline is observed, followed by an accelerated increase relative to the first PC, with proportionally greater responses upon subsequent challenges. It is believed this signal originates primarily from the microvascular bed. The initial decline in Hb_{oxy} is consistent with blood pooling, allowing for greater oxygen extraction. Following this, dilation occurs, perhaps in response to a buildup of local metabolic factors. The finding that the rate and extent of change in this signal increases in subsequent challenges suggests that insufficient time had elapsed between cycles to allow for complete washout of tissue factors. Support of this assignment is given in Fig. 13. In Fig. 13a it is seen that the signal associated with the first PC is mainly seen in the periphery (red regions), roughly in agreement with the location of near-surface veins. In contrast the signal associated with the second PC in Fig. 13b is found mainly in the ventral aspect of the forearm, which is dominated by well-perfused muscle.

Development of DOT technology is proceeding at a brisk pace, with new system designs being reported on a regular basis (91). As depicted here, TSI can be explored using time-series DOT. This, however, is not the only modality by which optical measures of peripheral vascular dynamics can be studied. Judging from advances made with photoacoustics (92), this approach may also prove useful (93,94). Additionally, the method could be extended further, to include use of injectable fluorescent probes or other forms of optical contrast media (95).

SUMMARY

Knowledge about the physiological and pathological state of the peripheral vasculature is extremely useful in the detection and staging of PVD, which occur with high prevalence and are associated with significant morbidity and mortality rates. Noninvasive measurements of those parameters are desirable because they can be performed more easily, are more cost effective, and are better suited for use as general screening tools than are invasive techniques. Volume-integrated dynamic flow measures can be obtained with a variety of available plethysmography methods. Measuring the flow response to occlusion maneuvers allows extraction of vessel flow and compliance values. In addition, plethysmography allows pulse volume recording in different locations on the extremities, and the interpretation of pulse wave contours and velocities permits drawing of inferences regarding peripheral, and even central, arterial health. Ultrasound methods, both structural and Doppler flow measurements, are valuable tools for diagnosing stenoses and thromboses. Laser Doppler flowmetry is a relatively new method for investigating the superficial microvessels.

Several imaging modalities are available for the assessment of blood flow and the vasculature. Most of these rely on contrast agents, and therefore can not strictly be considered noninvasive. The degree of invasiveness, however, seems small enough to justify their mentioning in the context of this article. Besides the conventional imaging methods, which include X ray CT, various MRI methods, and PET, the new noninvasive imaging technique of temporal spectral imaging, and in particular its

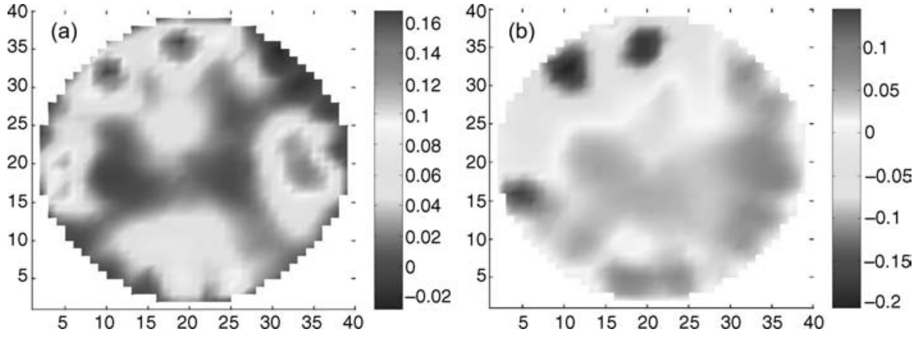


Figure 13. (a) Amplitude map of first principal component (PC) of total Hb (82% of total variance). (b) Amplitude map of second PC of total Hb (10% of total variance). [Reprinted with permission from R.L. Barbour et al., Functional imaging of vascular bed by dynamic optical tomography. Proceedings of SPIE, vol. 5369, in Physiology, Function, and Structure from Medical Images, 2004.]

implementation through diffuse optical tomography is introduced.

BIBLIOGRAPHY

Cited References

1. Drzewiecki G. Noninvasive arterial blood pressure and mechanics. Bronzino X, editor. The Biomedical Handbook. 2nd ed. Boca Raton, (FL): CRC Press and IEEE Press; 2000.
2. Peňáz J. Photoelectric measurement of blood pressure, volume, and flow in the finger. Digest 10th Internation Conference Medical Biological Engineering Dresden. 1973;104.
3. Wesseling KH, Settels JJ, De Wit B. The measurement of continuous finger arterial pressure non-invasively in stationary subjects. In: Schmidt TH, Dembroski TM, Bluemchen G, etitors. Biological and Psychological Factors in Cardiovascular Disease. Berlin: Springer-Verlag; 1986.
4. Parati G, et al. Non-invasive beat-to-beat blood pressure monitoring: new developments. Blood Press Mon 2003;8:31–36.
5. Bos WJW, et al. Reconstruction of brachial artery pressure from noninvasive finger pressure measurements. Circulation 1996;94:1870–1875.
6. Pressman GL, Newgard PM. A transducer for the continuous external measurement of arterial blood pressure. IEEE Trans Biomed Eng 1963;10:73–81.
7. Schäfer EA, Moore B. On the contractility and innervation of the spleen. J Physiol 1896;20:1–5.
8. Hewlett AW, Zwaluwenburg JG. The rate of blood flow in the arm. Heart 1909;1:87–97.
9. Whitney RJ. The measurement of volume changes in human limbs. J Physiol 1953;121:1–27.
10. McGivern RC, et al. Computer aided screening for deep venous thrombosis. Automedica 1991;13:239–244.
11. Schürmann M, et al. Assessment of the peripheral microcirculation using computer-assisted venous congestion plethysmography in post-traumatic complex regional pain syndrome type I. J Vasc Res 2001;38:453–461.
12. Leslie SJ, et al. Comparison of two plethysmography systems in assessment of forearm blood flow. J Appl Physiol 2004;96: 1794–1799.
13. Nyboer J, Bango S, Barnett A, Halsey RH. Radiocardiograms - the electrical impedance changes of the heart in relation to electrocardiograms and heart sounds. J Clin Invest 1940; 19:773.
14. Nyboer J. Regional pulse volume and perfusion flow measurements: Electrical impedance plethysmography. Arch Int Med 1960;105:264–276.
15. Nybor J. Electrical Impedance Plethysmography. 2nd ed, Springfield, (IL): Charles C Thomas; 1970.
16. Malmivuo J, Plonsey R. Bioelectromagnetism—Principles and Applications of Bioelectric and Biomagnetic Fields. New York: Oxford University Press; 1995.
17. Patterson R. Bioelectric impedance measurements. Bronzino X, editor. The Biomedical Handbook. 2nd ed, Boca Raton, (FL): CRC Press and IEEE Press; 2000.
18. Hertzman AB. Photoelectric plethysmography of the fingers and toes. Proc Soc Exp Biol Med 1937;37:529–534.
19. Hielscher AH, et al. Sagittal laser optical tomography for imaging of rheumatoid finger joints. Phys Med Biol 2004;49: 1147–1163.
20. De Trafford J, Lafferty K. What does photoplethysmography measure?. Med Biol Eng Comput 1984;22(5):479–480.
21. Jespersen LT, Pedersen OL. The quantitative aspect of photoplethysmography revised. Heart Vessels 1986;2:186–190.
22. Kamal AA, Harness JB, Irving G, Mearns AJ. Skin photoplethysmography — a review. Comput Methods Programs Biomed 1989;28(4):257–269.
23. Hashimoto J, et al. Pulse wave velocity and the second derivative of the finger photoplethysmogram in treated hypertensive patients: their relationship and associating factors. J Hypertens 2003;20(12):2415–2422.
24. Millasseau SC, et al. Noninvasive assessment of the digital volume pulse — Comparison with the peripheral pressure pulse. Hypertension W 2000;20(12):2415–2422.
25. Webb A. Introduction to Biomedical Imaging. Hoboken, (NJ): John Wiley & Sons Inc.; 2003.
26. Halpern JL. Evaluation of patients with peripheral vascular disease. Thrombosis Res 2002;106:V303–V311.
27. Hirsch AT, et al. Peripheral arterial disease detection, awareness, and treatment in primary care. JAMA 2001;286:317–1324.
28. Hooi JD, et al. Risk factors and cardiovascular diseases associated with asymptomatic peripheral arterial occlusive disease. The Limburg PAOD Study. Peripheral Arterial Occlusive Disease. Scand J Prim Health Care 1998;16(3): 177–182.
29. Weitz JI, et al. Diagnosis and treatment of chronic arterial insufficiency of the lower extremities: A critical review. Circulation 1996;94:3026–3049.
30. Carman TL, Fernandez BB. A primary care approach to the patient with claudication, Phys Am Fam 2000;61: 1027–1032. 1034.
31. Holland T. Utilizing the ankle brachial index in clinical practice. Ostomy Wound Manage 2002;48(1):38–40, 43–46, 48–49.
32. Sacks D, et al. Position statement on the use of the ankle brachial index in the evaluation of patients with peripheral vascular disease — A consensus statement developed by the standards division of the Society of Interventional Radiology. J Vasc Interv Radiol 2003;14:S389.
33. Donnelly R, Hinwood D, London NJM. ABC of arterial and venous disease: Non-invasive methods of arterial and venous assessment. studentBMJ 2000;8:259–302.
34. Vascular Technology Professional Performance Guideline — Lower Extremity Arterial Segmental Physiologic Evaluation, Society for Vascular Ultrasound; 2003.
35. Rietzschel E-R, et al. A comparison between systolic and diastolic pulse. Hypertension 2001;37:e15–e22.
36. Frank O. Die Grundform der Arteriellen Pulses. Acta Biol 1899;37:426–483.
37. Cohn JN, et al. Noninvasive pulse wave analysis for the early detection of vascular disease. Hypertension 1995;26:503–508.
38. Millasseau SC, Kelly RP, Ritter JM, Chowienczyk PJ. Determination of age-related increases in large artery stiffness by digital pulse contour analysis. Clin Sci 2002;103: 371–377.
39. Lekakis JP, et al. Arterial stiffness assessed by pulse wave analysis in essential hypertension: relation to 24-h blood pressure profile. Int J Cardiol 2005;102:391–395.
40. O'Rourke MF, Gallagher DE. Pulse wave analysis. J Hypertens Suppl 1996;14(5):S147–157.
41. Tsai W-C, et al. Association of risk factors with increased pulse wave velocity detected by a novel method using dual-channel photoplethysmography. Am J Hyper 2005;18:1118–1122.
42. Nigam A, Mitchell GF, Lambert J, Tardif J-C. Relation between conduit vessel stiffness (assessed by tonometry) and endothelial function (assessed by flow-mediated dilatation) in patients with and without coronary heart disease. Am J Cardiol 2003;92:395–399.

43. O'Rourke MF, Adji AA. An updated clinical primer on large artery mechanics: implications of pulse waveform analysis and arterial tonometry. Curr Opin Cardiol 2005;20(4): 275–81.
44. Gamble J, Gartside IB, Christ F. A reassessment of mercury in silastic strain gauge plethysmography for microvascular permeability assessment in man. J Physiol 1993;464: 407–422.
45. Maskell NA, et al. The use of automated strain gauge plethysmography in the diagnosis of deep vein thrombosis. Br J Radiol 2002;75:648–651.
46. Vigilance JE, Reid HL. Venodynamic and hemorheological variables in patients with diabetes mellitus. Arc Med Res 2005;36:490–495.
47. Goddard AJP, Chakraverty S, Wright J. Computer assisted strain-gauge plethysmography is a practical method of excluding deep venous thrombosis. Clin Radiol 2001;56: 30–34.
48. Larsson M, Nilsson H, Strömberg T. *In vivo* determination of local skin optical properties and photon path length by use of spatially resolved diffuse reflectance with applications in laser Doppler flowmetry. Appl Opt 2003;42(1):124–134.
49. Bonner R, Nossal R. Model for laser Doppler measurements of blood flow in tissue. Appl Opt 1981;20(20):2097–2107.
50. Serov A, Lasser T. High-speed laser Doppler perfusion imaging using an integrating CMOS image sensor. Opt Expr 2005;13(17):6416–6428.
51. Freccero C, et al. Laser Doppler perfusion monitoring of skin blood flow at different depths in finger and arm upon local heating. Microvasc Res 2003;66:183–189.
52. Eskandary H, Shahabi M, Asadi AR. Evaluation of carpal tunnel syndrome by laser Doppler flowmetry. Iran J Med Sci 2002;27(2):87–89.
53. Scharf A, et al. Evaluation of two-dimensional versus three-dimensional ultrasound in obstetric diagnostics: A prospective study. Fetal Diag Ther 2001;16:333–341.
54. Zotz RJ, Trabold T, Bock A, Kollmann C. *In vitro* measurement accuracy of three-dimensional ultrasound. Echocardiography 2001;18:149–56.
55. Hoksbergen AWJ, et al. Success rate of transcranial color-coded duplex ultrasonography in visualizing the basal cerebral arteries in vascular patients over 60 years of age. Stroke 1999;30:1450–1455.
56. Fürst G, et al. Reliability and validity of noninvasive imaging of internal carotid artery pseudo-occlusion. Stroke 1999;30: 1444–1449.
57. Rubin JM, et al. Power Doppler US: a potentially useful alternative to mean frequency-based color Doppler US. Radiology 1994;190:853–856.
58. Steinke W, et al. Power Doppler imaging of carotid artery stenosis: Comparison with color Doppler flow imaging and angiography. Stroke 1997;28:1981–1987.
59. Blomley MJK, et al. Liver microbubble transit time compared with histology and Child-Pugh score in diffuse liver disease: a cross sectional study. Gut 2003;52:1188–1193.
60. Della Martina A, Meyer-Wiethe K, Allémann E, Seidel G. Ultrasound contrast agents for brain perfusion imaging and ischemic stroke Therapy. J Neuroimaging 2005;15: 217–232.
61. Rofsky NM, Adelman MA. MR angiography in the evaluation of atherosclerotic peripheral vascular disease. Radiology 2000;214:325–338.
62. Baumgartner I, et al. Leg ischemia: Assessment with MR angiography and spectroscopy. Radiology 2005;234:833–841.
63. Kucharczyk W, et al. Intracranial lesions: flow-related enhancement on MR images using time-of-flight effects. Radiology 1986;161:767–772.
64. Whittemore AR, Bradley WG, Jinkins JR. Comparison of cocurrent and countercurrent flow-related enhancement in MR imaging. Radiology 1989;170:265–271.
65. Cebral JR, et al. Blood-flow models of the circle of Willis from magnetic resonance data. J Eng Math 2003;47:369–386.
66. Fatouraee N, Amini AA. Regularization of flow streamlines in multislice phase-contrast MR imaging. IEEE Trans Med Imag 2003;22:699–709.
67. Thomas DL, et al. The measurement of diffusion and perfusion in biological systems using magnetic resonance imaging. Phys Med Biol 2000;45:R97–R138.
68. Sorensen AG, et al. Hyperacute stroke: Simultaneous measurement of relative cerebral blood volume, relative cerebral blood flow, and mean tissue transit time. Radiology 1999;210:519–527.
69. Wang J, et al. Arterial transit time imaging with flow encoding arterial spin tagging (FEAST). Mag Reson Med 2003;50:599–607.
70. Mandeville JB, Marota JJA. Vascular filters of functional MRI: Spatial localization using BOLD and CBV contrast. Mag Reson Med 1999;42:591–598.
71. Ogawa S, Menon RS, Kim S-G, Ugurbil K. On the characteristics of functional magnetic resonance imaging of the brain. Annu Rev Biophys Biomol Struct 1998;27:447–474.
72. Lerman LO, Rodriguez-Porcel M, Romero JC. The development of x-ray imaging to study kidney function. Kidney Inte 1999;55:400–416.
73. Romero JC, Lilach LO. Novel noninvasive techniques for studying renal function in man. Semi Nephrol 2000;20:456–462.
74. Gobbel GT, Cann CE, Fike JR. Measurement of regional cerebral blood flow using ultrafast computed tomography: Theoretical aspects. Stroke 1991;22:768–771.
75. Horn P, et al. Xenon-induced flow activation in patients with cerebral insult who undergo xenon-enhanced CT blood flow studies. AJNR Am J Neuroradiol 2001;22:1543–1549.
76. Matsuda M, et al. Comparative study of regional cerebral blood flow values measured by Xe CT and Xe SPECT. Acta Neurolog Scand 1996;166:13–16.
77. Yonas H, et al. Side effects of xenon inhalation. J Comput Assist Tomogr 1981;5:591–592.
78. Roberts HC, et al. Multisection dynamic CT perfusion for acute cerebral ischemia: The "toggling-table" technique. AJNR Am J Neuroradiol 2001;22:1077–1080.
79. Röther J, et al. Hemodynamic assessment of acute stroke using dynamic single-slice computed tomographic perfusion imaging. Arch Neurol 2000;57:1161–1166.
80. Compare, for example, the corporate web pages http://www.anzai-med.co.jp/eigo/Xe-Per.htm and http://www.gehealthcare.com/usen/ct/case_studies/products/perfusion.html, which assert superiority for the Xe- and I-based approaches, respectively. One apparent reason for disagreement is the issue of the relative contributions to the image of the blood vessels' walls and of their contents.
81. Bambakidis NC, et al. Functional evaluation of arteriovenous malformations. Neurosurg Focus 2001;11: Article 2.
82. Schaefer WM, et al. Comparison of microsphere-equivalent blood flow (^{15}O-water PET) and relative perfusion (^{99m}Tc-tetrofosmin SPECT) in myocardium showing metabolism-perfusion mismatch. J Nucl Med 2003;44:33–39.
83. Barbour RL, et al. Functional imaging of the vascular bed by dynamic optical tomography. Proc SPIE 2004;5369: 132–149.
84. Barbour RL, Lubowsky J, Graber HL. Use of reflectance spectrophotometry (RS) as a possible 3-dimensional (3D) spectroscopic imaging technique. FASEB J 1988;2:A 1772.

85. Barbour RL, Graber H, Lubowsky J, Aronson R. Monte Carlo modeling of photon transport in tissue (PTT) [I. Significance of source-detector configuration; II. Effects of absorption on 3-D distribution (3DD) of photon paths; III. Calculation of flux through a collimated point detector (CPD); IV. Calculation of 3-D spatial contribution to detector response (DR); V. Model for 3-D optical imaging of tissue]. Biophy J 1990;57:381a–382a.
86. Grünbaum FA. Tomography with diffusion. Inverse methods in Action In: Sabatier PC, Editor. (Proceedings of 1989 Multicennials Meeting on Inverse Problems. New York: Springer-Verlag; 1990; 16–21.
87. Barbour RL, Lubowsky J, Aronson R. Method of Imaging a Random Medium, US pat. 5,137,355; awarded 8/11/92.
88. Wilson BC, Sevick EM, Patterson MS, Chance B. Time-dependent optical spectroscopy and imaging for biomedical applications. Proc IEEE 1992;80:918–930.
89. Barbour RL, et al. Temporal Imaging of Vascular Reactivity by Optical Tomography. In: Gandjbakhche AH, editor. Proceedings of Inter–Institute Workshop on *In Vivo* Optical Imaging at the NIH. (Optical Society of America, Washington, (DC); 1999), pp. 161–166.
90. Barbour RL, et al. Optical tomographic imaging of dynamic features of dense–scattering media. J Opt Soc Am A 2001;18: 3018–3036.
91. Schmitz CH, et al. Instrumentation for fast functional optical tomography. Rev Sci Instr 2002;73:429–439.
92. Wang X, et al. Noninvasive laser-induced photoacoustic tomography for structural and functional in vivo imaging of the brain. Nature Biotechnol 2003;21:803–806.
93. Kruger RA, Stantz KM, Kiser WL. Thermoacoustic CT of the Breast. Proc SPIE 2002;4682:521–525.
94. Ku G, Wang L-H. Deeply penetrating photoacoustic tomography in biological tissues enhanced with an optical contrast agent. Opt Lett 2005;30:507–509.
95. Weissleder R, Ntziachristos V. Shedding light onto live molecular targets. Nature Med 2003;9(1):123–128.

See also CUTANEOUS BLOOD FLOW, DOPPLER MEASUREMENT OF; IMPEDANCE PLETHYSMOGRAPHY.

PET SCAN. See POSITRON EMISSION TOMOGRAPHY.

PHANTOM MATERIALS IN RADIOLOGY

YOICHI WATANABE
C. CONSTANTINOU
Columbia University
New York, New York

INTRODUCTION

What is Phantom?

Radiation has both particle and electromagnetic wave nature. Radiation carries energy; hence, upon striking the human body, radiation deposits the energy in the human tissue. This interaction consequently can damage the tissue by causing strand breaks in genetic molecules called deoxyribonucleic acid (DNA) in nucleus of living cells. Such damages are considered a major cause of cancers.

Radiation, such as X rays, (γ rays, electrons (or β-particles), helium ions (or α-particles), and neutrons were discovered in late nineteenth century to early twentieth century. Since that time scientists and engineers invented and developed beneficial applications of radiation by taking advantages of the penetrating and damaging power of the radiation. Medical uses are the most noticeable applications of radiation. Radiation is used to diagnose and cure human illness.

Radiologists use X rays and other particulate radiation in hospitals and clinics to diagnose disease through imaging diseased sites. Radiation oncologists use X rays, electrons, and other forms of radiation available in radiation oncology centers to cure cancer.

While radiation is useful, careless use of radiation can lead to harmful effects on the health of people. Therefore, it is quite important to carefully evaluate the distribution of radiation energy absorbed by human tissue (or dose) during the radiological procedures. If the potential radiation damage is not well understood, clinical uses of new radiation sources without careful and thorough evaluation must be avoided. Placement of radiation measurement instrumentation in the human body is not easy, thus hampering precise dose measurements. Therefore, radiation scientists developed simulated human bodies or organs, herein called phantoms, to evaluate actual radiation doses. The phantoms are used to estimate radiation dose and transmission (or attenuation) of radiation in the human body for radiological studies.

Phantom materials for radiology should mimic the radiological characteristics of tissues. The homogeneity of radiological characteristics over the phantom is very important. Often the shape of the phantom should mimic the shape of a human body or a part of the body. Hence, the material should be easily made into various shapes and it should be easy to machine the material. The materials should maintain the mechanical integrity and the radiological characteristics for a long time.

Historical Background

To simulate radiation transport processes in human body, scientists developed phantoms made of tissue mimicking materials. The phantom should be made of material that absorbs and scatter radiation in the same way as the real tissue. Spires showed that the phantom material should have the same density as tissue and contain the same number of electrons per gram (1).

Water was the first material to be used as a tissue substitute in radiation measurements by Kienbock (2). Baumeister introduced wax in 1923 (3). The first formulated solid, called Siemen's Wax, composed of paraffin wax and magnesium oxide as a corrective filter, was reported by Ott in 1937 (4). Several similar wax-based products were subsequently introduced in Europe and North America, including MixD (5), Harris Wax (6), and M3 (7). Many phantoms comprised of either simple stacked sheets or more complex body-like structures have been constructed from these latter materials.

Plastics and rubbers have found an increasing application in the specialty of tissue simulation. From the polyethylene-based Markite (8) stemmed the conducting plastics of Shonka et al. (9) and polyurethane systems of Griffith et al. (10). The last three products have been used in the manufacture of elaborate anthropomorphic body phantoms with airways, simulated lungs, and embedded skeletons. An important elementally equivalent liquid system was introduced by Rossi and Failla (8). A mixture of water, glycerol, urea, and sucrose was used to match an approximate formula for soft tissue. This mixture was simplified by Goodman (11) and extended to more complex elemental formulas (12). Of the tissue substitutes introduced before 1970, only a handful had radiation absorption and scattering characteristics within ± 5% of those of the corresponding real tissues over extended energy range, and these included most of the above-mentioned phantom materials. The most important of them was water. Fortunately, it was readily available and cheap. An extensive program of research and development was initiated at St. Bartholomew's Hospital in London in 1970. Over 160 tissue substitutes were formulated, simulating a wide range of body tissues. Liquid, gel, solid, and particulate systems were produced for use with photon and particulate radiations (13–16). Other groups also developed tissue equivalent materials. Herman et al. used polyethylene to develop water-equivalent material, as well as fat and muscle materials in 1980s (17–19). Homolka et al. used polymer powders together with suitable additives to adjust photon attenuation (20,21). They created adipose, muscle, bone, and water equivalent materials, which simulate radiological characteristics of tissues for low energy photons, that is, energy < 100 keV for diagnostic X rays. Latest work includes development of tissue equivalent materials for pediatric radiology by Bloch et al. (22). Suess et al. manufactured a phantom material based on polyurethane resin for low contrast resolution measurements of computed tomography (CT) scanners (23). Iwashita used polyurethane resin mixed with $CaCO_3$ to create cortical and cancellous bones (24). Burmeister et al. made brain tissue equivalent conducting plastic for low-energy neutron applications (25).

Physics Background

Medical Radiation. Radiologists and radiation oncologists use radiation in several forms for diagnostics and therapy. The most common radiation is photons, which can be produced by X-ray generators and linear accelerators or are emitted by radioactive source. The photon energy used for medical applications ranges from 10 keV to 20 MeV. Electrons are another common form of radiation for medical uses. Positively charged electron or positrons are used for diagnostic purpose with positron emission tomography (PET) scanners. Heavier particles are also employed for therapeutic radiology. Protons, alpha particles, pi-mesons, neutrons, and heavy charged particles, such as carbons-ions were used in the past or are being introduced into clinic.

Interaction of Radiation with Matter. Photons interact with matter in three main physical processes: photoelectric absorption, Compton scattering, and pair production. Depending on the photon energy, one of three interactions play major role. Electrons in the energy range of interest collide with electrons in atoms–molecules of the material. Electrostatic force is the major interaction mechanism. Protons and heavy charged particles go through electrodynamic interactions similar to electrons. Neutrons do not carry electric charge; hence, those interact mostly with the nucleus of atoms.

Photon interaction probability is represented by the attenuation coefficient, which is the loss rate of photon particles per unit length from the original photon flight path. Electron scattering is quantified by stopping power, which represents electron energy loss rate per unit path. Energy absorption of radiation in tissue is considered per unit mass of tissue. The unit of radiation dose is joules per kilogram (J/kg) or gray (Gy). Mass attenuation coefficient and mass stopping power are often used to describe the effectiveness of material to attenuate photons and electrons.

Radiological Equivalence of Material to Tissue. Ideally, a phantom material should have the same mass attenuation coefficient for photons and the same mass stopping power for electrons as the tissue it simulates. If the phantom can have the same atomic composition as tissue, those parameters of the phantom are the same as those of tissue. However, making the atomic composition of phantom exactly the same as tissue is not easily achievable. Hence, as a guideline of tissue equivalent material, physicists expect that the material has a similar mass density, effective atomic number, and electron density as the real tissue. The reasoning for this approach is the following. As mentioned before, photons interact with matter through three physical mechanisms. The magnitude of the photoelectric interaction is approximately proportional to a certain power of the atomic number of the atom. The concept of the effective atomic number is introduced to present how close a material is to another material for photons in the energy range in which the photoelectric effect dominates as the main interaction process. Such energy range is generally < 100 keV. The Compton interaction and pair-production are essentially proportional to the number of electrons in the material. Electron stopping power is also proportional to the number of electrons since electrons directly interact with electrons in the material. Hence, the electron density of the material is another important parameter to represent the radiological characteristics of each material.

Outline

There is concern about the materials used to manufacture phantoms for radiology applications. The next section gives extensive discussion on the materials mainly developed by White, Constantinou, and there co-workers. More detailed discussion can be found in an ICRU report (26) and relevant references by those authors. The third section presents how those materials are used for radiation dosimetry, radiation therapy, and diagnostic radiology. The discussion on applications will be limited to photons and electrons, since most medical applications utilize those particles. The last section is devoted to speculative discussion on what types of phantom material will be developed in the near future.

PHANTOM MATERIALS: SIMULATED TISSUES AND CRITICAL TISSUE ELEMENTS

The tissue substitutes produced before 1970 were designed to simulate predominantly muscle, bone, lung, and fat. The sources of reliable data on elemental composition and mass densities of real tissues were limited. The main sources included the reports of Woodard (27), giving the elemental composition of cortical bone, and a report by the International Commission on Radiological Units and Measurements (28), giving the elemental composition of striated muscle and compact bone (femur). Unfortunately, there was a disagreement between the above sources on the composition of bone, which made bone simulation work more difficult.

The publication of the *Reference Man* data by the International Commission on Radiological Protection (ICRP) (29) in 1975 and the improvement in available equipment and technology enabled the formulation of new tissue substitutes for 15 different tissues that are described here. The *Reference Man* publication included tabulations of the concentrations of 51 elements in 81 organs, tissues, and tissue components. It also included the mass densities and the ratio of water/fat/protein contents in each one of them. Based on the above information, White and Constantinou developed substitutes for the following categories of tissues and body organs:

1. Principal soft tissues, namely, muscle, blood, adipose tissue, and skin. Adipose tissues are defined by ICRP as composed of 70% fat, 15% water, and 15% protein by mass.
2. Principal skeletal tissues, namely, cortical bone, inner bone, and red marrow. Yellow marrow is very close to adipose tissue and was not included. The formula given by Woodard (27) was adopted as more correct. This reference gives 55.8% Ca plus P, 12.5% water, 25.2% protein, and 6.5% sugar by mass for cortical bone.
3. Body organs, namely brain, kidneys, liver, lung, and thyroid. The elemental data for these organs were obtained from the ICRP Reference Man document.
4. Average tissues, which included average breast, total soft tissue, and total skeleton. The latter two formulas were derived from the ICRP source, while the formulas for average breast were based on 50% fat and 50% water by mass (13). Other formulas for average breast described in the literature (14,30) are based on 25% fat-75% muscle, 50% fat-50% muscle, and 75% fat - 25% muscle, referring to young, middle-aged, and older breast, respectively.

The percentage by mass for the elements H, C, N, O, Na, Mg, P, S, Cl, K, and Ca in each real tissue is given in Table 1 with information on mass densities and additional elements when appropriate. New tissue substitutes presented in this section were formulated so that, whenever possible, they have exactly the same elemental composition and mass density as the corresponding real tissue. In most of the solid substitutes where epoxy resin

Table 1. Elemental Compositions of the Principal Organs and Tissues, healthy adult[a]

	Elemental Composition (percentage by mass)												
Tissue	H	C	N	O	Na	Mg	P	S	Cl	K	Ca	Other Elements	Mass Density, kg·m^{-3}
Principal soft tissues													
Adipose tissue	11.2	51.7	1.3	35.5	0.1			0.1	0.1				970
Blood	10.2	11.0	3.3	74.5	0.1		0.1	0.2	0.3	0.2		Fe(0.1)	1060
Muscle	10.2	14.3	3.4	71.0	0.1		0.2	0.3	0.1	0.4			1050
Skin	10.0	20.4	4.2	64.5	0.2		0.1	0.2	0.3	0.1			1090
Principal skeletal tissues													
Cortical bone	3.4	15.5	4.2	43.5	0.1	0.2	10.3	0.3			22.5		1920
Inner bone (Spongiosa)	8.5	40.4	2.8	36.7	0.1	0.1	3.4	0.2	0.2	0.1	7.4	Fe(0.1)	1180
Red marrow	10.5	41.4	3.4	43.9			0.1	0.2	0.2	0.2		Fe(0.1)	1030
Body organs													
Brain	10.7	14.5	2.2	71.2	0.2		0.4	0.2	0.3	0.3			1040
Kidney	10.3	13.2	3.0	72.4	0.2		0.2	0.2	0.2	0.2	0.1		1050
Liver	10.2	13.9	3.0	71.6	0.2		0.3	0.3	0.2	0.3			1040
Lung	10.3	10.5	3.1	74.9	0.2		0.2	0.3	0.3	0.2			260
Thyroid	10.4	11.9	2.4	74.5	0.2		0.1	0.1	0.2	0.1		I(0.1)	1050
Average tissues													
Breast (whole)	11.5	38.7		49.8									960
Average soft tissue (male)	10.5	25.6	2.7	60.2	0.1		0.2	0.3	0.2	0.2			1030
Skelton (sacrum) (whole)	7.4	30.2	3.7	43.8		0.1	4.5	0.2	0.1	0.1	9.8	Fe(0.1)	1290

[a] See Ref. 31

systems, acrylics, or polyethylene were used as major components, a partial replacement of oxygen by carbon and vice versa had to be accepted. For this reason, an effort was made to determine which of the elements present in various tissues play a critical role in the energy deposition process when interacting with various radiation modalities. During this evaluation, basic interaction data have been calculated for photons and electrons from 10 keV up to 100 MeV, protons from 1 up to 1000 MeV, and neutrons from 100 eV up to 30 MeV. Detailed accounts of the computations are given in the literature (13,14,32) and only a summary is given here.

When a photon beam interacts with a tissue, photon energy absorption scattering depends primarily on the atomic number Z of the constituents and the electrons/kilogram of the tissue. Since hydrogen has double the electron density of other elements, hydrogen and the high Z constituents of a tissue are the critical elements. Consequently, their percentage by mass in the substitute must match that of a real tissue as accurately as possible. In order to evaluate the accuracy with which a substitute material simulated the corresponding real tissue, the mass attenuation coefficients (μ/ρ) and energy absorption coefficients (μ_{en}/ρ) were calculated at 33 energy points between 10 keV and 100 MeV, using the mixture rule:

$$\mu/\rho = \sum_i w_i(\mu/\rho)_i$$

where w_i is the proportion by mass of the ith element having a coefficient $(\mu/\rho)_i$.

The irradiation of tissues with beams of charged particles, such as electrons and protons, leads to energy deposition through collisional and radiative processes. Collisional interactions of incident particles with the electrons of the target material are the major cause of energy loss for electrons < 500 keV and protons < 1000 MeV. Radiative (bremsstrahlung) losses become important for higher energies. In order to evaluate the new tissue substitutes for electron and proton interactions, the collision stopping powers $(s/\rho)_{\text{coll}}$ and the radiative stopping powers $(s/\rho)_{\text{rad}}$ were calculated for both substitutes and the corresponding real tissues, and comparison was made between the total stopping powers $(s/\rho)_{\text{tot}}$ of the substitutes and those of the corresponding real tissues. A phantom material was accepted as a useful substitute only if its radiation characteristics were within $\pm$ 5% of those of the real tissue that it was designed to simulate.

In the case of tissues being irradiated with neutrons (10 eV–50 MeV), hydrogen was found to be the most critical element for all energies. Nitrogen was found to be the second most important element of neutron energies < 5 MeV. This is due to the significance of the elastic scattering of neutrons with hydrogen nuclei at higher neutron energies $^{1}H(n, n)^{1}H$ and the contribution of the capture process $^{1}H(n, \gamma)^{2}H$ and $^{14}N(n, p)^{14}O$ at the low and thermal neutron energies. Oxygen and carbon play a great role than nitrogen above 10 MeV. For neutron energies up to 14 MeV, the interactions with hydrogen account for 70–90% of the total dose in soft tissue (33,34). In view of the above finding, efforts were made to match the hydrogen, oxygen, carbon, and nitrogen contents of all the substitutes to those of the real tissues as accurately as possible.

The relative proportion of carbon and oxygen in tissue was found to be less critical in the attenuation and energy absorption from neutrons and high energy protons. Trace elements, with concentrations of $< 0.5\%$ by mass, were found to play no significant role in the absorption of energy from fast neutrons, high energy protons, and X rays > 100 keV. Detailed calculation and tabulation of the above-mentioned radiation interaction quantities for all the real tissues and their substitute material are available in the literature (13,14,30,35).

FORMULATION PROCEDURES

Three main methods were applied in the formulation of the tissue substitutes described in this article, namely, the elemental equivalence method, the basic data method, and the effective atomic number method. The following criteria formed the basis of the tissue simulation studies.

Criteria for Tissue Equivalence

Two materials will absorb and scatter any type of radiation in the same way, only if the following quantities are identical between them: (*1*) photon mass attenuation and mass absorption coefficients, (*2*) electron mass stopping powers and mass angular scattering powers, (*3*) mass stopping powers for heavy charged particles and heavy ions, (*4*) neutron interaction cross-sections and kerma factors, and (*5*) the mass densities of the two materials must be the same. A brief description of the formulation methods is now presented.

Method of Elemental Equivalence

Based on the above criteria, it is obvious that only material with the same elemental constituents and in the same proportion by mass as the corresponding real tissue can be termed tissue equivalent for all radiation modalities. A number of such materials were formulated, particularly in the liquid and gel phase (14,15,30,36). If a substitute is elementally correct and has the correct bulk density, the only source of error in the absorbed dose calculations from measurements in the phantom material will be phase differences due to differences in chemical binding. Such errors are difficult to evaluate because of lack of extensive data, but they have been found small and rather insignificant in conventional radiation dosimetry.

The method of elemental equivalence was first applied by Rossi and Failla (8) who tried to reproduce an approximate formula for soft tissue $(C_5H_{40}O_{16}N)_n$. They formulated a mixture of water–glycerol–urea and sucrose, which had the formula $C_5H_{37.6}O_{18}N_{0.97}$, but their publication did not explain how they arrived at their formulation. Frigerio et al. (37) used water as base material and then selected compound that could be dissolved in it in such proportions as to satisfy the CHNO molar ratio. They considered each compound as the sum of two components one of which was water; for example, glycerol $C_3H_2(H_2O)_3$ can be written as $C_3H_8O_3$. Using this approach, they produced a liquid system with elemental composition almost identical to that of muscle tissue.

The method of elemental equivalence was applied later with minor modifications (14,30), and as a result > 35 tissue equivalent liquids and gels were formulated. The following constraints were used during this work.

1. Once a base material was selected, the additives should be chosen from a library of compounds that are neither toxic nor corrosive, explosive, volatile, or carcinogenic.
2. The number of components should be kept to a minimum.
3. The proportion by mass of each constituents of a tissue substitute should be within 0.5% of that of the real tissue, except for hydrogen for which the agreement should be within 0.1%.

The basic steps followed in the formulation of elementally correct tissue substitutes have been reported in the references cited.

Basic Data Method

The second most accurate method of formulating tissue substitutes is the basic data method, which matches basic interaction data, for example, mass attenuation coefficients for photoelectric and Compton scattering, and mass stopping powers of the tissue substitutes to those for the body tissue over the required energy interval. This method was used by White (13,38) to formulate a large group of solid and liquid tissue substitutes for use with photons and electrons. The mathematical procedures developed enable two-component tissue substitutes (base material + filter) to be formulated for a given base material, with the most appropriate filler being selected from a library of compounds. Any degree of matching accuracy (e.g., 1% between μ/ρ values) can be specified. Recently, Homolka et al. (21) developed a computer program, which minimizes the difference between the linear attenuation coefficients of a phantom material and a tissue by considering the energy dependence of the attenuation coefficient. The program optimizes the components of base materials, such as polystyrene, polypropylene, and high density polyethylene together with admixtures of TiO_2, MgO, $CaCO_2$, and graphite. They showed that the measured Hounsfield number of the water equivalent phantom material agrees with those of water within eight Hounsfield units for X ray energy from 80 to 140 kV.

Effective Atomic Number Method

An indirect method of simulation is based on effective atomic number, $\hat{Z}$, which may be used to characterize a partial mass attenuation coefficient (τ/ρ, σ_e/ρ, κ/ρ, etc.) for a given group of elements and specified photon energy. The fundamental assumption for this method is that materials with the same value of the product of electron density and Z^x, where x is the Z exponent derived for a given partial interaction process, as a reference material shows the same photon and electron interaction characteristics as the reference material. A formulation technique similar to the basic data method can be derived, that is, the selection of an appropriate filler for a specific base material and the establishment of the relative proportions to achieve a specified degree of matching accuracy of the electron density and the effective atomic number between two materials. Accounts of this method were given by White et al. (39) and Geske (40).

MATERIALS AND METHOD OF MANUFACTURE OF THE NEW TISSUE SUBSTITUTES

Phantom materials currently in use can be grouped in-to four types depending on the base material. White et al. mainly developed epoxy-resin-based material (13,14,16, 38,41,42). Hermann et al. used polyethylene-based technique. ^{17}Homolka's group made phantoms based on fine-polymer powers, such as polyethylene, polypropylene, polystyrene or polyurethane (20). Suess, Iwashita and others used polyurethane resin (23,24). Since one of the current authors is very familiar with the epoxy-resin-based method and other methods use similar manufacturing techniques except the base material, more space is devoted to discussing the epoxy-resin-based phantom in this section than other methods.

Epoxy Resin-Based Method

Materials. The base materials used by White and Constantinou for the manufacturing of solid-phantom materials included four epoxy resin systems designated CB1, CB2, CB3, and CB4, respectively. The epoxy resin systems consist of a viscous resin and a lower viscosity liquid hardener (Diluents). The two are mixed in such proportions by mass as determined by the chemical reaction occurring during the curing process. The constitutes and elemental compositions of the epoxy-resin systems used in the manufacture of the new tissue substitutes were described in detail (14,38). These resin systems are rich in hydrogen (7.9–11.3% by mass) and nitrogen (1.60–65.62% by mass), but they are rather low in oxygen (13.15–20.57% by mass), compared to what is needed to match the oxygen content of the real tissue. As a result, in most solid substitutes, part of the oxygen needed was replaced by carbon, but an effort was made to have the sum of (C+O) in the substitute equal to that in the real tissue. Following the addition of the necessary powdered filler, low density ($\sim$ 200 $kg{\cdot}m^3$) phenolic microspheres (PMS) are also added in small precalculated quantities to make the bulk density of the mixture match that of real tissues. In the case of lung substitutes, the addition of a foaming agent (DC1107) in quantities of 1% by mass or less leads to sample with bulk densities as low as 200 $kg{\cdot}m^{-3}$ (43).

In the case of liquid substitutes, water was selected as the base material because it is an important component of real tissues and it is readily available. Various organic and inorganic compounds can be dissolved in it, in proportions necessary to satisfy the requirements for both the main elements C, H, N, O and the trace elements, such as Na, Mg, P, S, Co, K, and Ca.

The use of gelatin facilitated the formulation of many gel substitutes useful for short-term applications. For the production of elementally equivalent material, gelatin is preferred to other gelling agents such as agar (37) and Laponite used in the production of thyrotrophic gels (44), because it has an elemental composition very close to that of protein. Since real tissues are composed of varying proportions of water, carbohydrates, protein, and fat, it is easier to formulate elementally correct gel substitutes with it. By adding trace quantities of bacteriostatic agent (e.g., sodium azide) and sealing them into polyethylene base, gels can be preserved for longer periods.

Mixing Procedures. The manufacture of a solid substitute starts by adding first the appropriate quantity of the resin into a Pyrex reaction vessel followed by the lower viscosity hardener–diluent. The powder fillers are then added in order of decreasing mass density. Following a short manual mix, a ground glass lid is attached to the reaction vessel. This lid has one central and two peripheral glands (openings). A twin-bladed rotor is passed through the central gland and connected to an electric stirrer. One of the peripheral gland openings is connected to a vacuum pump while the third is used to control the air pressure inside the mixing vessel. During mixing, the system is evacuated to approximately 1.3 Pa (10^{-2} mmHg). The trapped air escapes as the rotor blades break the resulting foam. After ~ 20 min of stirring under reduced pressure, a homogeneous, air-free mix is obtained. The vacuum is then released and the mix is carefully poured into waxed metal, silicon, rubber, or Teflon molds. A more detailed description is found in the references listed above.

When mixing lung substitutes, no vacuum is applied. The components are mixed thoroughly under atmospheric pressure and then a liquid foaming agent is added (activator DC1107) and quickly stirred into the mix. The foaming action starts in ~ 30 s and the mixture must be poured in the mold to foam to the required mass density. The resulting bulk density depends on the volume of the activator added. For example, 170 kg mass of foaming agent will result in a lung substitute with a density of $\sim 250\ \text{kg}\cdot\text{m}^{-3}$.

The manufacture of water-based liquid and gel substitutes is relatively easy. The required quantity of distilled water is used and the inorganic compound necessary for the introduction of trace elements are stirred into solution one by one, ensuring that each is completely dissolved, before adding the next, thus avoiding the formation of intermediate precipitates. Urea, commonly used to satisfy the nitrogen requirements, is dissolved next, followed by any other organic liquid components. If a gel substitute is required, the water with the dissolved trace elements is heated up to $\sim 80\ ^{\circ}\text{C}$ before adding and dissolving the necessary quantity of gelatin. Once a clear uniform solution is obtained, it is left to cool to almost room temperature before the remaining components and the bacteriostatic agent are added. The mixture is usually added into polyethylene bags, heat sealed to inhibit water loss, and left to gel before use.

Polyethylene-Based Method

This method uses polyethylene powder and inorganic admixtures, $CaCO_3$ and MgO, in powdered form (17,18). The polyethylene powder has a melting point of 105 °C and density of $0.917\ \text{g}\cdot\text{cm}^{-3}$. A processing temperature is 200–240 °C. Dry mixing of the polyethylene powder and inorganic powder is performed in a long Plexiglas drum rotated on a lath, with internal Plexiglas shelves providing mixing. The mixture was then poured on iron plates that carried quadratic iron frames. Plastic plates are formed during melting at 180 °C. Homogeneous and smooth plastics are obtained with inorganic admixtures of up to 10 % of the total mass. Machining is easy to make different thickness of plates. For making thinner foils, a milling machine was provided with a vacuum fixing device.

Polymer Powder-Based Method

Homolka and Nowotny discusses the manufacturing technique of polymer powder-based phantom in a publication (20). The base materials for this method are polymer powders made of polyethylene (PE), polypropylene (PP), polystyrene (PS), and polyurethane (PU). All powders are particles of sizes much $< 100\ \mu\text{m}$. Typical additives were $CaCO_3$, MgO, TiO_2, calcium hydroxyapatite (bone mineral), and high purity graphite. These additives are available in a suitable grain size $< 100\ \mu\text{m}$. A base material is mixed with additives using a ball mill. The material then is sintered in an evacuated vessel at temperature above the melting point of the polymers. The melting temperature (softening temperature) of PE, PP, and PS are 107, 165, and 88 °C, respectively. To remove any air or other gases, a pressure of ~ 1 Pa was applied during the sintering process.

Polyurethane-Based Method

Polyurethane consists of a chain of organic units joined by urethane links. It can be made in a variety of textures and hardness by varying the particular monomers and adding other substances. It is most commonly used to produce foam rubber. Suess and Kalendar manufactured tissue equivalent phantom materials using low density polyurethane-resin (23). The resin has a high viscosity, leading to a homogeneous mixing of fillers. The stirring of the resin, hardening, and additives are done under vacuum conditions. Air bubbles are removed at pressures < 100 Pa. The ingredients are dehumidified since different degrees of humidity interfere with the polymerization and causes variations in the cured resin density. The temperature of the base materials are maintained at 20 °C before mixing and at 40 °C during curing. The density of the material is modified by adding small amounts of low density phenolic microspheres and high density poly(tetrafluorethylene) powder.

Quality Control

Quality control is necessary in order to maintain the quality of the manufactured substitutes. Two simple and effective types of investigations are usually performed, namely, mass density measurements and radiographic imaging. Casing or machining rigid solids into cubes or cylinders and measuring their mass and volume directly provides mass density data with an error of ± 0.5%. Density bottles are useful for mass density determinations of liquids and gels.

The use of X rays in the 20–50 keV energy range and computed tomography scans are simple and sensitive methods for homogeneity test on the tissue substitutes. With radiographic techniques, the smallest detectable size of high atomic number particulate fillers or trapped air pockets are ~ 100 μm. The high contract resolution of CT scanners is limited to ~ 0.6 mm, but the ability of the scanners to show low contrast differences can help in detecting unacceptably nonuniform macroscopic areas in the samples. Optical transmission microscopy of thin sample scan offers more sophisticated uniformity testing if a high degree of homogeneity is required

The uniformity of the manufactured solid substitutes may be tested by mass density determinations, multiple slice CT scanning, and conventional radiographic techniques as discussed above. The mass densities at different point in a well made sample were found to be within ±0.5% of the average value, except for lung substitutes, which show a density variation of up to ±3% of the average value.

CLASSIFICATION AND TESTING OF THE NEW TISSUE SUBSTITUTES

The available tissue substitutes were classified according to the magnitude of the discrepancy between their radiation characteristics and the radiation characteristics of the corresponding real tissues. A muscle substitute, for example, with mass attenuation and mass energy absorption coefficients within 5% of the same coefficient for real muscle, is considered as Class A substitute for photon interactions. If the discrepancy is between 5 and 20%, the substitute is called Class B material and if the error exceeds 20%, the substitute is termed Class C. In addition, material with discrepancy within 1% are called tissue equivalent and classified as A^*. A tissue substitute that is not elementally correct could be Class A for one radiation modality, but may be Class B or even Class C for another. Table 2 shows some of the recommended tissue substitutes and their components by mass, while Table 3 shows their classification for photon, electron, proton, and neutron interactions. The best results are obtained with Class A^* materials, which are elementally correct and have mass densities within ±1% of the real tissue densities. The

Table 2. Tissue Substitutes

Tissue Substitutes	Description	$Kg{\cdot}m^{-3}$	References
	Adipose Tissue		
AP/SF1	Flexible solid based on Epoxy CB3 with fillers of glucose, polyethylene, and phenolic microspheres; a four-component formula is available for trace elements	920	14
AP6	Rigid solid using low exotherm Epoxy CB4; fillers are Teflon, polyethylene, and phenolic microspheres	920	40
AP/LS	Water-based substitutes containing urea, propanol, and phospheoric acid; a four-component formula is available for trace elements	920	14
RF1	Polyethylene-based solid, fat equivalent	930	18
	Blood		
BL/L2	Water-based substitutes containing urea, ethylene glycol, and acetic acid; trace elements are available (five components)	1060	14
	Muscle		
A150	Polymer-based (electrically conducting) substitute comprising polyethylene, nylon, carbon, and calcium fluoride	1120	9, 47
Griffith urethane	Polyurethane-based material having calcium carbonate as filler	1120	10
MS/SR4	Rigid solid using Epoxy CB4 and fillers urea, polyethylene, and phenolic microspheres; five-component formula for trace elements are available	1060	14
MS20	Rigid end-product made up of Epoxy CB2 and fillers magnesium oxide, polyethylene, and phenolic microspheres	1000	40
Figerio liquid	Water-based substitute containing urea, ethylene glycol, and glycerol; a six-component formula for trace elements is available	1080	12
MS/L1	Water-based substitute containing urea, ethylene glycol, urea, and acetic acid; a six-component formula for trace elements is available	1070	14, 30
Water	H_2O	1000	2
MS/G1	A water–gelatin gel containing ethanol and, if required, a six-component formulation for trace elements	1060	14,30
MS/G2	As MS/G1, but urea and propanol replace ethanol.	1050	14,30
RM1	Polyethylene-based solid.	1030	18
	Cortical bone		
B110	A polymer-based electrically conducting, material made up of nylon, polyethylene, carbon, and calcium fluoride	1790	48
HB/SR4	Rigid end-product comprising Epoxy CB2, urea, calcium oxide, calcium hydrogen orthophosphate, magnesium sulfate, and sodium sulfate	1670	14
SB3	Rigid end-product comprising Epoxy CB2 and calcium carbonate	1790	42
Witt liquid	Saturated solution of dipotassium hydrogen orthophosphate in water	1720	49
BTES	Polymer-based material made up of Araldite GY6010, Jeffamine T403, silicon dioxide, and calcium carbonate	1400	22

Table 2. (*Continued*)

Tissue Substitutes	Description	Kg·m^{-3}	References
	Inner bone		
IB/SR1	Epoxy resin-based (CB2) solid having filers of calcium orthophosphate, polyethylene, and sodium nitrate	1150	14
IB7	Rigid solid based on Epoxy CB4; fillers are calcium carbonate polyethylene, and phenolic microspheres	1120	40
IB/L1	Water-based substitute comprising dipotassium hydrogen orthophosphate, sodium nitrate, phosphoric acid, urea, and ethylene glycol	1140	14
	Red marrow		
RM/SR4	Rigid solid using Epoxy CB4 and fillers of ammonium nitrate, polyethylene, and phenolic microspheres; a five-component formula for trace elements is available	1030	14
RM/L3	Water-based substitute containing urea and glycerol; a five-component formula for trace elements is available	1040	14
RM/G1	A water–gelatin gel containing glucose; trace elements may be added using a four-component formulation	1070	14
	Brain		
BRN/SR2	Epoxy resin-based (CB2) solid using fillers of acrylics and polyethylene; formula (five-components) for trace element is available	1040	14
BRN/L6	Water-based substitute containing urea, ethanol, and glycerol; a four-component formula for trace elements is available	1040	14, 30
A181	Polyethylene-based solid		25
	Kidney		
KD/L1	Water-based substitute containing sodium chloride, dipotassium hydrogen orthophosphate, urea, and ethylene glycol	1050	14, 30
	Liver		
LV/L1	Water-based substitute containing sodium chloride, dipotassium hydrogen sulfate, sodium chloride, urea, ethanol, and glycol	1060	14, 30
	Lung		
LN/SB4	Foamed rigid epoxy (CB4) system; fillers include urea, polyethylene and the foaming agent DC1107; a five-component formula is available for trace elements	300	14
LN10/75	Foamed rigid epoxy (CV2) system; fillers include polyethylene, magnesium oxide, phenolic, microspheres, surfactant DC200/50, and the foaming agent DC1107	310	42
LTES	Epoxy resin-based system; fillers include phenolic microspheres, surfactant, and foaming agent DC1107,	300	22
	Thyroid		
TH/L2	Water-based substitute containing urea, ethylene glycol, and acetic acid; a three-component formula for trace elements is available	1080	14, 30
	Average breast		
BR12	Rigid solid using low exotherm Epoxy CB4; fillers are calcium carbonate, polyethylene, and phenolic microspheres	970	40
AV.BR/L2	Water-based substitute containing ethanol and pentanediol	960	14
	Total soft tissue		
TST/L3	Water-based substitute containing urea, ethanol, and ethylene glycol; a five-component formula is available for trace elements	1040	14
	Total skeleton		
TSK/SF3	Flexible solid-based on Epoxy CB3; fillers include calcium hydrogen orthophosphate, calcium orthophosphate, and acrylics; a three-component formula is available for trace elements	1360	14
TSK/L1	Water-based substitute containing diammoonium hydrogen orthophosphate, dipotassium hydrogen orthophosphate, and glucose	1360	14

two-part code used indicates the type of tissue being simulated; for example, MS is muscle and BRN is brain, and whether the end product is solid flexible (SF), solid rigid (SR), liquid (L), and so on. Table 4 shows the elemental composition of the recommended substitutes.

As an example for agreement of photon interaction parameters between real tissue and a tissue mimicking material, the mass absorption coefficients for adult adipose tissue and AP6 for photon energies ranging from 10 keV to 10 MeV were calculated. The adult adipose tissue data were obtained from an ICRU report (31). The book compiles photon, electron, proton, and neutron data for body tissues. The AP6 data were calculated using the atomic composition given in Table 3 and the photon interaction data compiled in a NIST report (45). Figure 1 shows an excellent agreement of the interaction parameter between two materials. Table 3 indicates that AP6 is Class A material of the adipose tissue for the entire photon energy range.

Several experiments were carried out to verify the accuracy with which the various substitutes simulate

Table 3. Classification of Tissue Substitute

			Classification					
			Photons		Electrons	Protons	Neutrons	
Tissue being simulated	Substitute	Phase	10–99 keV	100 keV–100 MeV	10 keV–100 MeV	1–1000 MeV	1–99 keV	100 keV–30 MeV
Adipose tissue	AP/SF1	Solid	B	A	A*(A)	A	A*	A
	AP6	Solid	A	A	A	B	C	C
	AP/L2	Liquid	C	B	A(B)	A	A*	A
	AP/RF1	Solid	A					
Blood	BL/L2	Liquid	A	A	A*(a)	A*	A*	A*
Muscle	A150	Solid	C	B	B	B	A	B
	Griffith urethane	Solid	B	B	B	B	B	B
	MS/SR4	Solid	C	B	B	B	A	B
	MS20	Solid	A	A	A	B	B	B
	Figerio liquid	Liquid	A*	A*	A*	A*	A*	A*
	MS/L1	Liquid	A*	A*	A*	A*	A*	A*
	Water	Liquid	A	A(A*)	A*	A	B	B
	MS/G1	Gel	A	A*	A*	A*	A*	A*
	MS/G2	Gel	A	A*	A*	A*	A*	A*
	MS/RM1	Solid	A					
Cortical bone	B110	Solid	A*	A	A	B	B	B
	HB/SR4	Solid	B	A	A	B	C	C
	SB3	Solid	A	A	A	A	C	B
	Witt liquid	Liquid	A(B)	A	A	A	C	C
	BTES	Solid	A*	A*				
Inner bone	IB/SR1	Solid	B	B	B	B	A	B
	IB7	Solid	A	B(A)	A(B)	A	B	C
	IB/L1	Liquid	B	A(B)	A	A	A*	A
Red marrow	RM/SR4	Solid	C	B(A)	A(B)	A	A*	A
	RM/L3	Liquid	C	B	A(B)	B	A*	B
	RM/G1	Gel	C	B	A(B)	B	A*	B
Brain	BRN/SR2	Solid	C	B	B	B	A	B
	BRN/L6	Liquid	A*	A*	A*	A*	A*	A*
	A181	Solid					A	A
Kidney	KD/L1	Liquid	A	A*	A*	A*	A*	A*
Liver	LV/L1	Liquid	A*	A*	A*	A*	A	A*
Lung	LN/SB4	Solid	C	B	B	B	A	B
	LN1	Solid	A	A	A	B	C	C
	LN10	Solid	A	B(A)	A(B)	B	A	B
	LTES	Solid	A*	A*				
Thyroid	TH/L2	Liquid	A(B)	A*(B)	A*	A*	A*	A*
Average breast	BR12	Solid	A	A	A	B	C	C
	AV.BR/L2	Liquid	A*	A*	A*	A*	A*	A*
Total soft tissue	TST/L3	Liquid	A*	A*	A*	A*	A*	A*
Total skeleton	TSK/SF3	Solid	A	A	A	A	A	B
	TSK/L1	Liquid	B	A*(B)	A*(B)	A	B	B

the corresponding real tissues. In one such series of experiments, thin-walled cells with 10 × 10 cm^2 cross section were filled with real tissues and immersed in to appropriate tissue equivalent liquid to displace equal volume of that liquid. Central axis depth doses and beam profiles were then measured in the liquid behind the cells and the results compared with those obtained in the liquid alone. The material used or the construction of the cells was solid muscle substitute for the comparison with human muscle, beef stake, and pork, Brain substitute was used for the comparison with human brain. Co-60 γ source was used for the irradiation experiments. In no case did the readings at the depth differ by $> 1\%$ in each comparison. Similar tests were made with a 160 MeV synchrocyclotron proton beam and a neutron beam with average energy of 7.5 MeV (14,46). The results with the substitutes in place were generally within 0.5% of the readings for real tissues. In another series of tests, the relationship between the attenuation coefficients and the Hounsfield units measured with a computed tomography (CT) scanner was established first using 120 kVp X rays, and then the CT numbers were measured for a range of the new tissue substitutes. The attenuation coefficients derived from the measured CT number of each material was compared

Table 4. Elemental Compositions of the Tissue Substitute

	Elemental Composition (percentage by weight)											
Tissue Substitutes	H	C	N	O	Na	Mg	P	S	Cl	K	Ca	Other Elements
Adipose tissue												
AP/SF1	11.96	75.50	0.80	11.11	0.05		0.02	0.07	0.45	0.03	0.02	
AP6	8.36	69.14	2.36	16.94					0.14			F(3.07)
AP/L2	12.12	29.29	0.80	57.40	0.05	0.002	0.18		0.12	0.08	0.002	
AP/RF1	14.11	84.07		0.92		0.30					0.60	
Blood												
BL/L2	10.01	9.82	2.91	76.37	0.18	0.002		0.20	0.27	0.14	0.004	
Muscle												
A150	10.10	77.60	3.50	5.20							1.80	F(1.70)
Griffith urethane	9.00	60.20	2.80	26.60							1.72	Sn(0.01)
MS/SR4	9.50	70.28	3.48	15.55	0.08/	0.02	0.18	0.50	0.12	0.30	0.01	
MS20	8.12	58.35	1.78	18.64		13.03			0.09	0.39	0.01	
Figerio liquid	10.20	12.30	3.50	72.89	0.07	0.02	0.20	0.32	0.08	0.39	0.01	
MS/L1	10.20	12.30	3.50	72.90	0.07	0.02	0.20	0.32	0.09	0.39	0.01	
Water	11.19			88.81								
MS/G1	10.20	12.51	3.50	73.00	0.07	0.02	0.20		0.09	0.39	0.01	
MS/G2	10.35	12.31	3.50	73.04	0.07	0.02	0.20		0.09	0.39	0.01	
MS/RM1	12.24	73.36		6.37		6.03					2.00	
Cortical bone												
B110	3.70	37.10	3.20	4.80							26.29	F(24.39)
HB/SR4	4.45	29.09	3.88	31.93	0.06	0.21	10.00	0.32	0.06		19.99	
SB3	3.10	31.26	0.99	37.57					0.05		27.03	
Witt liquid	4.70			56.80			10.90			27.90		
BTES	4.0	37.8	1.5	35.3					0.1		9.4	Si(11.9)
Inner bone												
IB/SR1	8.73	63.19	2.36	17.83	0.06		2.62		0.12		5.09	
IB7	6.86	59.01	2.08	24.12					0.12		7.81	
IB/L1	8.65	17.27	2.58	60.83	0.06		2.49			4.99		
Red marrow												
RM/SR4	10.08	73.57	2.16	13.77	0.01	0.003	0.03	0.14	0.11	0.15		
RM/L3	10.17	12.77	2.22	74.24	0.08		0.03	0.15	0.17	0.17		
RM/G1	10.20	9.38	2.36	78.18	0.08		0.03	0.15	0.17	0.17		
Brain												
BRN/SR2	10.69	72.33	1.28	14.59	0.18	0.01	0.36		0.06	0.30	0.01	
BRN/L6	10.68	15.14	1.29	71.67	0.18		0.34	0.17	0.23	0.30		
A181	10.7	80.3	2.2	3.3							1.8	F(1.7)
Kidney												
KD/L1	10.40	11.35	2.74	74.50	0.18		0.19		0.28	0.25		
Liver												
LV/L1	10.18	14.40	2.83	71.80	0.11			0.24	0.18	0.29		
Lung												
LN/SB4	9.70	70.26	2.80	16.30	0.17	0.01	0.12	0.22	0.11	0.19	0.01	Si(0.50)
LN1	6.00	51.44	4.29	30.72								Al(7.55)
LN10	8.38	60.40	1.68	17.28		11.4			0.15			Si(0.84)
LTES	7.0	57.4	2.1	22.4		9.3	1.7			9.1		
Thyroid												
TH/L2	10.01	13.58	2.20	73.52	0.22		0.08		0.14	0.19		I(0.06)
Average Breast												
BR12	8.68	69.95	2.37	17.91					0.14		0.95	
AV.BR/L2	11.79	37.86		50.41								
Total soft tissue												
TST/L3	10.46	23.33	2.59	62.54	0.11	0.01	0.13	0.20	0.13	0.20	0.02	
Total skeleton												
TSK/SF3	7.16	45.50	3.08	26.12	0.31	0.12	7.02	0.16	0.47	0.15	10.03	
TSK/L1	7.45	4.64	2.94	66.93	0.32		7.00		0.13	10.15		

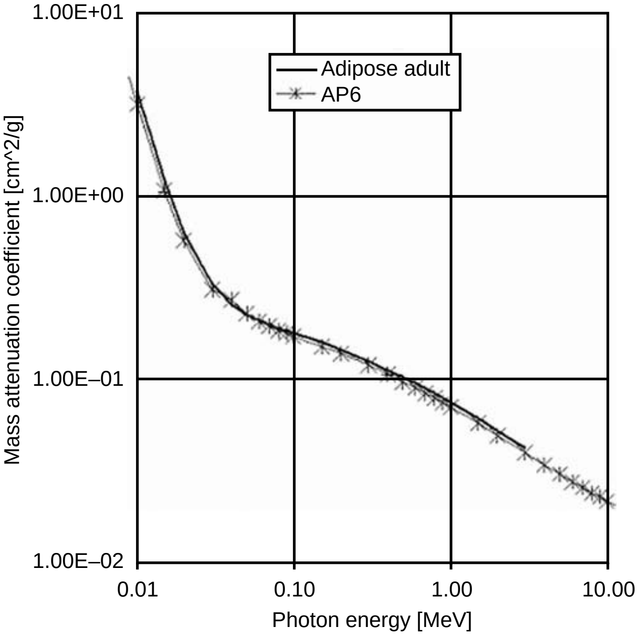


Figure 1. Comparison of mass attenuation coefficient for adult adipose tissue and tissue-mimicking material AP6.

to the calculated μ value of each material. The measured μ values were generally within 2% of the computed ones (14).

APPLICATIONS: RADIATION DOSIMETRY

Radiation exposure is equal to the number of electric charges liberated by interaction of photons with air. The gold standard for exposure measurement is a free-air chamber. The chamber size must be large enough to stop photons and associated secondary electrons. Since the size could be very large for practical uses, physicists developed small cylindrical chambers called thimble chambers by making the chamber wall with air-equivalent material (47–49). Graphite, Bakelite ($C_{43}H_{38}O_7$), or mixture of those is commonly used as the wall material. These have a smaller effective atomic number than that of air; but, it is accepted because the central electrode of the ionization chamber is usually made of aluminum, whose atomic number, 13, is much larger than air (or 7.67).

Radiation dose absorbed in tissue is the most important physical parameter for therapeutic applications of radiation. It can cause fatal effects on a person or may fail to kill the malignant cells of a patient unless the delivered dose is carefully monitored. The mail instrument for dose measurement is the ionization chamber. The most common ionization chambers for dose measurement are cylindrical with an outer diameter of $\sim$ 1 cm and a length of air cavity of $\sim$ 2 cm. When the ionization chamber is used in a solid phantom or water, some corrections are needed to estimate the dose in the medium because of differences in materials of air, the chamber wall, and the phantom material (50).

Absorbed dose in real patients can be measured by placing radiation detectors on or inside the patient during treatment. Thermoluminescent detector (TLD), a solid-state detector, is a well-established instrument for the *in vivo* dose measurements. The TLD is made of thermoluminescent material, which absorbs radiation energy. The electric charges created in the material can be measured by heating the chip after irradiation and used to estimate the absorbed dose. In addition to the thermoluminescent characteristics, TLD should be radiologically equivalent to tissue. The common TLD material is lithium fluoride (LiF) with some impurities such as Magnesium (Mg) or Titanium (Ti) to improve the property. The effective atomic number is kept close to the tissue, or 8.2, for LiF TLD to minimize the fluence disturbance due to a foreign material placed in tissue. Since the TLD material is not identical to tissue, the response to radiation is different for that of tissue. A great concern exists when absorbed dose for low energy photons, or energies $<$ 100 keV, must be measured because the radiation response is very different from the tissue in this energy range (47).

APPLICATIONS: RADIATION THERAPY

Accurate prediction of radiation dose delivered to patients is the most important step to generate effective radiotherapy treatment strategies. Medical physicists, who are responsible for physical aspects of the treatment planning, have to establish the physical data necessary for radiotherapy before anyone can be treated and even during actual treatment. For given radiation sources, medical physicists have to know how much radiation energy or dose is absorbed at any point in the patient's body. Generally, computers are used to predict the dose distribution. The computer can calculate doses using radiation source specific physical data incorporated in the software. The necessary data vary according to the calculation model used by the computer program. In general the absolute dose at any point or at any depth in the human body and the relative dose in the entire body are needed.

Medical physicists have developed many physics concepts since radiation sources were introduced into the clinic. For radiotherapy using electron accelerators, medical physicists introduced three important concepts: output factor, depth dose, and beam profile (48,49). The output factor is the radiation dose at a specific depth along the central beam axis or at a reference point for a given standard field size, (e.g., 10 × 10 cm). The depth dose gives the dose to a point at a given depth as a fraction of the dose at the reference point along the central beam axis. The beam profile shows the variation of dose along a line on a plane perpendicular to the central beam axis.

Physical models of the radiation source and human body are not perfect. Dose calculations in a real human body are difficult because of the variation in shape and tissue densities. Consequently, necessary data must be measured. The measurements are generally performed in a uniform and large medium, such as a water phantom.

Tissue Equivalent Phantom

Water. Water is a favorite medium for dose measurements for several reasons. Water is nearly tissue equivalent and inexpensive (or readily available). Furthermore, a radiation detector can be scanned through in water for dose

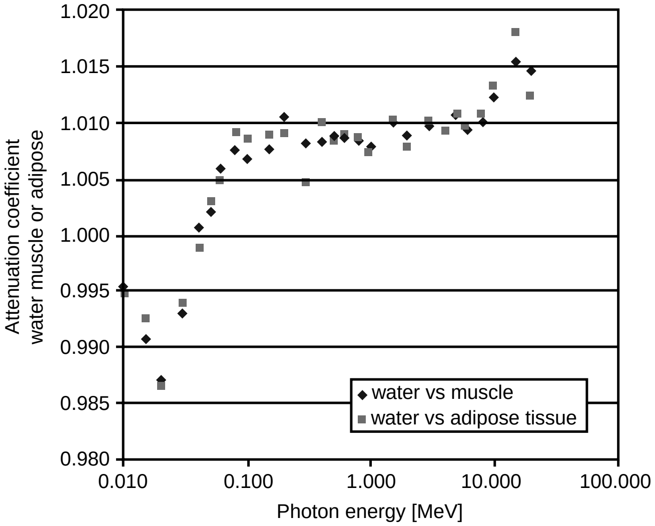


Figure 2. The ratio of mass attenuation coefficients between water and muscle or adipose tissue versus photon energy.

measurements at points spaced very closely. One should remember, however, that the equivalence of water to tissue depends on the photon energy and the tissue to which water is compared. Mass attenuation coefficients of water, muscle, and adipose tissue were evaluated for photons in the energy range between 1 keV and 20 MeV. The ratios of water to muscle and water to adipose tissue are plotted in Fig. 2. The mass attenuation coefficients of water are 1 % larger than those of muscle and adipose tissue for energies > 0.1 MeV. However, in the energy range between 0.01 and 0.1 MeV the mass attenuation coefficients of water are smaller as much as 1.5% with those of muscle and adipose tissue.

Solid Phantom. Liquid water is cumbersome to use in a high electric voltage environment such as in an accelerator room. Hence, medical physicists manufactured water-equivalent solid phantoms. There are a few solid phantom materials currently available commercially. The list of products is given in Table 5. Poly (methyl methacrylate) or acrylic is sold under names of Lucite, Plexiglas, or Perspex. Polystyrene has usually clear color and it is common phantom material for physics quality assurance programs because of relatively low price. Physicists discovered that PMMA and polystyrene leads to as much as 5% error when those are used for the absolute dose measurements of electron beams. Hence, an epoxy resin-based solid water material was developed as a substitute for water (16). It was made to simulate radiological characteristics of water more accurately than PMMA and polystyrene. Both mass attenuation coefficients and electron stopping powers of the solid water agree with those of water within 1% for the energy range between 10 keV and 100 MeV (51).

The new generation of solid water is being manufactured by Gammex-RMI (Middleton, WI). The phantoms are available in slabs of various sizes and thickness as seen in Fig. 3. Plastic Water was developed by CIRS (Computerized Imaging Reference Systems Inc., Norfolk, VA). It is a variation of the solid water. It is flexible. It was shown that measured outputs of electron beams in the Plastic Water agree with water within 0.5% for energy ranging from 4 to 20 MeV (52). Med-Tec, Inc. (Orange City, IA) is manufacturing what they call Virtual Water, which has the same chemical composition and density as solid water, but the manufacturing process is different. Hermann et al. developed polyethylene-based water-equivalent solid material (17). The material is manufactured by PTW (Freiburg, Germany) and it is sold as White Water or RW3. It contains titanium oxide as additive.

The dose at a depth (5 or 7 cm for photon beams and the depth of dose maximum for electron beams) in solid phantoms was measured with ionization chambers for various photon and electron energies. The results were compared with the dose measured in water. Figure 4 shows the ratio of measured doses in a solid phantom and water for plastic water (PW), white water (RW-3), solid water (SW-451 and SW-457), and virtual water (VW) (52,53). The horizontal axis of the figure indicates the photon energy. The larger the ionization ratio, the higher the photon energy. The beam energy ranges from Co-60 γ ray (1.25 MeV) to 24 MV. The solid water and virtual water show the best agreement

Table 5. Elemental Compositions of Common Water-Equivalent Phantom Materials

		Elemental Composition (percentage by mass)												
Tissue		H	B	C	N	O	Mg	Al	Cl	Ca	Other Elements	Mass Density, kg·m^{-3}	Z_{eff}	N, e·kg^{-1} × 10^{26}
PMMA[a] (Acrylic)	Lucite, Plexiglas, Perspex	8.0		60.0		32.0						1170	6.24	3.25
Polystyrene		7.7		92.3								1060	5.69	3.24
Solid water 457	GAMMEX-RMI	8.1		67.2	2.4	19.9			0.1	2.3		1042	8.06	3.34
Virtual water	MEDTEC	8.1		67.2	2.4	19.9			0.1	2.3		1070	8.06	3.48
Plastic water	CIRS	7.4	2.26	46.7	1.56	33.52	6.88	1.4	0.24			1030	7.92	3.336
White water RW3	PTW	7.61		91.38		0.14					Ti(0.78)	1045	5.71	3.383
Polymer gel	MGS	10.42		10.45	2.44	76.68						1050	7.37	3.49
Water		11.19				88.81						1000	7.42	3.343

[a] Poly (methyl methacrylate = PMMA.)

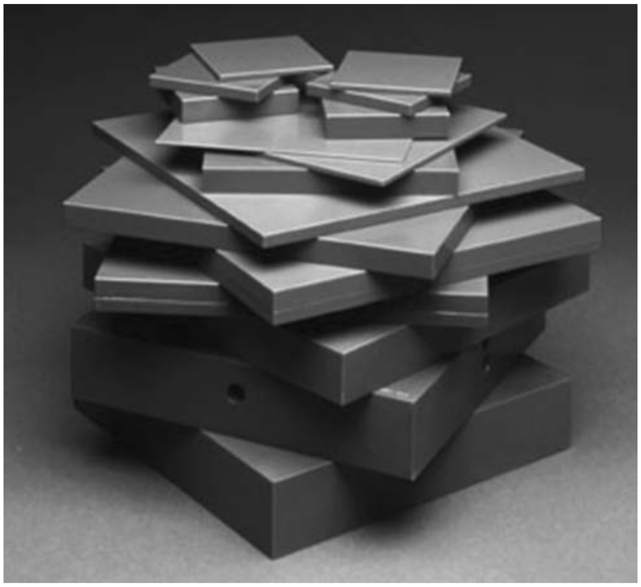


Figure 3. Solid water phantom manufactured by GAMMEX-RMI (Middleton, WI).

among the evaluated materials. The results for electron beams are given in Fig. 5, where the horizontal axis indicates the effective energy, which is an approximation of the beam energy. The solid water and virtual water also show a good agreement for the entire electron energy.

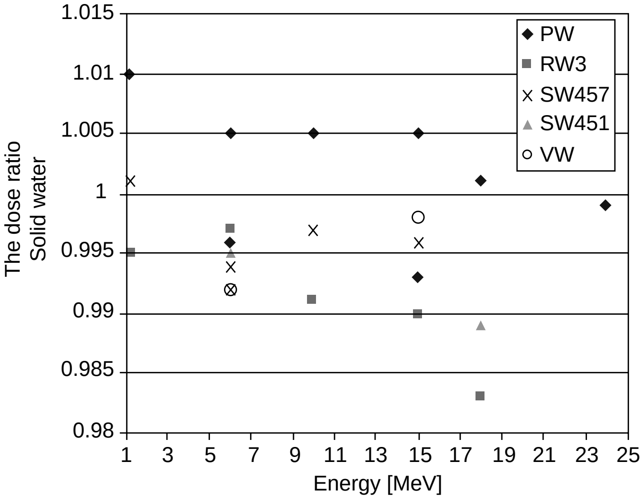


Figure 4. The ratio of measured output in solid phantom and water for photon beams.

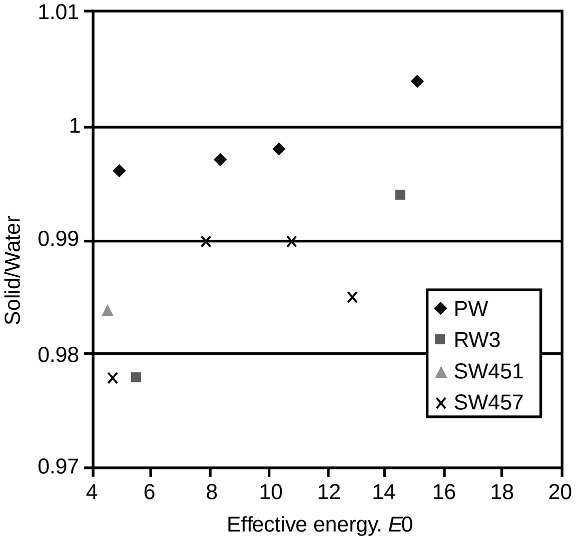


Figure 5. The ratio of measured output in solid phantom and water for electron beams.

Gel Phantom. A relatively new development is taking place on a tissue-equivalent material in gel form. The new material can record the radiation dose without additional instrument inserted in the phantom. The phantom can be made large enough to simulate the human body. Among many variations of gel phantoms, the most promising is polymer gel manufactured by MGS Research Inc. (Guilford, CT) (54). Originally it was made of acrylamide, *N*, *N*′-methylenebisacrylamide (bis), agarose, and water. The chemical composition was optimized over years. Currently, the polymer gel is sold as BANG3 with 80% water, 14% of gelatin, and 6% of methacrylic acid by mass. A typical atomic composition of the polymer gel as well as the mass density, the effective atomic number, and the electron density are given in Table 5. The effective atomic number and the electron density of the polymer gel agree with those of water very well. The polymer gel produces highly linked polymers when it is irradiated. The structural change causes a change in color and mass density. The recorded dose distribution can be indirectly read by measuring the photon attenuation of both visible light and X ray. The polymerization also causes a change in the magnetic property of the gel. The most popular method is currently to scan the irradiated phantom with a magnetic resonance imaging (MRI) scanner. It takes advantage of the change in the spin–spin relaxation rate, which increases with increasing absorbed dose.

Differing from more traditional dose measurement tools, polymer gel dosimeter enables medical physicists to obtain full three-dimensional (3D) dose distributions in a geometrically consistent way. Polymer gel phantom was used to measure 3D dose distributions of advanced radiotherapy treatment such as those for intensity modulated radiation therapy (IMRT) (55) and Gamma Knife stereotactic radiosurgery (56). Figure 6 shows an MRI image taken after the polymer gel was irradiated with a Gamma Knife system (Elekta AB, Stockholm, Sweden). The blighter color indicates higher dose.

Geometric Phantom

With the advent of rapid development of highly conformal radiation therapy, modern radiation therapy requires high geometrical precision of radiation delivery. At the same time, the complexity of treatment planning software has substantially increased. This has urged medical physicists to test the geometrical precision of the treatment planning software (57). Special phantoms are being manufactured to assure the quality of geometry created by the software.

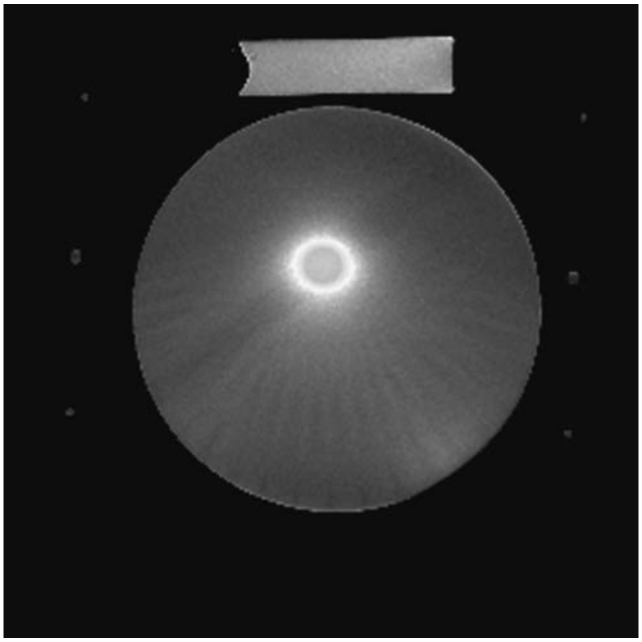


Figure 6. Dose distribution images obtained by polymer gel dosimeter for Gamma Knife irradiation.

Here we discuss two of those phantoms currently on market.

Lucy 3D Precision Phantom. Lucy phantom was developed by Toronto Sunny Brook Regional Cancer Center and Sandstrom Trade and Technology, Inc. (Ontario, Canada) (58). The acrylic phantom of a shape of head (Fig. 7) is used to test the image quality of CT, MRI, and X ray imaging modalities, which are used for stereotactic radiation therapy. It can verify imaging errors, image distortions, and the geometrical accuracy of treatment planning system. It serves also as routine machine QA equipment.

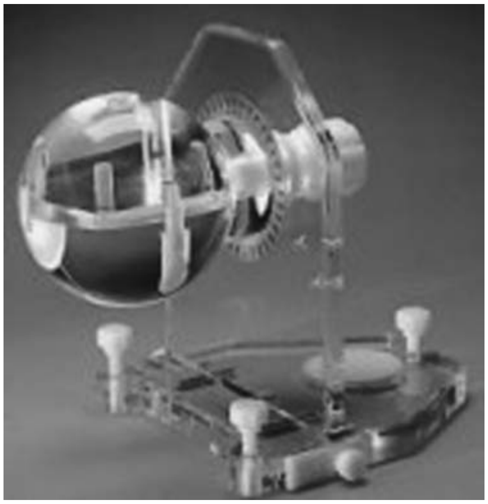


Figure 7. Lucy 3D precision phantom from the Sandstrom Trade and Technology Inc. (Ontario, Canada).

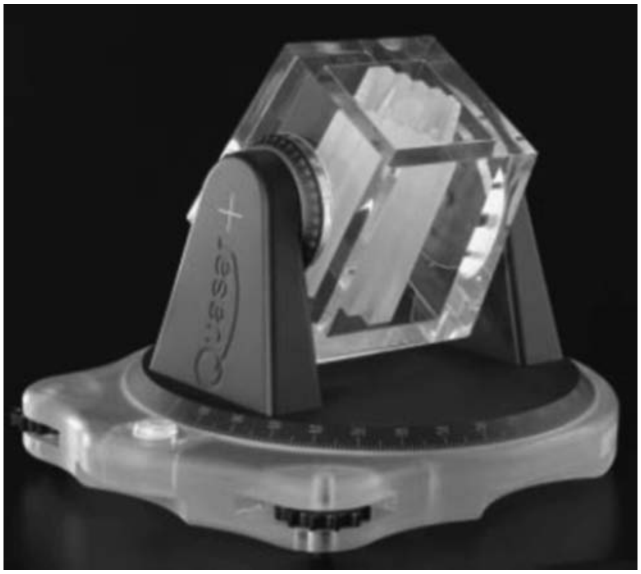


Figure 8. Quasar beam geometry MLC phantom from Modus Medical Devices, Inc. (Ontario, Canada).

Quasar Phantom. Another Canadian company, Modus Medical Device, Inc. (Ontario, Canada), developed phantoms to verify geometrical accuracy generated by radiation therapy treatment planning software (59). The phantoms are made of Lucite, cedar, and polystyrene. One of those, Quasar body phantom, can be used to test the accuracy of the digitally reconstructed radiography (DRR) images. DRR images are generated from a number of axial images taken by a CT scanner and the DRR images are compared with the images taken with an electronic portal imaging system or with radiographic port films before treatment to verify the patient position accuracy. Quasar beam geometry MLC phantom shown in Fig. 8 can verify the geometrical display accuracy of leaf position and the size of the multileaf collimator, which have replaced traditional blocks to define radiation fields.

Humanoid Phantom

Complex high precision radiation delivery techniques, such as IMRT and stereotactic radiation therapy, demand ever increasing accuracy of dose delivery in geometry close to the human body. Body phantoms such as Rando and Alderson were introduced many years ago. Those were used to predict to radiation exposure to humans during radiation therapy and diagnostic radiology procedures. The Rando phantom manufactured by the Phantom Laboratory (Salem, NY) is shown in Fig. 9. The body phantom is sliced into many 5 cm thick slabs. Radiographic films can be inserted between the slabs for dose measurements. The slabs can also hold TLD chips for absolute dose measurements. The phantom is made of soft-tissue equivalent material, which is manufactured with proprietary urethane formulation. The phantom uses natural human skeletons. Lungs and breast closely mimic the real tissues.

Recently, many phantoms which simulate a part of body were developed, for example, Gammex-RMI, Phantom

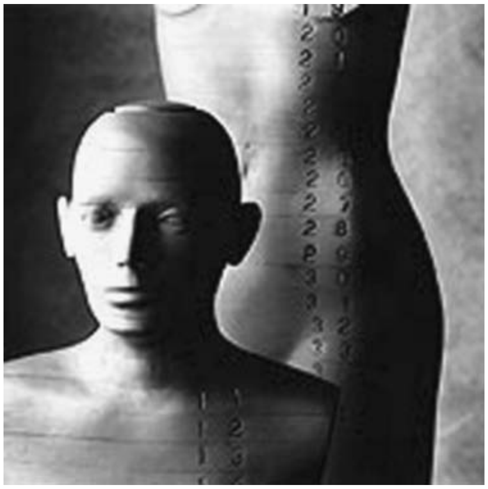


Figure 9. Rando phantom manufactured by the Phantom Laboratory (Salem, NY).

Laboratory, CIRS (60), and so on. Many of those phantoms are made to measure dose inside the phantom for verification of radiation therapy treatment. There are phantoms in various shapes for IMRT QA. Those phantoms can accommodate ionization chambers, radiographic films, and TLDs for dose measurements. Those may have special inserts for nonsoft tissue materials, such as lung and bone. The shapes of the phantoms are torso, head, thorax, pelvis, and neck.

DIAGNOSTIC IMAGING

X Rays

X rays are used to take images of parts of the body for diagnostic purposes (61). The oldest and most commonly used X ray imaging modality is X ray radiography. Images can be recorded on radiographic films. Recently, digital recording has become more common since the digital technique requires no wet-film processing and allows radiologists to manipulate and store the images more easily than hard-copy films.

Medical physicists use phantoms to test the quality of images. Some of common phantoms are dual-energy X ray absorptiometry phantom, anatomical phantoms (62), digital subtraction angiography (DSA) phantom, contrast detail phantom, and dental image QA phantoms.

Fluoroscopy

A regular X ray device can take static images of patients. A fluoroscopic unit, consisting of an X-ray tube, a camera, an image intensifier, and a TV monitor, can record images of moving parts of the body and of objects placed inside the body. This method of recording images is called fluoroscopy. Interventional radiologists use fluoroscopy to monitor the location of very thin wires and catheters going through blood vessels as those are being inserted during a procedure. Phantoms are an important instrument to assure the quality of fluoroscopy images. There are fluoroscopic phantoms, such as cardiovascular fluoroscopic benchmark phantom, fluoroscopy accreditation phantom, and test phantom.

Mammography

Mammography is a major diagnostic modality to detect breast cancer. The tumor size is often very small, that is, submillimeter diameter. Though small, early discovery of such small tumors is very important and can lead to better therapy outcomes, that is, higher cure rates and longer survival. Hence, the quality of X ray equipment used for mammography is a key for the success of this imaging modality and its performance is tightly controlled by federal and local governments. Consequently, medical physicists developed many phantoms to evaluate the imaging quality of mammography devices accurately and efficiently. Several phantoms are used, known as QA phantom, accreditation phantom, high contrast resolution phantom, contrast detail phantom, digital stereotactic breast biopsy accreditation phantom, phototimer consistency test tool, and collimator assessment test tool. A tissue-equivalent phantom manufactured by CIRS (Norfolk, VA) is shown in Fig. 10. The phantom is 4.5 cm thick and simulates the shape of a breast during the imaging. It is made of CIRS resin material mimicking the breast tissue. Objects with varying size within the phantom simulate calcifications, fibrous tissue in ducts, and tumor masses.

Computed Tomography

Computed tomography scanning systems were developed in the 1970s and rapidly deployed into clinics into the 1980s. Currently, this imaging modality is commonly used for diagnosis and radiation therapy in clinic and hospitals. Compute tomography provides patient's anatomy on planes transverse to body axis (from head to toe). The plane images called axial slices are taken every 0.05–1 cm. Depending on the axial length of the scan, the number of slices can vary from 20 to 200. Computed tomography

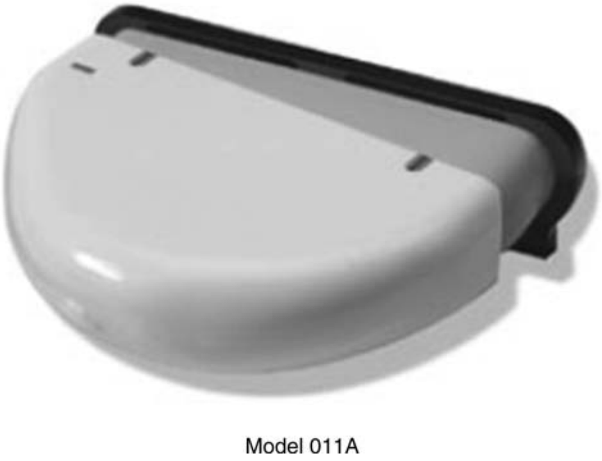


Figure 10. Tissue-equivalent mammography phantom from CIRS (Norfolk, VA).

enables radiologists to visualize the location of disease in 3D, leading to potentially more accurate diagnosis. Medical physicists use phantoms to estimate dose to patients during CT scan, to test the precision of scanning geometry, and verify the quality of CT images. Such phantoms are the bone mineral analysis phantom, spiral/helical CT phantom, CT dose phantom, orthopedic calibration phantom, spine phantom, electron density phantom, and CT performance phantom.

Each picture element (pixel) of a CT image is represented by a CT number (or Hounsfield units). The CT number for water is 0, while the CT numbers in the lung are below -200 and those in the bones are ~ 200. Since the CT number is found to depend on the electron density, among other parameters, it can be used to identify the material type and the distribution of the tissue inhomogeneities, which is very useful in radiation therapy treatment planning. One electron density phantom, developed in order to establish the relationship of CT number and tissue electron densities, is made of solid water and has cylindrical shape with a radial size of pelvis and 6 cm thickness. Small cylindrical inserts mimicking different tissues are plugged into the phantom. When it is CT scanned, the measured CT numbers of those materials can be plotted against the predetermined electron densities and the resulting data are used in treatment planning computer. This electron density CT phantom, which is now commercially available, shown in Fig. 11, was designed by Constantinou (63) and manufactured by GAMMEX-RMI (Middleton, WI). Available insert materials are lung (LN-300 and LN-450), adipose (AP6), breast, CT solid water, CB3 resin, brain, liver (LV1), inner bone, bone (B200, CB2-30% mineral, CB2-50% mineral), cortical bone (SB3), Titanium simulating Titanium implants for hip replacements, and true water.

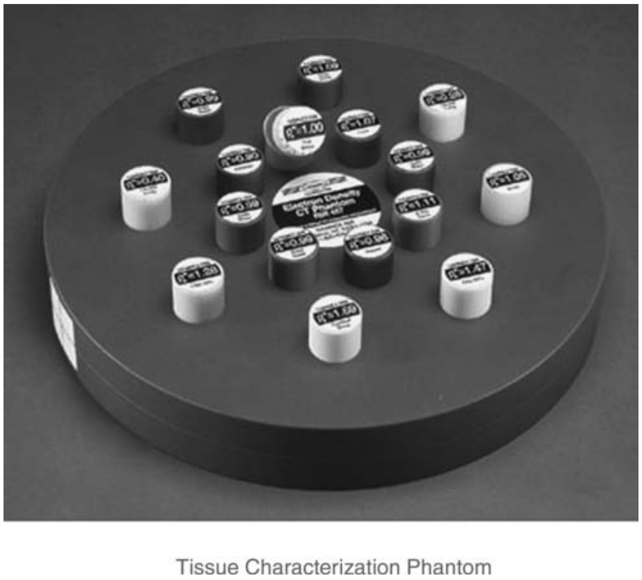


Figure 11. Electron density phantom RMI 467 from GAMMEX-RMI (Middleton, WI).

NUCLEAR MEDICINE

For nuclear medicine procedures, radioactive compound composed from Technetium, Iodine, Fluorine, etc. is injected into blood stream (64). The material is carried to a potential disease site by blood and stays there. Radioactive material emits photons or positively charged electrons called positrons. Since photons can easily escape the patient body, the photons can be detected by a photon detecting device. The photons are used to form images. Positron emission tomography (PET) uses positron emitting radioactive material. When a positron interacts with an electron in the body, both particles are converted into photons. The PET is more advanced nuclear medicine procedure and can generate 3D distributions of radioactive material. Phantoms are used for quality assurance of nuclear medicine systems. A NEMA (National Electrical Manufacturers Association) scatter phantom manufactured by CIRS Inc. (Norfolk, VA) is a circular polystyrene cylinder. Radioactive material with known activity is delivered to line source insert tubes. The image of the phantom is taken to test scatter fraction and count losses.

FUTURE

Tissue-equivalent materials for solid phantoms are well developed. There may be a need of incremental improvement in phantom materials possibly for lower price and more accurate representation of tissue types. Furthermore, phantom materials are needed for newer radiation types such as heavy ions. As new therapy and imaging modalities will be put into practice, new phantoms will be developed for efficient and accurate testing of those new tools. Modern radiology utilizes not only ionizing radiation such as photons, electrons, and heavy particles, but also electromagnetic radiation (or EM-waves) and ultrasound. Phantom materials exist for testing ultrasound devices and MRI scanners. This is an active area for development. For example, D'Souza recently developed a phantom used for quality testing of ultrasound, MRI, and CT (65). Readers interested in those phantoms should consult with the phantom manufactures, such as GAMMEX-RMI (Middleton, WI), CIRS, Inc. (Norfolk, VA) and Fluke/Cardinal Health (Cleveland, OH).

Here, the focus is on two aspects of interest for future development. It may be necessary to develop more biologically tissue equivalent phantom materials. Such materials simulate not only the radiological characteristics of radiation modalities, but also it can closely simulate the radiation effects on the tissues in realistic geometry. Biomedical engineers and scientists are vigorously working on artificial tissue, potentially replacing the real tissue. Such material could be also used as phantom material. Readers should refer to a comprehensive review on the current state of art in tissue engineering authored by Lanza et al. (66).

Computer modeling of human body is an active field of research and development. Noticeable example is the Visual Human Project sponsored by National Library of Medicine, NLM (67). Whole bodies of male and female cadavers were scanned with CT and MRI. The datasets are available for anyone who is interested in the information through licensing with NLM and with minimal cost. The datasets for the male consist of 12 bits axial MRI images of the head and neck and up to 1871 axial CT slices of the whole body. The female data set is 5000 axial CT images, with which one can reconstruct 3D images with $0.33 \times 0.33 \times 0.33$ mm cubic voxels. The data can be used to construct human body model for quality assurance of radiological systems, in particular, for testing the radiation therapy treatment planning system. In addition to real geometry, the data contain the detailed information of tissue heterogeneities in human body. Such data are indispensable for accurate assessment of new radiological technologies.

BIBLIOGRAPHY

1. Spires FW. Materials for depth dose measurement. Br J Radiol 1943;16:90.
2. Kienbock R. On the quantimetric method. Arch Roentgen Ray 1906;1:17.
3. Baumeister L. Roentgen ray measurements. Acta Radiol 1923;2:418.
4. Ott P. Zur Rontgenstrahlenbehandlung oberblachlich gelagerter Tumoren. Strahlentherapie 1937;59:189.
5. Jones DEA, Raine HC. Bri J Radiol 1949;22.
6. Harris JH, et al. The development of a chest phantom for radiologic technology. Radiology 1956;67:805.
7. Markus B. The concept of tissue equivalence and several water-like phantom substances for energies of 10 KeV to 100 MeV as well as fast electrons. Strahlentherapie 1956; 101:111–131.
8. Rossi HH, Failla G. Tissue equivalent ionization chamber. Nucleonics 1956;14:32.
9. Shonka FR, Rose JE, Failla G. Conducting plastic equivalent to tissue, air and polystyrene. Prog Nucl Energy 1958;12:184.
10. Griffith RV, Anderson AL, Dean PN. Further realistic torso phantom development. University of California Research Laboratory, UCRL-50007-76-1; 1976.
11. Goodman LJ. A modified tissue equivalent liquid. Health Phys 1969;16:763.
12. Frigerio NA, Sampson MJ. Tissue equivalent phantoms for standard man and muscle. Argonne National Laboratory, Argonne, IL. ANL-7635; 1969.
13. White DR. The formulation of substitute materials with predetermined characteristics of radiation absorption and scattering. Ph.D. dissertation University of London, London; 1974.
14. Constantinou C. Tissue substitutes for particulate radiations and their use in radiation dosimetry and radiotherapy. Ph.D. dissertation, University of London, 1978.
15. White DR, Constantinou C. Prog Med Radiat Phys 1982;1:133.
16. Constantinou C, Attix FH, Paliwal BR. A solid water phantom material for radiotherapy x-ray and gamma-ray beam calibrations. Med Phys 1982;9:436–441.
17. Hermann KP, Geworski L, Muth M, Harder D. Polyethylene-based water-equivalent phantom material for x-ray dosimetry at tube voltages from 10 to 100 kV. Phys Med Biol 1985;30: 1195–1200.
18. Hermann KP, et al. Muscle- and fat-equivalent polyethylene-based phantom materials for x-ray dosimetry at tube voltages below 100 kV. Phys Med Biol 1986;31:1041–1046.
19. Kalender WA, Suess C. A new calibration phantom for quantitative computed tomography. Med Phys 1987;14:863–886.
20. Homolka P, Nowotny R. Production of phantom materials using polymer powder sintering under vacuum. Phys Med Biol 2002;47:N47–52.
21. Homolka P, Gahleitner A, Prokop M, Nowotny R. Optimization of the composition of phantom materials for computed tomography. Phys Med Biol 2002;47:2907–2916.
22. Jones AK, Hintenlang DE, Bolch WE. Tissue-equivalent materials for construction of tomographic dosimetry phantoms in pediatric radiology. Med Phys 2003;30:2072–2081.
23. Suess C, Kalender WA, Coman JM. New low-contrast resolution phantoms for computed tomography. Med Phys 1999;26: 296–302.
24. Iwashita Y. Basic study of the measurement of bone mineral content of cortical and cancellous bone of the mandible by computed tomography. Dentomaxillofac Radiol. 2000;29:209–215.
25. Burmeister J, et al. A conducting plastic simulating brain tissue. Med Phys 2000;27:2560–2564.
26. ICRU. Tissue substitutes in radiation dosimetry and measurement. International Commission on Radiological Units and Measurements, Bethesda, MD, ICRU Report 44; 1989.
27. Woodard HQ. The elementary composition of human cortical bone. Health Phys 1962;8:513–517.
28. ICRU. Physical Aspects of Irradiation. International Commission on Radiological Units and Measurements, Bethesda, MD, Report 10b; 1964.
29. ICRP. Reference man: anatomical, physiological and metabolic characteristics. International Commission on Radiological Protection, Stockholm, Sweden, ICRP Publication 23; 1975.
30. Constantinou C. Phantom materials for radiation dosimetry. I. Liquids and gels. Br J Radiol, 1982;55:217–224.
31. ICRU. Photon, electron, proton and neutron interaction data for body tissues, International Commission on Radiological Units and Measurements, Bethesda, (MD): ICRU Report 46; 1992.
32. White DR, Fitzgerald M. Calculated attenuation and energy absorption coefficients for ICRP Reference Man (1975) organs and tissues. Health Phys 1977;33:73–81.
33. Bewley DK. Pre-therapeutic experiments with the fast neutron beam from the Medical Research Council cyclotron. II. Physical aspects of the fast neutron beam. Br J Radiol 1963; 36:81–88.
34. Jones TD. Distributions for the design of dosimetric experiments in a tissue equivalent medium. Health Phys 1974;27: 87–96.
35. White DR. Phantom materials for photons and electrons. Hospital Physicists Association, London, S. Rep. Ser. No. 20; 1977.
36. Frigerio NA, Coley RF, Sampson MJ. Depth dose determinations. I. Tissue-equivalent liquids for standard man and muscle. Phys Med Biol 1972;17:792–802.
37. Frigerio NA. Neutron penetration during neutron capture therapy. Phys Med Biol 1962;6:541–549.
38. White DR. The formulation of tissue substitute materials using basic interaction data. Phys Med Biol 1977;22:889–899.
39. White DR. An analysis of the Z-dependence of photon and electron interactions. Phys Med Biol 1977;22:219–228.
40. Geske G. The concept of effective density of phantom materials for electron dosimetry and a simple method of their measurement. Radiobiol Radiother 1975;16:671–676.
41. White DR, Martin RJ, Darlison R. Epoxy resin based tissue substitutes. Br J Radiol 1977;50:814–821.
42. White DR. Tissue substitutes in experimental radiation physics. Med Phys 1978;5:467–479.
43. White DR, Constantinou C, Martin RJ. Foamed epoxy resin-based lung substitutes. Br J Radiol 1986;59:787–790.

44. White DR, Speller RD, Taylor PM. Evaluating performance characteristics in computed tomography. Br J Radiol 1981;54: 221.
45. Hubbell JH, Seltzer SM. Tables of x-ray mass attenuation coefficients and mass energy-absorption coefficients 1 keV to 20 MeV for elements $Z = 1$ to 92 and 48 additional substances of dosimetric interest. National Institute of Standards and Technology, Gaithersburg, (MD): NISTIR 5632; 1995.
46. Constantinou C, et al. Physical measurements with a high-energy proton beam using liquid and solid tissue substitutes. Phys Med Biol 1980;25:489–499.
47. Attix FH. Introduction to Radiological Physics and Radiation Dosimetry. New York: Wiley-Interscience; 1986.
48. Johns HE, Cunningham JR. The Physics of Radiology. 4th ed. Springfield (IL): Charles C. Thomas Publisher; 1983.
49. Kahn FM. The Physics of Radiation Therapy. 2nd ed. Baltimore: Williams&Wilkins; 1994.
50. AAPM, A protocol for the determination of absorbed dose from high-energy photon and electron beams. Med Phys 1983;10:741–771.
51. Thomadsen B, Constantinou C, Ho A. Evaluation of water-equivalent plastics as phantom material for electron-beam dosimetry. Med Phys 1995;22:291–296.
52. Tello VM, Tailor RC, Hanson WF. How water equivalent are water-equivalent solid materials for output calibration of photon and electron beams? Med Phys 1995;22:1177–1189.
53. Liu L, Prasad SC, Bassano DA. Evaluation of two water-equivalent phantom materials for output calibration of photon and electron beams. Med Dosim 2003;28:267–269.
54. Maryanski MJ, Gore JC, Kennan RP, Schulz RJ. NMR relaxation enhancement in gels polymerized and cross-linked by ionizing radiation: a new approach to 3D dosimetry by MRI. Magn Reson Imaging 1993;11:253–258.
55. Low DA, et al. Evaluation of polymer gels and MRI as a 3-D dosimeter for intensity- modulated radiation therapy. Med Phys 1999;26:1542–1551.
56. Scheib SG, Gianolini S. Three-dimensional dose verification using BANG gel: a clinical example. J Neurosurg 2002;97:582–587.
57. Fraass B, et al. American Association of Physicists in Medicine Radiation Therapy Committee Task Group 53: quality assurance for clinical radiotherapy treatment planning Med Phys 1998;25:1773–1829,.
58. Ramani R, Ketko MG, O'Brien PF, Schwartz ML. A QA phantom for dynamic stereotactic radiosurgery: quantitative measurements. Med Phys 1995;22:1343–1346.
59. Craig T, Brochu D, Van Dyk J. A quality assurance phantom for three-dimensional radiation treatment planning. Int J Radiat Oncol Biol Phys 1999;44:955–966.
60. CIRS. Phantoms, Compurerized Imaging Reference Systems, Inc., Norfolk, (VA); 2005.
61. Curry TS, Dowdey JE, Murry RCJ. Chirstensen's Physics of Diagnostic Radiology. 4th ed. Philadelphia: Lea&Febiger; 1990.
62. Constantinou C, Cameron J, DeWerd L, Liss M. Development of radiographic chest phantoms. Med Phys 1986;13:917–921.
63. Constantinou C, Harrington JC, DeWerd LA. An electron density calibration phantom for CT-based treatment planning computers. Med Phys 1992;19:325–327.
64. Sorenson JA, Phelps ME. Physics in Nuclear Medicine. 2nd ed. Philadelphia: W.B.Saunders; 1987.
65. D'Souza WD, et al. Tissue mimicking materials for a multi-imaging modality prostate phantom. Med Phys 2001;28:688–700.
66. Lanza R, Langer R, Chick W. Principles of Tissue Engineering. New York: Academic Press; 1997.
67. NLM. The Visible Human Project. [Online]. Available at http://www.nlm.nih.gov/research/visible/visible_human.html. Accessed 2003 Sept 11.

See also COBALT 60 UNITS FOR RADIOTHERAPY; RADIATION DOSE PLANNING, COMPUTER-AIDED; RADIATION DOSIMETRY, THREE-DIMENSIONAL.

PHARMACOKINETICS AND PHARMACODYNAMICS

PAOLO VICINI
University of Washington
Seattle, Washington

INTRODUCTION

Drug discovery and development is among the most resource intensive private or public ventures. The Pharmaceutical Research and Manufacturers Association (PhRMA) reported that in 2001, pharmaceutical companies spent ~ $30.5 billion in R&D, 36% of which was allocated to preclinical functions (1). Given the expense and risk associated with clinical trials, it makes eminent sense to exploit the power of computer models to explore possible scenario of, say, a given dosing regimen or inclusion–exclusion criteria before the trial is actually run. If anything, this goes along the lines of what is already commonly done in the aerospace industry, for example. Thus, computer simulation is a relatively inexpensive way to run plausible scenarios *in silico* and try and select the best course of action before investing time and resources in a (sequence of) clinical trial(s). Ideally, this approach would integrate information from multiple sources, such as *in vitro* experiments and preclinical databases, and that is where the difficulties specific to this field start.

System analysis (in the engineering sense) is at the foundation of computer simulation. A rigorous quantification of the phenomena being simulated is necessary for this technology to be applicable. Against this background, the quantitative study of drugs and their behavior in humans and animals has been characterized as pharmacometrics, the unambiguous quantitation (via data analysis or modeling) of pharmacology (drug action and biodistribution). In a very concrete sense, pharmacometrics is the quantitative study of exposure-response (2), or the systematic relationship between drug dosage (or exposure to an agent) and drug effect (or the consequences of agent exposure on the organism). Historically, there have been two main areas of focus of pharmacometrics (3). Pharmacokinetics (PK) is the study of drug biodistribution, or more specifically of absorption, distribution, metabolism, and elimination of xenobiotics; it is often characterized as "what the body does to the drug". Pharmacodynamics (PD) is concerned with the effect of drugs, which can be construed both in terms of efficacy and toxicity. This aspect is often characterized as "what the drug does to the body".

It has to be kept in mind that both pharmacokinetic and pharmacodynamic systems are "dynamic" systems, in the sense that they can be modeled using differential equations (thus, to an engineer, this distinction may seem a bit unusual). In addition, a third aspect of drug action will be discussed, disease progression, in the rest of this article.

The joint study of the biodistribution and the efficacy of a compound, or a class of compounds, is termed PK-PD, to signify the inclusion of both aspects. While the PK subsystem is a map from the dosing regimen to the subsequent time course of drug concentration in plasma, the PD subsystem is a map from the concentration magnitudes attained in plasma to the drug effect. This paradigm naturally postulates that the time course in plasma is a mediator for the ultimate drug effect, that is, that there is a causal relationship between the plasma time course and the effect time course. As seen later, this causal relationship does not exclude the presence of intermediate steps between the plasma and effect time courses (e.g., receptor binding and delayed signaling pathways).

Both PK and PD are amenable to quantitative modeling (Fig. 1). In fact, there is a growing realization that the engineering principles of system analysis, convolution, and identification have always been (sometimes implicitly) an integral part of the study of new drugs, and that realization is now giving rise to strategies that aim at accelerating drug development through computer simulation of expected drug biodistribution and efficacy time courses. Interestingly, the intrinsic complexity of PK-PD systems has sometimes motivated the applications of generalizations of standard technology, (e.g., the convolution integral) (4), through sophisticated methodological contributions that have yet to impact the engineering mainstream. Lastly, the PK-PD model is currently viewed as a perfectly adequate approach not only to process, but to extract new, quantitative information from noisy measurements. As such, the model can be construed as a probe, or a measurement tool or medical instrument in its own right, with its own built-in confidence limits and design limitations (5).

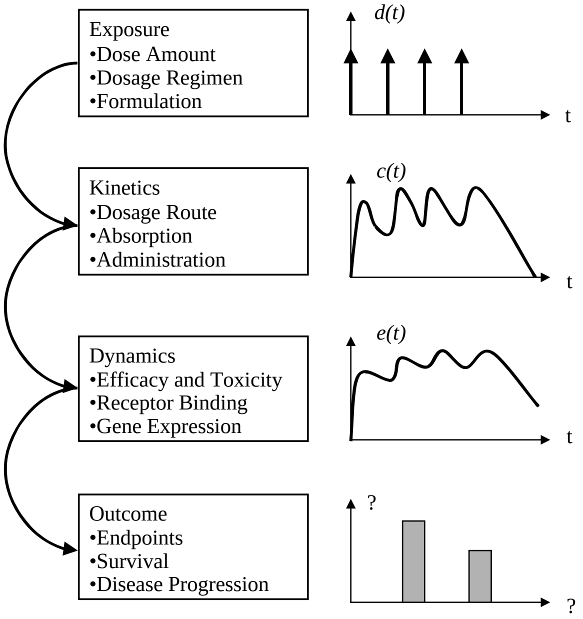


Figure 1. This figure briefly summarizes the functional pathways that exist from dosing (exposure) to outcome through drug biodistribution (pharmacokinetics) and drug effect (pharmacodynamics). it is not unusual to find a disconnect between the resolution available in the dynamic measurements and the actual clinical outcome. Also, while the convolution operator can be used to map the dosing regimen $d(t)$ to the concentration-time profile $c(t)$ and then to the effect $e(t)$, this is not necessarily true when going from the dynamic behavior to clinical measurements, which are often lower resolution and/or discrete.

Pharmacokinetics: What Shapes the Concentration–Time Curve?

The first application of mathematical models to biology is usually attributed to Teorell (6) and led to the development of the class of compartmental models whose usage is now so widespread. Teorell's work was conducted on exogenous substances, xenobiotics (e.g., drugs). Model-based analysis of exogenous substances is somewhat simpler than endogenous compounds (e.g., glucose, insulin and other hormones, where autoregulation is crucial to understanding), since endogenous fluxes that may be exquisitely sensitive to changes in circulating concentrations are absent.

Teorell's original motivation survives in modern PK-PD modeling. The PK model is typically a useful mathematical simplification of the underlying physicochemical and physiological processes. It is usually a lumped parameter, compartmental model, and describes the absorption, distribution, metabolism, and elimination (ADME) of a drug from the body (7). A practical review of simple PK models can be found, for example, in Ref. 8. As an example, the distribution of a drug in the plasma space following an intravenous injection may be well described by a monoexponential decay:

$$c(t) = C_0 e^{-\alpha t} \tag{1}$$

where $c(t)$ is the drug concentration at time t, C_0 is the concentration at time 0 and α is the decay rate. As it is straightforward to verify, this functional form of $c(t)$ is the solution to a first-order ordinary linear differential equation, or a single compartment model:

$$\begin{aligned} \dot{q}(t) &= -k_{\mathrm{el}} q(t) + D\delta(t) \\ q(0) &= 0 \\ c(t) &= \frac{q(t)}{V} \end{aligned} \tag{2}$$

where $q(t)$ is the amount of drug in the plasma space, D is the injected dose, $\delta(t)$ is the Dirac delta, and V is the drug's volume of distribution. It can be easily verified that

$$\begin{aligned} C_0 &= \frac{D}{V} \\ \alpha &= k_{\mathrm{el}} \end{aligned}$$

so that the elimination rate of the drug is equal to the decay rate of the concentration–time curve. An important pharmacokinetic parameter is the drug's clearance, or the volume cleared per unit time, which for the model above can be expressed as:

$$\mathrm{CL} = k_{\mathrm{el}} V$$

Another common pharmacokinetic parameter is the area under the concentration curve:

$$\mathrm{AUC} = \int_0^\infty c(t)dt = \frac{D}{\mathrm{CL}}$$

The AUC is probably the most common measure of exposure in pharmacokinetics, since it summarizes dosing and systematic information. The model in equation 2 could be extended to accommodate first-order absorption, for example, to model the plasma appearance of an oral or intramuscular dose as opposed to an intravenous dose:

$$\begin{aligned} \dot{q}_1(t) &= -k_\mathrm{a} q_1(t) + D\delta(t) \\ \dot{q}_2(t) &= +k_\mathrm{a} q_1(t) - k_\mathrm{el} q_2(t) \\ q_1(0) &= q_2(0) = 0 \\ c(t) &= \frac{q_2(t)}{V} \end{aligned} \tag{3}$$

where k_a is the absorption rate constant (again in units of inverse time), and $q_1(t)$ and $q_2(t)$ are the amounts in the absorption compartment and in the plasma compartment, respectively. It can be easily verified that this model also describes the convolution of a single-exponential absorption forcing function with the single exponential impulse response of the plasma space.

For the simple single compartmental model (eq. 2), the pharmacokinetic parameters can be readily estimated from the data (9), and the assumption of a mechanistic model does not affect their values. In other words, both compartmental (model-dependent) and noncompartmental (data-dependent) analyses provide the same result. However, this relatively straightforward interpretation of the eigenvalues of the system matrix in relation to the observed rate of decay is correct only for single-compartment systems such as this one. In the case when the drug diffuses in two compartments (e.g., a plasma and extravascular compartment), then its time course is described by a sum of two exponential functions:

$$c(t) = A_1 e^{-\alpha t} + A_2 e^{-\beta t} \tag{4}$$

but the corresponding ordinary linear differential equation is

$$\begin{aligned} \dot{q}_1(t) &= -(k_{10} + k_{12}) q_1(t) + k_{21} q_2(t) + D\delta(t) \\ \dot{q}_2(t) &= +k_{12} q_1(t) - k_{21} q_2(t) \\ q_1(0) &= q_2(0) = 0 \\ c(t) &= \frac{q_1(t)}{V} \end{aligned} \tag{5}$$

where the parameters are the same as before, except that now $q_1(t)$ and $q_2(t)$ are the amount of drug in the plasma and extraplasma compartments, respectively, k_{10} is the rate constant (in units of inverse time) at which the drug leaves the system, and k_{12} and k_{21} are the rate constants at which the drug is transported out of the plasma and extraplasma compartments, respectively. One can relatively easily estimate the α and β parameters describing the observed rates of decay (one slow and the other fast) of concentration; however, the algebraic relationship between those and the fractional rate constants that mechanistically describe plasma-extraplasma exchange is not trivial. The complexity of these expressions increases with increasing number of compartments (10) and rapidly grows to be daunting. The limitations of noncompartmental and compartmental analysis of pharmacokinetic data have been discussed elsewhere, so we will not cover them here (20). There are other sources of complexity in pharmacokinetics and drug metabolism, which turn into more complex expressions for the model equations required to describe the drug's fate. For example, the underlying compartmental model may be not linear in the kinetics. This could happen, for example, when the fractional rate of disappearance changes with the drug level and goes from zero order at high concentrations to first order at low concentrations, as it happens for phenytoin (11) or ethanol (12):

$$\begin{aligned} \dot{q}(t) &= -\frac{V_m}{K_m + q(t)} q(t) + D\delta(t) \\ q(0) &= 0 \\ c(t) &= \frac{q(t)}{V} \end{aligned} \tag{6}$$

where the elimination rate is not constant, rather it exhibits a saturative behavior that resembles the classic Michaelis–Menten expression from enzyme kinetics. In this case, the principle of superposition does not hold and the concentration–time curve cannot be expressed in algebraic (closed) form, and the time course is not exponential, except at values of $q(t)$ that are much smaller than K_m. There are many other types of pharmacokinetic nonlinearities, of which this is just an example. This is why most modern PK analyses directly use the differential equations when building a PK model.

Excellent historical reviews of the properties of compartmental models, especially with reference to tracer kinetics, can be found in Refs. 13–16. More modern viewpoints are available in Refs. 17,18. Perspectives from drug development are available in Refs. 7,19. A succinct and practical review of compartmental and noncompartmental methods can be found in Ref. 8. Lastly, as we mentioned, a comparison of the strengths and weaknesses of compartmental and noncompartmental approaches has been carried out in Ref. 20.

A PK model can be used to make informed predictions about localized drug distribution and dose availability to the target organ, especially with physiologically based pharmacokinetic (PBPK) and toxicokinetic (PBTK) models. An important application has always been to individualize dosing (21,22) and improve therapeutic drug monitoring (23,24), often by borrowing approaches from process engineering and automatic control theory (25,26). Physiologically based models of pharmacokinetics are becoming an integral part of many drug development programs (27), mainly because they provide a mechanistic way to scale dosing regimens between species and between protocols. This affords (at least in theory) a seamless integration between the preclinical (*in vitro* and *in vivo*, animal studies) and clinical (*in vivo*, human studies) aspects of a drug development program. Animal to human scaling is of increasing importance in several areas of pharmacotherapy (28) and has a long and illustrious history, from the early work by Dedrick and co-workers on methotrexate (29) to a more general application (30,31) to modern experiences (32) motivated by first time in man (FTIM) dose finding (33,34). The approach is, however, not without its critics (35), mainly due to the lack of statistical evaluation that often accompanies the scaling. It is noteworthy that between-species scaling of body mass (36,37) and metabolic rate (38) has been and is currently an object of investigation outside drug development, and mechanistic findings about the origin of allometric scaling (39) have lent scientific support to the empiricism of techniques used in drug development (40,41).

Now is a good time to note that, most often, these projections are accompanied by some statistical evaluation. From the very beginnings of PK-PD, the realization that one needed to account both for variation between subjects (due to underlying biological reasons, e.g., genetic polymorphisms and environmental factors) and variation within subject (due, e.g., to intrinsic measurement error associated with the quantification of concentration time courses and efficacy levels) was particularly acute. Statistical considerations have thus always been a part of PK-PD. The role of *population analysis*, or the explicit modeling of variability sources, in the analysis of clinical trial data is discussed in more detail later. As far as other examples, statistical applications to PK-PD have proven useful in aiding the estimation of rates of disease progression (42,43), determining individually tailored dosing schemes (23,44) and resolving models too complex to identify without prior information (45). As mentioned later, a common framework often exploited in PK-PD has to do with the incorporation of population-level information (e.g., statistical distributions of model parameters) together with individual-specific information (limited concentration–time samples, e.g.).

Pharmacodynamics: The Mechanistic Link Between Drug Exposure and Effect

The PD models (46–49) can link drug effect (characterized by clinical outcomes or intermediate pharmacological response markers) with a PK model (50), through biophase distribution, biosensor process and biosignal flux (51). Formally, the simplest PD model is the so-called $E_{\max}$ model, where effect plateaus at high concentrations:

$$e(t) = \frac{E_{\max} c(t)}{EC_{50} + c(t)} \qquad (7)$$

where $E_{\max}$ is maximal effect, and EC_{50} is the effective plasma concentration at which effect $e(t)$ is half-maximal. This model is sometimes modified to account for sigmoidal effect shapes by adding an exponent that varies from ~ 1 to around 2:

$$e(t) = \frac{E_{\max} c^{\mathrm{H}}(t)}{EC_{50}^{\mathrm{H}} + c^{\mathrm{H}}(t)} \qquad (8)$$

Often, the driver of pharmacological effect is not plasma concentration $c(t)$, but some concentration remote from plasma and delayed with respect to plasma. This representation is similar to the one used to represent delayed effect of glucose regulatory hormones such as insulin (52). The delayed effect site concentration is them modeled using the differential equation:

$$c_{\mathrm{e}}(t) = k_{\mathrm{eo}}[c(t) - c_{\mathrm{e}}(t)] \qquad (9)$$

whereas the PD model now contains effect site concentration, not plasma concentration:

$$e(t) = \frac{E_{\max} c_{\mathrm{e}}(t)}{EC_{50} + c_{\mathrm{e}}(t)} \qquad (10)$$

and the interpretation of the parameters is the same as before except that they are defined with reference to effect site concentration as opposed to plasma concentration. Other classes of models are the so-called indirect response models, where the drug modulates either production or degradation of the effect (response) variable. These models are particularly appealing due to their mechanistic interpretation, and have been proposed in Refs. (53) and (54), reviewed in Ref. (55), and applied in many settings, including pharmacogenomics (56).

In many ways, the PD aspect of drug development is more challenging than the PK aspects, for many different reasons. First of all, the process of gathering data to inform about the PD of a drug is more challenging. Moreover, PD measurements may be less sensitive than it would be desirable (pain levels, which are both categorical and subjective, are a good example). Lastly, the mechanism of action of the drug may not be entirely known, and thus the best choice of measurements for the PD time course may be open to debate. How should the effect of a certain drug be quantified? If the drug is a painkiller, are pain levels sufficient, or would the levels of certain chemical(s) in the brain provide a more sensitive and specific correlate of efficacy? These and other questions become very relevant, for example, when drugs are studied that exert their effect in traditionally inaccessible locations (e.g., the brain) (57).

This is the “best biomarker question”, as it relates to the now classical classification of biomarkers, surrogate endpoints and clinical endpoints (58). Basically, while a biomarker is any quantitative measure of a biological process (concentration levels, pain scores, test results and the like), a surrogate endpoint is a biomarker that substitutes for a clinical endpoint (e.g., survival or remission). In other words, surrogate endpoints, when unambiguously defined, are predictive of clinical endpoints, with the added advantage of being easier to measure and usually being characterized by a more favorable time frame (59). In the United States, the Food and Drug Administration (FDA) now allows the possibility of accelerated approval based on surrogate endpoints, provided certain conditions are met (60). Against this framework, it makes eminent sense for the PK-PD model to be focused on a relevant biomarker (or surrogate endpoint) as soon as possible in the drug development process. Often, the earlier in the process, the less influence the choice of biomarker will have. In other words, when the selection of lead compounds is just starting, proof of concept may be all that is needed, but the closer one gets to the clinical trial stage, the more crucial an appropriate choice of surrogate endpoint will be. Another way to think about this is the establishment of a causal link between therapeutic regimen and outcome. Where these concepts come together is in the (relatively novel) idea of disease progression, and how it can be monitored.

Disease Progression: How to Know Whether the Drug is Working

Recently, the contention has been made that an integral part of the understanding of PK and PD cannot prescind from, say, the background signal that is present when the drug is administered. In other words, not all patients will be subjected to therapy at the same point in time, or at the same stage in their individual progression from early to late disease stages. There is thus a growing realization that the therapeutic intervention is made against a constantly changing background of disease state, and that the outcome of the therapy may depend on the particular disease state at a given moment in time. Disease progression modeling (61) is thus the point of contact between PK and PD and the mechanistic modeling of physiology and pathophysiology that is advocated, e.g., by the Physiome Project (62,63), recently taken over by the International Union of Physiological Sciences' Physiome and Bioengineering Committee (64).

Interesting applications are starting to emerge. For example, Ref. 65 has shown that the rate of increase of bloodstream glucose concentration in Type 2 diabetic patients is $\sim 0.84\ \mathrm{mmol{\cdot}L^{-1}{\cdot}year^{-1}}$ (with a sizable variation between patients of 143%), thus providing a quantitative handle on the expected deteriorating trend of overall glycemic levels in this population of patients (together with a measure of its expected patient-to-patient variability). In Alzheimer's disease, another study (66) has demonstrated a natural rate of disease progression of 6.17 ADASC $\mathrm{units{\cdot}year^{-1}}$ (where ADASC is the cognitive component of the Alzheimer disease assessment scale). A word of caution: As often done in this area of application, the models are informed on available data. In other words, the model parameters are quantified (estimated) based on available measurements for the drug PK and PD. This poses the challenge of developing models that are not too detailed nor too simplistic, but that can be reasonably well informed by the data at hand. Clearly, a detailed mechanistic model of the disease system requires substantial detail, but a balance needs to be struck between detail and availability of independently gathered data. Visualization approaches are a recent addition to the arsenal of the drug development expert (67). This kind of mechanistic pharmacodynamics is being applied more and more often to a variety of areas: A good example of rapid development comes from anticancer agents (68), where applications of integrated, mechanistic PK-PD models that take into account the drug mechanism of action are starting to become more and more frequent (69).

Population Variation: Adding Statistical Variation to PK, PD and Disease Models

In drug development, it is of utmost interest to determine the extent of variation of PK-PD and disease progression among members of a population. Basically, this implies determining the statistics of the biomarkers of interest in a population of patients, not just in an individual, and provide these measures together with some degree of confidence. This requirement connects well with analogous epidemiological population studies (often not model-based). The estimation of variability coupled with the evaluation of the relative role of its sources (covariates), for example, demographics, anthropometric variables or genetic polymorphisms, is tackled by the discipline of population kinetics, mainly due to the pioneering work of Lewis Sheiner and Stuart Beal at UCSF (70,71). Often called also population PK-PD, it makes use of two-level hierarchical models characterized by nested variability sources, where the models' individual parameter values are not deterministic, but unknown: They instead arise from population statistical distributions (biological variation, or BSV, between subject variability). On top of this source of variability, measurement noise and other uncertainty sources are added (RUV, residual unknown variation) to the concentration or effect signals (Fig. 2). The pharmaco-statistical models that integrate BSV and RUV are often described in statistical journals as nonlinear mixed effects models.

An example that builds on those we have already presented earlier may suffice here to clarify the fundamental concept of mixed effects models. Let us extend the model just described for intravenous injection (eq. 2) to the situation when the concentration measurements are affected by noise:

$$\begin{aligned} \dot{q}(t) &= -k_{\mathrm{el}}q(t) + D\delta(t) \\ q(0) &= 0 \\ c(t) &= \frac{q(t)}{V} + \varepsilon(t) \end{aligned} \tag{11}$$

where the parameters are the same as in equation 2, except that now $\varepsilon(t)$ is a normally distributed measurement error with mean zero and variance σ^2, that is $\varepsilon(t) \in N[0, \sigma^2]$. If data about $c(t)$ are available and have been gathered in a single individual, the model parameters k_{el} and V (assuming D is known) can then be fitted to the data using weighted (9) or extended (72) least squares or some other variation of maximum likelihood, and thus individualized estimates (with confidence intervals) can be obtained (often the measurement error variance σ^2 is not known but it can also be estimated). This can be done even if only a single subject data are available, provided that the sampling is performed (at an absolute minimum) at three or more time points (since the model has two unknown structural parameters, k_{el} and V, plus σ^2).

The nonlinear mixed-effects modeling approach is an extension of what we have just seen. It takes the viewpoint that the single subject estimates are realizations of an underlying population density for the model parameters. As such, the individual values of k_{el} and V simply become realizations (samples) of this underlying density. For example, the assumption could be made that they are both distributed lognormally: in which case, $\log(k_{\mathrm{el}}) \in N(\mu_{\mathrm{kel}}, \omega^2_{\mathrm{kel}})$ and $\log(V) \in N(\mu_\nu, \omega^2_\nu)$, where the μ denote expected logarithmic values and the ω^2 denote population variances.

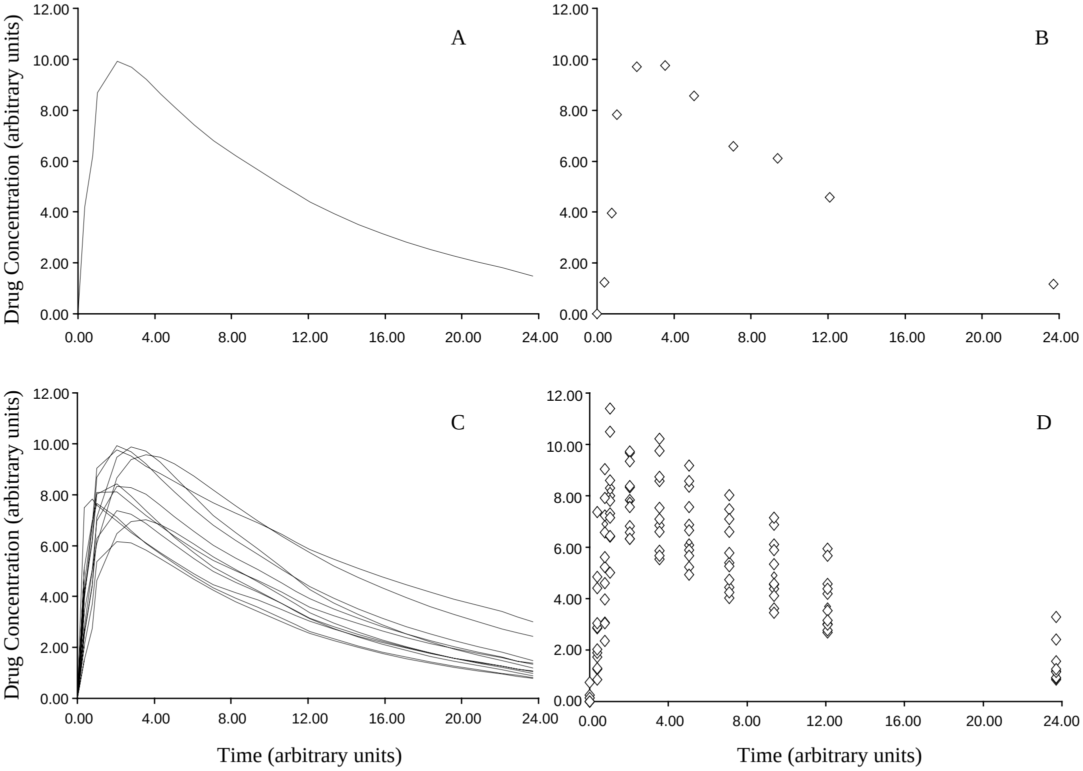


Figure 2. Nested uncertainty in pharmacokinetics. Panel A shows the true, but unknown, time course over 24 time units of a hypothetical drug administered orally to a single experimental subject. The true, but unknown, time course is a smooth function of time. Panel B shows the same smooth function sampled at discrete time points, which in practice may correspond to as many blood draws. Measurement uncertainty (currently termed RUV, residual unknown variability) is now superimposed to the true, but unknown time course in Panel A. Panel C shows the true, but unknown, time course over 24 time units of a hypothetical drug administered orally to several experimental subjects. Clearly, no time course is the same, since each subject will have a different absorption and elimination rate, together with a different volume of distribution. The time courses in Panel C are the result of BSV (between subject variability). Lastly, Panel D shows the same smooth functions in Panel c sampled at discrete time points. The RUV is again superimposed to the true, but unknown time courses, but this time the data spread is due to both BSV and RUV, and distinguishing between BSV and RUV requires a model of the system. See text for details.

The assumption is implicitly made here that k_{el} and V are uncorrelated (which may or may not be the case). In the nonlinear mixed-effects model procedure, the moments of the statistical distributions of model parameters become the new unknowns, and thus μ_{k01}, ω^2_{k01} and μ_ν, ω^2_ν are estimated by optimizing approximations of the maximum likelihood objective function expressed for the whole population of data. The value of σ^2 in the population (a composite value of the measurement error variance across subjects) can also be estimated, as before. The five distribution parameters are the fixed effects of the population (since they do not change between subject), while the individual values attained by the parameters k_{el} and V in separate individuals are random effects, since every subject has a different value (hence, the mixed-effects parlance). The approach requires data on more than one subject, ideally on many more subjects than there are fixed effects (the caveat is that it is easier to estimate expected values than it is to estimate variances or covariances, and the data needs consequently grow). Note also that, if k_{el} and V are both lognormal, clearance CL is *not* lognormal: which statistical model to choose for which parameter will depend on the available information. The main advantage of population kinetics is that, since it is estimating distributional parameters (moments), it can use relatively sparse and/or noisy data at the individual level, provided that there is a large number of population data (in other words, there can be few data for each subject, as long as there are many subjects).

This framework is quite general and powerful, and allows for modeling of complex events (e.g., adherence, or patient compliance to dosing recommendations) (73). Tutorials on this modeling approach can be found in papers (74–76), review articles (70,61,77,78), and textbooks (79). Software is also available, both for population (Aarons, 1999) and individual (Charles and Duffull, 2001) PK-PD analysis.

Mixed-effects models are used both to solve the forward problem (simulation of putative drug dosing scenarios) and the inverse problem (estimation of BSV and RUV statistics conditional on PK-PD models and clinical measurements). Which one is of interest depends on what is available and what the intent of the study is. If the intent is to analyze data and determine the underlying distribution of PK-PD and disease parameters, then one has an inverse, or estimation, problem (80). If the intent, on the other hand, is to explore possible dosing or recruitment scenarios, then this is a direct, or simulation, problem (81). As mentioned, mixed-effects models can be applied to sparse and noisy data, as often happens in therapeutic monitoring in the clinical setting (a situation that occurs both in drug development and in applied clinical research). Their use is so widespread that the FDA recently issued one of its guidances for industry to deal with their use for population pharmacokinetic analysis (82). Interestingly, a very similar framework is also applied to evolutionary genetics, in the study of "function-valued traits" (also called "infinite-dimensional characters"). The idea is to use mixed models to link genetic information to traits that are not constant, rather are functions of time (83–85). It is important to model population genetics both for polymorphisms of drug-metabolizing enzymes affecting ADME and for polymorphisms affecting the dynamics of response.

A Role for Modeling and Simulation

In summary, the overall goal of integrated PK-PD mathematical models is a better understanding of therapeutic intervention: Their contribution are often reflected in improved clinical trials designs. While evidence of the impact of modeling and simulation of PK-PD on the drug development process is often anectodal, many reviews of PK-PD in drug development have recently appeared (86–89), together with compelling examples of subsequent drug development acceleration (90,91).

While the fundamental algorithmic steps of computer simulation, especially Monte Carlo simulation, are well known, they require some adaptation to be used within the field of pharmacokinetics and pharmacodynamics and in the context of drug development (Fig. 3) A good practices document was issued in 1999 following a workshop held by the Center for Drug Development Science (92). The crafting of this document has contributed to clarifying the steps and tools necessary for carrying out a simulation experiment within the context of drug development. In particular, the report clarifies that simulation in drug development has an untapped potential which extends well beyond the pharmacokinetic aspect and into the pharmacodynamic and clinical domains (92).

Nowadays, pharmacometric technology (93) is used in industry and academia as a way to support and strengthen R&D. As an example, low signal/noise ratio routine clinical data obtained with sparse sampling may often be analyzed with pharmacometric techniques to determine whether a compound is metabolized differently because of phenotypic differences arising from genetic makeup, ethnicity, gender, age group (young, elderly), or concomitant medications causing drug–drug interactions (94). The role of this technology can only increase. A recently issued FDA report that focuses on the recent decrease in applications for novel therapies submitted to the agency is clear in this regard: "The concept of model-based drug development, in which pharmaco-statistical models of drug efficacy and safety are developed from preclinical and available clinical data, offers an important approach to improving drug development knowledge management and development decision making. Model-based drug development involves building mathematical and statistical characterizations of the time course of the disease and drug using available clinical data to design and validate the model. The relationship between drug dose, plasma concentration, biophase concentration (pharmacokinetics), and drug effect or side-effects (pharmacodynamics) is characterized, and relevant patient covariates are included in the model. Systematic application of this concept to drug development has the potential to significantly improve it" (95).

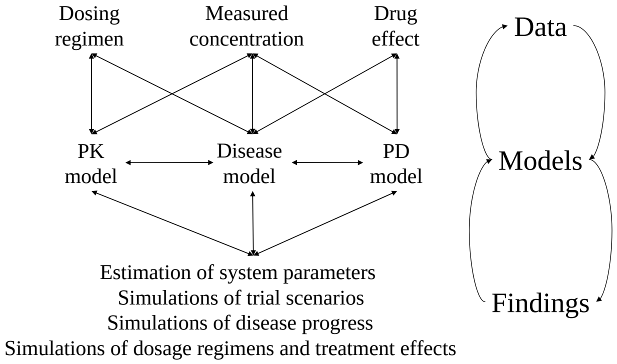

Dosing regimen
Measured concentration
Drug effect
PK model
Disease model
PD model
Estimation of system parameters
Simulations of trial scenarios
Simulations of disease progress
Simulations of dosage regimens and treatment effects
Data
Models
Findings

Figure 3. Information flow in a clinical trial simulation or data analysis. The doses, the measured concentration and drug effect are situated at the beginning or the end of the process (Data level). At the core of the simulation (Model level) lie models of PK, PD and disease processes. Trial designs are possible conditional on these other elements, or features like model parameters can be estimated from the trial data and measured doses and PK and PD data (Findings level). The picture is oversimplified and does not include, for example, adherence to dosing recommendations and protocol dropouts, which may be important to ensure realistic trial designs.

In summary, modern-day drug development displays a need for information integration at the whole-system, cellular, and genomic level (96) similar to that found in integrative physiology (97) and comparative biology (98,99). As mentioned earlier, simulation of clinical trials is a burgeoning discipline well-founded upon engineering and statistics (81): examples have appeared in clinical pharmacology (100,101) and pharmacoeconomics (102). Drug candidate selection is another application, possibly through PK models of varying complexity (103) and high throughput screening coupled with PK (104). Next, integrated models allow to link genomic information with disease biomarkers and phenotypes, such as in the Luo-Rudy model of cardiac excitation (105).

As a concluding remark, progress in the development of plausible, successful, and powerful data analysis methods has already had a substantial payoff, and can be substantially accelerated by encouraging multidisciplinary, multi-institutional collaboration bringing together investigators at multiple facilities and providing the infrastructure to support their research, thus allowing the timely and cost-effective expansion of new technologies. The need is more and more often voiced for increased training of quantitative scientists in biologic research as well as in statistical methods and modeling to ensure that there will be an adequate workforce to meet future research needs (106).

BIBLIOGRAPHY

1. Marasco C. Surge In Pharmacometrics Demand Leads to New Master's Program. JobSpectrum.org – a service of the American Chemical Society. Available at http://www.cen-chemjobs.org/job_weekly31802.html, 2002.
2. Abdel-Rahman SM Kauffman RE. The integration of pharmacokinetics and pharmacodynamics: understanding dose-response. Annu Rev Pharmacol Toxicol 2004;44:111–136.
3. Rowland M, Tozer TN. Clinical Pharmacokinetics: Concepts and Applications. 3rd Ed. Baltimore: Williams & Wilkins; 1995.
4. Verotta D. Volterra series in pharmacokinetics and pharmacodynamics. J Pharmacokinet Pharmacodyn. 2003;30(5): 337–362.
5. Vicini P, Gastonguay MR, Foster DM. Model-based approaches to biomarker discovery and evaluation: a multidisciplinary integrated review. Crit Rev Biomed Eng 2002;30 (4–6):379–418.
6. Teorell T. Kinetics of distribution of substances administered to the body. I. The extravascular modes of administration. II. The intravascular modes of administration. Arch Int Pharmacodyn Ther. 1937; 57: 205–240.
7. Gibaldi M, Perrier D. Pharmacokinetics, 2nd Ed. New York: Marcel Dekker; 1982.
8. Jang GR, Harris RZ, Lau DT. Pharmacokinetics and its role in small molecule drug discovery research. Med Res Rev 2001;21(5):382–396.
9. Landaw EM, DiStefano JJ 3rd. Multiexponential, multicompartmental, and noncompartmental modeling. II. Data analysis and statistical considerations. Am J Physiol 1984;246(5 Pt. 2):R665–77.
10. DiStefano JJ 3rd, Landaw EM. Multiexponential, multicompartmental, and noncompartmental modeling. I. Methodological limitations and physiological interpretations. Am J Physiol 1984;246(5 Pt. 2):R651–664.
11. Grasela TH, et al. Steady-state pharmacokinetics of phenytoin from routinely collected patient data. Clin Pharmacokinet 1983;8(4):355–364.
12. Holford NH. Clinical pharmacokinetics of ethanol. Clin Pharmacokinet, 1987;13(5):273–292.
13. Anderson DH. Compartmental Modeling and Tracer Kinetics. Lecture Notes in Biomathematics, Vol. 50, Berlin: Springer-Verlag; 1983.
14. Godfrey K. Compartmental Models and Their Application. New York: Academic; 1983.
15. Jacquez JA. Compartmental Analysis in Biology and Medicine, 2nd ed. Michigan: University of Michigan Press; 1985.
16. Carson ER, Cobelli C, Finkelstein L. The Mathematical Modeling of Endocrine-Metabolic Systems. Model Formulation, Identification and Validation. New York: Wiley; 1983.
17. Carson E, Cobelli C. Modelling Methodology for Physiology and Medicine. San Diego: Academic Press; 2000.
18. Cobelli C, Foster D, Toffolo G. Tracer Kinetics in Biomedical Research: From Data to Model. London: Kluwer Academic/ Plenum; 2001.
19. Atkinson AJ, et al., eds. Principles of Clinical Pharmacology. San Diego: Academic Press; 2001.
20. DiStefano JJ 3rd. Noncompartmental vs. compartmental analysis: some bases for choice. Am J Physiol 1982;243(1): R1–6.
21. Goicoechea FJ, Jelliffe RW. Computerized dosage regimens for highly toxic drugs. Am J Hosp Pharm 1974;31(1):67–71.
22. Sheiner LB, Beal S, Rosenberg B, Marathe VV. Forecasting individual pharmacokinetics. Clin Pharmacol Ther 1979;26(3):294–305.
23. (a) Jelliffe RW, et al. Individualizing drug dosage regimens: roles of population pharmacokinetic and dynamic models, Bayesian fitting, and adaptive control. Ther Drug Monit 1993;15(5):380–393. (b) Jelliffe RW. Clinical applications of pharmacokinetics and adaptive control. IEEE Trans Biomed Eng 1987;34(8):624–632.
24. (a) Jelliffe RW, et al. Adaptive control of drug dosage regimens: basic foundations, relevant issues, and clinical examples. Int J Biomed Comput 1994;36(1–2):1–23. (b) Jelliffe RW, Schumitzky A. Modeling, adaptive control, and optimal drug therapy. Med Prog Technol 1990;16(1–2):95–110.
25. Jelliffe RW. Clinical applications of pharmacokinetics and adaptive control. IEEE Trans Biomed Eng 1987;34(8):624–632.
26. Jelliffe RW, Schumitzky A. Modeling, adaptive control, and optimal drug therapy. Med Prog Technol 1990;16(1–2):95–110.
27. Parrott N, Jones H, Paquereau N, Lave T. Application of full physiological models for pharmaceutical drug candidate selection and extrapolation of pharmacokinetics to man. Basic Clin Pharmacol Toxicol 2005;96(3):193–199.
28. Gallo JM, et al. Pharmacokinetic model-predicted anticancer drug concentrations in human tumors. Clin Cancer Res 2004;10(23):8048–8058.
29. Dedrick R, Bischoff KB, Zaharko DS. Interspecies correlation of plasma concentration history of methotrexate (NSC-740). Cancer Chemother Rep 1970;54(2):95–101.
30. Dedrick RL. Animal scale-up. J Pharmacokinet Biopharm 1973;1(5):435–461.
31. Dedrick RL, Bischoff KB. Species similarities in pharmacokinetics. Fed Proc 1980;39(1):54–59.
32. Mahmood I, Balian JD. Interspecies scaling: predicting clearance of drugs in humans. Three different approaches. Xenobiotica 1996;26(9):887–895.
33. Mahmood I, Green MD, Fisher JE. Selection of the first-time dose in humans: comparison of different approaches based on interspecies scaling of clearance. J Clin Pharmacol 2003;43(7):692–697.
34. Iavarone L, et al. First time in human for GV196771: interspecies scaling applied on dose selection. J Clin Pharmacol 1999;39(6):560–566.
35. Bonate PL, Howard D. Prospective allometric scaling: does the emperor have clothes? J Clin Pharmacol 2000;40(6):665–670. discussion 671–676.
36. West GB, Brown JH, Enquist BJ. A general model for the origin of allometric scaling laws in biology. Science 1997;276(5309):122–126.
37. (a) Iavarone L, et al. First time in human for GV196771: interspecies scaling applied on dose selection. J Clin Pharmacol 1999;39(6):560–566. (b) West GB, Brown JH, Enquist BJ. The fourth dimension of life: fractal geometry and allometric scaling of organisms. Science 1999;284(5420):1677–1679.
38. Gillooly JF, et al. Effects of size and temperature on metabolic rate. Science 2001;293(5538):2248–2251. Erratum in Science 2001;294(5546):1463.
39. White CR, Seymour RS. Mammalian basal metabolic rate is proportional to body mass2/3. Proc Natl Acad Sci USA 2003;100(7):4046–4049.
40. Anderson BJ, Woollard GA, Holford NH. A model for size and age changes in the pharmacokinetics of paracetamol in neonates, infants and children. Br J Clin Pharmacol 2000;50(2): 125–134.

41. van der Marel CD, et al. Paracetamol and metabolite pharmacokinetics in infants. Eur J Clin Pharmacol 2003;59(3): 243–251.
42. Craig BA, Fryback DG, Klein R, Klein BE. A Bayesian approach to modelling the natural history of a chronic condition from observations with intervention. Stat Med 1999;18(11):1355–1371.
43. Mc Neil AJ. Bayes estimates for immunological progression rates in HIV disease. Stat Med 1997;16(22):2555–2572.
44. Sheiner LB, Beal SL. Bayesian individualization of pharmacokinetics: simple implementation and comparison with non-Bayesian methods. J Pharm Sci 1982;71(12):1344–1348.
45. Cobelli C, Caumo A, Omenetto M. Minimal model SG overestimation and SI underestimation: improved accuracy by a Bayesian two-compartment model. Am J Physiol 1999;277 (3 Pt.1):E481–488.
46. Segre G. Kinetics of interaction between drugs and biological systems. Farmaco [Sci] 1968;23(10):907–918.
47. Dahlstrom BE, Paalzow LK, Segre G, Agren AJ. Relation between morphine pharmacokinetics and analgesia. J Pharmacokinet Biopharm 1978 6(1):41–53.
48. Holford NH, Sheiner LB. Pharmacokinetic and pharmacodynamic modeling in vivo. CRC Crit Rev Bioeng 1981a;5(4): 273–322.
49. Holford NH, Sheiner LB. Understanding the dose-effect relationship: clinical application of pharmacokinetic-pharmacodynamic models. Clin Pharmacokinet 1981b; 6(6):429–453.
50. Holford NH, Sheiner LB. Kinetics of pharmacologic response. Pharmacol Ther 1982;16(2):143–166.
51. Jusko WJ. Pharmacokinetics and receptor-mediated pharmacodynamics of corticosteroids. Toxicology 1995;102(1–2): 189–196.
52. Bergman RN, Phillips LS, Cobelli C. Physiologic evaluation of factors controlling glucose tolerance in man: measurement of insulin sensitivity and beta-cell glucose sensitivity from the response to intravenous glucose. J Clin Invest 1981;68(6): 1456–1467.
53. (a) Dayneka NL, Garg V, Jusko WJ. Comparison of four basic models of indirect pharmacodynamic responses. J Pharmacokinet Biopharm 1993;21(4):457–478. (b) Ramakrishnan R, et al. Pharmacodynamics and pharmacogenomics of methylprednisolone during 7-day infusions in rats. J Pharmacol Exp Ther 2002;300(1):245–256.
54. Jusko WJ, Ko HC. Physiologic indirect response models characterize diverse types of pharmacodynamic effects. Clin Pharmacol Ther 1994;56(4):406–419.
55. Sharma A, Jusko WJ. Characteristics of indirect pharmacodynamic models and applications to clinical drug responses. Br J Clin Pharmacol 1998;45(3):229–239.
56. Ramakrishnan R, et al. Pharmacodynamics and pharmacogenomics of methylprednisolone during 7-day infusions in rats.J Pharmacol Exp Ther 2002;300(1):245–256.
57. Bieck PR, Potter WZ. Biomarkers in psychotropic drug development: integration of data across multiple domains. Annu Rev Pharmacol Toxicol 2005;45:227–246.
58. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 2001;69(3):89–95.
59. Rolan P, Atkinson AJ Jr., Lesko LJ. Use of biomarkers from drug discovery through clinical practice: report of the Ninth European Federation of Pharmaceutical Sciences Conference on Optimizing Drug Development. Clin Pharmacol Ther 2003;73(4):284–291.
60. The Food and Drug Modernization Act of 1997. Title 21 Code of Federal Regulations Part 314 Subpart H Section 314.500.
61. Chan PL, Holford NH. Drug treatment effects on disease progression. Annu Rev Pharmacol Toxicol 2001;41:625–659.
62. Bassingthwaighte JB. The macro-ethics of genomics to health: the physiome project. C R Biol 2003;326(10–11): 1105–1110.
63. Hunter PJ, Borg TK. Integration from proteins to organs: the Physiome Project. Nat Rev Mol Cell Biol 2003;4(3):237–243.
64. Hunter PJ. The IUPS Physiome Project: a framework for computational physiology. Prog Biophys Mol Biol 2004; 85(2–3):551–569.
65. Frey N, et al. Population PKPD modelling of the long-term hypoglycaemic effect of gliclazide given as a once-a-day modified release (MR) formulation. Br J Clin Pharmacol 2003;55(2):147–157.
66. Holford NH, Peace KE. Methodologic aspects of a population pharmacodynamic model for cognitive effects in Alzheimer patients treated with tacrine. Proc Natl Acad Sci USA 1992;89(23):11466–11470.
67. Bhasi K, Zhang L, Zhang A, Ramanathan M. Analysis of pharmacokinetics, pharmacodynamics, and pharmacogenomics data sets using VizStruct, a novel multidimensional visualization technique. Pharm Res 2004;21(5):777–780.
68. Friberg LE, Karlsson MO. Mechanistic models for myelosuppression. Invest New Drugs 2003;21(2):183–194.
69. Karlsson MO, et al. Pharmacokinetic/pharmacodynamic modelling in oncological drug development. Basic Clin Pharmacol Toxicol 2005;96(3):206–211.
70. Beal SL, Sheiner LB. Estimating population kinetics. Crit Rev Biomed Eng 1982;8(3):195–222.
71. Sheiner LB, Ludden TM. Population pharmacokinetics/dynamics. Annu Rev Pharmacol Toxicol 1992;32:185–209.
72. Peck CC, Beal SL, Sheiner LB, Nichols AI. Extended least squares nonlinear regression: a possible solution to the 'choice of weights" problem in analysis of individual pharmacokinetic data. J Pharmacokinet Biopharm 1984;12(5): 545–558.
73. Girard P, et al. A Markov mixed effect regression model for drug compliance. Stat Med 1998;17(20):2313–2333.
74. Sheiner LB. Analysis of pharmacokinetic data using parametric models-1: Regression models. J Pharmacokinet Biopharm 1984;12(1):93–117.
75. (a) Landaw EM, DiStefano JJ 3rd. Multiexponential, multicompartmental, and noncompartmental modeling. II. Data analysis and statistical considerations. Am J Physiol 1984; 246(5 Pt. 2):R665–77. (b) Sheiner LB. Analysis of pharmacokinetic data using parametric models. II. Point estimates of an individual's parameters. J Pharmacokinet Biopharm 1985; 13(5):515–540.
76. Sheiner LB. Analysis of pharmacokinetic data using parametric models. III. Hypothesis tests and confidence intervals. J Pharmacokinet Biopharm 1986;14(5):539–555.
77. Ette EI, Williams PJ. Population pharmacokinetics II: estimation methods. Ann Pharmacother 2004;38(11):1907–1915.
78. Ette EI, Williams PJ. Population pharmacokinetics I: background, concepts, and models. Ann Pharmacother 2004; 38(10):1702–1706.
79. Davidian M,and Giltinan DM. Nonlinear Models for Repeated Measurement Data. Boca Raton: Chapman and Hall/CRC; 1995.
80. Sheiner L, Wakefield J. Population modelling in drug development. Stat Methods Med Res 1999;8(3):183–193.
81. Holford NH, Kimko HC, Monteleone JP, Peck CC. Simulation of clinical trials. Annu Rev Pharmacol Toxicol 2000;40:209–234.

82. United States Food and Drug Administration, Department of Health and Human Services. Guidance for Industry: Population Pharmacokinetics. Availabel at http://www.fda.gov/cder/guidance/1852fnl.pdf, 1999.
83. Kirkpatrick M, Lofsvold D. The evolution of growth trajectories and other complex quantitative characters. Genome 1989;31(2):778–783.
84. Pletcher SD, Geyer CJ. The genetic analysis of age-dependent traits: modeling the character process. Genetics 1999; 153(2):825–835.
85. (a) Ma CX, Casella G, Wu R. Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. Genetics 2002;161(4):1751–1762. (b) Sheiner LB, Beal S, Rosenberg B, Marathe VV. Forecasting individual pharmacokinetics. Clin Pharmacol Ther 1979;26(3): 294–305.
86. Rolan P. The contribution of clinical pharmacology surrogates and models to drug development—a critical appraisal. Br J Clin Pharmacol 1997;44(3):219–225.
87. Sheiner LB, Steimer JL. Pharmacokinetic/pharmacodynamic modeling in drug development. Annu Rev Pharmacol Toxicol 2000;40:67–95.
88. Aarons L, et al. COST B15 Experts. Role of modelling and simulation in Phase I drug development. Eur J Pharm Sci 2001;13(2):115–122.
89. Blesch KS, et al. Clinical pharmacokinetic/pharmacodynamic and physiologically based pharmacokinetic modeling in new drug development: the capecitabine experience. Invest New Drugs 2003;21(2):195–223.
90. Piscitelli SC, Peck CC. Pharmacokinetic and pharmacodynamic methods in biomarker development and application. In: Downing GJ, editor. Biomarkers and Surrogate Endpoints: Clinical Research and Applications. New York: Elsevier; 2000. pp. 27–35.
91. Lesko LJ, Rowland M, Peck CC, Blaschke TF. Optimizing the science of drug development: opportunities for better candidate selection and accelerated evaluation in humans. Pharm Res 2000;17(11):1335–1344.
92. Holford NH, et al. Simulation in Drug Development: Good Practices. Draft Publication of the Center for Drug Development Science (CDDS) Draft version 1.0, July 23, 1999, Available at http://cdds.georgetown.edu/research/sddgp723.html
93. Sun H, et al. Population pharmacokinetics. A regulatory perspective. Clin Pharmacokinet 1999;37(1):41–58.
94. Krecic-Shepard ME et al. Race and sex influence clearance of nifedipine: results of a population study. Clin Pharmacol Ther 2000;68(2):130–142.
95. United States Food and Drug Administration, Department of Health and Human Services. Innovation/Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products. Available at http://www.fda.gov/oc/initiatives/criticalpath/, 2004.
96. Mamiya K, et al. The effects of genetic polymorphisms of CYP2C9 and CYP2C19 on phenytoin metabolism in Japanese adult patients with epilepsy: studies in stereoselective hydroxylation and population pharmacokinetics. Epilepsia 1998; 39(12):1317–1323.
97. Bassingthwaighte JB. Strategies for the physiome project. Ann Biomed Eng 2000;28(8):1043–1058.
98. Davidson EH, et al. A genomic regulatory network for development. Science 2002;295(5560):1669–1678.
99. von Dassow G, Meir E, Munro EM, Odell GM. The segment polarity network is a robust developmental module. Nature (London) 2000;406(6792):188–192.
100. Kimko HC, Reele SS, Holford NH, Peck CC. Prediction of the outcome of a phase 3 clinical trial of an antischizophrenic agent (quetiapine fumarate) by simulation with a population pharmacokinetic and pharmacodynamic model. Clin Pharmacol Ther 2000;68(5):568–577.
101. Nestorov I, et al. Modeling and stimulation for clinical trial design involving a categorical response: a phase II case study with naratriptan. Pharm Res 2001;18(8):1210–1219.
102. Hauber AB, et al. Potential savings in the cost of caring for Alzheimer's disease. Treatment with rivastigmine. Pharmacoeconomics 2000;17(4):351–360.
103. (a) Jang GR, Harris RZ, Lau DT. Pharmacokinetics and its role in small molecule drug discovery research. Med Res Rev 2001;21(5):382–396. (b) Roberts SA. High-throughput screening approaches for investigating drug metabolism and pharmacokinetics. Xenobiotica 2001;31(8–9):557–589.
104. Roberts SA. High-throughput screening approaches for investigating drug metabolism and pharmacokinetics. Xenobiotica 2001;31(8–9):557–589.
105. Rudy Y. From genome to physiome: integrative models of cardiac excitation. Ann Biomed Eng 2000;28(8):945–950.
106. De Gruttola VG, et al. Considerations in the evaluation of surrogate endpoints in clinical trials. Summary of a National Institutes of Health workshop. Control Clin Trials 2001; 22(5):485–502.
107. Charles BG, Duffull SB. Pharmacokinetic software for the health sciences: choosing the right package for teaching purposes. Clin Pharmacokinet 2001;40(6):395–403.

See also DRUG DELIVERY SYSTEMS; DRUG INFUSION SYSTEMS; RADIOPHARMACEUTICAL DOSIMETRY; TRACER KINETICS.

PHONOCARDIOGRAPHY

HERMAN VERMARIEN
Vrije Universiteit Brussel
Brussel, Belgium

INTRODUCTION

Mechanical heart action is accompanied by audible noise phenomena, which are easy to perceive when the ear is placed next to a person's chest wall. These cardiovascular sounds can be designated as being weak in comparison with other physiological sounds, such as speech, stomach and intestine rumbling, and even respiration noises. In fact, the latter can be heard at a certain distance from the subject, which is not true for heart noises (provided one overlooks cases of artificial heart valves). The frequency content of heart sounds is situated between 20 and 1000 Hz, the lower limit being set by the ability of human hearing; Mechanical valve prostheses may largely exceed the upper limit. Examination of cardiovascular sounds for diagnostic purposes through the human hearing sense, that is, auscultation, has been commonly practiced for a long time (1–5). The only technology involved is the stethoscope, establishing a closed air compartment between a part of the person's chest surface and the physician's ear orifice. This investigation method, however, being completely psychophysical and thus subjective, has proved its benefit and continues to be an important tool in cardiovascular diagnosis.

Phonocardiography (PCG) may simply be defined as the method for obtaining recordings of cardiovascular sound, that is, the phenomena perceivable by auscultation. The origins of the method are strongly anchored in auscultation. The recordings of sounds are evaluated, on paper or computer screen, possibly in the presence of other synchronous signals (e.g., the electrocardiogram) (ECG), partly psychophysically with another human sense, the eye, in examining waveform patterns and their relation with the other signals. Phonocardiographic signals are examined with respect to the occurrence of pathological patterns, relative intensities and intensity variations, timing and duration of events. Evidently more objective evaluation can be performed ranging from simple accurate timing of phenomena to advanced waveform analysis and comparing recorded results with waveforms from data banks. The importance of auscultation can be explained by the simplicity of the technique and by the strong abilities of the ear with respect to pattern recognition in acoustic phenomena. For obtaining equivalent information with phonocardiography, a single recording fails to be sufficient: A set of frequency filtered signals, each of them emphasizing gradually higher frequency components (by using high pass or band-pass filters), is needed. In this way, visual inspection of sound phenomena in different frequency ranges, adapted by a compensating amplification for the intensity falloff of heart sounds toward higher frequencies, is made possible, thus rendering the method equivalent with hearing performance: pattern recognition abilities and increasing sensitivity toward higher frequencies (within the above mentioned frequency range).

Laennec (1781–1826) was the first to listen to the sounds of the heart, not only directly with his ear to the chest, he also invented the stethoscope and provided the basis of contemporary auscultation. As physiological knowledge increased through the following decades, faulty interpretations of heart sounds were progressively eliminated. The first transduction of heart sounds was made by Hürthle (1895), who connected a microphone to a frog nerve-muscle preparation. Einthoven (1907) was the first to record phonocardiograms with the aid of a carbon microphone and a string galvanometer (6). Different investigators were involved in the development of filters to achieve a separation of frequency phenomena, as the vacuum tube, and thus electronic amplification became available. The evolution of PCG is strongly coupled with auscultatory findings and the development was predominantly driven by clinicians. A result of this situation is that a large variety of apparatus has been designed, mostly according to the specific needs of a clinic or the scientific interests of a medical researcher. During the 1960s, the necessity for standardization was strongly felt. Standardization committees made valuable proposals (7–9) but the impact on clinical phonocardiographic apparatus design was limited.

During the 1970s and the beginning of the 1980s, fundamental research on physical aspects of recording, genesis, and transmission of heart sound was performed (10–12) which, together with clinical investigations, improved the understanding of the heart sound phenomena. At the same time, ultrasonic methods for heart investigation became available and gradually improved. Doppler and echocardiography provided information closer related to heart action in terms of heart valve and wall movement, and blood velocity. Moreover, obtaining high quality recordings of heart sound with a high signal-to-noise ratio is difficult. Hampering elements are the inevitable presence of noise (background noise, respiration noise, muscle tremors, stomach rumbling), nonoptimal recording sites, weak sounds (obese patients), and so on. Thus interest in PCG gradually decreased.

In describing the state of the art, PCG is usually compared with ECG, the electrical counterpart, also a noninvasive method. The ECG, being a simple recording of electrical potential differences, was easily standardized, thus independent of apparatus design and completely quantitative with the millivolt scale on its ordinate axis. Phonocardiography has not reached the same level of standardization, remains apparatus dependent, and thus semiquantitative. Nowadays Doppler echocardiography and cardiac imaging techniques largely exceed the possibilities of PCG and make it redundant for clinical diagnosis. Whereas auscultation of cardiac sounds continues to be of use in clinical diagnosis, PCG is now primarily used for teaching and training purposes and for research. As a diagnostic method, conventional PCG has historical value. Nevertheless, the electronic stethoscope (combined with PC and software), as a modern concept for PCG, may gain importance for clinical purposes.

The generation of sounds is one of the many observable mechanical effects caused by heart action: contraction and relaxation of cardiac muscle, pressure rising and falling in the heart cavities, valve opening and closure, blood flowing and discontinuation of flow. Figure 1 shows a schematic representation of typical cardiac variables: the ECG, the logic states of the heart valves, low and high frequency phonocardiograms, a recording of a vessel pulse (carotid artery), and of the heart apex pulse (apexcardiogram). The heart cycle is divided into specific intervals according to the valve states of the left heart. The left ventricular systole is composed of the isovolumic contraction and the ejection period; The left ventricular diastole covers the isovolumic relaxation and the left ventricular filling (successively, the rapid filling, the slow filling, and the atrial contraction). A similar figure could be given for the right heart; Valve phenomena are approximately synchronous with those of the left heart. Small time shifts are typical: Mitral valve closure precedes tricuspid closure and aortic valve closure precedes pulmonary closure. The low frequency PCG shows the four normal heart sounds (I, II, III, and IV); In the high frequency, trace III and IV have disappeared and splitting is visible in I and in II. In the next sections details are given on the physiological significance, the physical aspects and recording methods, processing and physical modeling of heart sounds. Special attention is given to the electronic stethoscope.

HEART SOUNDS AND MURMURS

The sounds of the normal heart can be represented by a simple onomatopoeic simulation: "…lubb-dup…" (1–5). Two sounds can clearly be identified, the first being more dull than the second. A heart sound or a heart sound component is defined as a single audible event preceded and followed by a pause. As such, "splitting of a sound" occurs as one can clearly distinguish two components separated by a small pause. The closest splitting that

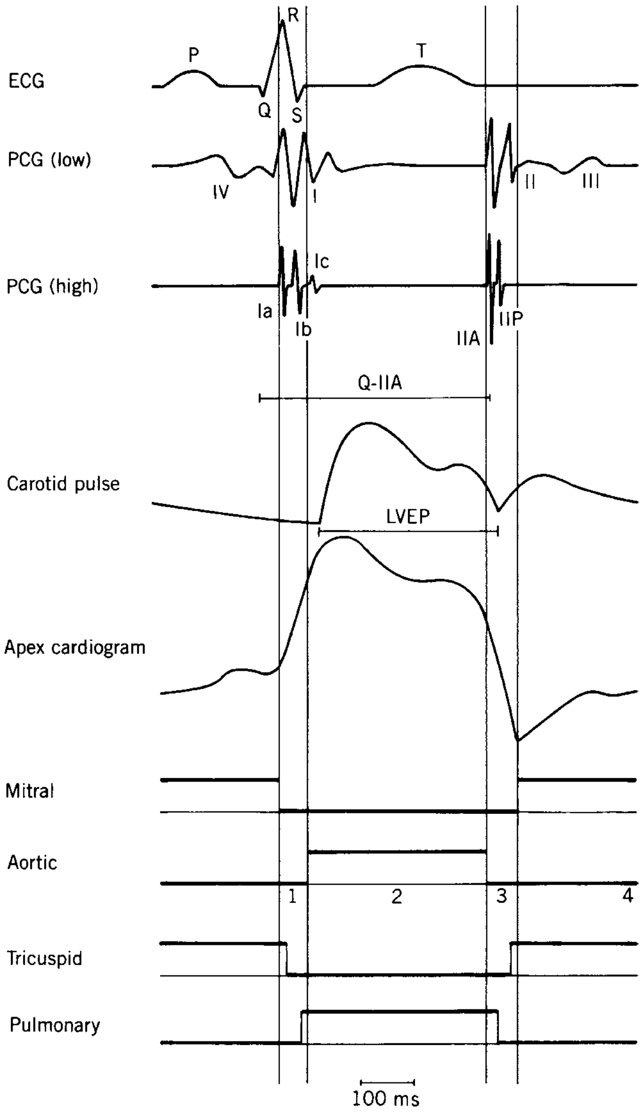


Figure 1. The ECG, PCG (low and high filtered), carotid pulse, apexcardiogram, and logic states (high = open) of left heart valves, mitral and aortic valve, and right heart valves, tricuspid and pulmonary valve. Left heart mechanical intervals are indicated by vertical lines: isovolumic contraction (1), ejection (2), isovolumic relaxation (3), and filling (4) (rapid filling, slow filling, atrial contraction). The low frequency PCG shows the four normal heart sounds (I, II, III, and IV); In the high frequency trace III and IV have disappeared and splitting is visible in I [Ia and Ib (and even a small Ic due to ejection)] and in II [IIA (aortic valve) and IIP (pulmonary valve)]. Systolic intervals LVEP (on carotid curve) and Q-IIA (on ECG and PCG) are indicated.

can be appreciated is ~20–30 ms. Similar guidelines are followed for the identification of phonocardiographic recordings: A sound is a complex of succeeding positive and negative deflections alternating with respect to the baseline, preceded and followed by a pause. A sound is said to be split if a small pause between the components can be perceived. At this point, the effect of frequency filtering may be important: Splitting, being invisible on a low frequency recording, may become recognizable on a high frequency recording (Fig. 1). Summarizing, we can state that in clinical PCG primarily the envelope of the recorded signal is regarded, not the actual waveform as, for example, in ECG, blood pressure, and velocity recordings. As spectral performance of phonocardiography may exceed the possibilities of human hearing inaudible low frequency phenomena can be recorded; They are also indicated as "(inaudible) sounds".

Acoustic phenomena originated by the heart are classified into two categories: heart sounds and heart murmurs (1–5,10–12). Although the distinction between them is not strict, one can state that heart sounds have a more transient, musical character (cf. the touching of a string) and a short duration (Fig. 1), whereas most murmurs have a predominantly noisy character and generally (but not always) a longer duration (e.g., a "blowing" murmur, a "rumbling" murmur) (Fig. 2). It is also believed that the genesis of both types is different: Heart sounds are indicated as types of resonant phenomena of cardiac structures and blood as a consequence of one or more sudden events in the cardiohemic system (such as valve closure), and most heart murmurs are said to be originated by blood flow turbulence. Many aspects of the problem of the genesis of these phenomena are still being discussed, including the relative importance of the valves and of the cardiohemic system in the generation of heart sounds (valvular theory versus cardiohemic theory).

Four normal heart sounds can be described (Fig. 1): I, II, III, and IV (also indicated as S1, S2, S3, S4). The two having the largest intensity, that is, the first (I, S1) and the second (II, S2) sound, are initially related to valve closure. The third (III, S3) and the fourth (IV, S4) sound, appearing extremely weak and dull and observable only in a restricted group of people, are not related to valve effects. The so-called closing sounds (I and II) are not originated by the coaptation of the valve leaflets (as the slamming of a door). On the contrary, it is most probably a matter of resonant-like interaction between two cardiohemic compartments suddenly separated by an elastic interface (the closed valve leaflets) interrupting blood flow: Vibration is generated at the site of the valve with a main direction perpendicular to the valve orifice plane and dependent on the rapid development of a pressure difference over the closed valve. In the case of the first sound, this phenomenon is combined with the effect of a sudden contraction of cardiac ventricular muscle. Pathologies of the cardiohemic system can affect the normal sounds with respect to intensity, frequency content, timing of components (splitting) (1).

The first heart sound (I) occurs following the closing of the mitral valve and of the tricuspid valve, during the isovolumic contraction period, and, furthermore, during the opening of the aortic valve and the beginning of ejection. In a medium or high frequency recording, a splitting of the first sound may be observed. Components related to the closing of the mitral valve (Ia, M1), the closing of the tricuspid valve (Ib, T1) and the opening of the aortic valve may be observed. There is a direct relation between the intensity of I and the heart contractility, expressed in the slope of ventricular pressure rising; with high cardiac output (exercise, emotional stress, etc.) sound I is enhanced. The duration of the PR-interval (electrical conduction time from the physiological pacemaker in the right

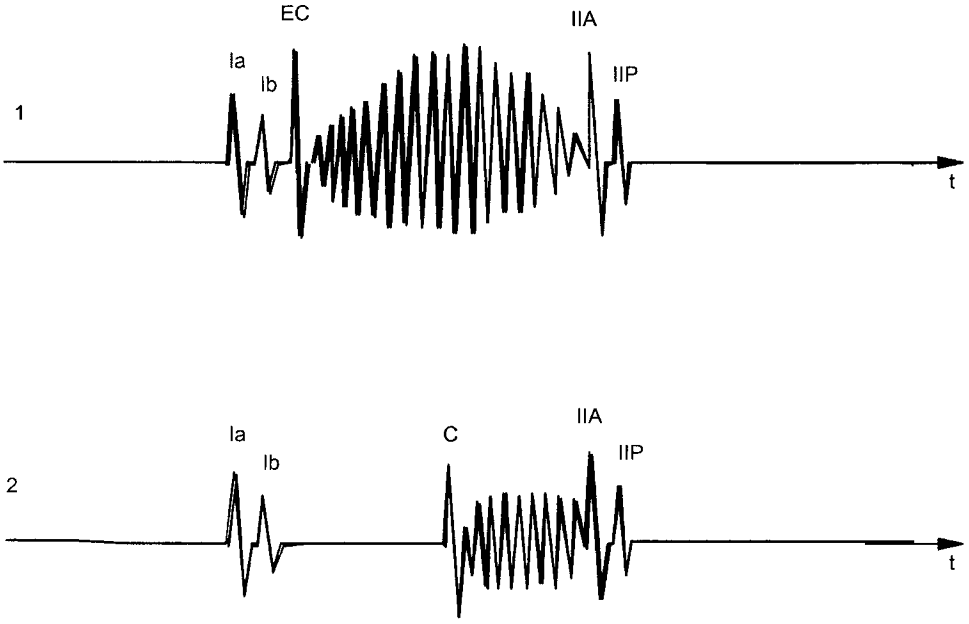


Figure 2. Examples of pathological sounds and murmurs. (1). A systolic murmur (ejection murmur, crescendo, decrescendo) as a consequence of aortic valve stenosis preceded by a clear aortic ejection click (EC). (2). Midsystolic click (C) as a consequence of mitral valve prolapse followed by a systolic murmur due to mitral valve regurgitation.

atrium to the ventricles) is a determining factor: The shorter the time between the atrial and ventricular contraction and, consequently, the larger the distance between the mitral valve leaflets, the larger the intensity of the first sound appears. With a long PR-interval mitral valve leaflets have evolved from wide open during atrial contraction to a state of partially open to almost closed when ventricular contraction starts; the result is a weak first sound. Cardiovascular pathologies can have an effect on timing and intensities of the first heart sound components. Wide splitting is observed in right bundle branch block, tricuspid stenosis, and atrial septal defect due to a delayed tricuspid component (Ib). In left bundle branch block Ia and Ib can coincide resulting in a single sound I. A diminished sound I is found in cases of diminished contractility (myocardial infarction, cardiomyopathy, heart failure), in left bundle branch block, mitral regurgitation and aortic stenosis; An intensified sound I is found in mitral stenosis with mobile valve leaflets and in atrial septal defect.

The second sound (II) is associated with the closure of the aortic valve and, following, the closure of the pulmonary valve. Splitting of the sound in an aortic (IIA, A2) and a pulmonary (IIP, P2) component is often observed. Splitting increases during inspiration as a consequence of increased difference in duration of left and right ventricular systole caused by increased right and decreased left ventricular filling; both components may fuse together at the end of expiration. Paradoxical splitting (the pulmonary component preceding the aortic one) is pathological. The pulmonary component normally has a lower intensity; an increased intensity with respect to the aortic component is generally abnormal. There is a direct relation between the intensity and the frequency of II and the slope of ventricular pressure falling during isovolumic relaxation. Stiffening of the valve leaflets results in a reduction of II. A higher valve radius or a lowered blood viscosity gives rise to an increased second sound. Cardiovascular pathologies can have an effect on timing and intensities of the second heart sound components. Wide splitting of sound II can be due to delayed pulmonary valve closure or advanced aortic valve closure. Delayed pulmonary valve closure can be caused by right bundle branch block, pulmonary stenosis, pulmonary hypertension, atrial septal defect; advanced aortic valve closure can result from mitral regurgitation and ventricular septal defect. Paradoxical splitting of sound II can be due to delayed aortic valve closure or advanced pulmonary valve closure. Delayed aortic valve closure can be caused by left bundle branch block, aortic stenosis and arteriosclerotic heart disease. Advanced pulmonary valve closure can be caused by tricuspid regurgitation and advanced right ventricular activation. IIA, respectively, IIP, can be absent in severe aortic, respectively, pulmonary, valve stenosis. IIA is decreased in aortic regurgitation and in pathologically diminished left ventricular performance.

The third sound (III) occurs during the rapid passive filling period of the ventricle. It is believed that III is initiated by the sudden deceleration of blood flow when the ventricle reaches its limit of distensability, causing vibrations of the ventricular wall. It can often be heard in normal children and adolescents, but can also be registered in adults (although not heard) in the low frequency channel. It is a weak and low pitched (low frequency) sound. Disappearance of III is a result of aging as a consequence of increasing myocardial mass having a damping effect on vibrations. High filling rate or altered physical properties of the ventricle may cause an increased third sound. If III reappears with aging (beyond the age of 40 years) it is pathological in most cases. A pathological sound III is found in mitral regurgitation, aortic stenosis, ischemic heart disease.

The fourth sound (IV) coincides with the atrial contraction and thus the originated increased blood flow through the mitral valve with consequences as mentioned for the third sound. It is seldom heard in normal cases, sometimes in older people, but is registered more often in the low frequency channel. The sound is increased in cases of augmented ventricular filling or reduced ventricular distensability. A pathological sound IV is found in mitral regurgitation, aortic stenosis, hypertensive cardiovascular disease, and ischemic heart disease.

Besides these four sounds, some pathological heart sounds may be present (Fig. 2). Among the systolic sounds there is the ejection sound and the nonejection systolic click. The ejection sound can be found in different pathological conditions such as congenital aortic or pulmonary valvular stenosis where opening of the cusps is restricted. A nonejection systolic click may be associated with a sudden mitral valve prolapse into the left atrium. An opening snap, a diastolic sound, may occur at the time of the opening of the mitral valve, for example, in cases with valve stenosis.

Heart murmurs are assumed to be caused by different mechanisms as compared to heart sounds. In fact, most murmurs result from turbulence in blood flow and occur as random signals. In normal blood vessels at normal velocity values blood flow is laminar, that is, in layers, and no turbulence is observed. In a normal resting human, there may be turbulent flow only in the vicinity of the aortic and pulmonary valves. As flow turbulence, a phenomenon that is generally irregular and random, is associated with pressure turbulence and, consequently, vessel wall vibration, acoustic phenomena may be observed. For flow in a smooth straight tube, the value of the Reynolds number, a dimensionless hydrodynamic parameter, determines the occurrence of turbulence. This number is proportional to the flow velocity and the tube diameter, and inversely proportional to the viscosity of the fluid. If this number exceeds a threshold value, laminar flow becomes turbulent. According to this theory, so-called innocent murmurs can be explained: They are produced if cardiac output is raised or when blood viscosity is lowered; they are generally early or midsystolic, have a short duration, and coincide with maximum ventricular outflow. Turbulence and thus intensity of the murmur increase with flow velocity. Pathological murmurs may be originated at normal flow rate through a restricted or irregular valve opening (e.g., in cases of valve stenosis) or by an abnormal flow direction caused by an insufficient (leaking) valve or a communication between the left and the right heart. As such systolic, diastolic, or even continuous murmurs may be observed. Systolic ejection murmurs occur in aortic and in pulmonary stenosis (valvular or non-valvular), diastolic filling murmurs in mitral and tricuspid stenosis. Aortic and pulmonary regurgitation cause diastolic murmurs; mitral and tricuspid regurgitation cause systolic murmurs. A systolic murmur and a diastolic murmur can be observed in ventricular septal defect. Continuous murmurs occur in patent ductus arteriosus (a connection between pulmonary artery and aorta). Musical murmurs occur as deterministic signals and are caused by harmonic vibration of structures (such as a valve leaflet, ruptured chordae tendinae, malfunctioning prosthetic valve) in the absence of flow turbulence; these are seldom observed.

The location of the chest wall where a specific sound or murmur is best observed (in comparison with the other phenomena) may help in discriminating the source of the sound or the murmur (1). These locations are dependent, not only on the distance to the source, but also on the vibration direction. Sounds or murmurs with an aortic valve origin are preferably investigated at the second intercostal space right of the sternum and those of pulmonary origin left of the sternum. The right ventricular area corresponds with the lower part of the sternum at the fourth intercostal space level, the left ventricular area between the sternum and the apex point of the heart (at the fifth intercostal space level). Furthermore, specific physiological maneuvers influencing cardiac hemodynamics may be used for obtaining better evaluation of heart sounds and murmurs.

In conclusion, the existence, timing, location at the chest wall, duration, relative intensity and intensity pattern, and frequency content of murmurs and/or pathological sound complexes form the basis of auscultatory, and/or phonocardiographic diagnosis of cardiac disease.

FUNDAMENTAL ASPECTS OF HEART VIBRATIONS

Mechanical heart action can be indicated by a set of time signals, which can be measured by invasive means: Most important variables are blood pressure in heart cavities and in blood vessels, myocardial and vessel wall tension, ventricular volume, blood flow velocity, heart wall deformation, and movement. At the chest surface only kinematic information is available: the movement of the chest surface as a result of mechanical heart action. As, in general, the movement of a material point can be indicated by a vector and as this vector appears to differ at various chest wall sites, one can state that mechanical information available at the chest wall is described by a spatiotemporal vector function. As far as the effect of the heart is concerned the movement is defined with respect to an equilibrium position; thus one can speak of a vibratory phenomenon. This movement of the chest surface, surrounded with air, gives rise to acoustic pressure in air; the latter is generally so weak that nothing can be perceived by hearing at a distance from the chest wall (except for artificial valve cases). Only if closed air cavities are used is a sound effect observable by the ear: The closed cavity (such as the stethoscope) prevents dispersion of acoustic energy and thus attenuation of acoustic pressure. It is out of the spatiotemporal kinematic vector function that phonocardiography takes a sample in order to evaluate cardiac activity.

As there is a vector function involved, three components should be taken into account. In practice, only the component perpendicular to the chest surface is measured and the two tangential components are disregarded. A kinematic function may be represented by different time representations, for example, displacement (m), velocity (m/s), acceleration (m/s^2), or even higher time derivatives (m/s^n, n representing the order of time derivative). Fundamentally, each representation contains identical information as they are all connected by a simple mathematical operation, that is, time derivation, but for visual inspection or time signal processing they reveal different vibratory patterns. Speaking in terms of the frequency domain, time derivation implies multiplication of the amplitude of an harmonic with its frequency: Time derivation is thus an operation of emphasizing higher frequencies in the signal with respect to the lower ones. According to linear system theory, a similar effect is obtained with high pass filtering, the effect of filtering being described by the Nth time derivative of the signal in the attenuation band (i.e., for frequencies well below the cutoff frequency). The number N represents the order of the filter and determines the slope in the attenuation band of the amplitude characteristic ($N \times 20$ dB/decade). High pass filtering, order N, is theoretically identical with the corresponding low pass filtering of the Nth time derivative of the signal.

In biomedical signal processing, it is relatively uncommon to consider different time representations. From auscultation, we learned that in case of PCG it is rather beneficial. In a chest wall displacement curve, no sounds can be perceived, for example, at the site of the apex of the heart one can measure the apexcardiogram (Fig. 1), which is essentially a recording of the displacement of the skin surface and the ordinate axis could have a millimeter scale. Nevertheless, at the site of the apex sounds can be recorded if time derivation or high pass filtering is applied. In practice, transients corresponding to heart sounds become clearly visible in the acceleration recording. The ear cannot sense displacements such as the apexcardiogram; this can simply be explained if one regards the ear's sensitivity curve. In the range below < 1000 Hz, the sensitivity increases with increasing frequency, equivalent with the effect of high pass filtering.

The kinematic effect of heart action at the chest surface is not completely covered by phonocardiography. Historically, the frequency spectrum is divided into two parts: the low frequency part (up to 20 Hz) is handled under the title mechanocardiography and the second part beyond 20 Hz under PCG. The reason of separation lies in the nature of auscultation (frequencies beyond 20 Hz, according to hearing performance) and, additionally, palpation (frequencies < 20 Hz, according to tactile sensitivity). According to this, displacement recording belong to the domain of mechanocardiography, which studies arterial and venous pulse tracings, and the apexcardiogram. The carotid pulse (Fig. 1) is a recording of skin surface displacement at the site of the neck where the carotid artery pulsation is best palpated. The curve reflects local volume changes, and consequently pressure changes in the artery at the measurement site. It thus reflects changes in aortic pressure after a time delay determined by the propagation time of the blood pressure wave (10–50 ms). The beginning of the upstroke of the graph corresponds to the opening of the aortic valve and the dicrotic notch corresponds to its closing. As such, the left ventricular ejection period (LVEP) can be derived. The apexcardiogram (Fig. 1) is a recording of skin surface displacement of the chest wall at the site of the heart apex; in this case there is no propagation delay. The abrupt rise of the graph corresponds to the isovolumic contraction, the fall with the isovolumic relaxation. The minimum point occurs at the time of opening of the mitral valve. These displacement curves have been used for identifying heart sounds.

Chest wall kinematics are not exclusively caused by heart action. The most important movement in a resting human is originated by the respiration act. Two phenomena should be mentioned: a low frequency event corresponding with the breathing movement itself, and having its fundamental frequency at the breathing rhythm (~0.2–0.4 Hz), and a high frequency phenomenon corresponding to breathing noises, "lung sounds", due to the turbulent air stream in airways and lungs. The latter may cause great disturbances in high frequency heart sound recording. To these effects one can add the result of stomach and intestine motility and, moreover, environment noise picked up by the chest wall. From the standpoint of PCG, these effects are merely disturbing and thus to be minimized.

In PCG, one discriminates between heart sounds and murmurs (based on auscultation). It has already been indicated that the source types are different as well as the acoustic impressions they provoke. From the standpoint of signal analysis heart sounds correspond better with transients originated by a sudden impact, whereas murmurs, except for the musical types, have a random character. In fact, if one considers a set of subsequent heart cycles, one may find that heart sounds are more coherent compared to murmurs. For example, averaging of sounds of subsequent heart cycles with respect to an appropriate time reference gives a meaningful result, but the same fails for murmurs as a consequence of their random character.

Conventional heart sound recording is executed at the chest wall. Some exceptions are worth mentioning. Intracardiac phono signals are obtained from cardiac blood pressure curves during catheterization. If a catheter-tip pressure transducer with sufficient bandwidth is used, intracardiac sound recordings can be obtained by submitting the pressure signal to high pass filtering or time derivation. It is also possible to get closer to the heart in a noninvasive way by measuring pressure or kinematics in the esophagus. In this way, the posterior part of the heart, lying close to the esophagus, is better investigated.

RECORDING OF HEART SOUNDS

In auscultation, the physician uses a stethoscope as a more practical alternative for putting the ear in close contact to the chest wall. Recording of heart sounds is a problem of vibration measurement (13), more specifically on soft tissue. It implies the need of a sensor, appropriate amplification and filtering, storage and visualization on paper (14,15), or by using information technology. The useful bandwidth is ~20–1000 Hz. The sensor needs to be a vibration transducer (vibration pickup), in this case also called a heart sound microphone; an alternative is a stethoscope provided with a microphone: the electronic stethoscope. Except for the sensor, virtual instrumentation technology can be used in the measuring chain: This implies a PC with a data acquisition card and signal processing software (such as Labview).

The Transducer

In a normal situation the chest wall vibrates only surrounded with air, which exerts an extremely low loading effect. Consequently, force at the surface may be neglected and only the kinematic variable is important. This ceases to be true when a transducer is connected to the chest wall, exerting a significant loading effect. This loading influence is described by the mechanical impedance of the pickup (force/velocity) or by the dynamic mass (force/acceleration). The loading effect is also dependent on the chest wall tissue properties; chest wall mechanical impedance is determined by tissue elasticity, damping, and mass. In general, heart sound microphones provoke a large and difficult to quantify loading effect on the chest wall, so that in no case the standard unloaded vibration is recorded. The same is not true in electrocardiography, for example, where the apparatus is designed with a sufficiently high input impedance to record the unaffected electrical potential at the electrode site. The latter is hard to achieve in phonocardiography (16).

Heart sound transducers can be divided into two types: the absolute pickup and the relative pickup. The absolute pickup measures the vibration at the site of application, averaged over the area of application. In general, these are contact pickups that are rigidly connected to the chest wall; the measuring area is thus identical to the contact area. These types are similar to the ones used for industrial measurements on mechanical structures or in seismography. The relative pickup measures the vibration averaged over a certain area with respect to a reference area; it is thus a kind of differential pickup. Air-coupled pickups are differential pickups. Essentially they consist of an air-filled cavity, generally with a circular shape, the edge of the cavity rigidly and air-tight-connected to the chest wall. It is thus the difference of the displacement under the cavity (the measuring area) and the displacement under the edge of the cavity (the reference area) giving rise to an acoustic air pressure within the cavity that is measured. The electronic stethoscope can be considered as a relative pickup. Figure 3 shows the principles.

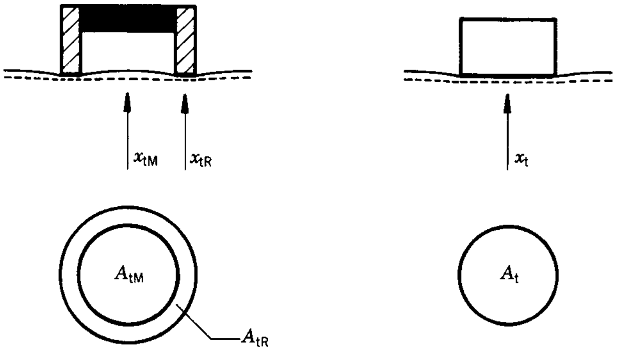


Figure 3. A schematical representation of heart vibration pickups: left, the relative type; right, the absolute type; above, the pickup positioned at the vibrating chest wall; below, the corresponding areas relating the pickup with the chest wall. The absolute pickup measures the kinematics x_t at its contact area A_t. The relative pickup, presented as an air cavity with a pressure-sensing device at the top of the cavity (black part), measures the difference between the kinematics under the cavity x_{tM}, that is, of the measuring area A_{tM}, and the one under the edge of the cavity x_{tR}, that is, the reference area A_{tR}.

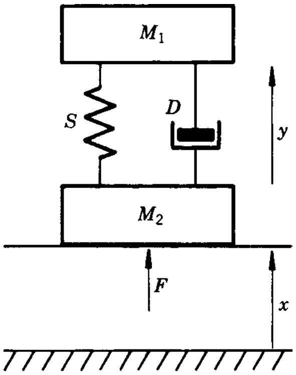


Figure 4. The seismic system, the mechanical model for a contact pickup (i.e., an accelerometer). The system is attached to the vibrating surface (F, x) with its contact mass M_2; The seismic mass M_1 is coupled to the contact mass via a stiffness S and a damping D. The displacement y between the contact and the seismic mass represents the measuring value.

With the contact vibration pickup the average kinematics of the measuring area in the loaded situation is recorded. The most generally applied transducing principle is the seismic type: A seismic mass is connected via a spring-damping system to the contact mass coupled with the vibrating surface (Fig. 4). The relative displacement between the seismic mass and the contact mass is measured with the aid of a mechanoelectric transducing device. The latter can be a piezoelectric crystal that generates an electrical charge proportional to its deformation (Fig. 5); The complete device behaves as an accelerometer for frequencies (f) below its acceleration resonance frequency (f_1); it measures displacement above f_1. Measuring acceleration is generally the normal function. The acceleration charge sensitivity s_Q (charge per acceleration unit, pC/m·s^{-2}) is thus

$$s_Q = B/(2\pi f_1)^2 \tag{1}$$

for $f < f_1$. The parameter B stands for a mechanoelectric transducing efficiency and depends on the crystal type and

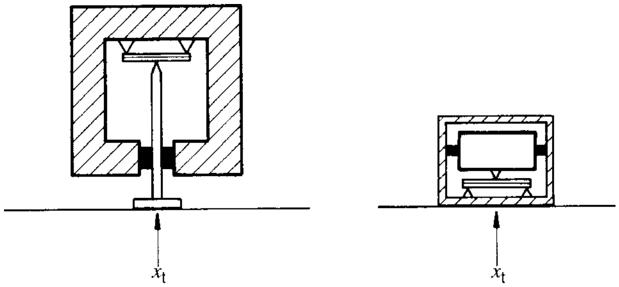


Figure 5. A schematic representation of two types of phonocardiographic contact pickups: at left, the heavy, and at right, the low weight type. The mechanoelectric transducing device (here presented as a bilaminar bending crystal) is built in to measure the displacement between the seismic mass and the material in contact with the vibrating surface. The black parts indicate elastic material. The heavy type (possibly held by hand at the massive case, representing the seismic mass) makes contact with the chest wall via a coin-shaped disk connected to the crystal by an extension rod. For the low weight type, the case makes contact with the vibrating surface and all remaining parts (including the seismic mass attached to the crystal) are at the inside. *Note*: The heavy type is mentioned because of its historical value.

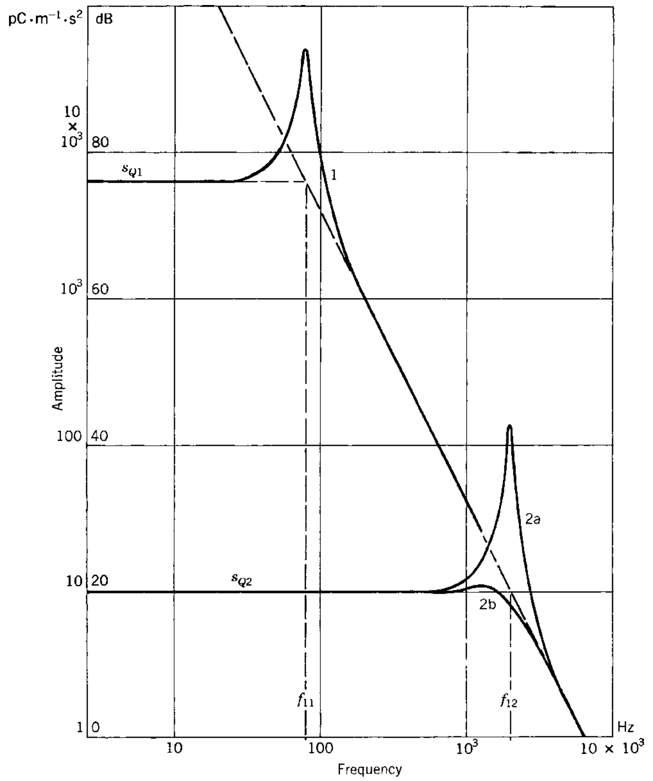


Figure 6. The acceleration amplitude frequency characteristic of a seismic pickup (double logarithmic scale). Curves presented can be obtained using the same mechanoelectric transducing device, but with a different mechanical layout of the seismic system. The oblique broken line represents a sensitivity limit for the specific mechanoelectric transducing device. The characteristic of a seismic system corresponds with a second-order low pass filter. The resonant frequency f_1 determines the bandwidth (flat part), but also the sensitivity (Eq. 1). Two different types are shown; the low resonant frequency f_{11} can be found in heavy phonocardiographic pickups; the high one, f_{12}, in low weight types. For the latter, the effect of damping is shown: Curve 2b corresponds to a higher damping in the system, as can be seen by the decreased height of the resonance peak.

on its mounting; f_1 is a construction parameter determined by the seismic mass and the stiffness incorporated between the seismic and the contact mass. The complete amplitude frequency characteristic corresponds with a low pass second-order filter, with f_1 also representing the cutoff frequency (Fig. 6), thus determining the measuring bandwidth and the sensitivity. Evidently other then piezoelectric elements can be used for measuring displacement: piezoresistance, variable capacitance (both of them needing a polarization voltage), electret elements, and so on.

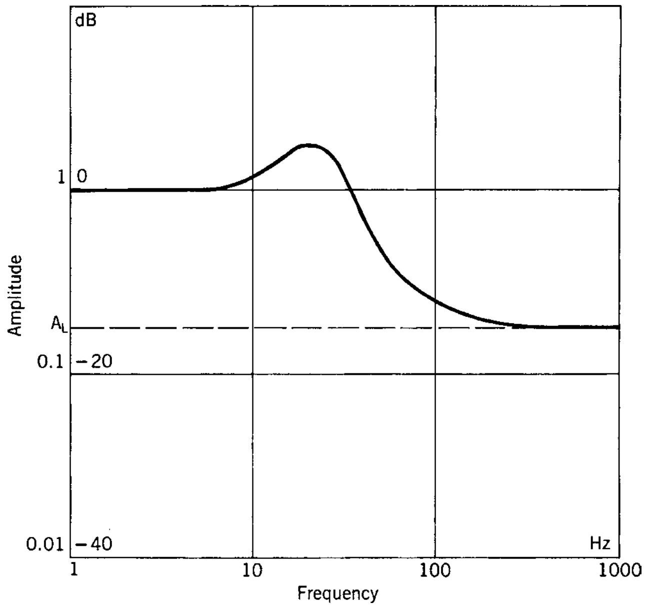


Figure 7. Amplitude frequency characteristic of the loading effect caused by the coupling of a rigid element to the soft tissue of the chest wall (as e.g., a contact pickup) (drawing for a 10 g, 20 mm diameter element). Attenuation is about constant beyond 100 Hz (Eq. 2).

The loading influence of the pickup can be presented by its dynamic mass. In comparing contact pickups with different masses with an ultralow weight type (Fig. 7) it was found that distortion and attenuation is caused by the mass (16): Beyond 100 Hz, the amplitude ratio (loaded compared to unloaded) approximates a value (A_{L}):

$$A_{\mathrm{L}} = M_{\mathrm{t}}/(M_{\mathrm{t}} + M_{\mathrm{l}}) \qquad (2)$$

with $M_{\mathrm{t}} = a\, d_{\mathrm{t}}^3$, $a = 280$ kg/m^3 (typical); M_{l} is the loading mass (of the pickup), M_{t} is the thorax wall output mass, and d_{t} is the contact diameter. According to this formula, a 10-g pickup with a 20 mm diameter would result in an attenuation to 18% of the original unloaded amplitude. For quantitative purposes, ultralow weight pickups can thus be recommended. It must also be emphasized that not the weight per se, but the weight divided by the third power of the contact diameter is the parameter to be minimized (according to Eq. 2).

In the case of the air-coupled vibration pickup (17), the average kinematics of the measuring area (under the cavity) with respect to the reference area (under the edge of the cavity) in the loaded situation is recorded (Fig. 3). Air pressure in the cavity as a result of the relative displacement of the chest wall is registered with a built-in sensor measuring acoustic pressure (a microphone). The movement of the membrane of this microphone is transformed into an electrical signal, for example, by the moving coil principle (dynamic type) and variable capacitance (condenser type with a polarizing voltage, electret type). As such, the measuring characteristics of the air-coupled pickup are determined by the dimensions of the air cavity and by the features of the included microphone. If the microphone membrane is rather stiff as compared with air and the height (l) of the cavity small as compared with the wavelength of heart vibrations in air, the pressure (p) generated at the site of the membrane is simply proportional to the relative displacement of the chest wall ($x_{\mathrm{tMR}} = x_{\mathrm{tM}} - x_{\mathrm{tR}}$):

$$p = (c^2/l)\, x_{\mathrm{tMR}} \qquad (3)$$

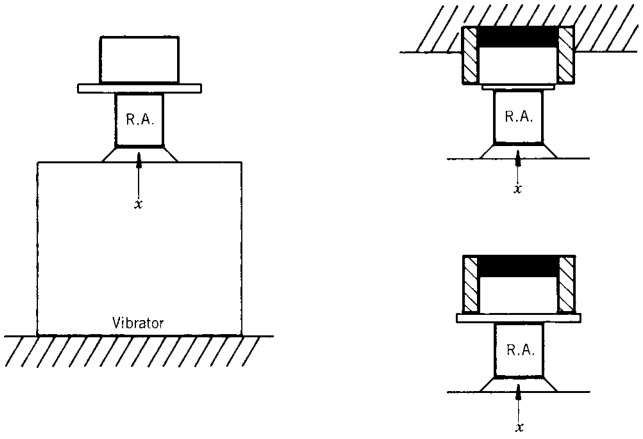


Figure 8. Calibration of heart vibration pickups. The pickup to be tested is compared with a reference accelerometer (RA): left, the contact pickup; right, the relative pickup. The contact pickup is rigidly connected with the reference accelerometer, mounted on a vibrator (left). For the relative pickup two measurements can be performed (right, vibrator not displayed). In a first test, the differential characteristic is determined (right, above): With the housing fixed, a movement is generated at the orifice of the cavity, which is closed airtight with (e.g., an elastic membrane). In a second test the common mode sensitivity is determined, the pickup fixed at a stiff plate. An ideal relative pickup features zero common mode sensitivity.

where ρ is the air density (kg/m^3) and c is the wave propagation velocity in air (m/s). The loading effect under the cavity is given by a stiffness term (S_{tM}, N/m) and

$$S_{\mathrm{tM}} = d_{\mathrm{tM}}^2 \rho c^2/(4l) \tag{4}$$

where d_{tM} is the diameter of the measuring area. At the reference area, the pickup exerts a mass loading; the combined loading effect is hard to describe.

For calibration of a phonocardiographic pickup one can use a vibrator, a reference accelerometer having an ideal frequency response, and two amplifiers (Fig. 8). The contact pickup to be tested should be rigidly attached to the reference pickup. In case of an air-coupled pickup, one has to investigate the differential and the common mode characteristics, the first one with the housing fixed, a displacement being generated at the input of the cavity (e.g., airtight sealed with an elastic membrane), and the second one with the complete microphone fixed at the rigid vibration table (the relative displacement between inner and edge of the cavity being zero). For an ideal relative pickup, the common mode term should be zero; In practice, it is to be small with respect to the differential sensitivity. Mechanical impedance of a pickup can similarly be measured by applying a force transducer combined with the reference accelerometer (an impedance head). It should be emphasized that in this way only the properties of the pickup can be obtained, but no information is acquired on the ultimate distortion effect due to the loading of the chest wall, as this is also dependent on the tissue parameters. This effect can only be measured *in situ*, that is, on the patient themself; the procedure is rather complicated.

In conclusion, accurate quantitative recording of the unloaded chest wall vibration is extremely difficult. Industrial accelerometers (equivalent to the described contact types) show the same disadvantages with respect to loading on the soft tissue of the chest wall. Moreover, as a rule they are less sensitive and have an unnecessary broad bandwidth. Some noncontact methods have been reported, but only for infraphonocardiographic frequencies.

Preamplification and Signal Preprocessing

Without going into details regarding electronics, we must emphasize the importance of the preamplifier, as this part of the recording unit, in combination with the pickup, determines the electronic noise level of the apparatus. It is adequate to resume possible disturbing influences at this point. First, there are the physiological vibrations generated by organs other than the heart. Respiration sounds may be the most inconvenient disturbances, especially with patients suffering from lung disease. Recording in an expired resting state can thus be advised. Furthermore there are environmental vibrations; Air-transmitted noises, especially the higher frequencies at which cardiac vibration intensity is weak, can be very inconvenient. Therefore, the air-coupled pickup should be attached in an airtight manner to the chest wall. This does not solve the problem completely, as environmental noises seem to be picked up by the chest wall and, in this way, transmitted to the pickup (air-coupled as well as contact type). Theoretically, these disturbing vibrations can be minimized, but for a given apparatus electronic noise cannot be affected and thus sets a limit to noise diminishing and determines a threshold under which no heart sound or murmur can be recognized.

Besides the noise level, the frequency characteristic of the preamplifier connected to the pickup should be regarded. For example, in the case of a piezoelectric pickup, its electrical output property is capacitive and one must bear in mind that the combination of this capacitance with the input resistance of a voltage preamplifier gives rise to a first-order high pass effect. To avoid this problem, charge or current amplifiers can be used. Whereas the charge amplifier measures acceleration, the current amplifier measures its time derivative.

Whether or not signals are digitized after preamplification, a high pass filtering (or band-pass filtering) process is necessary (18,19). High pass filtering and appropriate amplification (analogue or digital) of the filter channels compensate for the fact that visual inspection of a single recorded phonocardiogram (even when optimally chosen) does not reveal the same amount of information as gained from the acoustical impression during auscultation. Furthermore, the amplification has to compensate for the decreasing amplitude of heart vibrations at increasing frequencies. Conventionally, a set of about four high pass filters are used, each characterized by a gradually increasing cutoff frequency and/or increasing slope in the attenuation band. For example, Maass Weber high pass filters are used with a common cutoff frequency of 1 kHz and slopes of 20, 30, 40, and 60 dB/octave (EEC, see the section Hardware and Software). Generally, filter sets have been determined qualitatively by applying a range of subsequent filters and recording a set of normal subjects and patients with different heart diseases: A filter providing information also perceivable in another one was eliminated from the set. Furthermore, filtering must permit discrimination between phenomena having a different physiological or pathological origin. Clear splitting between slightly asynchronous phenomena (or minimal overlapping) is thus desired for vibrations having similar frequency content. Now discrimination between phenomena having a different frequency content in adjacent filter channels is expected. The chosen set is not uniquely optimal: It depends on the preceding elements in the measuring chain (the vibration pickup, including its loading effect, and the preamplifier). As such, a filter set chosen for a contact pickup is not evidently optimal for an air-coupled type.

Different transducers, the mostly unknown distortion effect due to loading and different filter sets, might seem remarkable to physicists and engineers from the viewpoint of measuring quality. Nevertheless, for (semiquantitative) phonocardiography the use of filtering with adaptable amplification compensates in some degree for microphone, loading, and preamplification characteristics. For example, attenuation due to loading in a specific frequency band may be partly compensated by increased amplification of the corresponding channel.

Storage and Visualization

In older apparatus intended for recording of ECG, PCG, and pulses, a paper recorder (strip chart recorder) was included. The first was an analog type [as the galvanometric pen writer, having a limited bandwidth (~100 Hz), but equipped with special techniques for recording high frequency sounds], and later it was a digital type as the thermal array recorder. The latter, also available as a general purpose paper recorder, functions completely digitally: It sets a dot at the point corresponding to the instantaneous value of the signal to be recorded: No moving mechanical parts are present except for the paper drive. The recording technique is characterized by a sampling and a writing frequency. The latter may be lower than the first: During the writing interval all points between the maximum and the minimum value of the signal samples are dotted. As such, the recording is a subsequence of vertical lines: For visual inspection of the overall vibration pattern no information is lost.

Furthermore, data can be handled by common information technology: a (portable) personal computer with appropriate data-acquisition possibilities, virtual instrument software for signal conditioning and processing, visualization, archiving, and hard copy generation.

PROCESSING OF HEART SOUNDS AND PHYSICAL MODELING

Physical modeling aims at the localization of a specific sound source in the heart and, by analyzing the externally recorded vibration signals, at the quantification of the constitutive properties of the cardiac structures involved (e.g., stiffness of a valve leaflet, myocardial contractility) and of the driving forces, which set these structures into vibration. The physical situation is extremely complicated. The vibration source is situated within the cardiac structures (having viscoelastic properties) containing and driving blood (a viscous fluid). The transmission medium, the tissues between the heart and the chest wall, is viscoelastic and inhomogeneous.

Transmission in such a viscoelastic medium implies compression and shear waves, which both contribute to the vibrations at the chest wall (20). It is not simply a problem of acoustic pressure as in a perfect fluid. Distortion due to transmission seems obvious. In order to study transmission and to relate chest wall vibrations to properties of cardiac structures and hemodynamic variables, advanced signal processing techniques are used. A broad review is given by Durand et al. (21).

As the chest wall vibratory phenomenon is represented by a spatiotemporal kinematic function, it can principally be approached in two ways: by sampling in time, as a set of images of chest wall movement, or by sampling in space by a set of time signals obtained with multisite recording. Multisite heart sound recording implies a large set of pickups (preferably light weight, thus inducing minimal loading). In this way, spatial distribution of vibration waveforms on the chest wall can be derived. Based on the results of such a method a physical model for heart sound genesis has been presented that can analytically be solved in a viscoelastic medium: a sphere vibrating along the axis of the valve orifice (20). This mechanical dipole model agrees to the idea of sound generation as a resonant-like vibration of the closed elastic valve leaflets and the surrounding blood mass. With this model a typical inversion of vibration waveforms on the chest wall could be explained: The phase reversal is expressed most for the second sound, according to the anatomical position and direction of the aortic orifice. The model has been used to calculate source functions (the inverse problem). Spatial parameters on vibration waveforms have been formulated (22–25).

Physical modeling aims at the quantification of the constitutive properties of cardiac structures (e.g., of the valve leaflets) and the driving forces (e.g., blood pressure). For example, with respect to the second sound, the aortic valve was modeled as a circular elastic membrane, it was allowed to vibrate in interaction with the surrounding blood mass, with as a driving force the slope of the development of the transvalvular pressure difference during isovolumic relaxation (11,26). Typical characteristics of IIA and IIP could thus be explained. For example, the reduction of amplitude and frequency shift (toward higher frequencies) as a consequence of valve stiffening, the diminishing of amplitude in patients with poor ventricular performance (characterized by a slow pressure drop in the ventricle during the isovolumic relaxation), and the augmentation of amplitude in cases of anemia (implying reduced blood viscosity and thus reduced damping in the resonant system). In another model, the ventricle is modeled as a finite thick-walled cylinder and the amplitude spectra of computed vibration waveforms contain information concerning the active elastic state of muscular fibers that is dependent on cardiac contractility (27).

Transmission of vibrations by comparing vibrations at the epicardial surface and at the chest wall has been studied (21). Esophageal PCG proved to be beneficial for recording vibrations originated at the mitral valve (28). The disappearance of the third sound with aging was explained with the ventricle modeled as a viscoelastic oscillating system with increasing mass during growth (29). Spectral analysis of the pulmonary component of the second sound reveals information on the pressure in the pulmonary artery (30).

Frequency content and timing of heart vibrations is of major importance; Time–frequency analysis of signals is thus performed. Classical Fourier analysis uses harmonic signals (sine and cosine waves) as basic signals. The frequencies of the harmonics are multiples of the fundamental frequency and the signal can be composed by summing the sine and cosine waves multiplied with the Fourier coefficients. Sine waves have an infinite duration and the method is thus beneficial for periodic functions. A phonocardiogram can be considered as a periodic function, but it is composed of a number of phenomena shifted in time with specific frequency content (heart sound components and murmurs). When applying classical Fourier analysis, information on timing is lost. Thus Fourier analysis has to be performed on shorter time intervals (by dividing the heart cycle into subsequent small intervals) resulting in time and frequency information. To minimize errors resulting from calculating in these small intervals, mathematical techniques have to be applied. Wavelet analysis calculates wavelet coefficients based on transient-like irregular signals with limited duration, called wavelets. Wavelets are derived from a mother wavelet and obtained by scaling in time (subsequently with a factor 2) and by shifting in time. As in Fourier analysis, the signal can be composed by summing shifted and scaled wavelets multiplied with their wavelet coefficients. The waveform of the mother wavelet can be chosen. As scaling in time corresponds to frequency, this method also gives time and frequency information, but it performs better for analyzing signals of a nonstationary nature, such as heart sounds and murmurs. Sudden changes or discontinuities in the signal can better be identified and located in time. A large number of studies has been executed with respect to time–frequency analysis of heart sounds and murmurs and different calculation methods have been compared (21). Spectral analysis of heart murmurs appeared to be useful to estimate transvalvular pressure difference in patients with aortic valve stenosis (31,32). Spectral analysis was used to monitor the condition of bioprosthetic valves and mechanical valve prostheses (33,34). Wavelet transform (35) and a nonlinear transient chirp signal modeling approach (36) were used to detect the aortic and the pulmonary component of the second sound. The matching pursuit method was used to identify the mitral and the tricuspid component in the first sound (37). A tailored wavelet analysis has been used to automatically detect the third heart sound (38). Time–frequency analysis was applied for classification of heart murmurs produced by bioprosthetic valves (39), for studying the first heart sound (40), and for automated detection of heart sounds (41).

THE ELECTRONIC STETHOSCOPE

Clinical interest in PCG in its classical form has been decreasing during the last decade, but there seems to be an increasing interest in heart sound recording with the aid of electronic stethoscopes (1), combined with information technology allowing easy data-acquisition, visualization, data handling and parameter extraction, playback, telemedicine applications, waveform recognition, and diagnosis with the aid of databanks. Also, virtual instrumentation technology (such as Labview) is used for heart sound acquisition and processing.

The modern acoustic stethoscope comprises a binaural headset and a chest piece connected by elastic tubing (1). The headset is composed of ear tubes and ear tips; the chest piece can consist of a bell and a diaphragm part. The ear tips should fit in the ear orifice, preventing air leakage, with the tube properly aligned to the hearing canal, that is slightly directed forward. The tube connecting the headset and the chest piece should not be to long to restrict attenuation of acoustic pressure, especially of higher frequencies, generated at the chest piece. With the diaphragm part provided with a stiff membrane (diameter ~4 cm), applied firmly to the skin, the high frequency sounds are better observed. With the bell part of the chest piece, applied with low pressure to the skin (enough to prevent air leaks between skin and bell edge) low frequency vibrations are best picked up. The bell diameter should be large enough to span an intercostal space (~2.5 cm for adults). Firm application of the bell makes the skin act as a membrane thus diminishing its low frequency performance. Some stethoscopes have only one part with a specially mounted membrane, which can function in the “bell mode” or in the “membrane mode” by altering applied pressure for the purposes cited above. As such, in the application of the stethoscope, frequency filtering (as in phonocardiography) is performed by using a specific shape and mechanical coupling to the chest wall.

The electronic stethoscope (e-stethoscope) combines the simplicity of the acoustic stethoscope with the benefits of electronics and information technology. Essentially, the e-stethoscope is an acoustical type provided with a built-in microphone; as such, it can be indicated as an air coupled vibration pick-up. In its simplest form, sounds are transmitted to the ears by tubing as in the acoustical one. The more advanced type has the microphone built within the chest piece, with adjustable amplification and filtering, mode control with easy switching between bell and diaphragm modes, generation of processed sound by miniature speakers to the air tips, cable or wireless connection to a personal computer for further processing. Adjustment of stethoscope performance can be executed during auscultation. Most stethoscopes are intended for observation (and recording) of heart sounds and murmurs, and for lung and airway sounds as well. A special type, the esophageal stethoscope, can be used for monitoring heart and lung sounds during anesthesia (42).

User-friendly software is available for diagnostic and for training purposes. Recorded signals can be printed, visualized, adapted by digital filtering and scaling, improved by elimination of artifacts and disturbances, and combined with synchronously recorded ECG. Processed sounds can be reproduced and played back with speakers with a sufficient bandwidth (in the low frequency range down to 20 Hz). Spectral analysis is also possible: Frequency content as a function of time can be displayed. Automated cardiac auscultation and interpretation can be useful in supporting diagnosis (43–45). Sounds recorded by a local general physician can be sent via internet to the cardiologist for accurate diagnosis (46).

Educational benefits are obvious. Heart sounds recorded with the e-stethoscope or obtained from a databank can be visually observed and listened to. CD–ROMs with a collection of typical heart sounds and murmurs are available for training purposes. Multimedia tools were found to contribute to the improvement of quality of physical examination skills (47,48).

HARDWARE AND SOFTWARE

In this paragraph, some practical details are given with respect to available hard and software. A conventional form of a phonocardiograph (heart sound transducer, amplifier and filters, audio, output connectable to recorder, also fit for lung sound recording) can be obtained from EEC (http://www.eeconnet.com). ADInstruments provides a heart sound pickup (http://www.adinstruments.com). Colin (http://www.colin-mt.jp/eng) provides a phonocardiograph together with ECG and noninvasive blood pressure measurement for noninvasive assessment of arteriosclerosis. Electronic stethoscopes can be purchased at Cardionics (http://www.cardionics.com), 3M (Litttmann) (http://www.3M.com/product/index.jhtml), Meditron (http://www.meditron.no/products/stethoscope), Philips (http://www.medical.philips.com/main/products/), EEC (http://www.eeconnet.com). Software supporting the physician in the evaluation of heart sounds recorded with an electronic stethoscope is provided by Zargis (http://www.zargis.com), Stethographics (http://www.stethographics.com/index. html). Software intended for training in heart sound auscultation (heart sounds recorded with an electronic stethoscope or from data banks) can be obtained from Biosignetics (http://www.bsignetics.com), Zargis (http://www.zargis.com), Cardionics (http://www.cardionics.com).

EVALUATION

Evaluation of heart sounds and murmurs remains an important method in the diagnosis of abnormalities of cardiac structures. Conventional PCG, however, essentially the graphic recording of sounds for visual inspection, has lost interest as a result of a number of reasons. First, the vibration signals are complex and thus difficult to interpret; they are characterized by a broad frequency range and, as such, different time representations present specific information (low and high frequencies). Obtaining high quality recordings having a high signal-to-noise ratio is difficult. Genesis and transmission of vibrations is difficult to describe and insufficiently known. A variety of waveforms are observable at the chest surface; Multisite recording and mapping are useful with respect to the solving of the genesis and transmission problem but are difficult to execute and result in a large amount of data to be analyzed. The recording technique is not standardized; The ordinate axis of a phonocardiographic waveform does not have a physical unit as, for example, the millivolt in electrocardiography. The latter is due to the different transducer types, unquantified loading effect of the transducer on the chest wall, different frequency filter concepts. Thus, the method remains bound to a specific recording method and is semiquantitative. No guidelines for universal use have been developed and proposed to the clinical users. The most important reason evidently is found in the availability of technologies like echocardiography, Doppler, and cardiac imaging techniques, which provide more direct and accurate information concerning heart functioning. The latter, however, have the disadvantages of being costly and restricted to hospitals. Nevertheless, knowledge of heart sounds and murmurs has been greatly increased with the PCG technique and research is still going on. Signal analysis, more specifically time–frequency analysis, has proven to be very useful in the identification and classification of heart sound components and murmurs and their relation to cardiac structures and hemodynamic variables.

Conventional PCG has lost interest. Nevertheless, the historical value of the method has to be stressed. Auscultation, being simple, cheap, and not restricted to the hospital environment, held its position as a diagnostic tool for the general physician and for the cardiologist as well. However, this technique requires adequate training. Recording and processing of heart sounds remain beneficial for training and for supporting diagnosis. Electronic stethoscopes coupled to a laptop with suitable software and connected to the internet for automated or remote diagnosis by a specialist may grow in importance in the coming years.

BIBLIOGRAPHY

1. Tilkian AG. Understanding heart sounds and murmurs with an introduction to lung sounds. Philadelphia: W.B. Saunders; 2001.
2. Salmon AP. Heart sounds made easy. London: Churchill Livingstone; 2002.
3. Wartak J. Phonocardiology: Integrated Study of Heart Sounds and Murmurs. New York: Harper & Row; 1972.
4. Luisada AA. The Sounds of the Normal Heart. St. Louis, MO: Warren H. Green; 1972.
5. Delman AJ, Stein E. Dynamic Cardiac Auscultation and Phonocardiography. A Graphic Guide. Philadelphia, PA: W.B. Saunders; 1979.
6. Einthoven W. Die Registrierung der Menschlichen Hertztone mittels des Saitengalvanometers. Arch Gesamte Physiol Menschen Tiere 1907;117:461.
7. Mannheimer E. Standardization of phonocardiography. Am Heart J 1957;54:314–315.
8. Holldack K, Luisada AA, Ueda H. Standardization of phonocardiography. Am J Cardiol 1965;15:419–421.
9. Groom D. Standardization of microphones for phonocardiography. Biomed Eng 1970;5:396–398.
10. Rushmer RF. Cardiovascular Dynamics. Philadelphia, PA: Saunders; 1976.
11. Stein PD. Physical and Physiological Basis for the Interpretation of Cardiac Auscultation. Evaluations Based Primarily on the Second Sound and Ejection Murmurs. New York: Futura Publishing Co.; 1981.

12. Luisada AA, Portaluppi F. The Heart Sounds. New Facts and Their Clinical Implications. New York: Praeger; 1982.
13. Harris CM. Shock and Vibration Handbook. 5th ed. New York: McGraw-Hill; 2001.
14. van Vollenhoven E, Suzumura N, Ghista DN, Mazumdar J, Hearn T. Phonocardiography: Analyses of instrumentation and vibration of heart structures to determine their constitutive properties. In: Ghista DN, editor. Advances in Cardiovascular Physics. Vol. 2, Basel: Karger; 1979. pp. 68–118.
15. Verburg J, van Vollenhoven E. Phonocardiography: Physical and technical aspects and clinical uses. In: Rolfe P, editor. Non Invasive Physiological Measurements. London: Academic Press; 1979. pp. 213–259.
16. Vermariën H, van Vollenhoven E. The recording of heart vibrations: A problem of vibration measurement on soft tissue. Med Biol Eng Comput 1984;22:168–178.
17. Suzumura N, Ikegaya K. Characteristics of air cavities of phonocardiographic microphones and the effects of vibration and room noise. Med Biol Eng Comput 1977;15:240–247.
18. Maass H, Weber A. Herzschallregistrierung mittels differenzierende filter. Eine Studie zur Herzschallnormung. Cardiologia 1952;21:773–794.
19. van Vollenhoven E, Beneken JEW, Reuver H, Dorenbos T. Filters for phonocardiography. Med Biol Eng 1967;5:127–138.
20. Verburg J. Transmission of vibrations of the heart to the chest wall. In: Ghista DN, editor. Advances in Cardiovascular Physics. Volume 5, Part III, Basel: Karger; 1983. pp. 84–103.
21. Durand LG, Pibarot P. Digital signal processing of the phonocardiogram: review of the most recent advancements. Crit Rev Biomed Eng 1995: 23(3–4):163–219.
22. Vermariën H. Mapping and vector analysis of heart vibration data obtained by multisite phonocardiography. In: Ghista DN, editor. Advances in Cardiovascular Physics. Volume 6, Basel: Karger; 1989. pp. 133–185.
23. Wood JC, Barry DT. Quantification of first heart sound frequency dynamics across the human chest wall. Med Biol Eng Comput 1994;32(4 Suppl):S71–78.
24. Baykal A, Ider YZ, Koymen H. Distribution of aortic mechanical prosthetic valve closure sound model parameters on the surface of the chest. IEEE Trans Biomed Eng 1995;42(4): 358–370.
25. Cozic M, Durand LG, Guardo R. Development of a cardiac acoustic mapping system. Med Biol Eng Comput 1998;36(4): 431–437.
26. Blick EF, Sabbah HN, Stein PD. One-dimensional model of diastolic semilunar valve vibrations productive of heart sounds. J Biomech 1979;12:223–227.
27. Lewkowicz M, Chadwick RS. Contraction and relaxation-induced oscillations of the left ventricle of the heart during the isovolumic phases. J Acoust Soc Am 1990;87(3):1318–1326.
28. Chin JGJ, van Herpen G, Vermariën H, Wang J, Koops J, Scheerlinck R, van Vollenhoven E. Mitral valve prolapse: a comparative study with two-dimensional and Doppler echocardiography, auscultation, conventional and esophageal phonocardiography. Am J Noninvas Cardiol 1992;6:147–153.
29. Longhini C, Scorzoni D, Baracca E, Brunazzi MC, Chirillo F, Fratti D, Musacci GF. The mechanism of the physiologic disappearance of the third heart sound with aging. Jpn Heart J 1996;37(2):215–226.
30. Chen D, Pibarot P, Honos G, Durand LG. Estimation of pulmonary artery pressure by spectral analysis of the second heart sound. Am J Cardiol 1996;78(7):785–789.
31. Nygaard H, Thuesen L, Hasenkam JM, Pedersen EM, Paulsen PK. Assessing the severity of aortic valve stenosis by spectral analysis of cardiac murmurs (spectral vibrocardiography). Part I: Technical aspects. J Heart Valve Dis 1993; 2(4):454–467.
32. Nygaard H, Thuesen L, Terp K, Hasenkam JM, Paulsen PK. Assessing the severity of aortic valve stenosis by spectral analysis of cardiac murmurs (spectral vibrocardiography). Part II: Clinical aspects. J Heart Valve Dis 1993;2(4): 468–475.
33. Sava HP, Grant PM, Mc Donnell JT. Spectral characterization and classification of Carpentier-Edwards heart valves implanted in the aortic position. IEEE Trans Biomed Eng 1996;43(10):1046–1048.
34. Sava HP, Mc Donnell JT. Spectral composition of heart sounds before and after mechanical heart valve implantation using a modified forward-backward Prony's method. IEEE Trans Biomed Eng 1996;43(7):734–742.
35. Obaidat MS. Phonocardiogram signal analysis: techniques and performance comparison. J Med Eng Technol 1993; 17(6):221–227.
36. Xu J, Durand LG, Pibarot P. Extraction of the aortic and pulmonary components of the second heart sound using a nonlinear transient chirp signal model. IEEE Trans Biomed Eng 2001;48(3):277–283.
37. Wang W, Guo Z, Yang J, Zhang Y, Durand LG, Loew M. Analysis of the first heart sound using the matching pursuit method. Med Biol Eng Comput 2001;39(6):644–648.
38. Hult P, Fjallbrant T, Wranne B, Ask P. Detection of the third heart sound using a tailored wavelet approach. Med Biol Comput 2004;42(2):253–258.
39. Debiais F, Durand LG, Guo Z, Guardo R. Time-frequency analysis of heart murmurs, Part II: Optimisation of time-frequency representations and performance evaluation. Med Biol Eng Comput 1997;35(5):480–485.
40. Chen D, Durand LG, Lee HC, Wieting DW. Time-frequency analysis of the first heart sound. Part 3: Application to dogs with varying cardiac contractility and to patients with mitral mechanical prosthetic heart valves. Med Biol Eng Comput 1997;35(5):455–461.
41. Sava H, Pibarot P, Durand LG. Application of the matching pursuit method for structural decomposition and averaging of phonocardiographic signals. Med Biol Eng Comput 1998;36(3):302–308.
42. Manecke GR, Jr., Poppers PJ. Esophageal stethoscope placement depth: its effect on heart and lung sound monitoring during general anesthesia. Anesth Analg 1998;86(6):1276–1279.
43. Thompson WR, Hayek CS, Tuchinda C, Telford JK, Lombardo JS. Automated cardiac auscultation for detection of pathologic heart murmurs. Pediatr Cardiol 2001;22(5): 373–379.
44. Hayek CS, Thompson WR, Tuchinda C, Wojcik RA, Lombardo JS. Wavelet processing of systolic murmurs to assist with clinical diagnosis of the heart disease. Biomed Instrum Technol 2003;37(4):263–270.
45. Pavlopoulos SA, Stasis AC, Loukis EN. A decision tree—based method for the differential diagnosis of aortic stenosis from mitral regurgitation using heart sounds. Biomed Eng Online 2004;3(1):21.
46. Guo Z, Moulder C, Zou Y, Loew M, Durand LG. A virtual instrument for acquisition and analysis of the phonocardiogram and its internet-based application. Telemed J E Health 2001;7(4):333–339.
47. Stern DT, Mangrulkar RS, Gruppen LD, Lang AL, Grum CM, Judge RD. Using a multimedia tool to improve cardiac auscultation knowledge and skills. J Gen Intern Med 2001; 16(11):763–769.
48. Woywodt A, Herrmann A, Kielstein JT, Haller H, Haubitz M, Purnhagen H. A novel multimedia tool to improve bedside teaching of cardiac auscultation. Postgrad Med J 2004; 80(944):355–357.

See also ELECTROCARDIOGRAPHY, COMPUTERS IN; GRAPHIC RECORDERS.

PHOTOTHERAPY. See ULTRAVIOLET RADIATION IN MEDICINE.

PHOTOGRAPHY, MEDICAL

JACKIE K. CHAN
EDWARD K. FUNG
Columbia University
New York

INTRODUCTION

Photography is widely used in many areas of medicine for the documentation and treatment of diseases. Photography involves making pictures by capturing light reflected from objects onto a sensitive medium (e.g., film or the more recent technique of light-sensitive chips from a digital camera). In ophthalmology, the transparency of the living eye allows photographs to image diseases as far back as the retina. In dermatology, traditional methods of photography are used to document and track skin lesions. Every type of medical imaging comes with its own technical challenges. The medical photographer plays a vital part in promoting and supporting quality healthcare by providing services in photography. Furthermore, imaging has served as an important research tool. Photomicrography involves taking images in the laboratory of tissue or culture specimens in both the gross and cell level. The goals of photography, in general, may include characterizing the basic anatomy and physiology of the body, understanding changes caused by aging or disease, and discovering disease mechanisms.

OPHTHALMIC PHOTOGRAPHY

Photographing the living eye poses some challenging issues despite its transparency. The eye is sensitive to light and can easily be bleached after a certain number of flashes and intensity. The human retina is also designed for capturing light rather than reflecting it. Images may result in poor contrast and may affect the performance of diagnostic procedures. The main absorbing pigments in the eye are blood, hemoglobin, photo pigments, macular pigments, and water. Moreover, research has shown that many sight-threatening diseases are embedded deep in the retina, where conventional tools of photography cannot be used (1,2). Fortunately, specialized instruments and image enhancement processes have been developed to obtain better images (Fig. 1).

The Fundus Camera

The instrument widely used by ophthalmologists to view the posterior segment of the eye is the fundus camera. The fundus is photographed using a white light source to provide high resolution images at the micron range. Fundus photography can also create stereographics images that provide depth. These qualities make fundus photography the established standard for clinical studies of macular diseases. Newer models now take digital images and can be directly stored to a computer database.

Many fundus camera models are currently available in the United States: the German Zeiss, the Topcon fundus camera, the Olympus fundus camera, and the Nikon fundus camera. In choosing an instrument, one should compare the engineering and more importantly, the photo quality. Some instruments will be more expensive than its competitors, but takes excellent photographs while others have a wider view and good quality images (3) (Fig. 2).

General Methodology

A full-time photographer is specialized to operate the equipment in a clinic. While gaining technical skills, the photographer is often familiar with the clinical pathology of the fundus. This is advantageous to the ophthalmologist in situations where photos need to be interpreted. In operating a fundus camera, there are some general guidelines to follow (3):

1. The pupil of the patient must be dilated. Dilation of the pupil may take 20–30 min. An 8 mm diameter pupil is ideal, but even a pupil much smaller may be acceptable.
2. The eyepiece must be carefully focused to avoid getting out-of-focus photographs. Young children will have accommodation problems and may pose extra attention. The settings should be checked before each set of photographs on the patient is taken.
3. Check the shutter speed for the proper electronic flash synchronization. Shutter speeds in the range of 1/30 to 1/60 of a second can be used.
4. Take an identification photograph of the patient's name, date photographed, and other pertinent information. Properly labeling photos will avoid confusion later.

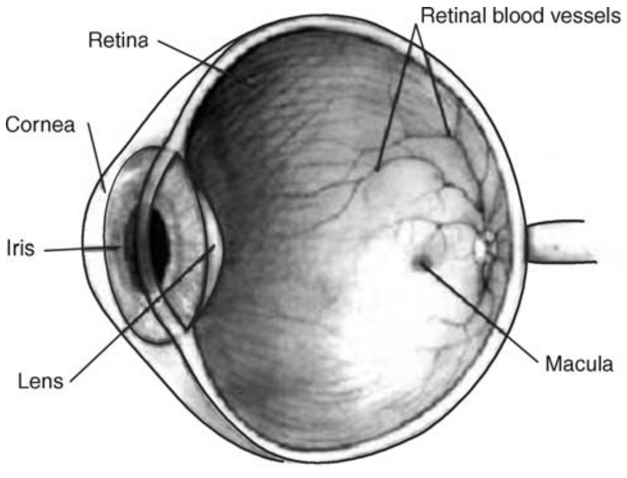


Figure 1. Cross-section of the eye. The transparency of the eye permits the ability to receive light from the external world. Light enters the cornea and goes through the iris and the lens until it reaches the retina. The macula contains many rods and cones and is the area of greatest visual acuity.

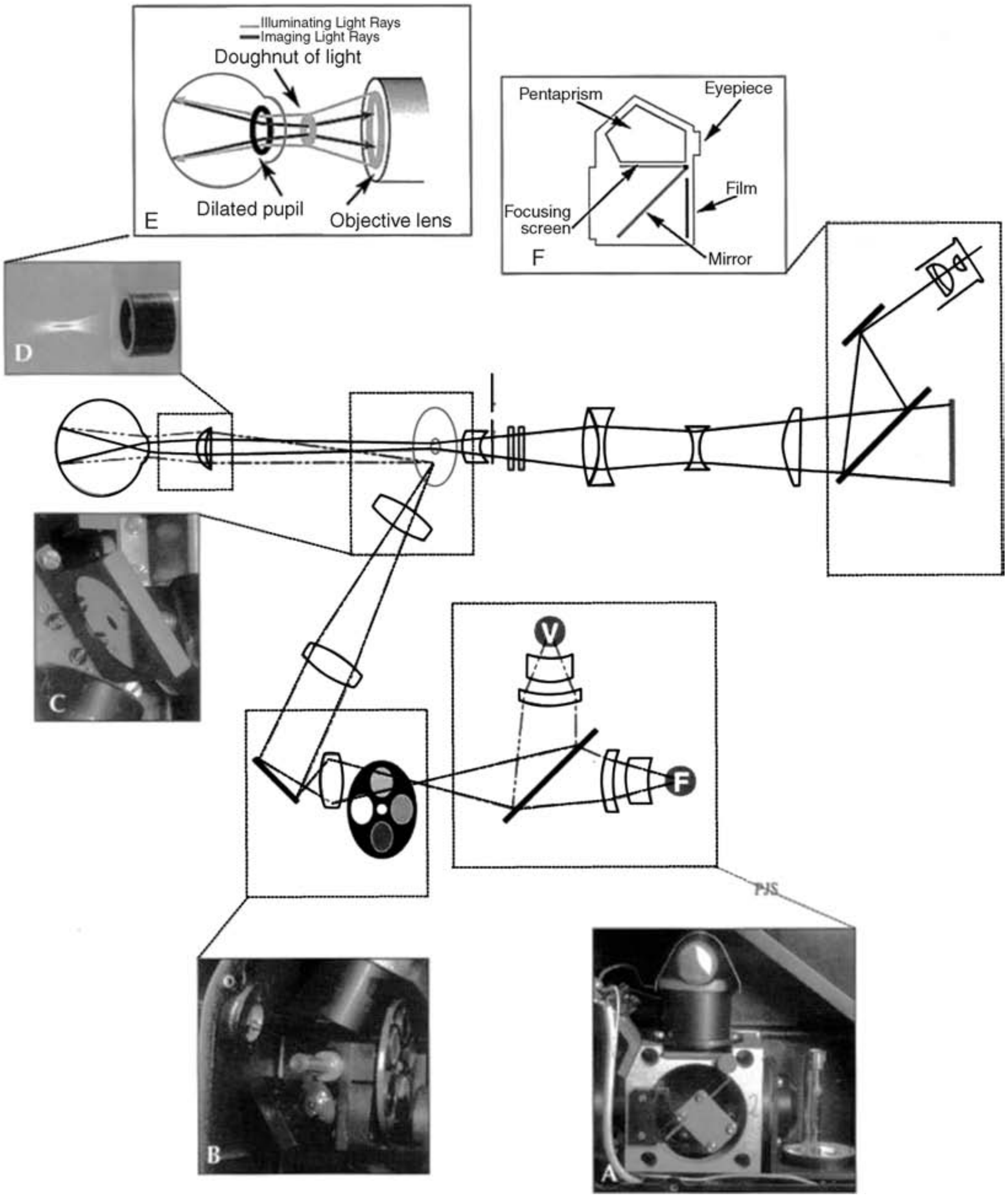


Figure 2. The optical pathway of the fundus camera. Light generated from either the viewing lamp (V) or the electronic flash (F) (A) is projected through a set of filters and onto a round mirror (B). This mirror reflects the light up into a series of lenses that focus the light. A mask on the uppermost lens shapes the light into a doughnut. The doughnut shaped light is reflected onto a round mirror with a central aperture (C), exits the camera through the objective lens, and proceeds into the eye through the cornea. When the illuminating light is seen from the side, one can appreciate the complexity and precision of the optical scheme (D). Assuming that both the illumination system and the image are properly aligned and focused, the retinal image exits the cornea through the central, unilluminated portion of the doughnut (E). The light continues through the central aperture of the previously described mirror, through the astigmatic correction device and the diopter compensation lenses, and then back to the single lens reflex camera system. (From Saine PJ, Taylor ME, Ophthalmic Photography (2nd ed.), Butterworth-Heinemann, 2002 with permission.)

5. When the patient is comfortably seated at the instrument with their chin on the chin rest and their forehead against the headrest, instruct the patient to look at the fixation device.
6. After properly focusing the filament of the viewing lamp, look through the eyepiece and bring the retinal vessels into focus. Release the shutter.

Stereo Fundus Photos

Fundus photos in stereo pair give a perception of depth that greatly improves reading performance. For example, a stereo photo can document blurred disk margins, optic nervehead cupping, and the degree of retinal elevation from conditions (e.g., serous or solid detachments). When evaluating patients with age-related macular degeneration (AMD), graders can view the elevation of drusen under the retina much more effectively.

The most popular method of taking stereo photographs was the cornea-induced parallax method, advocated by Bedell (4). To take the stereo pair, the first photograph is taken through the temporal side of the pupil and the second through the nasal side. A lateral shift of 3.5 mm is recommended for optimum stereopsis, but any lateral shift will create a stereo photograph. The photographer aligns and focuses the camera through the center of the pupil, then uses the joystick to move the camera slightly to the right, takes an image, and then slightly to the left, and takes the second image. The translation of the camera sideways changes the angle of view. The amount of camera shift can vary from image pair to image pair, making apparent depth a variable. Attachments are available for some cameras that allow the photographer to shift the camera through a consistent distance. Specialized cameras are commercially available to take stereo-pair fundus photos simultaneously, though the resolution is reduced.

Digital Imaging

Until recently, fundus photos were developed in film and the photographer had to wait hours or even days to see the results. The resolution was limited by the grain size of the film. Today, fundus cameras take digital photos that can be evaluated instantaneously and stored digitally. A typical image is in 24-bit, red, green, and blue (RGB), true color with a resolution size of 2000×2000 pixels. A digital imaging system ensures that the original quality of the image is preserved and will produce flawless duplication.

Images can be stored in a number of different formats, but with digital files comes the requirement for an efficient digital storage system. Tagged image file format (TIFF) images are not compressed, and therefore require large storage spaces. Recent advances in computer engineering have provided large storage in affordable costs. Nevertheless, it has been shown by Lee (5) that joint photographic experts group (JPEG) formats allow for varying degrees of image compression without compromising the resolution of fundus photos necessary for image analysis. In image compression, images are applied to a lossy algorithm, which permit the image to be reconstructed based on partial data. The result is not an exact restoration of the image, but is sufficient for diagnostic purposes.

In the evaluation of clinical AMD, Lee reported that TIFF images and low compression (30:1) JPEG images were virtually indistinguishable (5). Digital images of AMD patients were analyzed in software for drusen identification and quantification. Drusen detection in the conventional stereo fundus slides using a manual protocol was highly comparable to the digital format (Figure 3).

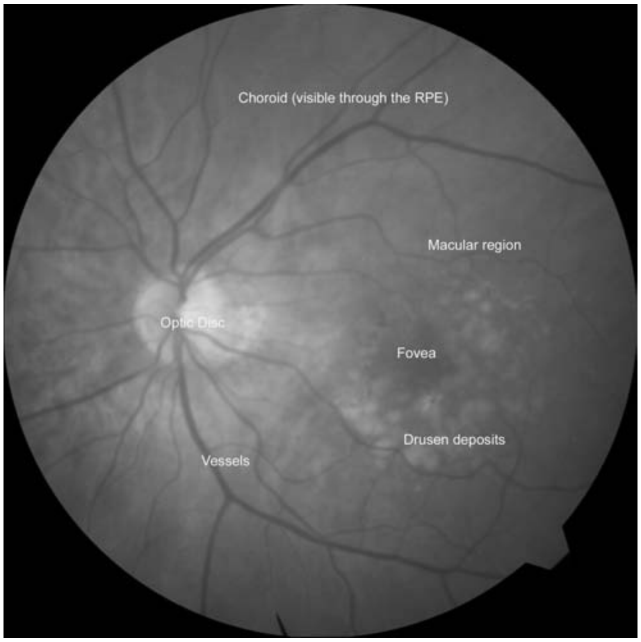


Figure 3. Fundus image of an AMD patient. The fundus camera is used to photograph the retina. The macula corresponds to an area of the retina 5–6 mm in diameter, centered on the fovea. Drusen are the white deposits surrounding the fovea that are characteristic of AMD, a sight-threatening disease. The retina has many layers, including the retinal pigment epithelium (RPE) and the choroid. Drusen are usually found embedded below the RPE layer.

Clinical Evaluation of AMD in Fundus Photos

The fundus camera has been routinely employed for diagnostic purposes (e.g., the clinical study of patients with AMD) (6,7), which is the leading cause of blindness in the developed world (8). The natural history of AMD is hallmarked by a subretinal pathology known as drusen (9–17). The identification and measurement of drusen are central to clinical studies of early AMD. The current standard of grading AMD is through manual viewing of stereographic fundus slides on a light box. However, this method involves time-consuming analysis of drusen size, number, area, and morphology in several subcategories (18).

Despite the movement to digital fundus imaging, a digital and automated method of quantification of macular pathology in AMD has yet to gain widespread use. Computer-assisted image analysis offers the potential for greater accuracy, objectivity, and reproducibility, but designing algorithms for that purpose is nontrivial. Many methods have been attempted in the last two decades with unsatisfactory results (19–25).

One major obstacle stems from the clinical appearance of the normal macula, which is a composite of complex light reflections from and absorption by multiple layers of the retina and associated structures. The application of segmenting any pathology superimposed on the macula in an automated fashion is a difficult task by the nonuniform reflectance of the normal macular background. For example, absorption by luteal pigment in the central macula (fovea and parafovea) contributes to the darker central appearance. The nerve fiber layer, conversely, is highly reflectant. It is thickest in the arcuate bundles along the arcades and thinnest at the central fovea. This makes the arcade regions relatively brighter, hence also contributes to the macula appearing darker centrally (26). Therefore, simply choosing a global threshold would not be equally effective in segmenting or identifying drusen in the darker central regions as it would in the relatively brighter regions in the periphery, and vice versa.

Shin et al. (19) used adaptive thresholding techniques to handle the nonuniform macular background reflectance. They divided the image into separate windows of variable sizes. Within each window, a local histogram was applied to check for skewness, and to determine if drusen was present. However, the method was often misleading if either a large area of background or a large drusen dominated the region, which resulted in an incorrect threshold. Moreover, windows containing vessels would sometimes be incorrectly interpreted in the bimodal distribution. Thus, operator supervision and some postprocessing steps were added.

Shin's method was improved by Rapantzikos et al. (20), which used morphological operators (e.g., kurtosis and skewness) to predict whether drusen was present in the local window. Their idea gave better results, but was not infallible as a completely automated system of drusen segmentation. For example, many different combinations of image features (drusen and background) can yield the same histogram.

An alternative method was presented by Smith et al. (27,28) that aims to level the macular background reflectance, which can change significantly over distances of 50–100 μm. In previous methods, inadequate segmentation centrally and overinclusive segmentation in the peripheral macula was an indication that the background variability of the macula had not been resolved.

Smith found that the background reflectance of a normal fundus image could be modeled geometrically (29–31). It has been shown that a partial normal background containing drusen provided enough information to model the entire background by a elliptical contour graph (27,32). After leveling the nonhomogeneity of the background reflectance, they overcame the challenges posed in purely histogram-based methods of other researchers. A combined automated histogram technique and the analytic model for macular background presented a completely automatic method of drusen measurement.

Their algorithm is briefly explained. First, an initial correction of the large-scale variation in brightness found in most fundus photographs was applied. This is achieved by calculating a Gaussian blur of the image and subtracting it from the original image in each of the three RGB color channels. Further processing was carried out in this preprocessed image. The main idea was to level the background such that the reflectance was uniform over the entire macula. They proposed a multizone math model to reconstruct the macula. Each zone divided the macula into different annular and grid-like regions. The pixel gray levels were used as input for fitting into the custom software employing least-squares methods. After the geometric model of the macula was created, it was subtracted from the original image to obtain a leveled image. The process was automatically iterated until a sufficient leveling has been achieved. In other words, the range of gray levels in the image is minimized to an acceptable level.

The drusen, which is superimposed on the image, appears brighter than the regular intensity of the background. The final threshold was obtained by applying a histogram analysis approach to the final leveled image (33). An optimal threshold defined the separation of background areas from drusen areas.

Smith's technique has been validated successfully with the current standard of fundus photo grading by stereo pair viewing in the central 1000 and middle 3000 μm diameter subfields (28). One advantage of automation is portability of the software to use at other institutions. The work to create an automated algorithm will provide a useful, cost-effective tool in clinical trials of AMD.

Despite the advances in automated quantification of drusen, some obstacles still remain. Drusen identification may be confounded by other objects present in the image. A computer ultimately must be taught how to differentiate drusen from other lesions (e.g., as retinal pigment epithelial hypopigmentation, exudates, and scars). Implementation with neural networks or morphological criteria may prove effective in eliciting unique features in the lesions. For now, some cases may still require supervision in digital automated segmentation (Fig. 4).

SCANNING LASER OPHTHALMOSCOPE

Photography is often associated with taking a white light source to obtain an image. However, monochromatic sources of light at a specific wavelength are available with the scanning laser ophthalmoscope (SLO). Originally used as a research tool, it is increasingly favored with other clinical imaging standards now. Moreover, pupil dilation is unnecessary, and light levels are dim during acquisition (34). The reflectance and absorbance spectra of the structures of interest can be seen. For example, the SLO provides better penetration and give views of the choroids layer. By the same token, things of less interest on the retina (e.g., drusen) will not be seen (26).

In autofluorescence (AF) imaging, a 488 nm laser source with a 500 nm barrier filter is used (26). This light source reveals structures that intrinsically fluoresce, primarily due to lipofuscin and its main fluorophore, A_2E (35,36). Focal changes in AF or the changes in their spatial distribution are means of studying the health of the RPE. Previous studies have been made on correlations of change in AF distribution with pathological features (35,37–41). Focally increased AF (FIAF) refers to increased fluorescence in an area with respect to the rest of the background. This is abnormally high in patients with Stargardt disease and may be a marker for RPE disease in ARMD. In actual RPE death, as in geographic atrophy, there is focally decreased AF (FDAF), seen as a blackened area in the AF image (26).

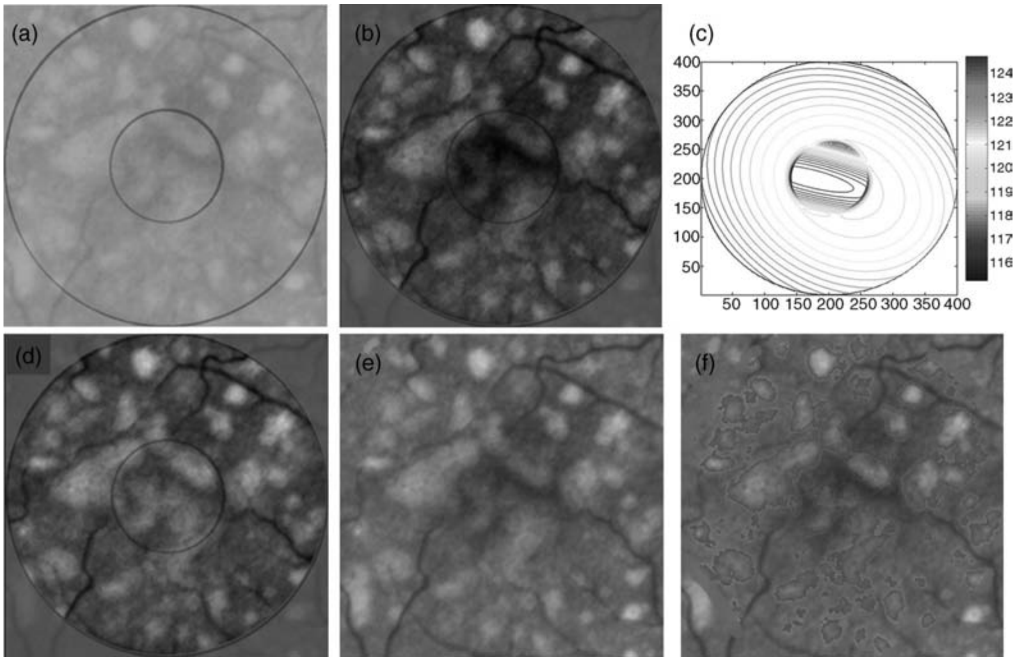


Figure 4. Automated digital photo grading. Color fundus photo with circular grading template overlaid on top (a). Green channel (grayscale) of fundus photo (b). Using an algorithm that finds the lower and upper thresholds, the macula is divided into three sections: vessels and darker perivasculature, normal background, and bright drusen. The normal background was used to calculate the mathematical model, displayed as a contour graph (c). The model was subtracted from the original image to yield a background leveled image of uniform intensity in (d). Contrast-enhanced image of a to show the boundaries of drusen (e). The algorithm then uses the leveled image (d) to calculate a threshold that determines the drusen areas (segmented in green) and overlaid on top of the contrast enhanced layer (f).

TELEMEDICINE

As the average age of populations in developed countries continue to rise, the number of elderly people needing an eye exam will boom. In less populated areas, an ophthalmologist may not be readily accessible. In the United States, less than one-half of the diabetic population received an annual eye examination (42). One of the goals of telemedicine is to bridge the gap between places where patients can be evaluated and where service is rendered. An examination usually involves taking the patient's fundus photos. With the advent of telemedicine and digital photography, the patient and doctor do not even have to be in the same room. Data can be efficiently and securely transferred across different institutions.

Telemedicine Framework

The goal of telemedicine is to establish a screening system with the ability to reach out to millions of unexamined patients for primary prevention. The following guidelines can be used to set up a generalized framework for telemedicine. We will use an ophthalmology setting as an example:

First, a patient will come into a primary care office to have their retina imaged. The camera itself should be operator-friendly, one that is easy to use by either a doctor's assistant or a technician. The disease to be screened should have identifiable pathology on the image. Examples would be diabetic retinopathy or AMD. The fundus photograph is then sent electronically to a reading center, which can be established at a far away university or hospital. Photos should be in a compressed digital format (e.g., JPEG).

In a paper by Sivagnanavel et al. (43), it was suggested that custom software for detecting AMD could be used at two different institutions. A grader at each institution independently ran the software and performed drusen quantification successfully, thus demonstrating portability and potential for automated software examinations. The results of the grading with software are comparable to grading manually by stereo-viewing. The only drawback is that the software may not be suitable in $\sim 20\%$ of the images taken. For the percentage of cases that need supervision, a trained reader can manually examine the images. Images that were disqualified had poorly identified areas or multiple types of lesions that were difficult to distinguish.

At the reading center, a patient report is generated and sent electronically back to the primary care physician (PCP). Finally, a good network referral base must be established such that the PCP can refer an ophthalmologist if the patient screening comes up positive.

Screening Instrument

Built differently from the classic fundus camera, a telemedicine instrument is meant for screening more than

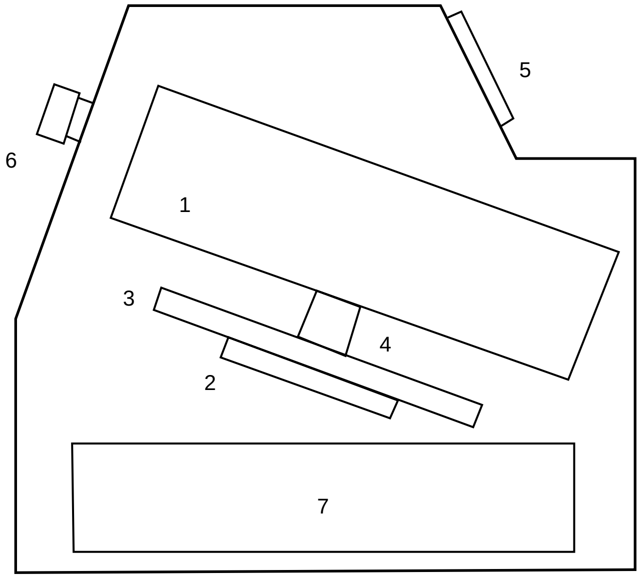


Figure 5. Layout of the DigiScope. The imaging head (1) is mounted on an *XYZ* motorized and computer-controlled stage. The *X*, *Y*, and *Z* components (2, 3, and 4, respectively) allow movements in the superior-inferior direction, to and from the eye, and in the nasal-temporal direction, respectively. The operator interfaces with the camera by a touch screen (5). The subject leans the head and places it against the nosepad (6) to view the imaging system inside. The electronics and computer (7) that control the system and store the images are embedded in the machine and hidden from view. (used with permission for R. Zeimer, IOVS 43:5,2002.)

careful clinical diagnosis. In a telemedicine fundus camera (e.g., the Digiscope) (44), the goal was to create a good model for providing ophthalmic screening to the primary care office.

Keeping the instrument at low cost for production greatly increased its attractiveness for the primary doctor's office. Therefore, a video camera with a resolution of 50 pixels per degree provided enough detail comparable to a fundus photograph. Such cameras typically yield images with 930 pixels diagonally, or a field of 19°.

Figure 5 shows a simplified diagram of the DigiScope. The eye to be imaged is oriented by an internal light-emitting diode. Fixation and imaging covers the entire posterior pole and takes about eight images. The illumination is generated by a halogen bulb. Infrared light is eliminated by the first filter and the visible spectrum is limited to green by the second filter, which passes light between 510 and 570 nm. The beam then expands and illuminates the fundus.

There are two modes of imaging with the DigiScope. The first mode takes four frames per shutter, with the shutter lasting 130 ms. The four frames differ by the fine focus, such that the sharpest image can be chosen out of each set. The second mode is meant for stereoscopic effect. Four frames are acquired while the optical head moves horizontally along the x axis. A pair of images can be selected to generate a stereo effect.

The operator interfaces with the machine by a simple touch-screen. The duties of the operator are limited to basic tasks (e.g., explaining the procedure, encouraging the patient to fixate at the blinking light, and checking the quality of the image).

A non-mydriatic camera is in consideration to provide greater patient comfort. Poor imaging quality and lack of stereo may cause a tendency to include RPE atrophy as drusen. However, newer generation cameras taking higher resolution images, coupled with the ability for digital stereo, may eliminate such drawbacks in the future.

PHOTOMICROGRAPHY

Photomicrography in biomedicine is the creation of images of biomaterial at magnification. Tissue or culture samples can be photographically recorded at both the gross level and at the level of cells. Images can be used for archival purposes or for analysis. Photomicrography replaced earlier camera lucida apparatus that projected microscope images through a beam splitter onto a plane for manual tracing.

Still photographs of biological specimen have traditionally been captured on 35 mm film. The development and popularization of digital camera technology, however, has had a drastic impact on the field of photomicrography. The relative ease of use has spurred the application of photomicrography in many areas, especially pathology and basic research.

Digital Cameras

Digital cameras offer several advantages over their film variants for most uses. Most digital cameras include some type of LCD (liquid-crystal display) screen. This screen can be used to display a photograph right after it is taken, allowing the user to evaluate the result of different camera settings and adjust as needed. This is a great convenience given the unusual light conditions under which photomicrographs are generally taken and the fact that most scientific users will probably not have professional photography training. On consumer models, the LCD can also display a real-time image of the scene to be photographed. Furthermore, digital cameras store images in a digital format, usually on a removable memory card. Images can be directly transferred to a personal computer for analysis or processing without having to use a scanner to digitize as with 35 mm film. The formats come in JPEG or TIFF. Still others cameras have the option of outputting in the RAW format. This is the raw data from the CCD (charge-coupled device) array that forms the core of the camera. The CCD cameras have also proven particularly useful in extremely low light microscopy, such as in certain fluorescence microscopy situations. Quantum efficiency, a measure of light sensitivity, can reach 90% in a back illuminated electron multiplying CCD, compared to ~ 20% for a conventional video camera (45,46) (Fig. 6).

Photomicroscope Description

The attachment of the camera to the microscope is essentially the same for both digital and film cameras though different adapters may be used. A camera can simply be held with the lens pressed flush against the eyepiece of a microscope or mounted with a special adapter for the

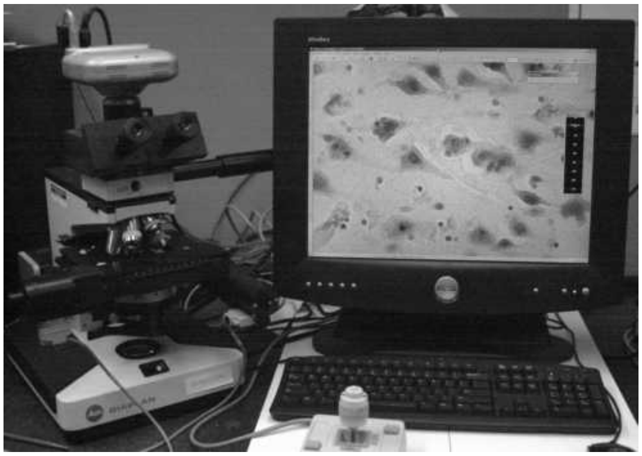


Figure 6. Retrofitted computerized microscope system with motorized stage and CCD digital imager directly connected to computer by Firewire cable. (used with permission from the laboratories of Victoria Arango, Ph.D., J. John Mann, M.D., and Mark Underwood, Ph.D. at NYSPI and Columbia University.)

eyepiece (46,47). This method can be used for many consumer digital cameras. Alternatively, a camera can be attached to the phototube or trinocular port of a microscope or through a special beam splitting adapter (47,48). Typically, SLRs (single lens reflex) and cameras specifically designed for photomicrography are attached in this way by their lens mounts. The tube thread standards commonly used in professional cameras to attach the lens are C-, T-, and F-mounts. Relay lens adapters may be necessary to correctly focus the image onto the film or CCD array in the camera as the normal lens is removed to use the C-, T-, or F-mount. Many of the major camera brands as well as third-party manufacturers produce these adapter kits. The primary concerns when attaching a camera are ensuring correct focus and avoiding vignetting. Vignetting refers to the darkening of the edges of an image (47). Proper Koehler's illumination of the microscope, as well as focusing and zooming should correct this effect.

The CCD technology has also been applied successfully to real-time imaging through the microscope. Dedicated computer microscopy systems with CCD cameras have been made available by the major microscope manufacturers (e.g., Leica and Zeiss). The CCD video cameras can also be attached to existing microscope setups through the phototube as with still cameras. Traditional CCTV cameras can usually output in NTSC (National Television System Committee) or PAL (Phase Alternating Line), the common video standards in the United States and Europe, respectively. An image capture board is necessary to convert the TV signal for use with a computer. The CCD digital imagers specifically designed for scientific or microscopic use often can be connected by means of more conventional computer ports [e.g., Firewire (IEEE 1394) and USB (universal serial bus)]. These are preferred due to their resistance to radio frequency (RF) interference and higher resolutions (49). In both cases, specialized software is needed to manipulate and display the video image. Some of these imagers can be used to capture high resolution still images as well.

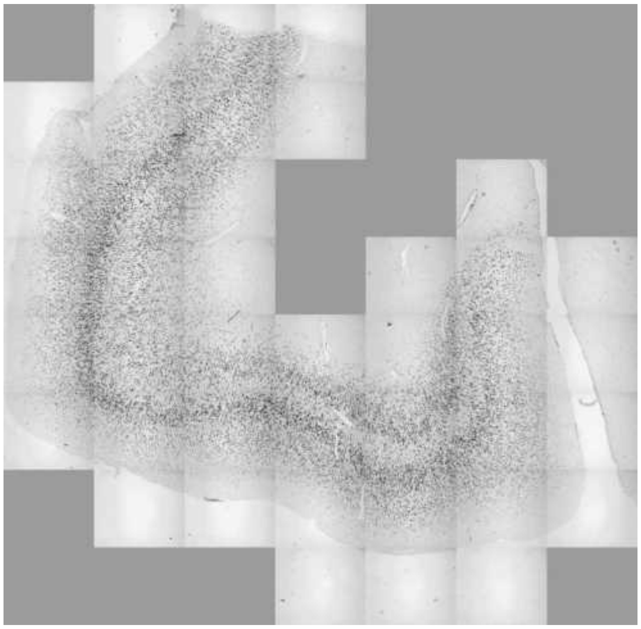


Figure 7. Montage of photomicrographs taken with 40× objective and stitched together with Neurolucida software (Microbrightfield, Inc., Williston, VT). Brodman area 24 of the prefrontal cortex is depicted. (used with permission from the laboratories of Victoria Arango, Ph.D., J. John Mann, M.D., and Mark Underwood, Ph.D. at NYSPI and Columbia University.)

For video applications, a more expensive three-chip, one-shot CCD system is preferred (46). These cameras use an array of photodetectors to record an image. One such array is referred to as a chip. Currently, cameras are available in one-chip and three-chip designs. A one-chip design produces color images by switching three colored filters over the array. This arrangement is referred to as one-chip, three-shot. Due to the switching of RGB (red, green, blue) filters, the framerate is necessarily reduced. A three-chip, one-shot camera splits the individual color channels and directs each to its own CCD array, imaging them simultaneously. Thus a higher framerate is maintained.

Camera computer systems integrated with a motorized microscope stage make possible more complicated photomicrograph sets. A computer controlled stage and camera can be programmed to take many overlapping digital images. With special software, these can be stitched together to form high magnification photomicrographs of large tissue sections (50). This represents an interesting alternative to low magnification macroscopic photography of tissue specimen. Systems have also been developed for 3D reconstruction using stacks of properly registered two-dimensional (2D) photomicrographs (Fig. 7).

BIBLIOGRAPHY

1. Elsner AE. Reflectometry with a Scanning Laser Ophthalmoscope. Appl Opt 1992;31:3697–3710.
2. Figueroa MS, Regueras A, Bertrand J. Laser photocoagulation to treat macular soft drusen in age-related macular degeneration. Retina 1994;14(5):391–396.

3. Yannuzzi LA, Gitter KA, Schatz H. The Macula: A Comprehensive Text and Atlas. Fundus Photography and Angiography. In: Justice JJ, editor. Baltimore: Williams & Wilkins; 1979.
4. Bedell AJ. Photographs of the Fundus Oculi. Philadelphia: F.A. Davis; 1929.
5. Lee MS, Shin DS, Berger JW. Grading, image analysis, and stereopsis of digitally compressed fundus images. Retina 2000;20(3):275–281.
6. Bird AC, et al. An international classification and grading system for age-related maculopathy and age-related macular degeneration. The International ARM Epidemiological Study Group. Survey Ophthal 1995;39(5):367–374.
7. Klein R, et al. The Wisconsin age-related maculopathy grading system. Ophthalmology 1991;98(7):1128–1134.
8. Klaver CC, et al. Age-specific prevalence and causes of blindness and visual impairment in an older population: the Rotterdam Study. Arch Ophthal 1998;116(5):653–658.
9. Smiddy WE, Fine SL. Prognosis of patients with bilateral macular drusen. Ophthalmology 1984;91(3):271–277.
10. Bressler SB, et al. Relationship of drusen and abnormalities of the retinal pigment epithelium to the prognosis of neovascular macular degeneration. The Macular Photocoagulation Study Group. Arch Ophthal 1990;108(10):1442–1447.
11. Bressler NM, et al. Drusen characteristics in patients with exudative versus non-exudative age-related macular degeneration. Retina 1998;8(2):109–114.
12. Holz FG, et al. Bilateral macular drusen in age-related macular degeneration. Prognosis and risk factors. Ophthalmology 1994;101(9):1522–1528.
13. Abdelsalam A, Del Priore L, Zarbin MA. Drusen in age-related macular degeneration: pathogenesis, natural course, and laser photocoagulation-induced regression. Survey Ophthalmol 1999;44(1):1–29.
14. Little HL, Showman JM, Brown BW. A pilot randomized controlled study on the effect of laser photocoagulation of confluent soft macular drusen [see comments]. Ophthalmology 1997;104(4):623–631.
15. Frennesson IC, Nilsson SE. Effects of argon (green) laser treatment of soft drusen in early age-related maculopathy: a 6 month prospective study. Br J Ophthalmol 1995;79(10):905–909.
16. Bressler NM, et al. Five-year incidence and disappearance of drusen and retinal pigment epithelial abnormalities. Waterman study. Arch Ophthalmol 1995;113(3):301–308.
17. Bressler SB, et al. Interobserver and intraobserver reliability in the clinical classification of drusen. Retina 1988;8(2): 102–108.
18. Age-Related Eye Disease Study Research Group, T. The Age-Related Eye Disease Study System for Classifying Age-related Macular Degenration From Stereoscopic Color Fundus Photographs: The Age-Related Eye Disease Study Report Number 6. Am J Ophthalmol 2001;132(5):668–681.
19. Shin DS, Javornik NB, Berger JW. Computer-assisted, interactive fundus image processing for macular drusen quantitation [see comments]. Ophthalmology 1999;106(6):1119–1125.
20. Rapantzikos K, Zervakis M, Balas K. Detection and segmentation of drusen deposits on human retina: Potential in the diagnosis of age-related macular degeneration. Med Image Analysis 2003;7(1):95–108.
21. Sebag M, Peli E, Lahav M. Image analysis of changes in drusen area. Acta Ophthalmol 1991;69(5):603–610.
22. Morgan WH, et al. Automated extraction and quantification of macular drusen from fundal photographs. Aust New Zealand J Ophthalmol 1994;22(1):7–12.
23. Kirkpatrick JN, et al. Quantitative image analysis of macular drusen from fundus photographs and scanning laser ophthalmoscope images. Eye 1995;9(Pt 1):48–55.
24. Goldbaum MH, et al. The discrimination of similarly colored objects in computer images of the ocular fundus. Invest Ophthalmol Vis Sci 1990;31(4):617–623.
25. Peli E, Lahav M. Drusen measurement from fundus photographs using computer image analysis. Ophthalmology 1986;93(12):1575–1580.
26. Smith RT. Retinal Imaging and Angiography, Basic Science Course. New York: Eye Institute, Columbia University; 2005.
27. Smith RT, et al. A method of drusen measurement based on reconstruction of fundus reflectance. Br J Ophthalmol 2005;89(1):87–91.
28. Smith RT, et al. Automated detection of macular drusen using geometric background leveling and threshold selection. Arch Ophthalmol 2005;123:200–207.
29. Smith RT, et al. Patterns of reflectance in macular images: representation by a mathematical model. J Biomed Opt 2004;9(1):162–172.
30. Smith RT, et al. The fine structure of foveal images. Invest Ophthalmol Vis Sci 2001;42(Mar. Suppl.):153.
31. Smith RT, et al. A two-zone mathematical model of normal foveal reflectance in fundus photographs. Invest Ophthalmol Vis Sci 2003;44:E-365.
32. Chan JWK, et al. A Method of Drusen Measurement Based on Reconstruction of Fundus Background Reflectance. Invest Ophthalmol Vis Sci 2004;45(5):E-2415.
33. Otsu N. A threshold selection method from gray-level histograms. IEEE Trans Systems Man Cybernetics 1979;9(1):62–66.
34. Elsner AE, Weiter JJ, Jalkh AE. New Devices for Retinal Imaging and Functional Evaluation. In: Freeman WR, ed. Practical Atlas of Retinal Disease and Therapy. New York: Raven Press; 1993. pp. 19–35.
35. Delori FC, et al. In vivo fluorescence of the ocular fundus exhibits retinal pigment epithelium lipofuscin characteristics. Invest Ophthalmol Vis Sci 1995;36(3):718–729.
36. von Ruckmann A, Fitzke FW, Bird AC. In vivo fundus autofluorescence in macular dystrophies. Arch Ophthalmol 1997;115(5):609–615.
37. Feeney-Burns L, Berman E, Rothman H. Lipofuscin of the human retinal pigment epithelium. Am J Ophthalmol 1980;90:783–791.
38. von Ruckmann A, Fitzke FW, Bird AC. Distribution of fundus autofluorescence with a scanning laser ophthalmoscope [comment]. Br J Ophthalmol 1995;79(5):407–412.
39. von Ruckmann A, Fitzke FW, Bird AC. Fundus autofluorescence in age-related macular disease imaged with a laser scanning ophthalmoscope. Invest Ophthal Vis Sci 1997;38(2):478–486.
40. Feeney-Burns L, Hildebrand E, Eldridge S. Aging human RPE: Morphometric analysis of macular, equatorial, and peripheral cells. Invest Ophthalmol Vis Sci 1984;25:195–200.
41. Wing G, Blanchard G, Weiter J. The topography and age relationship of lipofuscin concentrations in the RPE. Invest Ophthalmol Vis Sci 1978;17:600–607.
42. Mukamel D, et al. Barriers to compliance with screening guidelines for diabetic retinopathy. Ophthal Epidemiol 1999;6:61–72.
43. Sivagnanavel V, et al. An Interinstitutional Comparative Study and Validation of Computer-Aided Drusen Quantification. Br J Ophthalmol 2005;89:554–557.
44. Zeimer R, et al. A fundus camera dedicated to the screening of diabetic retinopathy in the primary-care physician's office. Invest Ophthalmol Vis Sci 2002;43(5):1581–1587.
45. Coates CG, et al. Optimizing low-light microscopy with back-illuminated electron multiplying charge-coupled device: enhanced sensitivity, speed, and resolution. J Biomed Opt 2004;9:1244–1252.

46. Riley RS, et al. Digital photograpy: a primer for pathologists. J Clin Lab Anal 2004;18:91–128.
47. Hamza SH, Reddy VVB. Digital image acquisition using a consumer-type digital camera in the anatomic pathology setting. Adv Anat Pathol 2004;11:94–100.
48. Haynes DS, et al. Digital microphotography: a simple solution. Laryngoscope 2003;113:915–919.
49. Hand WG. Practical guide to digital imaging for microscopy. Biores Online; 2000.
50. Berggren K, et al. Virtual slice: a novel technique for developing high-resolution digital brain atlases. Society for Neuroscience Annual Meeting, Miami; 2002.

See also ENDOSCOPES; FIBER OPTICS IN MEDICINE; IMAGING DEVICES; MEDICAL EDUCATION, COMPUTERS IN; PICTURE ARCHIVING AND COMMUNICATION SYSTEMS; RADIOLOGY INFORMATION SYSTEMS.

PHYSIOLOGICAL SYSTEMS MODELING

N. TY SMITH
University of California,
San Diego, California
KENTON R. STARKO
Point Roberts
Washington

INTRODUCTION

First, physiological systems need to be characterized. The word “system” means an interconnected set of elements that function in some coordinated fashion. Thus, the heart, with its muscles, nerves, blood, and so on, may be regarded as a physiological system. The heart, however, is a subsystem of the circulatory system, which is in turn a subsystem of the body. Dynamic systems are time-varying systems, and most physiological systems fall into this category. This article deals with classical, or macro, physiological systems, usually ignoring genetic, cellular and chemical systems, for example.

It should come as no surprise that the term super system has been used to describe the brain and the immune system (1). The cardiovascular system (CVS) and central nervous system (CNS) are closer to being super systems, however.

Correct terminology is vital. To paraphrase George Bernard Shaw, engineering and medicine are two disciplines separated by the same language. The same word may have two different meanings (system) or two different words may have the same meaning (parameter and variable; model and analogue). The latter is particularly pertinent. Medical researchers working with animals because of their resemblance to humans use the term “animal model”, while engineers would use the more descriptive “animal analog”.

Unless otherwise stated, model, We means mathematical model. Thus, these models are generally ignore: animal, *In vitro*, chemical, structural (Harvey’s observations on the circulation), and qualitative (Starling’s law of the heart).

What is a mathematical model? A mathematical model consists of elements each describing in mathematical terms the relation between two or more quantities. If all the descriptions are correct, the model will simulate the behavior of real-life processes. Especially when many different subprocesses are involved, the models are useful in helping understand all the complex interactions of these subsystems. A useful characteristic of a mathematical model is that predictions of the outcome of a process are quantitative. As Bassingthwaighte et al. (2) point out, these models are therefore refutable, thereby facilitating the entire process of science.

The difference between modeling and simulation is important. Modeling attempts to identify the mechanisms responsible for experimental or clinical observations, while with simulation, anything that reproduces experimental clinical, or educational data is acceptable. For this article, when you run a model, you are performing a simulation. Thus, simulation is not necessarily an overworked word. Most of the models described here are not simulators, however. That word should be reserved for those models that allow a nontechnologically oriented user to interact with the model and run a simulation. The words simulation or simulator are used when appropriate.

Having said that, the backgrounds and connotations of simulate and model are strikingly different. Reading the etymology of simulate and simulation is a chastening process, emphasizing as it does the unsavory past of the terms. In contrast, model has a more virtuous past. The Latin simulare can mean to do as if, to cheat, to feign or to counterfeit. Hence, a set of its meanings, from the OED, includes To assume falsely the appearance or signs of (anything); to feign, pretend, counterfeit, imitate; to profess or suggest (anything) falsely. The action or practice of simulating, with intent to deceive; false pretence, deceitful profession.

The Middle French modelle, on the other hand, implies perfect example worthy of imitation. Thus, one definition, again from the OED, is Something which accurately resembles or represents something else, esp. on a small scale; a person or thing that is the likeness of another. Freq. in the (very) model of.... Another definition is “A simplified or idealized description or conception of a particular system, situation, or process, often in mathematical terms, that is put forward as a basis for theoretical or empirical understanding, or for calculations, predictions, etc.; a conceptual or mental representation of something.” Our final definition is “A person or thing eminently worthy of imitation; a perfect exemplar of some excellence. Also: a representative specimen of some quality.”

This remarkable difference between the two terms can be transmuted to our basic philosophy: A simulation is no better than the model that drives it. Put another way, the model is arguably the most important part of the simulation. It helps prevent the simulation from being something that feigns.

Model Reduction versus Model Simplification

We were asked to consider the introduction of reduction, to distinguish from simplification, of the size of available models by weeding out minor effects. These concepts are indeed essential for the physiome project (at end of this article). Unfortunately, the literature that we have found usually fails to distinguish between the two terms, and too often uses them interchangeably. The following fragment of a definition was particularly discouraging: "Model reduction is the simplification or reduction of a mathematical model. . ." Not only is it a partial circular definition, but it also uses the term simplification as a way of achieving reduction of a model.

Dr. George Hutchinson, of GE Healthcare, has provided considerable help, and the remainder of this section comes from his ideas, as well as from the literature. He suggests that the difference between the two terms is in the purpose and the effect rather than in the action. Simplification of a model will eliminate elements, knowing that accuracy or validity may be significantly compromised. This simplification may be necessary because of (1) limitations in the platform running the model, (2) a decision to delay the work to model this portion of the system, or (3) this part of the model's not being germane to the focus of the model's intent. BODY Simulation does not model a real electrocardiogram (it is constructed from a look-up table) and that is a simplification of the model. This does compromise the overall validity of the model, but it was a simplification needed to make the rest of the model available to the public; this currently unnecessary submodel would have overwhelmed the rest of the model.

On the other hand, reduction is the elimination of some elements of a model in a way that should not significantly affect the accuracy or validity of the overall model. This could be done to streamline the code, to improve the faster-than-real-time performance, or to allow the model to run on a lesser platform. Again, with BODY Simulation as an example, deciding to limit the modeling of the arterial tree to large peripheral arteries, but not to extend it to the digits or the capillaries, is a reduction of the overall model, but it does not affect the intended use of BODY. The arterial waveforms are still good enough and the action of the MAP is unaffected.

This suggests that reduction and simplification are relative. Relative morality may be unacceptable, but relative definitions are reasonable. In a model built for one purpose, elimination of elements is reduction. In the same model built for a different purpose, such elimination is simplification. The problem of model reduction is to find a smaller system such that the number of components is much smaller than the original and the transfer function of the new system is close to the original transfer function.

Models, almost by definition, are necessarily simplifications and abstractions, to some degree, of the reality of the system. The modeler attempts to extract the important elements of the system (not all the variables, as that is not possible) and represent them so that they are simple enough to be understood and manipulated, yet realistic enough to portray the essential constructs and parameters of the situation. If either simplification or reduction is used, the techniques must be identified and justified, and the magnitude of their effect quantified.

Uncertainty analysis and sensitivity analysis are prerequisites for model building. The former allows quantifying the precision associated with the model response as a result of problems in the model input. Sensitivity analysis is aimed at establishing how the variation in the model output can be apportioned to different sources of variation, so that the modeler can establish how the given model depends on the information fed into it. Both are essential to assess the impact of simplification or reduction.

Here are some of the techniques that have been proposed for simplification and for reduction. Although these methods do overlap in their intended use, we have chosen to include them only with one or the other term. Also, lack of space precludes our going into detail about any of them.

Simplification

Eliminate short-term changes, when long-term ones only are important.
Make sure that the parameters and variables are important to the model.
Use a coarser (simpler) finite-element model.
Use intermediate variables, that is, reduce the number of input and output variables.
Use the reduced basis technique.
Simplify equations.
Simplify assumptions.
Use algebraic equations, instead of differential equations.

Reduction

Reduce the number of dimensions in a model.
Narrow down the search space among the input parameters.
Considers only significant inputs, as opposed to all inputs, of the full model.
Eliminate redundant information.
Use a numerical model reduction technique that assumes no knowledge of the system involved.

With such a large topic, one must draw the line somewhere, and the line used is rather large. For example, CNS is an enormous topic, and we were have to be content with a small subset: cerebral circulation. In general, pathology, diseases, or their effects are rarely addressed, although normal aging is discussed. Also usually avoided are the following systems and topics: GI, immune, CNS, peripheral nervous, hepatic, enzyme, cellular physiology, receptors and coagulation, as well as the effects of altitude, temperature changes, diving, microgravity, exercise, conditioning, hypertension, or pregnancy, for example. There will be little discussion of pediatrics, obstetric and other specialties. The ANS and acid–base regulation will be included with some extensive models, but not discussed as a separate topic. However, physiological PKPD models, which have been influential in the development of whole-body physiological models are included.

Since many physiological models involve control systems, this feature will be emphasized, including a special kind of control system known as autoregulation. Much of physiological systems modeling involves control systems. The body has many control systems, and only a few of them can be addressed here. Those discussed will be embedded in larger models, of the CVS, for example.

This is not intended to be a critical review, but rather a review resource for those wishing to use models. A few of the uses for a few of the models are mentioned briefly, in addition to that of incorporating them into one's own model or simulation. An attempted is made to differentiate between where a model is described and where the author presents material, especially equations, that could be used in a model.

Data to construct a model can come from many sources, including the literature and an author's own adhoc experiments. Unless otherwise stated, assume that data generated from the usually performed validating experiments fit a given model reasonably well. Weinstein (3) contends, however, that this form of validation may not be enough. Arguably, one may endorse a standard of model presentation in which the model builder shows not only what works, but also where the model fails, or where it makes novel predictions that have yet to be tested. While Weinstein suggests a good paradigm, published models are offered to the world to test out. Models are hypotheses, and no one expects final proof of an hypothesis in a paper: or series of papers. Some hypotheses, like some of Einstein's, seem to be eternally hypotheses. Modeling is a perpetual process.

The following lists a few uses of models:

1. Education is certainly the primary one.
2. Compression of enormous amounts of data.
3. An hypothesis.
4. Simulation and simulators.
    (a) Manikin-based simulators, such as an operating-room simulator.
    (b) Part-task trainer. The trainer could include a model and a ventilator, to teach students how to use the ventilator.
    (c) Screen-based simulator
5. Construct pharmacological models.
6. Simplify concepts to the user, especially in a simulation or simulator. The user of a model does not need to understand the mathematics or software, no more than one must understand how a car works before one drives it.
7. Suggest counterintuitive concepts for further exploration.
8. Suggest experiments to perform.
    (a) Fill missing gaps of knowledge.
    (b) Answer a posited question.
    (c) Try to explain ostensible modeling anomalies.
    (d) Provide more data for the model.
    (e) Any combination.
9. Substitute for animals or people in experiments.
10. Control part or all of an experiment.
11. Serve as an important component in adaptive control systems. A generic patient model learns about the patient and thereby helps improve the closed-loop control of an infused agent.

The most practical use to the reader, researcher, or educator is to operate a model, that is, run one's own simulation with it. Although most published models can not do that without considerable effort on the user's part, several models or sets of models make the process much easier. Werner et al. (4) state that their excellent model is "available for everybody". Levitt, Don Stanski, and Stephen Shafer all have large amounts of freely available software and data. The physiome project has more resources available than most of us can handle. The BODY Simulation model (see PBPM) is available as asimulator or as a dynamic linked library (dll). Because of the structure of this model, the latter means that the model can be connected to other software, such as simulation interfaces, or to a piece of electronically driven equipment, such as a ventilator, a patient monitor or an anesthesia machine. All of these are described in further detail, below.

RESOURCES

Several books stand out as resources for the reader. The classic is the late Vince Rideout's scholarly monograph *Mathematical and Computer Modeling of Physiological Systems* (5). It covers modeling for the CV, respiratory, and thermal regulatory systems, as well as transport and multiple modeling and parameter estimation. Baan, Noordergraaf, and Raines edited the proceedings of a symposium on cardiovascular system dynamics (6). The participants represented the international leaders in their fields: cardiology, physiology, engineering and physics. The postpresentation discussions alone are worth finding the book. The book comprises 62 papers, and it is difficult to choose which to emphasize. Some of them, however, will be discussed in the appropriate section. Most articles are detailed and worth exploring, for many reasons. A book edited by Papper and Kitz (7), again the result of a symposium, is described below.

TYPES OF MODELS

Mathematical models in our area fall into various categories, including mechanistic, black box; physiological, pharmacological, pharmacokinetic, pharmacodynamic; multiple, transport; analogue, hybrid and digital. Each one except physiological, is briefly described.

Mechanistic models can be understood in contrast to black box models in which only the input–output relations are important and not how these are realized. Though sometimes very useful, black box models have the limitation that they can only be used for descriptive purposes. Mechanistic models, for example, the prediction of flow through a vessel on the basis of diameter and length and pressure fall over it, allow the test of hypotheses. In case of the vessel example, one can test whether the viscosity assumed is the right one. Mechanistic models are important in testing hypotheses and formulating new ones.

Pharmacological is divided into two parts. Pharmacokinetic refers to the uptake, distribution and elimination of an agent, while pharmacodynamic refers to the action of the agent on the body. This is best remembered by the

Table 1. What, where, and how can be transported[a]

What	What (cont.)	Where	Route
Mass	Toxins	Atmosphere	Blood
Momentum	Chemical warfare agents	Anesthesia machine/ventilator	Lungs
Energy	Bacteria	Organs	Nerves
Gases	Viruses	Tissues	Membranes
Drugs	Genes	Blood	
Electrolytes	Molecules	Cells	
Proteins	Atoms	Receptors	
Hormones			
Endocrines			
Heat			
Information			

[a] There is no connection among the terms in each row.

following: pharmacokinetic is what the body does to the agent, while pharmacodynamic refers to what the agent does to the body.

Most of the complex models that are described are modular. Each of these modules can be a model, and the entire model is called a multiple model. A model with two or more modules is called a multiple model. If a model has CVS and respiratory components, each of those is a model. Essentially, anything that can be transported (see Table 1 and the next paragraph) can be a submodel in a multiple model. This includes anything from O_2 to a bacterial species, to a specific atom. In addition, the physiological systems themselves are submodels, for example CV, respiratory, and liver. Thus, the model for BODY Simulation (see PBPM) has almost 100 modules, or submodels. Ideally, each model can be sufficient unto itself, and can therefore be tested before it is incorporated into the main model.

Transport models are accurately named. Essentially, they transport something from one place to another. That something, we call an agent. In the case of the circulation, they, of course, transport agents around and around. The concept of transport modeling was developed by chemical engineers (8), whose models transported mass, momentum, and energy. Transport modeling is incredibly powerful. One has to be optimistic and creative about transport modeling. Assume, until proven otherwise, that a model can transport anything, anywhere and by any route. Table 1 lists a few of the agents that have been or could be transported in a model.

The following describes the various types of transport.

Momentum transport. In the wave equations describing blood flow, the concern is with momentum, blood viscosity, and the elasticity of the vessel walls in determining pressures and flows of the system.

Mass transport. Blood and lungs carry many important substances, such as O_2, CO_2, and pharmacological agents. The diffusion of these substances into or out of tissue is often essential to a model.

Energy transport. The blood in vessels, as well as the air we breathe, carries heat energy. This heat may diffuse through tissues, although in a way different from the mass diffusion mentioned above. Also, energy transformation, as well as transport, occurs in some tissues, including muscle, heart, brain, kidney, and liver.

Information transport. Information is carried throughout the body via nerve fibers, much as messages are transmitted in a communication system. Hormones also carry information, moving mostly with the aid of the bloodstream.

Models can also be categorized by the type of computer that implements them: analogue,hybrid or digital. This is not trivial, because even the most powerful digital computer cannot approach the speed and power of an analogue or hybrid computer, and we have yet to implement the entire original Fukui hybrid model (9) on a digital computer.

Whole-Body Models

A whole-body model is arbitrarily defined as one that includes the circulatory and respiratory systems, plus at least two other major systems. There is no model that includes all the major systems, much less any one system in detail. The good whole-body models have not been used extensively in education, while most simulators have used adhoc models assembled to meet the perceived needs.

The earliest and still archetypal whole-body model is that developed by Guyton and co-workers (4,10–13). To understand Guyton's contributions is to understand his model, and vice versa. One of his most important legacies was his application of principles of engineering and systems analysis to CV regulation. He used mathematical and graphical methods to quantify various aspects of circulatory function before computers were widely available. He built analogue computers and pioneered the application of large-scale systems analysis to modeling the CVS before the advent of digital computers. As digital computers became available, his CV models expanded dramatically in size to include many aspects of cardiac and circulatory functions. His unique approach to physiological research preceded the emergence of biomedical engineering, a field that he helped establish and promote in physiology, leading the discipline into a much more quantitative science.

The Guyton model has five main empirically derived physiological function blocks, with many subcomponents. The model has nearly 400 parameters and is remarkable in its scope. It comprises the following physiological subsystems, or modules:

Circulatory dynamics.
Nonmuscle oxygen delivery.
Muscle blood control and PO_2.
Vascular stress relaxation.
Kidney dynamics and excretion.
Thirst and drinking.
Antidiuretic hormone control.
Angiotensin control.
Aldosterone control.
Electrolytes and cell water.
Tissue fluids, pressures and gel.
Heart hypertrophy or deterioration.
Red cells and viscosity.
Pulmonary dynamics and fluids.
Heart rate and stroke volume.
Autonomic control.
Nonmuscle local blood flow control.

The Guyton model is alive and healthy. Werner et al. (4) coupled the Guyton model, which does not have a beating heart, to a pulsatile model. The new model comprises the hemodynamics of the four cardiac chambers, including valvular effects, as well as the Hill, Frank–Starling, Laplace, and ANS laws. They combined the two models because, with few exceptions, the extant published models were optimized for either studying short-term mechanics of blood circulation and myocardial performance—pulsatile models, for example—or mid- and long-term regulatory effects, such as exercise, homeostasis, and metabolism. In a gesture of intellectual philanthropy, the authors state that the program is written in the "C" language and is available to everybody (14). Fukui's model (9), actually, could do both short- and long-term studies. This was possible because it was implemented on a hybrid computer, and neither speed nor power was a consideration.

Another whole-body model, the Nottingham Physiology Simulator, uses a combination of CV, acid–base, respiratory, cerebrovascular, and renal models. Hardman et al. (15) partially validated this model for examining pulmonary denitrogenation, followed by apnea, by reproducing the methods and results of four previous clinical studies. They then used the model to simulate the onset and course of hypoxemia during apnea after pulmonary denitrogenation (replacing the nitrogen in the lung with O_2) (16). Several parameters were varied to examine their effects: functional residual capacity, O_2 consumption, respiratory quotient, hemoglobin concentration, ventilatory minute volume, duration of denitrogenation, pulmonary venous admixture, and state of the airway (closed versus open) The Nottingham group used their simulator for two other purposes. First, they assessed the accuracy of the simulator in predicting the effects of a change in mechanical ventilation on patient arterial blood–gas tensions. Second, they compared two methods of venous admixture estimation: one using the simulator and data commonly available in the intensive care unit, and the other using an isoshunt style calculation that incorporated assumed values for the physiological variables (17,18).

To read about the fascinating quest for a true whole-body model, (see Physiome), at the end of this article, plus (http://nsr.bioeng.washington/PLN) Other whole-body models will be discussed in the next section (physiologically based pharmacokinetic/pharmacodynamic models). Also, there are several whole-body models in the section on models of systems.

PHYSIOLOGICALLY BASED PHARMACOLOGICAL MODELS

It would seem outside the remit of this article to discuss pharmacological models, but some of them represent a source of detailed whole-body models. Many whole-body models were developed primarily as pharmacological models, and these models help flesh out what would otherwise be a scanty topic. Whole-body pharmacological models were originally called uptake-and-distribution models, but are now called a more respectable sounding physiological PKPD models. The term is broad, and very few physiological PKPD models are whole-body models, however.

It also turns out that pharmacology can be useful in physiological models, helping create realism. For example, by using epinephrine and norepinephrine, BODY Simulation simulates the CV response to pain or to hypercapnia simply by internally injecting one or both of these agents. Part of the realism relates to the fact that it takes time for circulating epinephrine, and its effects, to fade away: after the stimulus has vanished. The agent must be metabolized and redistributed, as in real life.

The history of these models is briefly explored. The senior author's interest in uptake and distribution was stimulated in the mid-1950s during Avram Goldstein's course on pharmacology, which highlighted the subject in a lucid, prescient way.

The desire for more accurate pharmacological modeling has had a significant impact on the development of better physiological models, especially whole-body models. Part of this improvement came from the need for more compartments, which were often physiological systems. The history of whole-body physiological PKPD models is embedded in the history of uptake and distribution, which goes back to the mid nineteenth century, when John Snow, the first epidemiologist and the first scientific anesthesiologist, made observations on the uptake and elimination of several agents that he was testing in patients, including chloroform (19,20). In the late nineteeth and early twentieth centuries, Frantz (21) and Nicloux (22) observed that among all the tissues, the brain had the highest tissue concentration of diethyl ether after inhalation of that agent. In 1924, Haggard (23) correctly surmised that this phenomenon was related to the brain's small size and relatively large blood flow, as well as ether's brain–blood partition coefficient, which he uncannily estimated to be 1.11, compared with the currently accepted 1.14.

Less than four decades after Haggard's papers, whole-body anesthetic models were appearing in the literature

(7,24–26). Because of the therapeutic and physicalchemical characteristics of anesthetic agents, especially inhaled, the early models usually contained fat, muscle and brain, plus low and high perfused compartments. The seminal book in the area was *Uptake and Distribution of Anesthetic Agents*, edited by Papper and Kitz (7). This multiauthored book summarized the extant modeling knowledge in clear detail. Most of the literature, and the book, were based on the uptake and distribution of inhaled anesthetic agents, with the notable exception of Henry Price (26), whose model on thiopental was apparently the first uptake and distribution model. Modeling pioneers featured in the book include Eger and Severinghaus.

As exciting and useful as the early models were, they suffered from at least four problems. (1) They were not pulsatile, that is, the heart did not beat and the lungs did not breathe. Incorporating pulsatility can solve many physiological and pharmacological problems. (2) They were linear. Neither CO_2, along with its regional distribution, nor ventilation, for example, changed as a function of the concentration of the agent—or vice versa. (3) They usually only dealt with one agent at a time. Thus, for example, one could not determine the interactions among two or more agents. [A brave exception was Rackow et al. (27). (4) They were not true physiological models, partly because of concern No. 1 and partly because none of the physiology of the organs was incorporated: they simply acted as agent capacitors.

Little changed from the 1960s until the early 1970s, when Ashman et al. (28), plus Zwart, Smith and Beneken. (29–31), published the first nonlinear models in this area. In our model, cardiac output, regional circulation, and ventilation changed as a function of agent concentration. These changes, in turn, affected the uptake of the agent. Our model used both mass transport and multiple modeling and was implemented on an analog–hybrid computer.

Another breakthrough also occurred in the 1970s: the incorporation of our uptake and distribution model into the Fukui hybrid-computer pulsatile model (9,32–35). Physiological and pharmacological modeling were united. Among other things, venous return, the effects of arrhythmias, and some drug interactions were now possible. Our original model (29–31) used the inhaled anesthetic halothane for the agent; Fukui's (9) used CO_2 and O_2. The second Fukui-Smith model (35) incorporated halothane, plus CO_2. All of these models were multiple, transport models, implemented on analog–hybrid or hybrid computers.

One major problem remained. Only about half a dozen computers, and people, in the world could run these models, and the wide spread use of simulation was not possible. In 1983 our laboratory translated the hybrid code into digital, on a VAX, in FORTRAN. In 1985, Charles Wakeland, from Rediffusion Simulation Corporation, transferred the code into C on a Sigmagraphics Iris 2300. Two processors were used, one for running the model and one for the graphics display. Finally, there was a portable simulator, portable if one were strong. More importantly, Sleeper could be used for teaching (36).

The next goal was to incorporate the model and simulator into a PC. This was accomplished in 1989 and was a major step. Now it was possible to bring simulation to a larger audience. Unfortunately, the PCs available then were not quite ready. In 1992, Starko completely revised the code, in DOS assembly language, making a simulator available on a laptop. Starko achieved an amazing fourfold increase in efficiency. He used many of the techniques used in flight simulation, his main work, to create stunning graphics and interfaces. In 1998, the code was converted to Windows in C++. The conversion allowed many new features that were not possible in DOS, however, DOS was much faster, and running the simulation in Windows slowed it down noticeably.

The model currently has 37 compartments, >90 agents, >80 parameters to set for each patient, and 45 parameters to set for each agent. The parameters are user settable. The latest additions to the agents have included cyanide and sarin (37,38), the latter a nerve gas. We were able to model cyanide toxicity because each organ and tissue has a changeable O_2 consumption. The nerve gases are actually physiological, involving as they do cholinergic and muscarinic effects. We could implement both agents because BODY Simulation has receptors, with agent concentration-effect curves that have user adjustable slopes (gamma) and amplitudes (IC50). All of the agents involved in the therapy for both toxins have also been incorporated, including the therapeutic byproducts methemoglobin, cyanmethemoglobin, and thiocyanate. Five equations relating to cyanide therapy have been incorporated. In keeping with our definition of a simulator (interactivity) the user can change the rate and, when appropriate, the equilibrium constants of the equations.

Because of the detail in BODY Simulation, it has been possible to incorporate the normal aging process (39). Over 60 user-changeable patient parameters are changed, directly or indirectly, to implement each elderly patient. Four patients were constructed, aged 65, 75, 85, and 95, although patients of any age can be implemented. To assess these patients, three scenarios were run: (1) administration of an anesthetic agent (thiopental) with CV and respiratory depressant properties, (2) hemorrhage of 1000 mL in 10 min, and (3) nonfatal apnea. The results confirm that the elderly generally have a decreased physiological reserve: the older the patient, the less the reserve.

More details on BODY Simulation can be found online (40).

Oliver et al. (41) used a whole-body physiologically based PK model that incorporated dispersion concepts. In whole-body PBPK models, each tissue or organ is often portrayed as a single well-mixed compartment with limited distribution and perfusion rate. However, single-pass profiles from isolated organ studies are more adequately described by models that display an intermediate degree of mixing. One such model is the dispersion model. A salient parameter of this model is the dispersion number, a dimensionless term that characterizes the relative axial spreading of a compound on transit through the organ. The authors described such a model, which has closed boundary conditions.

PKQuest (42–44), as the abbreviation implies, is also a pharmacokinetic model only, with no pharmacodynamics. Thus, the nonlinearities that some PKPD models possess are lost. Levitt's whole-body models have 12 compartments, including the essential brain and heart. The models, as well as files of tissue and agent parameters, are easily and freely accessible online (45) or from links in the papers, which are published on BioMed Central. The other models described above, including BODY Simulation, use only intravenous or pulmonary administration of an agent, while PKQuest explores the gastrointestinal, intramuscular, and subcutaneous routes. These models have been used to explore many areas that not only elucidate the PK of agents, but also give insight into body compartments, thus enhancing our knowledge of the compartments and suggesting further areas to explore physiologically. These compartments include the interstitial space.

The Stanford group, led by Donald Stanski and Steven Shafer have developed an extensive array of PKPD models, some of them incorporating whole-body models. Details and considerable software, much of it public domain, can be found online (46). These models usually do not incorporate the nonlinearities associated with agent-induced or physiologically induced changes in patient physiology.

Many anesthesia simulators incorporate whole-body models, although the completeness of the model varies considerably. Sometimes, little detail is available to the public concerning these models. A description of some of these models can be found in the section Uses of Models.

Even with the computational resources now available, most PK and many PKPD models contain only two to four compartments. Compare BODY Simulation's 37 compartments. Whole-body models are complex and difficult. Pharmacologists have seriously asked, Why do we need all those compartments? Can't I achieve the same results from an uncluttered three-compartment model? Aside from losing the physiological essence of any simulation that is run, one also loses nonlinearity, the impact of physiology on drug kinetics, the important feature of drug interactions, the influence of patient condition, such as aging, and the effect of stresses, such as hemorrhage or apnea. As just one example, any agent that decreases hepatic blood flow will affect the concentration, and therefore the action, of any agent that is metabolized by the liver.

REGIONAL CIRCULATION AND AUTOREGULATION

Regional Circulation

Only whole-body models can have regional circulation, of course. In fact, regional circulation is crucial in this type of model, to connect the various physiological modules. Many factors control regional circulation, as well as the relation among the circulations, but only some of them are discussed, briefly, primarily because few models take these factors into account. The BODY Simulation model (see PBPM) is one of the few exceptions.

One of the interesting factors controlling regional circulation is the importance to the body of a given organ or tissue. How well is an organ's circulation maintained in the face of an acute stress, rapid hemorrhage, for example? Organs and tissues can be divided, simplistically, into four types/classes.

1. Essential in the short term (a few minutes) (brain and heart).
2. Essential in the medium term (a few hours) (hepatic, renal and splanchnic).
3. Essential in the longer term (several hours) (skeletal muscle).
4. Essentially nonessential (skin, cartilage, bone).

The overall regulation of regional circulation reflects this hierarchy. In addition, some agents can influence it. Few models, except for Guyton's model (see Whole-body Models) and some the whole-body models described above, incorporate detailed regional circulation. BODY Simulation (see PBPM) and one of Ursino's many models (47) take this hierarchy into account. In BODY Simulation (see PBPM), regional circulation is impacted by several factors, including O_2, CO_2, the baroceptors, strength of the heart, blood volume, hemorrhage, extracellular fluid, pharmacological agents, the presence of circulating epinephrine (pain or hypercapnia), and so on.

The regional-circulation hierarchy impacts the various regional circulation control mechanisms, and autoregulation is often subservient to this pecking order (except for cerebral). If severe shock occurs, the hypoxic muscle will not get its allotted supply for example. The cerebral and coronary circulations come first. In some ways, this is a form of regional steal, a term often used in physiology, but not in this context.

General Autoregulation

When sudden alterations in perfusion pressure are imposed in most types of arterial beds, the resulting abrupt changes in blood flow are only transitory, with flow returning quickly to the previous steady-state level. The exception is, of course, a sudden arterial occlusion. The ability to maintain perfusion at constant levels in the face of changing driving pressure is termed autoregulation. Autoregulation only occurs between certain pressure limits (if the pressure drops too low or soars too high, autoregulation fails, and organ perfusion is compromised) at low pressures, perfusion decreases, and at high pressures, excessive flow occurs. Autoregulation keeps tissue or organ flow essentially constant between an MAP of 60–180 mmHg (7.9–23.9 kPa), the limits varying with the tissue or organ.

Much of autoregulation may occur in the microcirculation. Groebe (48) outlined gaps in the understanding of how the microvasculature maintains tissue homeostasis, as of the mid-1990s: (1) integration of the potentially conflicting needs for capillary perfusion and hydrostatic pressure regulation, (2) an understanding of signal transmission pathways for conveying information about tissue energetic status from undersupplied tissue sites to the arterioles, (3) accounting for the interrelations between precapillary and postcapillary resistances, and (4) an explanation of how the body achieves local adjustment of perfusion to metabolic demands. Using mathematical modeling, Groebe argued that precapillary pressure regulation combined with postcapillary adjustment of perfusion to tissue metabolic status helped clarify the understanding of microvascular control.

Mechanisms of autoregulation vary substantially between organs. Coronary autoregulation is, for example, quite different from brain autoregulation. During hemorrhagic shock in pigs, microcirculatory blood flows in the stomach, colon, liver, and kidney decrease in concert with decreasing systemic blood flow; flow in the jejunal mucosa is preserved, and pancreatic blood flow is selectively impaired 180-(49). Differential autoregulation has also been shown in response to increases in blood pressure. For example, in response to infusion of pressor agents, renal autoregulation is almost unimpaired, while regulation in the mesenteric vascular bed is less adequate, and differs with different pressor agents. Muscle autoregulation is mediated partly by the metabolic byproducts of exercise. The main site of autoregulation in the kidney is, however, the afferent glomerular arteriole. There are two main factors that affect vascular tone in the afferent arteriole: stretch-activated constriction of vessels (as for the brain) and tubulo-glomerular feedback. The autoregulatory response and the factors causing it can also vary in the same organ. For example, in exercising dogs, increased O_2 demand of the LV is met primarily by increasing coronary flow, while increased O_2 extraction makes a greater contribution to RV O_2 supply (50).

Needless to say, autoregulation is complex, and difficult to model. The following section gives some examples, in different circulatory beds.

Cerebral Autoregulation

Because the brain resides in a rigid box, very small changes in CBF can lead to catastrophic changes in intercranial pressure (ICP). Rapid CBF autoregulation is therefore vital, and cerebral flow normally remains constant within MAP ranges of 50–150 mmHg (6.6–19.9 kPa). Any sudden changes in MAP are transmitted to the cerebral circulation, inducing similar changes in CBF, but fortunately, under normal conditions the CBF tends to return to its original value within a few seconds. The main regulator of brain blood flow is pressure-dependent activation of smooth muscle in the arterioles of the brain. The more the arteriole is stretched, the more it contracts, and this lasts as long as the stretch occurs.

Paneraiy (51) critically reviewed the concepts of cerebral autoregulation, including the role of mathematical models. The most common approach to evaluating cerebral autoregulation tests the effects of changes in MAP on CBF, and is known as pressure autoregulation. A gold standard for this purpose is not available and the literature shows considerable disparity of methods and criteria. This is understandable because cerebral autoregulation is more a concept than a physically measurable entity. Static methods use steady-state values to test for changes in CBF (or velocity) when MAP is changed significantly. This is usually achieved with the use of drugs, or from shifts in blood volume, or by observing spontaneous changes. The long time interval between measurements is a particular concern. Concomitant changes in other critical variables, such as PCO_2, hematocrit, brain activation, and sympathetic tone, are rarely controlled for. Proposed indies of static autoregulation are based on changes in cerebrovascular resistance, on parameters of the linear regression of flow–velocity versus pressure changes, or only on the absolute changes in flow. Methods of dynamic assessment are based on transient changes in CBF (or velocity) induced by the deflation of thigh cuffs, Valsalva maneuvers (a bearing down, that induces an increase in airway pressure with resultant complex, but easily understood, hemodynamic changes), tilting, and induced or spontaneous oscillations in MAP. Classification of autoregulation performance using dynamic methods has been based on mathematical modeling, coherent averaging, transfer function analysis, cross-correlation function, or impulse response analysis.

Cerebral autoregulation has been modeled independently and in concert with other factors. The most nearly complete model of the combination of factors is by Lu et al. (52). The goal of their work was to study cerebral autoregulation, brain gas exchange, and their interaction. Their large model comprised a model of the human cardiopulmonary system, which included a whole-body circulatory system, lung, and peripheral tissue gas exchange, and the CNS control of arterial pressure and ventilation, and central chemoreceptor control of ventilation, as well as a detailed description of cerebral circulation, CSF dynamics, brain gas exchange, and CBF autoregulation. Two CBF regulatory mechanisms were included: autoregulation and CO_2 reactivity.

Three groups have mathematically examined cerebral autoregulation by itself (53–56). This allows insight into the process, as well as the ability to incorporate the concepts into one's own model. In addition, Czosnyka et al. (57) constructed a model that helps the clinician interpret bedside tests of cerebrovascular autoregulation.

A simple model (authors' term) was described by Ursino and Lodi (58). The model includes the hemodynamics of the arterial–arteriolar cerebrovascular bed, CSF production and reabsorption processes, the nonlinear pressure–volume relationship of the craniospinal compartment, and a Starling resistor mechanism for the cerebral veins. Moreover, arterioles are controlled by cerebral autoregulation mechanisms, which are simulated by a time constant and a sigmoidal static characteristic. The model is used to simulate interactions between ICP, cerebral blood volume, and autoregulation.

Hyder et al. (59) have developed a model that describes the autoregulation of cerebral O_2 delivery *In vivo*. According to the model, the O_2 diffusivity properties of the capillary bed,which can be modified in relation to perfusion, play an important role in regulating cerebral O_2 delivery *In vivo*. Diffusivity of the capillary bed, in turn, may be altered by changes in capillary PO_2, hematocrit, and/or blood volume.

Kirkham et al. (60) presented a mathematical model representing dynamic cerebral autoregulation as a flow-dependent feedback mechanism. They introduced two modeling parameters: the rate of restoration, and a time delay. Velocity profiles were determined for a general MAP, allowing the model to be applied to any experiment that uses changes in MAP to assess dynamic cerebral autoregulation. The comparisons yielded similar estimates for the rate of restoration and the time delay, suggesting that these parameters are independent of the pressure change stimulus and depend only on the main features of the dynamic cerebral autoregulation process. The modeling also indicated that a small phase difference between pressure and velocity waveforms does not necessarily imply impaired autoregulation. In addition, the ratio between the variation in maximum velocity and pressure variation can be used, along with the phase difference, to characterize the nature of the autoregulatory response.

Coronary Autoregulation

The range of coronary autoregulation is 60–130 mmHg (7.9–17.3 kPa). Demonstration of autoregulation in the coronary bed is difficult in intact animals because modification of coronary perfusion pressure also changes myocardial oxygen demand and the extrinsic compression of the coronary vessels. However, when perfusion pressure is altered, but ventricular pressure, cardiac contractility, and heart rate (the principal determinants of myocardial oxygen demand) are maintained constant, autoregulation is clearly evident.

The coronary circulation has a different set of problems from other systems, as well as many paradoxes. These problems include the following: (1) It is caught in a unique situation: it must feed and cleanse the very organ that generates it. (2) The heart is a high oxidative organ with a high demand for O_2 and a very high O_2 consumption. (3) The av O_2 difference is much wider in the coronary circulation, implying that less reserve is available. (4) As myocardial function increases (increased HR or contractility, e.g.), the demand for O_2 increases dramatically. (5) Intramyocardial flow actually decreases as contractility increases, however, because of compression and shearing effects; in other words, the more the heart works, the more it impedes the flow that it needs. (6) Because of the impairment of myocardial blood flow during systole, nearly 80% of coronary flow occurs during diastole. (7) Although most arterial flow occurs during diastole, most venous flow occurs during systole, both phenomena because of the compression effects. (8) Even though increased HR places an increased energy demand on the myocardium, there is less coronary flow as HR increases, since the valuable diastolic time decreases out of proportion to systolic time during tachycardia.

As in any vascular bed, blood flow in the coronary bed depends on the driving pressure and the resistance offered by this bed. Coronary vascular resistance, in turn, is regulated by several control mechanisms: myocardial metabolism (metabolic control), endothelial (and other humoral) control, autoregulation, myogenic control, extravascular compressive forces, and neural control. Each control mechanism may be impaired in a variety of conditions and each can contribute to the development of myocardial ischemia.

Control of coronary blood flow also differs depending on the type of vessel being considered: arteries, large arterioles, or smaller arterioles. Coronary capillaries also appear to make a significant contribution to coronary vascular resistance!

Several factors play a role in coronary autoregulation itself, including myogenic responses, resistance distribution in various size vessels, O_2 consumption, capillaries, flow-dependent dilation, and direct metabolic control. Regarding the last, for a substance to be considered a mediator of local metabolic control of coronary flow it should (1) be vasoactive, (2) found, in appropriate concentrations, to affect vascular tone, and (3) be of variable concentration in response to changes in metabolism. Myocardial O_2, along with CO_2 and other waste products are likely mediators of local metabolic control. Several models have examined one or more of these factors, although we could find no model that put everything together (61–68). One can, however, begin to separate out the contribution of each factor. For example, the synergistic interaction between PO_2 and PCO_2 accounts for about one-fourth of the change in coronary vascular conductance during autoregulation (62).

Cardiac Autoregulation

Cardiac autoregulation is briefly mentioned, although it is not circulatory autoregulation. The topic was one of the first to be studied, namely, the Frank–Starling mechanism, the dependence of individual stroke volume on end-diastolic volume. This mechanism helps ensure that what comes into the heart goes out, so that the heart does not blow up like a balloon. The functional importance of the Frank–Starling mechanism lies mainly in adapting left to right ventricular output. Just as with vascular autoregulation, many other factors can override the Frank–Starling mechanism, but it can become particularly important in CV disease.

In the past, the study of mechanical and electrical properties of the heart has been disjointed, with minimal overlap and unification. These features, however, are tightly coupled and central to the functioning heart. The maintenance of adequate cardiac output relies upon the highly integrated autoregulatory mechanisms and modulation of cardiac myocyte function. Regional ventricular mechanics and energetics are dependent on muscle fiber stress–strain rate, the passive properties of myocardial collagen matrix, adequate vascular perfusion, transcapillary transport and electrical activation pattern. Intramural hydraulic “loading" is regulated by coronary arterial and venous dynamics. All of these components are under the constant influence of intrinsic cardiac and extracardiac autonomic neurons, as well as circulating hormones.

MODELS OF SYSTEMS

Cardiovascular–Circulation

The CVS is essential to the body as the means for carrying substances, such as O_2, nutrients, and hormones to the tissues where they are needed and for bringing xenobiotics or waste products, such as CO_2 and acids, to the lungs, kidneys, or liver to be eliminated.

A pioneer in CV models was Bram Noordergraaf and his group, starting in the 1950s. His model began with the systemic circulation and later expanded to the pulmonary circulation. Over the period of 15 years, they evolved into models of great size, complexity, and sophistication. The model has a heart, and the systemic circulation has >115 segments. It included adjustable peripheral resistances, plus viscoelasticity, stiffness, radius and wall thickness of each segment, viscous properties of the blood, tapering in radius and elasticity, and frictional losses. Abnormalities, such as aortic stenosis, aortic insufficiency, idiopathic hypertrophic subaortic stenosis, and atrial septal defect have been simulated. We give two later, easily accessible references (69,70).

Another pioneer, in the Dutch school of modeling, was Jan Beneken. His models have also been sophisticated and rigorous (71,72). His models and the impressive ones of Karel Wesseling will be discussed later.

Two CV models are available for those who want to run one. The first is an interactive tutorial and mathematical model described by Rothe and Gersting (73). In addition, one can play with the 52-compartment CV model of Wesseling, as implemented by Starko, Haddock, and Smith (unpublished data). The inclusion of the two atria has made this model an especially unique and useful tool. Eventually, this model will be incorporated into BODY Simulation (see PBPM). Currently, however, it has no transport model(s), nor any control system, for example, an autonomic nervous system or baroceptors. In essence, it is a heart–lung preparation, without the lungs, but with pulmonary circulation. By itself, however, it is very useful in the study of the basics of the heart and circulation, allowing one to examine in great detail the subtle and not so subtle interactions of the CVS. This is made possible by the detail involved, as well as the ability to change any of hundreds of parameters and to study the time plots of any of hundreds of variables. The model allows the user to adjust the contractility and stiffness of the four chambers, as well as the pressure, volume, compliance and, when appropriate, resistance of each of the 52 compartments. In addition, heart rate, SVR, and total blood volume can be adjusted. Some of these changes involve changes in preload and afterload. Pressure, volume, and flow waveforms can be displayed for each compartment, and pressure–volume plots are available for the four chambers. Thirteen values, such as HR, MAP, and end-diastolic ventricular volumes, are numerically displayed. A small manual describes the model and how to use it.

Most cardiac models incorporate only a freestanding chamber, or chambers. The two ventricles, however, affect each other's dynamics. In addition, the pericardium, the stiff membrane that surrounds almost the entire heart, affects each chamber. The model of Chung et al. (74) describes the dynamic interaction between the LV and the RV over the complete cardiac cycle. The pericardium-bound ventricles are represented as two coupled chambers consisting of the left and right free walls and the interventricular septum. Timevarying pressure–volume relationships characterize the component compliances, and the interaction of these components produces the globally observed ventricular pump properties (total chamber pressure and volume). The model (1) permits the simulation of passive (diastolic) and active (systolic) ventricular interaction, (2) provides temporal profiles of hemodynamic variables (e.g., ventricular pressures, volumes and flows), and (3) can be used to examine the effect of the pericardium on ventricular interaction and ventricular mechanics. The model also yields qualitative predictions of septal and free wall displacements.

Similarly, few models have addressed the mechanical interaction between the CV and pulmonary systems, for example, how the combined cardiopulmonary system responds to large amplitude forcing (change in an important variable). To address this issue, Lu et al. (75), developed a human cardiopulmonary system model that incorporates important components of the cardiopulmonary system and their coupled interaction. Included in the model are descriptions of atrial and ventricular mechanics; hemodynamics of the systemic and pulmonary circulations; baroreflex control of arterial pressure; airway and lung mechanics; and gas transport at the alveolar-capillary membrane. They applied this model to the analysis of the Valsalva maneuver. The model could predict the hemodynamic responses to markedly increased intrathoracic (pleural) pressures during a Valsalva maneuver. In short, this model can help explain how the heart, lung, and autonomic tone interact during the Valsalva maneuver.

BODY Simulation (see PBPM) also incorporates an interaction between intrapleural pressure (as reflected by airway pressure) and venous return. Problems with a patient ventilator, for example, can create disastrous depression of the CVS (76). Combining computational blood flow modeling with 3D medical imaging provides a new approach for studying links between hemodynamic factors and arterial disease. Although this provides patient-specific hemodynamic information, it is subject to several potential errors. A different approach, developed by Moore et al. (77) can quantify some of these errors and identify optimal reconstruction methodologies.

For several reasons, modeling the pulmonary circulation presents challenges. Huang et al. (78) described a mathematical analogue-circuit model of pulsatile flow in cat lung based on existing morphometric and elastic data. In the model, the pulmonary arteries and veins were treated as elastic tubes, whereas the pulmonary capillaries were treated as two-dimensional (2D) sheets. Input impedances of the pulmonary blood vessels of every order were calculated under normal physiological conditions. The pressure-flow relation of the whole lung was predicted theoretically. Comparison of the theoretically predicted input impedance spectra with their experimental results showed that the modulus spectra were well predicted, but significant differences existed in the phase angle spectra between the theoretical predictions and the experimental results. The authors state that the current model cannot explain this latter discrepancy.

Occlusion experiments yield time–pressure and time–flow curves that are related to the longitudinal distribution of compliances and resistances in the pulmonary circulation. The standard approach to the analysis of these curves involves the observation of relevant features of their graphs, which may directly reflect model parameter values. De Gaetano and Cremona (79) considered five possible models of pulmonary vascular pressure dynamics and the relative (nonlinear) least-squares parameter estimation from experimental data, making simultaneous use of all available information. The five models included two linear models without inductance units, one linear model with inductance units, one nonlinear model with variable resistance, and one nonlinear model with variable compliance. In all cases, parameter estimation for the numerically integrated model was performed by unweighted least squares, using a variable-metric minimization technique.

Potentially, the more detailed the model, the more accurate it is. Karamanoglu et al. (80) simulated the effects of wave travel and wave reflection with a mathematical model of the whole arterial tree, which comprised 142 uniform transmission-line segments. The arterial model was partitioned into three separate segments: upper limbs, trunk, and lower limbs. Aging was simulated by increasing average pulse wave velocities of these segments. Reflection coefficients at the terminal elements were altered to simulate vasodilation and vasoconstriction.

Karamanoglu and Feneley (81) also used a linear mathematical model of the entire human arterial tree to derive realistic impedance patterns by altering (1) Young's modulus of the arterial wall of the individual branches, (2) peripheral reflection coefficients, and (3) distal compliances at the terminations. These calculated impedance patterns were then coupled to realistic LV outflow patterns determined by unique (1) end-diastolic and endsystolic pressure–volume relationships, (2) preload-recruitable stroke work relationships, and (3) shortening paths simulated by altered aortic flow contours. Left ventricular outflow patterns were as important as impedance parameters in determining late systolic pressure augmentation, at least in this model.

Cardiac valvular modeling and simulation are important, especially given the common use of echocardiography. Sun et al. (82) examined this area. The transmitral and pulmonary venous flow velocity patterns were related to the physiological state of the left heart with an electrical analogue model. Filling of the LV through the mitral valve was characterized by a quadratic Bernoulli's resistance in series with an inertance. Filling of the LA through the pulmonary veins was represented by a lumped network of linear resistance, capacitance, and inertance. The LV and LA were each represented by a time-varying elastance. A volume dependency was incorporated into the LV model to produce physiological PV loops and Starling curves. The model accurately reflected the expected effects of aging and decreasing LV compliance, and could serve as a useful theoretical basis for echocardiographic evaluation of LV and LA function.

Yellin et al. (83) examined the mechanisms of mitral valve motion in mid-diastole, diastole and at closure by simultaneously measuring mitral flow (electromagnetic), valve motion (echo), and AV pressures. Large variations in peak flow were accompanied by small variations in valve excursion. They concluded that the valve overshoots its equilibrium position and that the chordae produce tension on the valve during diastole. Their model offered a valve-closure theory unifying chordal tension, flow deceleration, and vortices, with chordal tension as a necessary condition for the proper functioning of the other two.

The Doppler transmitral velocity curve is commonly used to assess LV diastolic function. Thomas et al. (84) developed a mathematical formulation to study the physical and physiological determinants of the transmitral velocity pattern for exponential chamber PV relationships with active ventricular relaxation (2187 combinations investigated). They showed that transmitral velocity is fundamentally affected by three principal physical determinants: the transmitral pressure difference, the net AV compliance, and the impedance characteristics of the mitral valve. These physical determinants in turn are specified by certain compliance and relaxation parameters. They found that the peak mitral velocity is most strongly related to initial LA pressure but decreased by prolonged relaxation, low atria1 and ventricular compliance, and systolic dysfunction. Peak acceleration varies directly with atria1 pressure and inversely with the time constant of isovolumic relaxation, with little influence of compliance, whereas the mitral deceleration rate is approximately valve area divided by AV compliance.

A moderate reduction in coronary blood flow results in decreased myocardial O_2 consumption, accelerated glycolysis, decreased pyruvate oxidation, and lactate accumulation. To quantitatively understand cardiac metabolism during ischemia, Salem et al. (85) demonstrated a mechanistic, mathematical model based on biochemical mass balances and reaction kinetics in cardiac cells. Computer simulations showed the dynamic responses in glucose, fatty acid, glucose-6-phosphate, glycogen, triglyceride, pyruvate, lactate, acetyl-CoA, and free-CoA, as well as CO_2, O_2, phosphocreatine/creatine, nicotinamide adenine dinucleotide (reduced form)/nicotinamide adenine dinucleotide (oxidized form) (NADH/NAD1), and adenosine diphosphate/adenosine triphosphate (ADP/ATP). When myocardial ischemia was simulated by a 60% reduction in coronary blood flow, the model generated myocardial concentrations, uptakes, and fluxes that were consistent with experimental data from *in vivo* pig studies. With the onset of ischemia, myocardial lactate concentration increased and the myocardium switched from a net consumer to a net producer of lactate.

Olufsen et al. (86) modeled blood flow in large systemic arteries by using one-dimensional (1D) equations derived from the axisymmetric Navier–Stokes equations for flow in compliant and tapering vessels. The arterial tree is truncated after the first few generations of large arteries with the remaining small arteries and arterioles providing outflow boundary conditions for the large arteries. By modeling the small arteries and arterioles as a structured tree, a semianalytical approach based on a linearized version of the governing equations can be used to derive an expression for the root impedance of the structured tree in the frequency domain. In the time domain, this provides the proper outflow boundary condition. The structured tree is a binary asymmetric tree in which the radii of the daughter vessels are scaled linearly with the radius of the parent vessel. Blood flow and pressure in the large vessels are computed as functions of time and axial distance within each of the arteries.

The CVS is an internal flow loop with multiple branches circulating a complex liquid. The hallmarks of blood flow in arteries are pulsatility and branches, which cause wall stresses to be cyclical and nonuniform. Normal arterial flow is laminar, with secondary flows generated at curves and branches. Arteries can adapt to and modify hemodynamic conditions, and unusual hemodynamic conditions may cause an abnormal biological response. Velocity profile skewing can create pockets in which the wall shear stress is low and oscillates in one direction. Atherosclerosis tends to localize to these sites and creates a narrowing of the artery lumen: a stenosis. Plaque rupture or endothelial injury can stimulate thrombosis, which can block blood flow to heart or brain tissues. The small lumen and elevated shear rate in a stenosis create conditions that accelerate platelet accumulation and occlusion. The relationship between thrombosis and fluid mechanics is complex, especially in the poststenotic flow field. New convection models have been developed to predict clinical occlusion from platelet thrombosis in diseased arteries (87).

Cardiovascular Regulation

Vasquez et al. (88) presented a lucid review of the overall coordination of three of the major mechanisms involved in CVS control: the baroreceptors, chemoreceptors, and cardiopulmonary reflexes. The central chemoreceptors are the main body CO_2 (actually, pH acts as a surrogate for CO_2) sensors, the main sensors in the feedback loop. They are located in the medulla, near the ventricle and bathed in brain extracellular fluid, which is close to the composition of CSF. The normal pH of the fluid is 7.32 and is poorly buffered. Thus, CSF pH is regulated more rapidly than in the rest of the body. Although CO_2 diffuses rapidly into all tissues, H^+ does not penetrate into CSF. However, if PCO_2 increases in the blood, it diffuses into CSF and lowers the pH. Ventilation responds to changes in CSF pH.

Peripheral chemoreceptors are located in the carotid and aortic bodies and in clumps along the route of the abdominal vagus. Their role in the regulation of the CVS cannot be understood without knowing the various factors that can change chemoreceptor afferent activity. Furthermore, changes in chemoreceptor activity not only exert primary reflex effects on the CVS, but they evoke changes in the central drive for ventilation that secondarily affect the CVS. For an excellent, comprehensive review of this subject, see Marshall (89).

The arterial baroreflex contributes significantly to the short-term regulation of blood pressure and CV variability. Several factors, including reflex, humoral, behavioral, environmental and age, may influence gain and effectiveness of the baroreflex, as well as CV variability. Many central neural structures are also involved in the regulation of the CVS and contribute to the integrity of the baroreflex. For those who wish a good summary of the subject, read the review by Lanfranchi and Somers (90).

Continuous blood pressure recordings usually show a surprising variability in MAP. Short-term variability includes components with periods of a few seconds to many hours, and can be spontaneous or in response to a maneuver or activity. Blood pressure can increase 50 mmHg (6.6 kPa) during painful stimuli in a matter of a minute, or decrease 20–30% in a few seconds during an orthostatic maneuver. Often, blood pressure shows oscillatory fluctuations in an ~ 10 s rhythm. All this variability has been demonstrated in normal subjects, especially by Wesseling and co-workers (91–93). This short-term blood pressure variability challenges the concept of an effective, stabilizing baroreflex. Wesseling has called this phenomenon the baroreflex paradox and has proposed a baromodulation hypothesis (91–93). The baromodulation hypothesis states that the baroreflex gain can be modulated to have high gain in some situations, low gain in others. A decrease in baroreflex gain by itself causes blood pressure to increase, while an increase in gain causes it to fall (91). A gain change does not change baroceptor function itself. No changing baroceptor sensitivity is postulated, and no baroceptor resetting need occur. Theoretically, the physiological location where modulation occurs could be anywhere in the reflex loop, even in the individual efferent-pathways, separately. However, the logical site would be the vasomotor center in the medulla.

We should point out that this insight was made possible by using a noninvasive monitoring device that is enhanced by several models (see Finapres, under uses for Models), and that yet another model was used to test the hypothesis. Several investigators have based their work on this remarkable insight (references on request).

When blood volume is expanded, CVP increases, stimulating cardiopulmonary receptors in the atria and ventricles and perhaps arterial baroreceptors in the aortic arch and carotid sinus. Volume expansion or stimulation of atrial receptors can inhibit vasopressin release, decrease sympathetic nerve activity, and attenuate drinking. Atrial distension stimulates secretion of atrial natriuretic peptide (ANP) from the atria leading to natriuresis. Thus it is possible that ANP may inhibit vasopressin release, reduce blood pressure, decrease drinking, and lead to natriuresis and diuresis via central pathways.

Rose and Schwaber (94) described a model that included only HR in the baroceptor modulation of arterial pressure, since the vagus HR response is the most rapid responder to changes in arterial pressure. They observed that vagal induced changes in HR do not influence arterial pressure, except when certain initial conditions of the CVS are met. It may be that vagally mediated alterations in an inotropic and dromotropic state, which are not included in this model, play important roles in the fast reflex control of blood pressure or that the vagal limb of the baroreflex is of rather limited effectiveness. Had the authors decreased heart rate >50%, they may have observed profound CV changes arising from a heart rate change.

Ursino and Magosso (95) detailed a mathematical model of the acute CV response to isocapnic hypoxia. The model includes a pulsating heart, the systemic and pulmonary circulation, a separate description of the vascular bed in organs with higher metabolic need, and the local effect of O_2 on these organs. The model also includes the action of several reflex regulatory mechanisms: the peripheral chemoreceptors, the lung stretch receptors, the arterial baroreceptors, and the hypoxic response of the CNS. The early phase of the biphasic response (8–10 s), caused by activation of peripheral chemoreceptors, exhibits a moderate increase in MAP, a decrease in HR, a relatively constant CO, and a redistribution of blood flow to the organs with higher metabolic need, at the expense of other organs (see the hierarchy scheme, in Regional Circulation and Autoregulation). The later phase (20 s) is characterized by the activation of lung stretch receptors and by the CNS hypoxic response. During this phase, CO_2 and HR increase, and blood flow is restored to normal levels, in organs with lower metabolic need.

These authors performed an extensive validation of this model (96). The role of the different mechanisms involved in the CV response to hypoxia (chemoreceptors, baroreceptors, lung stretch receptors, and CNS hypoxic response) was analyzed in different physiological conditions. The simulation results revealed the following: (1) the model can reproduce the CV response to hypoxia very well between 100 and 28 mmHg (13.3 and 3.7 kPa) PO_2. (2) Sensitivity analysis of the impact of each individual mechanism underlines the role of the baroreflex in avoiding excessive derangement of systemic arterial pressure and CO_2 during severe hypoxia and suggests the existence of significant redundancy among the other regulatory factors. (3) With chronic sinoaortic denervation (i.e., simultaneous exclusion of baroreceptors, chemoreceptors, and lung stretch receptors), the CNS hypoxic response alone is able to maintain reasonably normal CV adjustments to hypoxia, although suppression of the CNS hypoxic response, as might occur during anesthesia, led to a significant arterial hypotension. (4) With controlled ventilation, a significant decrease in HR that can only partly be ascribed to inactivation of lung stretch receptors. (5) When maintaining a constant CO_2 during severe hypoxia, the chemoreflex can produce a significant decrease in systemic blood volume.

As an extension of the group's isocapnic hypoxia model, Magosso and Ursino (97) studied the effect of CO_2 on the CVS. The previous model (95) had already incorporated the main reflex and local mechanisms triggered by O_2 changes. The new features covered by the model were the O_2–CO_2 interaction with the peripheral chemoreceptors, the effect of local CO_2 changes on peripheral resistances, the direct CNS response to CO_2, and the control of central chemoreceptors on minute ventilation and tidal volume. The model could simulate the acute CV response to changes in blood gas content in a variety of conditions (normoxic hypercapnia, hypercapnia during artificial ventilation, hypocapnic hypoxia, and hypercapnic hypoxia). The model ascribes the observed responses to the complex superimposition of many mechanisms simultaneously working (baroreflex, peripheral chemoreflex, CNS response, lung-stretch receptors, local gas tension effect), which may be variably activated depending, on the specific stimulus under study. However, although some experiments can be reproduced using a single basal set of parameters, reproduction of other experiments requires a different combination of the mechanism strengths (particularly, a different strength of the local CO_2 mechanism on peripheral resistances and of the CNS response to CO_2).

Melchior et al. (98) used as an example the short-term response of the human CVS to orthostatic stresses to develop a mathematical model. They reviewed the physiological issues involved and how these issues have been handled in previous CV models for simulation of the orthostatic response. Most extant models were stimulus specific with no apparent ability to simulate the responses to orthostatic stimuli of different types. They suggest that a comprehensive model incorporating all known phenomena related to CV regulation is needed. The paper represents a good start in providing a framework for future efforts in mathematical modeling of the entire CVS, and the review of issues is outstanding.

Lerma et al. (99) combined parts of two systems: the baroceptor reflex in the CVS and the renal system in chronic renal failure (CRF). They developed a model of baroreflex control of MAP, in terms of a delay differential equation, and used it to predict the adaptation of short-term CV control in CRF patients. The model predicts stable and unstable equilibria close to steady-state MAP. Their results suggest that the cardiac pump has a more restricted response in CRF patients. The model quantifies the CV adaptations to CRF, including increased SVR and baroreflex delay, as well as decreased arterial compliance, cardiac period, and stroke volume.

In yet another paper, Ursino and Magosso (100) examined the response to O_2 and CO_2 changes mediated by one CV regulator mechanism, the carotid body chemoreceptor. The model assumes that the static chemoreceptor characteristic depends on O_2 saturation in the arterial blood and on CO_2 arterial concentration. The values of O_2 saturation and of CO_2 concentration are computed, from pressure, using blood dissociation curves, which include both the Bohr and Haldane effects. The dynamic response includes a term depending on the time derivative of CO_2 concentration and a low pass filter, which accounts for the time required to reach the steady-state level. With a suitable choice of parameters, the model reproduced the carotid chemoreceptor response under a variety of combined O_2 and CO_2 stimuli, both in steady-state conditions and in the transient period following acute CO_2 or O_2 pressure changes. During transient conditions, the effect of CO_2 pressure changes prevail over the effect of O_2 changes, due to the intrinsic derivative component of the response to CO_2.

Ursino et al. (101) explored one of the most important regulator effectors of the CVS: venous capacitance. To elucidate the role of venous capacity active changes in short-term CV homeostasis, they developed a mathematical model of the carotid-sinus baroreflex system. In the model, the CVS was represented as a series arrangement of six lumped compartments, which synthesized the fundamental hemodynamic properties of the systemic arterial, systemic venous, pulmonary arterial, and pulmonary venous circulations as well as of the left and right cardiac volumes. Cardiac outputs from the left and right ventricles were computed as a function of both downstream (afterload) and upstream atrial pressure (preload). Four distinct feedback regulatory mechanisms, working on SVR, HR, systemic venous unstressed volume, and systemic venous compliance, were assumed to operate on the CVS in response to carotid sinus pressure changes. The model was used to simulate the pattern of the main hemodynamic quantities in the short time period (1–2 min) after acute carotid sinus activation in vagotomized subjects. Simulations indicated that the model can reproduce experimental data quite well, with reference both to open-loop experiments and to acute hemorrhage performed in closed-loop conditions. Computer simulations also indicated that active changes in venous unstressed volume are very important in regulating CO_2 and MAP during activation of the carotid sinus baroreflex.

Modeling HR and blood pressure spontaneous variability can contribute to the understanding of both normal and pathologic CVS physiology. The observed fluctuations in HR and blood pressure are meaningful rhythmical fluctuations that reflect useful information about autonomic regulation. These rhythmical fluctuations, known as heart rate variability and blood pressure variation, are normally grouped into three major components: (1) the high frequency component, ~0.25 Hz, in synchrony with respiratory rate; (2) the low frequency component, generally centered ~0.1 Hz, which is attributed to the sympathetic activity and the closed-loop controlling action of cardiovascular regulation; and (3) The very low frequency component, ~0.04 Hz, which is probably due to the vasorhythmicity thermoregulatory system or to humoral regulations. Cohen and Taylor (102) nicely reviewed the subject and constructed their own model. Seydnejad and Kitney (103), as well as Cavalcanti and Belardinelli (104) have also modeled these areas. Of particular interest are the studies of Magosso et al. (105). Their findings include the following: (1) A significant increase in the gains and time delays (>9 s) of all the arterial baroreflex sympathetic mechanisms is required to induce instability (see Baromodulation and Wesseling, below). In this condition, systemic arterial pressure exhibits spontaneous oscillations with a period of ~20 s, similar to Mayer waves. The control of peripheral resistance seems more important than the venous volume control in the genesis of these oscillations. (2) An increase in the gain and time delay (~3 s) of the arterial baroreflex vagal mechanism causes the appearance of unpredictable fluctuations in heart period, with spectral components in the range 0.08–0.12 Hz 3) The cardiopulmonary baroreflex plays a less important role than does the arterial baroreflex in the genesis of these instability phenomena.

Aljuri and Cohen (106) took yet another approach to the problem. They emphasized the analytic algebraic analysis of the systemic circulation composed of arteries, veins, and its underlying physiological regulatory mechanisms of baroreflex and autoregulatory modulation of SVR, where the behavior of the system can be analytically synthesized from an understanding of its minimal elements. As a result of their analysis, they presented a mathematical method to determine short-term SVR fluctuations, which account for observed MAP fluctuations, and proposed a new CVS identification method to delineate the actions of the physiological mechanisms responsible for the dynamic couplings between CO_2, MAP, RA pressure, and SVR.

Ben-Haim et al. (107), using a finite-difference equation to model cardiac mechanics, simulated the stable action of the LV. Their model described the LV end-diastolic volume as a function of the previous end-diastolic volume and several physiological parameters describing the mechanical properties and hemodynamic loading conditions of the heart. Their simulations demonstrated that transitions (bifurcations) can occur between different modes of dynamic organization of the isolated working heart as parameters are changed. Different regions in the parameter space are characterized by different stable limit cycle periodicities. They proposed that mechanical periodicities of the heart action are an inherent part of its nonlinear nature. Although their model predictions and experimental results were compatible with previous experimental data, they may contradict several hypotheses suggested to explain the phenomenon of cardiac periodicities.

Ursino (108) made some interesting observations with his model on short-term carotid baroregulation and the pulsating heart. The model includes an elastance variable description of the left and right heart, the systemic (splanchnic and extrasplanchnic) and pulmonary circulations, the afferent carotid baroreceptor pathway, the sympathetic and vagal efferent activities, and the action of several effector mechanisms. The latter mechanisms work, in response to sympathetic and vagal action, by modifying systemic peripheral resistances, systemic venous unstressed volumes, heart period, and endsystolic elastances. The model is used to simulate the interaction among the carotid baroreflex, the pulsating heart, and the effector responses. Experimental data on HR control can be explained fairly well by assuming that the sympathetic-parasympathetic systems interact linearly on the heart period. The carotid baroreflex can significantly modulate the cardiac function curve. This effect, however, is masked *In vivo* by changes in arterial and atrial pressures. During heart pacing, CO_2 increases with frequency at moderate levels of heart rate and then fails to increase further because of a reduction in stroke volume. Shifting from nonpulsatile to pulsatile perfusion of the carotid sinuses decreases the overall baroreflex gain and significantly modifies operation of the carotid baroreflex. Finally, sensitivity analysis suggests that venous unstressed volume control plays the major role in the early hemodynamic response to acute hemorrhage, whereas systemic resistance and heart rate controls are slightly less important.

Short-term regulation of arterial blood pressure is accomplished by complex interactions between feedback and feed-forward information from the arterial and cardiopulmonary baroreceptors that combine with other local and neural factors to modulate CO_2 (HR and stroke volume) and SVR. Hughson et al. (109) used transfer function analysis and autoregressive moving average analysis to explore the interrelationships between CVP as an input to the cardiopulmonary baroreflex and MAP as an input to the arterial baroreflex in the regulation of SVR.

O'Leary et al. (110) used transfer function analysis to study the HR and vascular response to spontaneous changes in blood pressure from the relationships of systolic blood pressure to heart rate, MAP to SVR, and cerebrovascular resistance index, as well as stroke volume to SVR in healthy subjects in supine and 45° head-up tilt positions. Their data, which showed changes in MAP preceded changes in SVR as well as a possible link between stroke volume and SVR are consistent with complex interactions between the vascular component of the arterial and cardiopulmonary baroreflexes and intrinsic properties such as the myogenic response of the resistance arteries.

Toska et al. (111) used their mathematical model of baroreflexes and a simple circulation to analyze data from a previous study on humans (112). They modeled the heart, vascular bed, baroceptor reflexes and the ANS.

Microcirculation

Ostensibly, the microcirculation is the ultimate mediator of the major purposes of the circulation: delivering O_2 and cleansing the body of waste products, including CO_2. In addition to the mission of mass transport, the microcirculation and its endothelial cells have the role of regulation, signal transduction, proliferation, and repair. Lee (113) suggests, in addition, that the microcirculation is distensible and contains 40–50% of the total blood volume. In his Distinguished Lecture, he emphasized the integrative role of the microcirculation on circulatory control and its therapeutic role on blood volume compensation. He discussed shifts of volume from the microcirculation to the macrocirculation. It is possible that the microcirculation can play a more important role as a reservoir to compensate for blood volume loss than the venous system, and models are needed to investigate this intriguing concept.

Almost as an anomaly, it appears that hematocrit and nodal pressures can oscillate spontaneously in large microvascular networks in the absence of biological control. Carr and Lacoin (114) developed a model that not only explains the phenomenon, but also demonstrates how well-known phenomena explain it.

Schmidt-Schönbein (115) reviewed the mathematics of the microcirculation, including microvascular network topology, growth, and fluid mechanics; viscoelasticity and shape of microvessels; myogenic response; microvascular pressure–flow relationships; microvascular flow during pulsatile pressures; and non-Newtonian properties of blood in the microcirculation.

Pries and Secomb (116), in a paper that is part of the physiome project (see below), suggest a paradigm for attacking the complexity of the microcirculation. Terminal vascular beds exhibit a high degree of heterogeneity. Pertinent parameters are nonlinearly related, and their distributions are not independent. The classical typical vessel approach using averaged values for different vessel classes may not lead to a correct understanding of the physiology and pathophysiology of terminal vascular beds. Such problems can be avoided by studying microcirculatory functions at the network level using a combination of experiments and theoretical models. In this approach, distributions and relationships of pertinent parameters are measured *In vivo*, leading to the development of comprehensive databases. Such databases can be analyzed and complemented by suitable mathematical models, permitting estimation of parameters that are difficult to measure, and critical assessment of quantitative theories and hypotheses for microvascular function. This collaborative process between experimentally and theoretically oriented investigators may be facilitated in the future by the development of Web-based repositories of experimental data and theoretical models.

Beard and Bassingthwaighte (117) used a realistic geometric model for the 3D capillary network geometry as a framework for studying the transport and consumption of oxygen in cardiac tissue. The nontree-like capillary network conforms to the available morphometric statistics and is supplied by a single arterial source and drains into a pair of venular sinks. They explored steady-state O_2 transport and consumption in the tissue using a mathematical model that accounts for advection in the vascular network, nonlinear binding of dissolved oxygen to hemoglobin and myoglobin, passive diffusion of freely dissolved and protein-bound oxygen, and Michaelis–Menten consumption in the parenchymal tissue. The advection velocity field is determined by solving the hemodynamic problem for flow throughout the network. The resulting system is described by a set of coupled nonlinear elliptic equations, which are solved using a finite-difference numerical approximation. They found that coupled advection and diffusion in the 3D system enhance the dispersion of O_2 in the tissue compared with the predictions of simplified axially distributed models, and that no lethal corner, or oxygen-deprived region occurs for physiologically reasonable values for flow and consumption.

Cerebral

Considerable space has been devoted to this physiologically and personally important subject.

Clark and Kufahl (118) described a rigid-vessel model of the Circle of Willis, the complex circulatory system at the base of the brain. The circle is essential for redistributing blood flow after the sudden occlusion of a major cerebral vessel.

Lakin et al. (119) described a whole-body mathematical model for intracranial pressure dynamics. The model does not satisfy our definition of whole-body model, however, and we are including it in this section. Having said that, the model does avoid not simply presenting an isolated model of cerebral circulation. The model incorporates the dynamics of intracranial pressures, volumes, and flows. In addition to vascular connections with the rest of the body, the model incorporates a spinal-subarachnoid CSF compartment that bridges intracranial and extracranial physiology, allowing explicit buffering of ICP fluctuations by the spinal theca. The model contains cerebrovascular autoregulation, regulation of systemic vascular pressures by the sympathetic nervous system, regulation of CSF production in the choroid plexus, a lymphatic system, colloid osmotic pressure effects, and descriptions of CO_2.

Olufsen et al. (120) used a similar approach to study CBF during posture change from sitting to standing. Their model described pulsatile blood flow velocity and pressure in several compartments representing the systemic circulation. The model included compartments representing the trunk and upper extremities, the lower extremities, the brain, and the heart. They used physiologically based control mechanisms to describe the regulation of CBF velocity and arterial pressure in response to orthostatic hypotension resulting from postural change.

Sato et al. (121) first studied dynamic cerebrovascular responses in healthy humans during repetitive stepwise upward tilt (SUT) and stepwise downward tilt (SDT) maneuvers. The tilt maneuvers produced stepwise changes in both cerebral perfusion pressure and mean CBF velocity. The latter's response to SUT was well characterized by a linear second-order model. However, that to SDT demonstrated a biphasic behavior that was described significantly better by the addition of a slowly responding component to the second-order model. This difference may reflect both different CV responses to SUT or SDT and different cerebrovascular autoregulatory behaviors in response to decreases or increases in cerebral perfusion pressure.

The brain not only needs flow and O_2 regulation, it also needs temperature regulation, because of so many critical chemical and physical-chemical reactions. To study this phenomenon, Zhu (122) modeled selective brain cooling during hyperthermia. They developed a theoretical model to describe the effects of blood flow rate and vascular geometry on the thermal equilibration in the carotid artery based on the blood flow and the anatomical vascular geometry in the human neck. The potential for cooling of blood in the carotid artery on its way to the brain by heat exchange with the jugular vein and by radial heat conduction loss to the cool neck surface was evaluated. They showed that the temperature of the arterial blood can be as much as $1.1\,^{\circ}\mathrm{C}$ lower than the body core temperature, an observation in agreement with the difference between tympanic and body core temperatures. The model also evaluates the relative contributions of countercurrent heat exchange and radial heat conduction to selective brain cooling.

Does O_2 directly regulate CBF, or are there mediators? Ursino et al. (123,124) modeled the production and diffusion of vasoactive chemical factors involved in CBF regulation. Their model comprises two submodels. In the first, transport from capillary blood to cerebral tissue was analyzed to link changes in mean tissue PO_2 with CBF and arterial O_2 concentration changes. The second submodel described the production of vasoactive metabolites by cerebral parenchyma, arising from to a lack of O_2, and their diffusion toward pial perivascular space. They simulated the time dynamics of mean tissue PO_2, perivascular adenosine concentration and perivascular pH with changes in CBF. With their model, they concluded that the time delay introduced by diffusion processes is negligible compared with the other time constants in their system.

A second model (124) incorporated more submodels, each closely related to a physiological event. Thus, they could simulate the role played by the chemical factors described in the paragraph above, in the control of CBF during several different physiological and pathological conditions associated with the O_2 supply to cerebral tissue. These conditions included changes in autoregulation to changes in arterial and venous pressure, reactive hyperemia following cerebral ischemia and hypoxia. Their results suggest that adenosine and pH play a significant, but not exclusive, role in the regulation of the cerebrovascular bed.

Ursino and Magosso (125) presented a mathematical model of cerebrovascular regulation, in which emphasis was given to the role of tissue hypoxia on CBF. In the model, three different mechanisms are assumed to work on smooth muscle tension at the level of large and small pial arteries: CO_2 reactivity, tissue hypoxia, and a third mechanism necessary to provide good reproduction of autoregulation to cerebral perfusion pressure changes. The model is able to reproduce the pattern of pial artery caliber and CBF under a large variety of physiological stimuli, either acting separately (hypoxia, cerebral perfusion pressure changes, CO_2 pressure changes) or in combination (hypercapnia + hypoxia; hypercapnia + hypoten-hypotension). Furthermore, the model can explain the increase in CBF and the vasoconstriction of small pial arteries observed experimentally during hemodilution, ascribing it to the decrease in blood viscosity and to the antagonistic action of the flow-dependent mechanism (responsible for vasoconstriction) and of hypoxia (responsible for vasodilation). The interaction between hypoxia and ICP turns out to be quite complex, leading to different ICP time patterns, depending on the status of the CSF outflow pathways and of intracranial compliance.

Wakeland et al. (126) described a computer model of ICP dynamics that evaluated clinical treatment options for elevated ICP during traumatic brain injury. The model used fluid volumes as primary state variables and explicitly modeled fluid flows as well as the resistance, compliance, and pressure associated with each intra- and extracranial compartment (arteries and arterioles, capillary bed, veins, venous sinus, ventricles, and brain parenchyma). The model evaluated clinical events and therapies, such as intra- and extraparenchymal hemorrhage, cerebral edema, CSF drainage, mannitol administration, head elevation, and mild hyperventilation. The model was able to replicate observed clinical behavior in many cases, including elevated ICP associated with severe cerebral edema following subdural, epidural, or intraparynchemal hematoma. The model also mimics cerebrovascular regulatory mechanisms that are activated during traumatic brain injury.

Lodi and Ursino (127) demonstrated a mathematical model of cerebral hemodynamics during vasospasm. The model divided arterial hemodynamics into two cerebral territories: with and without vasospasm. It also included collateral circulation between the two territories, cerebral venous hemodynamics, CSF circulation, ICP, and craniospinal storage capacity. Moreover, the pial artery circulation in both territories was affected by CBF autoregulation mechanisms. First, the model was used to simulate some clinical results reported in the literature, concerning the patterns of middle cerebral artery flow velocity, CBF and pressure losses during vasospasm. Second, they performed a sensitivity analysis on certain model parameters (severity of caliber reduction, longitudinal extension of the spasm, autoregulation gain, ICP, resistance of the collateral circulation, and MAP) to clarify their influence on hemodynamics in the spastic area. The results suggested that the clinical impact of vasospasm depends on several concomitant factors, which should be simultaneously taken into account to reach a proper diagnosis.

Pasley et al. (128) used a mathematical model to test two hypotheses: (1) cyclic extravascular compressional modulation of the terminal venous bed occurs with positive pressure inhalation; and (2) the degree of modulation is diminished with increasing vascular dilation induced by increasing the level of $PaCO_2$. They made two modifications of Ursino's model of CSF dynamics (129–131):(1) terminal venous bed resistance was synchronously modulated with the ventilation cycle; and (2) both the depth of modulation and cerebrovascular resistance were progressively reduced with increasing levels of $PaCO_2$. Simulated and experimental correlation values progressively increased monotonically as the level of PCO_2 increased. Their results suggested that dilation of the cerebral vasculature reduces the influence of positive pressure ventilation on ICP by increasing venous pressure and thus diminishing the likelihood of vascular compression.

Increased ICP, which can result from etiologies ranging from tumors to trauma, can produce devastating results, of which death may be one of the more merciful. Modeling the phenomenon is critically important. Ursino and Lodi (132) used a mathematical model to characterize the relationships among CBF, cerebral blood volume, ICP, and the action of cerebrovascular regulatory mechanisms (autoregulation and CO_2 reactivity). The model incorporated CSF circulation, the ICP–volume relationship, and cerebral hemodynamics. The latter is based on three assumptions. (1) The middle cerebral arteries behave passively following transmural pressure changes. (2) The pial arterial circulation includes two segments (large and small pial arteries) subject to different autoregulation mechanisms. (3) The venous cerebrovascular bed behaves as a Starling resistor. A new aspect of this model relates to the description of CO_2 reactivity in the pial arterial circulation and in the analysis of its nonlinear interaction with autoregulation. Simulations obtained at constant ICP using various combinations of MAP and CO_2 support data on CBF and velocity concerning both the separate effects of CO_2 and autoregulation and their nonlinear interaction. Simulations performed in dynamic conditions with varying ICP suggest a significant correlation between ICP dynamics and cerebral hemodynamics in response to CO_2 changes. The authors believe that the model can be used to study ICP and blood velocity time patterns in neurosurgical patients, so that one can gain a deeper insight into the pathophysiological mechanisms leading to intracranial hypertension and resultant brain damage.

Loewe et al. (133) used a mathematical model to simulate the time pattern of ICP and of blood velocity in the middle cerebral artery in response to maneuvers simultaneously affecting MAP and end-tidal CO_2. First, they performed a sensitivity analysis, to clarify the role of some important model parameters (CSF outflow resistance, intracranial elastance coefficient, autoregulation gain, and the position of the regulation curve) during CO_2 alteration maneuvers performed at different MAP levels. Next, the model was applied to the reproduction of real ICP and velocity tracings in neurosurgical patients. They concluded that the model could be used to give reliable estimates of the main factors affecting intracranial dynamics in individual patients, starting from routine measurements performed in neurosurgical intensive care units.

Ursino et al. (134) analyzed changes in cerebral hemodynamics and ICP evoked by MAP and $PaCO_2$ challenges in patients with acute brain damage. The study was performed using a simple mathematical model of intracranial hemodynamics, particularly aimed at routine clinical investigation. The parameters chosen for the identification summarize the main aspects of intracranial dynamics, namely, CSF circulation, intracranial elastance, and cerebrovascular control.

By using a mathematical model, Ursino et al. (135) also studied the time pattern of ICP in response to typical clinical tests, namely, a bolus injection or withdrawal of small amounts of saline in the craniospinal space in patients with acute brain damage. The model included the main biomechanical factors assumed to affect ICP, CSF dynamics, intracranial compliance, and cerebrovascular dynamics. The simulation results demonstrated that the ICP time pattern cannot be explained simply on the basis of CSF dynamics, but also requires consideration of the contribution of cerebral hemodynamics and blood-volume alterations.

Sharan et al. (136) explored an interesting O_2-related phenomenon with a mathematical model. CBF increases as arterial O_2 content falls with hypoxic (low PO_2), anemic (low hemoglobin), and carbon monoxide (CO) (high carboxyhemoglobin) hypoxia. Despite a higher arterial PO_2, CO hypoxia provokes a greater increase in CBF than hypoxic hypoxia. They analyzed published data using a compartmental mathematical model to test the hypothesis that differences in PO_2 in tissue, or a closely related vascular compartment, account for the greater response to CO hypoxia. Calculations showed that tissue, but not arteriolar, PO_2 was lower in CO hypoxia because of the increased oxyhemoglobin affinity with CO hypoxia. Analysis of studies in which oxyhemoglobin affinity was changed independently of CO supports the conclusion that changes in tissue PO_2 (or closely related capillary or venular PO_2) predict alterations in CBF. They then sought to determine the role of tissue PO_2 in anemic hypoxia, with no change in arterial and little, if any, change in venous PO_2. Calculations predicted a small fall in tissue PO_2 as hematocrit decreases from 55 to 20%. However, calculations showed that changes in blood viscosity can account for the increase in CBF in anemic hypoxia over this range of hematocrits. It would have been interesting if the authors had tested hypoxia from cyanide poisoning, which blocks the utilization of O_2 at the enzyme cytochrome oxidase; blood and tissue actually increases.

Exploring well-defined physical phenomena is one thing; exploring fuzzy, undefined concepts, like consciousness, is quite another. Cammarota and Onaral (137) realized that complex physiological systems in which the emergent global (observable) behavior results from the interplay among local processes cannot be studied effectively by conventional mathematical models. In contrast to traditional computational methods, which provide linear or nonlinear input–output data mapping without regard to the internal workings of the system, complexity theory offers scientifically and computationally tractable models that take into account microscopic mechanisms and interactions responsible for the overall input–output behavior. The authors offered a brief introduction to some of the tenets of complexity theory and outlined the process involved in the development and testing of a model that duplicates the global dynamics of the induction of loss of consciousness in humans due to cerebral ischemia. Under the broad definition of complexity, they viewed the brain of humans as a complex system Successful development of a model for this complex system requires careful combination of basic knowledge of the physiological system both at the local (microscopic) and global (macroscopic) levels with experimental data and the appropriate mathematical tools. It represents an attempt to develop a model that can both replicate human data and provide insights about possible underlying mechanisms. They presented a model for complex physiological systems that undergo state (phase) transitions. The physiological system modeled is the CNS, and the global behavior captured by the model is the state transition from consciousness to unconsciousness. Loss of consciousness can result from many conditions such as ischemia (low blood flow), hypoxia (low oxygen), hypoglycemia, seizure, anesthesia, or a blow to the head, among others. Successful development of a model for this complex system requires careful combination of basic knowledge of the physiological system both at the local (microscopic) and global (macroscopic) levels with experimental data and the appropriate mathematical tools. Due to the wealth of human research and data available, the specific focus of the model is unconsciousness that results from the cerebral ischemia experienced by aircrew during aggressive maneuvering in high-performance aircraft.

Coronary

In the section Coronary Autoregulation, the problems and paradoxes of the coronary circulation are described. One of them was that myocardial perfusion was decreased in some areas by mechanical and shearing effects, and the harder the contraction, the greater the impairment. Smith (138) used an anatomically based computational model of coronary blood flow, coupled to cardiac mechanics to investigate the mechanisms by which myocardial contraction inhibits coronary blood flow. From finite deformation mechanics solutions the model calculates the regional variation in intramyocardial pressure (IMP) exerted on coronary vessels embedded in the ventricular wall. This pressure is then coupled to a hemodynamic model of vascular blood flow to predict the spatial–temporal characteristics of perfusion throughout the myocardium. The calculated IMP was shown to vary approximately linearly between ventricular pressure at the endocardium and atmospheric pressure at the epicardium through the diastolic loading and isovolumic contraction phases. During the ejection and isovolumic relaxation phases, IMP values increased slightly above ventricular pressure. The average radius of small arterial vessels embedded in the myocardium decreased during isovolumic contraction (18% in LV endocardium) before increasing during ejection (10% in LV endocardium) due to a rise in inflow pressure. Embedded venous vessels show a reduction in radius through both phases of contraction (35% at left ventricular endocardium). Calculated blood flows in both the large epicardial and small myocardial vessels show a 180° phase difference between arterial and venous velocity patterns with arterial flow occurring predominantly during diastole and venous flow occurring predominantly during systole. Their results confirm that the transmission of ventricular cavity pressure through the myocardium is the dominant mechanism by which coronary blood flow is reduced during the isovolumic phase of contraction. In the ejection phase of contraction, myocardial stiffening plays a more significant role in inhibiting blood flow.

Also illustrating this problem of impaired coronary flow during systole is a mathematical model that was based on an *In vitro* mechanical model consisting mainly of collapsible tubes (67). The pressure and flow signals obtained from both models were similar to physiological human coronary pressure and flow, both for baseline and hyperemic conditions.

Smith et al. (139) developed a discrete anatomically accurate finite element model of the largest six generations of the coronary arterial network. Using a previously developed anatomically accurate model of ventricular geometry, they defined the boundaries of the coronary mesh from measured epicardial coronaries. Network topology was then generated stochastically from published anatomical data. Spatial information was added to the topological data using an avoidance algorithm accounting for global network geometry and optimal local branch-angle properties. The generated vessel lengths, radii and connectivity were consistent with published data, and a relatively even spatial distribution of vessels within the ventricular mesh was achieved.

Pulmonary–Respiratory

The respiratory system is important as a means of O_2 uptake and CO_2 elimination. It may be compared with the CVS because gases are carried in it by pulsatile fluid flow, somewhat as gases and many other substances are carried by pulsatile blood flow in the CVS. The respiratory system is anatomically simpler, since it has but one branching out of the airflow passages, whereas the CVS fans out to the many body capillaries from the aorta, then fans in to the vena cavae, and repeats this pattern in the pulmonary circulation. However, analysis and modeling are far more complex in the respiratory system because air is a compressible fluid and because flow of air in the lungs is a tidal, or back-and forth, flow, in contrast to the one-way flow with superimposed pulsatility in the CVS. Also, although the respiratory system does not have valves as the CVS does, there are some important and rather difficult nonlinearities.

Good reviews are worth their weight in gold. Grotberg (140) reviewed respiratory fluid mechanics and transport processes. This field has experienced significant research activity for decades. Important contributions to the knowledge base come from pulmonary and critical care medicine, anesthesia, surgery, physiology, environmental health sciences, biophysics, and engineering. Several disciplines within engineering have strong and historical ties to respiration, including mechanical, chemical, civil–environmental, aerospace and, of course, biomedical engineering. Grotberg's review draws from the wide variety of scientific literature that reflects the diverse constituency and audience that respiratory science has developed. The subject areas covered include nasal flow and transport, airway gas flow, alternative modes of ventilation, nonrespiratory gas transport, aerosol transport, airway stability, mucus transport, pulmonary acoustics, surfactant dynamics and delivery, and pleural liquid flow. Within each area are several subtopics whose exploration can provide the opportunity of both depth and breadth for the interested reader.

The pioneer of computer modeling of respiratory control was Jim Defares (141,142). The one with the most impact, however, has been Fred Grodins (143), and several of the papers in this section acknowledge his significant contributions.

Chiari et al. (144) presented a comprehensive model of O_2 and CO_2 exchange, transport, and storage. The model comprises three compartments (lung, body tissue, and brain tissue) and incorporates a controller that adjusts alveolar ventilation and CO_2 by dynamically integrating stimuli coming from peripheral and central chemoreceptors. A realistic CO_2 dissociation curve based on a two-buffer model of acid–base chemical regulation is included. In addition, the model considers buffer base, the nonlinear interaction between the O_2 and CO_2 chemoreceptor responses, pulmonary shunt, dead space, variable time delays, and Bohr and Haldane effects. Their model fit the experimental data of ventilation and gas partial pressures in a very large range of gas intake fractions. It also provided values of blood concentrations of CO_2, HCO_3^-, and hydrogen ions in good agreement with more complex models characterized by an implicit formulation of the CO_2 dissociation curve.

Good sensitivity analysis can be difficult, at best. The tools in the paper by Hyuan et al. (145) are general and can be applied to a wide class of nonlinear models. The model incorporates a combined theoretical and numerical procedure for sensitivity analyses of lung mechanics models that are nonlinear in both state variables and parameters. They applied the analyses to their own nonlinear lung model, which incorporates a wide range of potential nonlinear identification conditions including nonlinear viscoelastic tissues, airway inhomogeneities via a parallel airway resistance distribution function, and a nonlinear block-structure paradigm. Model nonlinearities motivate sensitivity analyses involving numerical approximation of sensitivity coefficients. Examination of the normalized sensitivity coefficients provides insight into the relative importance of each model parameter, and hence the respective mechanism. More formal quantification of parameter uniqueness requires approximation of the paired and multidimensional parameter confidence regions. Combined with parameter estimation, they used the sensitivity analyses to justify tissue nonlinearities in modeling of lung mechanics for healthy and constricted airway conditions, and to justify both airway inhomogeneities and tissue nonlinearities during broncoconstriction. Some of the variables, parameters and domains included pressures, flows, volumes, resistances, compliances, impedances, pressure-volume and frequency.

A model of breathing mechanics (146) was used to interpret and explain the time course of input respiratory resistance during the breathing cycle, observed in ventilated patients. The authors assumed a flow-dependent resistance for the upper extrathoracic airways and volume-dependent resistance and elastance for the intermediate airways. A volume-dependent resistance described the dissipative pressure loss in the lower airways, while two constant elastances represented lung and chest wall elasticity. Simulated mouth flow and pressure signals obtained in a variety of well-controlled conditions were used to analyze total respiratory resistance and elastance estimated by an on-line algorithm based on a time-varying parameter model. These estimates were compared with those provided by classical estimation algorithms based on time-invariant models with two, three, and four parameters. The results confirmed that the difference between the end-expiration and end-inspiration resistance increases when obstructions shift from the upper to the lower airways.

Ursino et al. (147) presented a mathematical model of the human respiratory control system. It includes three compartments for gas storage and exchange (lungs, brain, tissue, and other body tissues), and various types of feedback mechanisms. These comprise peripheral chemoreceptors in the carotid body, central chemoreceptors in the medulla and a central ventilatory depression. The last acts by reducing the response of the central neural system to the afferent peripheral chemoreceptor activity during prolonged hypoxia of the brain tissue. The model also considers local blood flow adjustments in response to O_2 and CO_2 arterial pressure changes. Sensitivity analysis suggests that the ventilatory response to CO_2 challenges during hyperoxia can be almost completely ascribed to the central chemoreflex, while, during normoxia, the peripheral chemoreceptors also provide a modest contribution. By contrast, the response to hypercapnic stimuli during hypoxia involves a complex superimposition among different factors with disparate dynamics. Results suggest that the ventilatory response to hypercapnia during hypoxia is more complex than that provided by simple empirical models, and that discrimination between the central and peripheral components based on time constants may be misleading.

The phenomena collectively referred to as periodic breathing (including Cheyne Stokes respiration and apneustic breathing) have important medical implications. The hypothesis that periodic breathing is the result of delays in the feedback signals to the respiratory control system has been studied since the work of Grodins et al. (148) in the early 1950s. Batzel's dissertation (149) extended a model developed by Khoo et al. (150), to include variable delay in the feedback control loop and to study the phenomena of periodic breathing and apnea as they occur during quiet sleep in infants. The nonlinear mathematical model consists of a feedback control system of five differential equations with multiple delays. Numerical simulations were performed to study instabilities in the control system, especially the occurrence of periodic breathing and apnea in infants ~4 months of age. This time frame is important, since during it there is a high incidence of Sudden Infant Death Syndrome. Numerical simulations indicate that a shift in the controller ventilatory drive set point during sleep transition is an important factor for instability. Analytical studies show that delay-dependent stability is affected by controller gain, compartment volumes and the manner in which changes in minute ventilation are produced (i.e., by deeper breathing or faster breathing). Parenthetically, the increased delay resulting from congestive heart failure can induce instability at certain control gain levels.

The dimensions, composition, and stiffness of the airway wall are important determinants of airway cross-sectional area during dynamic collapse in a forced expiration or when airway smooth muscle is constricted. This can occur with asthma or COPD (emphysema). Under these circumstances, airway caliber is determined by an interaction between the forces acting to open the airway (parenchymal tension and wall stiffness) and those acting to close it (smooth-muscle force and surface tension at the inner gasliquid interface). Theoretical models of the airway tube law (relationship between cross-sectional area and transmural pressure) allow simulations of airway constriction in normal and asthmatic airways (151).

An excellent mathematical model of neonatal respiratory control (152) consists of a continuous plant and a discrete controller. Included in the plant are lungs, body tissue, brain tissue, a CSF compartment, and central and peripheral receptors. The effect of shunt is incorporated in the model, and lung volume and dead space are time varying. The controller uses outputs from peripheral and central receptors to adjust the depth and rate of breathing, and the effects of prematurity of peripheral receptors are included in the system. Hering–Breuer-type reflexes are embodied in the controller to accomplish respiratory synchronization. See also the Nottingham Physiology Simulator for a similar approach (Whole-body models).

Lung gas composition affects the development of anesthesia-related atelectasis, by way of differential gas absorption. A mathematical model (153) examines this phenomenon by combining models of gas exchange from an ideal lung compartment, peripheral gas exchange, and gas uptake from a closed collapsible cavity. The rate of absorption is greatest with O_2, less with NO_2 and minimal with N_2. BODY Simulation (see PBPM) achieves the same results.

Most respiratory models are limited to short-term (minutes) control. Those wishing to model longer term (days) control and adjustments should start with the detailed review by Dempsy and Forster (154). They discuss several important areas, including central chemoreception, cerebral fluids and chemoreceptor environment, physiologically important stimuli to medullary chemoreception, medullary chemoreceptor contributions to ventilatory drive, metabolic acid–base derangements, ventilatory response, mediation of ventilatory adaptation, ventilatory acclimatization to chronic hypoxia, ventilation during acute hypoxia, acclimatization during short-term hypoxia, acclimatization during long-term hypoxia, physiological significance of short- and long-term ventilatory acclimatization, ventilatory acclimatization to chronic CO_2 exposure, ventilation during chronic CO_2 inhalation, and whole-body CO_2 and H^+ during CO_2 exposure.

Renal

The kidneys have important physiological functions including maintenance of water and electrolyte balance; synthesis, metabolism and secretion of hormones; and excretion of the waste products from metabolism. In addition, the kidneys play a major role in the excretion of hormones, as well as drugs and other xenobiotics. The story of fluids and solutes is the story of the kidney, and vice versa.

The understanding of renal function has profited greatly from quantitative and modeling approaches for a century (155). One of the most salient examples is the concentration and dilution of the urine: a fundamental characteristic of the mammalian kidney. Only in the last three decades have the necessary components of this and other renal mechanisms been confirmed at the molecular level, but there have also been surprises. In addition, the critical role played by the fine regulation of Na^+ reabsorption in the collecting duct for the maintenance of normal blood pressure presents challenges to our understanding of the integrated interaction among systems. As a first step in placing the kidney in the physiome paradigm (see below), Schafer suggests (1) integrating currently restricted mathematical models, (2) developing accessible databases of critical parameter values together with indices of their degrees of reliability and variability, and (3) describing regulatory mechanisms and their interactions from the molecular to the interorgan level.

By now, you will have guessed our enthusiasm for good reviews. An excellent one to start with in this area is by Russell (156), on Na-K chloride cotransport. Obligatory, coupled cotransport of Na^+, K^+, and Cl^- by cell membranes has been reported in nearly every animal cell type. Russell's review examines the status of the knowledge about this ion transport mechanism.

In another review (the Starling Lecture, actually), DiBona (157) describes the neural control of the kidney. The sympathetic nervous system provides differentiated regulation of the functions of various organs. This differentiated regulation occurs via mechanisms that operate at multiple sites within the classic reflex arc: peripherally at the level of afferent input stimuli to various reflex pathways, centrally at the level of interconnections between various central neuron pools, and peripherally at the level of efferent fibers targeted to various effectors within the organ. In the kidney, increased renal sympathetic nerve activity regulates the functions of the intrarenal effectors: the tubules, the blood vessels, and the juxtaglomerular granular cells. This enables a physiologically appropriate coordination between the circulatory, filtration, reabsorptive, excretory, and renin secretory contributions to overall renal function. Anatomically, each of these effectors has a dual pattern of innervation consisting of a specific and selective innervation, in addition to an innervation that is shared among all the effectors. This arrangement permits maximum flexibility in the coordination of physiologically appropriate responses of the tubules, the blood vessels, and the juxtaglomerular granular cells to a variety of homeostatic requirements.

Physiologists have developed many models for interpreting water and solute exchange data in whole organs, but the models have often neglected key aspects of the underlying physiology to present the simplest possible model for a given experimental situation. Kellen and

Bassingthwaighte (158) developed a model of microcirculatory water and solute exchange and applied it to diverse observations of water and solute exchange in the heart. The key model features that permit this diversity are the use of an axially distributed blood-tissue exchange region, inclusion of a lymphatic drain in the interstitium, and the independent computation of transcapillary solute and solvent fluxes through three different pathways.

Endocrine

The insulin–glucose subsystem will be used as a paradigm of the endocrine system. Because diabetes is such a clinically complex and disabling disease, as well as a major and increasing personal and public health problem, many attempts at modeling have been made. Most of these models have examined various parts of the overall process, and one or two have tried to put it all together.

The normal blood glucose concentration in humans lies in the range of 70–110 mg·dL^{-1}. Exogenous factors that affect this concentration include food intake, rate of digestion, exercise, and reproductive state. The pancreatic hormones insulin and glucagon are responsible for keeping glucose concentration within bounds. Insulin and glucagon are secreted from beta-cells and alpha-cells respectively, which are contained in the islets of Langerhans, which are scattered in the pancreas. When blood glucose concentration is high, the beta-cells release insulin, resulting in lowering blood glucose concentration by inducing the uptake of the excess glucose by the liver and other cells (e.g., muscles) and by inhibiting hepatic glucose production. When blood glucose concentration is low, the alpha-cells release glucagon, resulting in increasing blood glucose concentration by acting on liver cells and causing them to release glucose into the blood.

Glucose concentrations outside the range 70–110 m·dL^{-1} are called hyperglycemia or hypoglycemia. Diabetes mellitus is a disease of the glucose–insulin regulatory system that is characterized by hyperglycemia. Diabetes is classified into two main categories: type 1 diabetes, juvenile onset and insulin dependent; and type 2 diabetes, adult onset and insulin independent.

Makroglu et al. (159) presented an extensive overview of some of the available mathematical models on the glucose–insulin regulatory system in relation to diabetes. The review is enhanced with a survey of available software. The models are in the form of ordinary differential, partial differential, delay differential and integrodifferential equations.

Tibell et al. (160) used models to estimate insulin secretion rates in patients who had undergone pancreas-renal transplant procedures.

When the complex physiology goes awry, manmade control systems can be used to understand and perhaps deal with the problem. These control systems require models. Ibbini et al. (161) have recently developed one such system. Parker et al. (162) set up a model-based algorithm for controlling blood glucose concentrations in type I diabetic patients.

Bequette (163) examined the development of an artificial pancreas in the context of the history of the field of feedback control systems, beginning with the water clock of ancient Greece, and including a discussion of current efforts in the control of complex systems. The first generation of artificial pancreas devices included two manipulated variables (insulin and glucose infusion) and nonlinear functions of the error (difference between desired and measured glucose concentration) to minimize hyperglycemia while avoiding hypoglycemia. Dynamic lags between insulin infusion and glucose measurement were relatively small for these intravenous-based systems. Advances in continuous glucose sensing, fast-acting insulin analogues, and a mature insulin-pump market bring closer the commercial realization of a closed-loop artificial pancreas. Model predictive control is discussed in-depth as an approach that is well suited for a closed-loop artificial pancreas. A major remaining challenge is handling an unknown glucose disturbance (meal), and an approach is proposed to base a current insulin infusion action on the predicted effect of a meal on future glucose values. Better meal models are needed, as a limited knowledge of the effect of a meal on the future glucose values limits the performance of any control algorithm.

One of the fascinating features of the insulin–glucose subsystem, and with other endocrine subsystems, is that it either oscillates or releases its hormones in an intermittent fashion. The terms oscillations, rhythms, ultradian, (relating to biologic variations or rhythms occurring in cycles more frequent than every 24 h [cf. circadian, about every 24 h])pulses, biphasic and bursting appear frequently in regards to insulin secretion. The following describes some of the models that study these phenomena.

Insulin is secreted in a sustained oscillatory fashion from isolated islets of Langerhans, and Berman et al. (164) modeled this phenomenon. Straub and Sharp (165) reviewed some models that have tried to explain the well-documented biphasic secretory response of pancreatic beta-cells to abrupt and sustained exposure to glucose.

Lenbury et al. (166) modeled the kinetics of insulin, using a nonlinear mathematical model of the glucose–insulin feedback system. The model has been extended to incorporate the beta-cells' function in maintaining and regulating plasma insulin concentration in humans. Initially, a gastrointestinal absorption term for glucose is used to effect the glucose absorption by the intestine and the subsequent release of glucose into the bloodstream, taking place at a given initial rate and falling off exponentially with time. An analysis of the model was carried out by the singular-perturbation technique to derive boundary conditions on the system parameters that identify, in particular, the existence of limit cycles in the model system consistent with the oscillatory patterns often observed in clinical data. They then used a sinusoidal term to incorporate the temporal absorption of glucose to study the responses in patients during ambulatory-fed conditions. They identified the ranges of parametric values for which chaotic behavior can be expected, leading to interesting biological interpretations.

An interesting electrophysiological phenomenon occurs in the islets of Langerhans: The release of insulin is controlled in these islets by trains of action potentials occurring in rapid bursts followed by periods of quiescence. This bursting behavior occurs only in intact islets: single cells do not display such bursting activity. Chay and Keizer (167) were the first attempt to model this phenomenon quantitatively. Sherman et al. (168) sought to explain the absence of bursting in single beta-cells using the idea of channel sharing. Keizer (169) modified this model by substituting an ATP- and ADP-dependent K channel instead of the Ca-dependent K channel. This model was then further improved by Keizer and Magnus (170). Sherman et al. (171) constructed a domain model to examine the effect of Ca on Ca-channel inactivation. Further refinements have been made by Keizer and Young (172).

Sherman (173) also reviewed mechanisms of ionic control of insulin secretion. He focused on aspects that have been treated by mathematical models, especially those related to bursting electrical activity. The study of these mechanisms is difficult because of the need to consider ionic fluxes, Ca handling, metabolism, and electrical coupling with other cells in an islet. The data come either from islets, where experimental maneuvers tend to have multiple effects, or from isolated cells, which exhibit degraded electrical activity and secretory sensitivity. Modeling aids in the process by integrating data on individual components such as channels and Ca handling and testing hypotheses for coherence and quantitative plausibility. The study of a variety of models has led to some general mathematical results that have yielded qualitative model-independent insights.

Endocrine systems often secrete hormones in pulses. Examples include the release of growth hormone and gonadotropins,, as well as insulin. These hormones are secreted over intervals of 1–3 h and 80–150 min, respectively. It has been suggested that relative to constant or stochastic signals, oscillatory signals are more effective at producing a sustained response in the target cells. In addition to the slow insulin oscillations, there are more rapid pulses that occur every 8–15 min. The mechanisms underlying both types of oscillations are not fully understood, although it is thought that the more rapid oscillations may arise from an intrapancreatic pacemaker mechanism. One possible explanation of the slow insulin oscillations is an instability in the insulin-glucose feedback system. This hypothesis has been the subject many studies, including some that have developed a mathematical model of the insulin–glucose feedback system. Tolic (174) reviewed several models that investigated oscillations, with glucose and the pancreas. Tolic's model (175) and others are available on the CellML site (176).

Shannahoff-Khalsa et al. (177), in Gene Yates' group, have expanded the purview of these fascinating phenomena by comparing the rhythms of the CV, autonomic, and neuroendocrine systems. Their results, from a time-series analysis using a fast orthogonal search method, suggested that insulin secretion has a common pacemaker (the hypothalamus) or a mutually entrained pacemaker with these three systems.

Fortunately, there is an excellent glucose–insulin model that helps one understand some of these concepts. Erzen et al. (178) have posted GlucoSim, a web-based simulator that runs on almost all computer platforms. GlucoSim is a program for simulating glucose–insulin interaction in a healthy person and in a type 1 diabetes patient. It has a flexible data output structure that can be used as an input into most postprocessing programs such as spreadsheet and graphics programs. Simulations can be performed by changing initial conditions or meal and insulin injection times. Simulation results can be plotted in 2D directly from the simulator by choosing appropriate buttons. The software is freely available to all users. This model is also on the CellML Site (176), but not the running model.

GlucoSim is one of the more nearly complete models that we have reviewed, and it satisfies our criteria for a whole-body model. It is a good example of the usefulness of a whole-body model with many compartments. This model is actually two whole-body models combined, with food ingestion in the glucose model and subcutaneous injection in the insulin model.

Action Potentials

In 1952, Hodgkin and Huxley published a paper showing how a nonlinear empirical model of the membrane processes could be constructed (179). In the five decades since their work, the Hodgkin–Huxley paradigm of modeling cell membranes has been enormously successful. While the concept of ion channels was not established when they performed their work, one of their main contributions was the concept that ion-selective processes existed in the membrane. It is now known that most of the passive transport of ions across cell membranes is accomplished by ion-selective channels. In addition to constructing a nonlinear model, they also established a method to incorporate experimental data into a nonlinear mathematical membrane model. Thus, models of action potentials comprise some of the oldest of nonlinear physiological models.

Action potential is a term used to denote a temporal phenomenon exhibited by every electrically excitable cell. The transmembrane potential difference of most excitable cells rests at some negative potential called the resting potential, appropriately enough. External current or voltage inputs can cause the potential of the cell to deviate in a positive direction, and if the input is large enough, the result is an action potential. An action potential is characterized by a depolarization that typically results in an overshoot >0 mV, followed by repolarization. Some cells may actually hyperpolarize before returning to the resting potential. After the membrane has been excited, it cannot be reexcited until a recovery period, called the refractory period, has passed.

When excitable cells are depolarized from their resting potential beyond a certain level (threshold), they respond with a relatively large, stereotyped potential change. It is the action potential's propagating away from the site of origin that constitutes impulse conduction in nerve, muscle and heart.

Action potentials are everywhere in life; if there's electricity, there are action potentials. This topic deserves an entire book, much less an encyclopedia entry. Accordingly, much of the information in this comes from articles in other books. Unless one defines electricity as a system, there seem to be no models that fit our definition of physiological systems models.

Having said that, electric phenomena are an integral part of every biological system, no matter what one's definition of the latter is. Action potentials keep our heart beating, our mind thinking—and make possible our seeing, hearing, smelling, feeling, tasting, digesting and moving. They are everywhere (plants and animals) and taking place all the time. We tried to estimate the number of action potentials occurring per second in one part of the human body (the brain) and may be underestimating by orders of magnitude. There are a trillion neurons in the human brain alone, and 10 quadrillion synapses, more than there are stars in the universe. The rate of action potentials can be from zero to >1000 action potentials per second. Let us estimate $10^{18}\ \mathrm{s}^{-1}$ ($10^{16} \times 10^{2}$). We have not even started with muscles or sensory receptors, for example. We repeat, it's a large topic.

Varghese (180) tells us that the key concept in the modeling of excitable cells is that of ion channel selectivity. A particular type of ion channel will only allow certain ionic species to pass through; most types of ion channels are modeled as being permeant to a single ionic species. In excitable cells at rest, the membrane is most permeable to K. This is because only K channels (i.e., channels selective to K) are open at the resting potential. For a given stimulus to result in an action potential, the cell has to be brought to threshold, that is, the stimulus has to be larger than some critical size. Smaller, subthreshold, stimuli will result in an exponential decay back to the resting potential. The upstroke, or fast initial depolarization of the action potential, is caused by a large influx of Na ions as Na channels open (in some cells, entry of Ca ions though Ca channels is responsible for the upstroke) in response to a stimulus. This is followed by repolarization as K ions start flowing out of the cell in response to the new potential gradient. While responses of most cells to subthreshold inputs are usually linear and passive, the suprathreshold response (the action potential) is a nonlinear phenomenon. Unlike linear circuits where the principle of superposition holds, the nonlinear processes in cell membranes do not allow responses of two stimuli to be added. If an initial stimulus results in an action potential, a subsequent stimulus administered at the peak voltage will not produce an even larger action potential; indeed, it may have no effect at all. Following an action potential, most cells have a refractory period, during which they are unable to respond to stimuli. Nonlinear features, such as these make modeling of excitable cells a nontrivial task. In addition, the molecular behavior of channels is only partially known and, therefore, it is not possible to construct membrane models from first principles.

Most membrane models discussed in Varghese's excellent article (180) involve the time behavior of electrochemical activity in excitable cells. These models are systems of ordinary differential equations where the independent variable is time. While a good understanding of linear circuit theory helps understand these models, most of the phenomena of interest involve nonlinear circuits with time-varying components. Electrical activity in plant and animal cells is caused by two main factors: (1) differences in the concentrations of ions inside and outside the cell; and (2) molecules embedded in the cell membrane that allow these ions to be transported across the membrane. The ion concentration differences and the presence of large membrane-impermeant anions inside the cell result in a polarity: the potential inside a cell is typically 50–100 mV lower than that in the external solution. Almost all of this potential difference occurs across the membrane itself; the bulk solutions both inside and outside the cell are usually at a uniform potential. This transmembrane potential difference is, in turn, sensed by molecules in the membrane, and these molecules control the flow of ions. The lipid bilayer, which constitutes the majority of the cell membrane, acts as a capacitor with a specific capacitance. Because the membrane is thin ($\sim$ 7.5 mm), it has a high capacitance, $\sim 1\ \mu\mathrm{F{\cdot}cm^{-2}}$. The rest of the membrane comprises large protein molecules that act as (1) ion channels, (2) ion pumps, or (3) ion exchangers. The flow of ions across the membrane causes changes in the transmembrane potential, which is typically the main observable quantity in experiments.

Barr (181), in another excellent article, expounds on bioelectricity, which has its origin in the voltage differences present between the inside and outside of cells. These potentials arise from the specialized properties of the cell membrane, which separates the intracellular from the extracellular volume. Much of the membrane surface is made of a phospholipid bilayer, an electrically inert material. Electrically active membranes also include many different kinds of *integral proteins*, which are compact, but complex structures extending across the membrane. Each integral protein comprises a large number of amino acids, often in a single long polypeptide chain, which folds into multiple domains. Each domain may span the membrane several times. The multiple crossings may build the functional structure, for example, a *channel*, through which electrically charged ions may flow. The structure of integral proteins is loosely analogous to that of a length of thread passing back and forth across a fabric to form a buttonhole. Just as a buttonhole allows movement of a button from one side of the fabric to the other, an integral protein may allow passage of ions from the exterior of a cell to the interior, or vice versa. In contrast to a buttonhole, an integral protein has active properties, for example, the ability to open or close. Cell membranes possess several active characteristics important to the cell's bioelectric behavior. (1) Some integral proteins function as *pumps*. These pumps use energy to transport ions across a membrane, working against a concentration gradient, a voltage gradient, or both. The most important pump moves Na ions out of the intracellular volume and K ions into that volume. (2) Other integral proteins function as channels, that is, openings through the membrane that open and close over time. These channels can function selectively so that, for a particular kind of channel, only Na ions may pass through, for example. Other kinds of channel may allow only K ions or Ca ions. (3) The activity of the membrane's integral proteins is modulated by signals specific to its particular function. For example, some channels open or close in response to photons or to odorants; thus they function as sensors for light or smell. Pumps respond to the concentrations of the ions they move. Rapid electrical impulse transmission in nerve and muscle is made possible by changes that respond to the transmembrane potential itself, forming a feedback mechanism. These active mechanisms provide ion-selective means of current's crossing the membrane, both against the concentration or voltage gradients (pumps) or in response to them (channels). While the pumps build up the concentration differences (and thereby the potential energy) that allow events to occur, channels use this energy actively to create the fast changes in voltage and small intense current loops that constitute nerve signal transmission, initiate muscle contraction, and participate in other essential bioelectric phenomena.

Action potentials, of course, are responsible for the external voltages that are measured all over the body: the electrocardiograph, the electroencephalograph, and the electromyography, for example.

There are two excellent sources for action-potential models, one a book article and the other a Web site. Varghese (180), in his article Membrane Models, (see above) reviewed a large number of models relating to action potentials (we counted 110 models). A listing will give an idea of the magnitude of the scope of this topic. In addition, it points out the fragmentation of the research in this area. Varghese's article nicely demonstrates the disparateness of the topic, as well as the lack of putting the models together into subsystems, much less systems.

Nerve Cells
Sensory Neurons
Efferent Neurons
Skeletal Muscle Cells
Endocrine Cells
Cardiac Cells
Epithelial Cells
Smooth Muscle
Plant Cells
Simplified Models

Under the heading 'Sensory Neurons', for example, are described.

Rabbit Sciatic Nerve Axons
Myelinated Auditory-Nerve Neuron
Retinal Ganglion Cells
Retinal Horizontal Cells,
Rat Nodose Neurons
Muscle Spindle Primary Endings
Vertebrate Retinal Cone Photoreceptors
Primary and Secondary Sensory Neurons of the Enteric Nervous System
Invertebrate Photoreceptor Neurons
Fly Optic Lobe Tangential Cells
Rat Mesencephalic Trigeminal Neurons
Primary Afferents and Related Efferents
Myelinated I Primary Afferent Neurons

Lloyd and the CellML site (176) have assembled an online repository of physiological system models, many of them related to action potentials. The models all conform with the CellML Specification (182). They are based on published mathematical models taken from peer-reviewed journals, from conference proceedings, and from textbook-defined metabolic pathways. The group has remained true to the original publications and has not assumed any reaction kinetics or initial values if they were not included in the original publication. All sources of information have been referenced in the model documentation.

These models represent several types of cellular processes, including models of electrophysiology, metabolism, signal transduction and mechanics. To facilitate the process of finding a particular model of interest, the models are grouped into broad subject categories.

The models on the site have been validated to a certain degree. Current validation processes include comparing the equations in the original paper with a PDF of the equations used in the CellML description. As tool development continues, both by the CellML team and by international collaborators, the groups expect to be able to carry a validation of the intent of the models by running simulations and comparing these results with those of the original publication. Presumably, validation of the model itself will ultimately take place.

With hundreds of models available on this site, it is impossible to go into detail. The topics include the following. Note that many of them deal with action potentials. GlucoSim (see above) is on this site, also: Signal Transduction Pathway Models, Metabolic Pathway Models, Cardiac Electrophysiological Models, Calcium Dynamics Models, Immunology Models, Cell Cycle Models, Simplified Electrophysiological Models, Other Cell Type Electrophysiological Models, Smooth and Skeletal Muscle Models, Mechanical Models and Constitutive Laws.

The following are examples of the formats in which a model may be obtained/downloaded.

1. The raw XML.
2. An HTML version for browsing online.
3. A PDF version suitable for printing.
4. A gzipped tarball with the XML and the documentation on the site.
5. A PDF of the equations described in the model generated directly from the CellML description using the MathML Renderer.

Thermal

Temperature control is one of the more impressive achievements in life. In higher forms of life, internal body temperature is maintained at a rather constant level, despite changes in the internal and external environment, to help provide conditions favorable to metabolism and other necessary processes within blood and tissue. Thus, in the human, an internal temperature close to 37 °C is normally maintained by a thermoregulatory feedback system despite large changes in ambient temperature and other environmental factors. It is also a daunting system to model, for many reasons.

Kuznetz (183) was one of the first to use more than 1D in thermal modeling. Previous models had failed to account for temperature distribution in any spatial direction other than radially outward from the body centerline. The models therefore could not account for nonuniform environmental conditions or nonuniform heat generation from muscles or organs within the body. However, these nonuniform conditions are commonplace and can lead to disparate skin temperatures and heat loss rates on different sides of the same body compartment. Kuznetz' mathematical model of human thermoregulation could predict transient temperature variations in two spatial dimensions, both radially and angularly, as measured from the body centerline. The model thereby could account for nonuniform environments and internal heat generation rates.

Downey and Seagrave (184) developed a model of the human body that integrates the variables involved in temperature regulation and blood gas transport within the CV and respiratory systems. It describes the competition between skin and muscles when both require increased blood flows during exercise and/or heat stress. After a detailed study of the control relations used to predict skin blood flow, four other control relations used in the model were tweaked. Dehydration and complete water replacement were studied during similar environmental and exercise situations. Control relations for skin blood flow and evaporative heat loss were modified, and water balance was added to study how the loss of water through sweat can be limiting. Runoff from sweating as a function of relative humidity was introduced, along with evaporation.

In two papers, Fiala et al. (185,186) developed a dynamic model predicting human thermal responses in different environmental temperatures. The first paper was concerned with the passive system: (1) modeling the human body, (2) modeling heat-transport mechanisms within the body and at its periphery, and (3) the numerical procedure. Emphasis was given to a detailed modeling of heat exchange with the environment: local variations of surface convection, directional radiation exchange, evaporation and moisture collection at the skin, and the nonuniformity of clothing ensembles. Other thermal effects were also modeled: the impact of activity level on work efficacy and the change of the effective radiant body area with posture. From that passive model, they developed an active mathematical model for predicting human thermal and regulatory responses in cold, cool, neutral, warm, and hot environments (186). The active system simulates the regulatory responses of shivering, sweating, and peripheral vasomotion of unacclimatized subjects. The rate of change of the mean skin temperature, weighted by the skin temperature error signal, was identified as governing the dynamics of thermoregulatory processes in the cold. Good general agreement with measured data was obtained for regulatory responses, internal temperatures, and the mean and local skin temperatures of unacclimatized humans for a wide spectrum of climatic conditions and for different activity levels.

Boregawda (187) used a novel approach based on the second law of thermodynamics to investigate the psychophysiology of and to quantify human stress level. Two types of stresses (thermal and mental) were examined. Using his own thermal and psychological stress indices, he implemented a human thermal model based on a finite element method. With this model, he examined thermal, mental and CV stress in the human body.

Gardner and Martin (188) developed a model of the human thermoregulatory system for normal subjects and burned patients. The human body was split into eleven segments (The Rule of Nines suggests that one can estimate surface areas in the cranial, abdominal, thoracic and extremity regions by multiplying the patient's body surface area by 9%, or a multiple thereof.), each having core, muscle, fat and skin layers. Heat transport through blood flow and conduction were simulated, and surface heat loss was separated into radiative, convective and evaporative components. The model was refined to fit the data through manipulation of heat flow commands and temperature set points controlling sweating and shivering. The model also described the responses of burn patients to skin layer destruction, increased body metabolism and fluid loss. The model shows that the ambient temperature at which sweating occurs increases with the area of burn injury. It has been used to predict optimum environmental temperatures for treatment of patients with burn wounds of varying extent, a critical need.

Aging of Physiological Systems

Geriatric medicine is becoming increasingly important, due to the aging of our population. Physiologists need to understand that physiological function continually changes with age, and that they need to take aging into account in their studies and in their models. Healthcare givers need methods that can teach efficiently and painlessly the complexities involved with aging. Modeling and simulation are ideal for this area. We implemented the anatomy and physiology of aging in BODY Simulation (see PBPM). In the BODY model, we changed >50 patient parameters. The categories changed included anatomical, cardiovascular, and respiratory, as well as hepatic and renal function. Four patients were created: normal, but elderly, patients, aged 65, 75, 85, and 95 years. To evaluate the new patients, we imposed three stresses in the elderly patients and in a young, healthy patient: anesthesia induction, hemorrhage, and apnea. We observed an age-related response to these stresses. In general, we saw a reduced physiological reserve. Again, this model is available as an interactive simulation, no knowledge of mathematics required.

USES FOR MODELS

Please see a list of uses, in the Introduction. One use for models is to substitute as an animal or person in experiments, with the goal of reducing the use of either type of subject in experiments. A control system was developed (189), using three models (190–192). When a control system had been developed with the models, we performed one animal experiment to test the control system. Noting any problems or discrepancies, we returned to the model, tuned up the control system, and repeated an experiment, reiterating between model and animal several times. We estimated that the decrease in animal experiments was 82.5%, from a projected 120–21. Furthermore, the dangerous experiments were performed on the models, and no animal was sacrificed. Noninvasive methods are less painful and dangerous in patients, and methods that render more accurate the data obtained from these methods are clinically and experimentally valuable. Kjaergaard et al. (193) used a model of pulmonary shunt and ventilationperfusion mismatch to help them quantify those two variables in patients, using noninvasive methods. The equations and diagrams can be found in the supplementary material to the paper, on the Web (click in the appropriate link, below the abstract).

Mesic et al. (194) used a mechanical lung simulator to simulate specific lung pathologies, to test lung-function equipment, and to instruct users about the equipment. A mathematical model of the respiratory system was interfaced with a model of physical equipment (the simulator, actuators, and the interface electronics) so that one can simulate the whole system. The control system, implemented on a personal computer, allows the user to set parameters.

Our noninvasive measurement theme continues, as we explore the work of Wesseling and his group. To the word noninvasive, we can add another important term, continuous, or continual. For areas where a patient's status can change in a matter of seconds (operating room, intensive care unit, and emergency room, e.g.), monitoring data should be available continuously or continually, and without delay. The work began with the Finapres, a device for noninvasively quantifying an arterial pressure waveform (195). They also developed a method for measuring continual cardiac output from an (invasive) aortic pressure waveform (196,197). Using a model, they could combine the two methods to achieve noninvasive, continual cardiac output from the radial artery (invasive) (198) or finger (noninvasive) (199). The pulse contour method has been extensively and carefully tested with aortic pressures, and it seemed prudent to use this pressure, if at all possible. To convert the less invasive radial pressure or the noninvasive finger pressure into an aortic pressure, Wesseling's group uses a three-element model of aortic input impedance that includes nonlinear aortic mechanical properties and a self-adapting systemic vascular resistance (200). Using the model enhances the accuracy of the less invasive methods (198,201). Another model increases the accuracy of the noninvasive technique even further (202). An inverse model of the averaged distortion models corrects for the pulse-wave distortion occurring between brachial and finger arteries.

Education should be one of the primary uses of physiological systems modeling. To adapt a complex whole-body model to a manikin-based simulator, however, requires considerable time and expense. As far as we can tell, two extant model- and manikin-based simulators exist. Another manikin-based simulator, which did not use a model, did not survive. Several disparate models have been published on the educational uses of one of the simulators (METI). The METI manikin-type simulator uses models based on the excellent models of Beneken, although how much, how detailed and to what extent is difficult to determine. The published models include some intracranial dynamics (203,204), pharmacokinetics and pharmacodynamics (205) and obstetric cardiovascular physiology (206). The PK-PD model is not a whole-body model, however, and the obstetric model omits the fetal circulation and hence is not realistic. BODY Simulation (see PBPM) uses a very detailed model for education, both clinical and in the basic sciences, and is a screen-based simulator. The best simulator from a purely educational point of view is Gas-Man, another screen-based simulator, developed by Jim Philip (207). As the name implies, the simulation deals only with inhaled anesthetic agents, but the manual and the teaching exercises are superb. The simulation is available for a Macintosh or a PC.

THE PHYSIOME PROJECT

What's next in physiological systems modeling? Hundreds of physiological models have been developed, some very small, some surprisingly large and complex. No one, however, is anywhere near describing human physiology in any reasonable completeness and detail. The physiome project may point the way. Essentially, the physiome project is the successor to the genome project, albeit many times more difficult. The following description of the project is adapted from Bassingthwaighte (208); the original was written in 2000. For MUCH more information, go to (209).

The physiome is the quantitative description of the functioning organism in normal and pathophysiological states. The human physiome can be regarded as the virtual human. Think genome, or proteome. The physiome is built upon the morphome, the quantitative description of anatomical structure, chemical and biochemical composition, and material properties of an intact organism, including its genome, proteome, cell, tissue, and organ structures up to those of the whole intact being. (We understand that the morphome, except for gross anatomy, is still in an early stage.) The physiome project is a multicenter integrated program to design, develop, implement, test and document, archive and disseminate quantitative information, and integrative models of the functional behavior of molecules, organelles, cells, tissues, organs, and intact organisms from bacteria to humans. A fundamental and major feature of the project is the databasing of experimental observations for retrieval and evaluation. Technologies that allow many groups towork together are rapidly being developed. Given a project that is so huge and complex, a particular working group can be expert in only a small part of the overall project. The strategies to be worked out must therefore include how to bring models composed of many submodules together, even when the expertise in each is scattered among diverse institutions, departments, talents and constituencies. Developing and implementing code for large-scale systems has many problems. Most of the submodules are complex, requiring consideration of spatial and temporal events and processes. Submodules have to be linked to one another in a way that preserves mass balance and gives an accurate representation of variables in nonlinear complex biochemical networks with many signaling and controlling pathways. Microcompartmentalization vitiates the use of simplifiedmodel structures. The stiffness of the systems of equations is computationally costly. Faster computation is neededwhen using models as thinking tools and for iterative data analysis. Perhaps the most serious problem is the current lack of definitive information on kinetics and dynamics of systems, due in part to the almost total lack of databased observations, but also because, although we are nearly drowning in new information being published each day, either the information required for the modeling cannot be found, has never been obtained or is totally irrelevant. Simple things like tissue composition, material properties, and mechanical behavior of cells and tissues are not generally available.

Currently, there are several subprojects, the magnitude of each of which boggles the mind. As the *Economist* put it, Computer organs are not for the technologically faint-hearted (210). The subprojects include the Cardiome Project (211), the Microcirculatory Project (212), the Pulmonary/Respiratory Project (213,214), The Kidney Project (215), and the Coagulation Project (210).

How rapidly and completely can all of this take place? This is an enormously ambitious project, many times more difficult than the genome project. The genome project succeeded relatively quickly for at least two reasons. First, to fill well-defined gaps in knowledge, a huge number of devices automatically churned out mountains of carefully and precisely specified data. Second, industry, sensing that there was gold in the genome, invested (and is still investing) enormous sums of money.

We have several questions and comments regarding the physiome project, questions that might apply to any very large project. The questions are restricted mainly to normal physiology. Disease and normal aging will add to the number and difficulty of the questions. These questions are intended to help sort out the potential problems, difficulties and considerations while the project is still at an early stage.

1. The questions in the physiome project are much more difficult to formulate than those of the Genome Project.
   (a) Should these questions be formulated formally?
   (b) If so, who will formulate them?
2. Can or should the physiome project control the data, including the content and structure, back to and including the planning of the experiment that generates the data?
3. Who is going to determine whether the whole-body model works? We worry that it may become like some software (no single person knows every detail about it) what it can do and what it cannot do.
4. How does one handle the flood of new information?
   (a) How often will a large, multicenter submodel be updated: and who decides?
   (b) Do new methods for handling the flood of information need to be developed?
   (c) If so, what?
   (d) What are the methods for validating models and for sensitivity analysis?
   (e) How will validation and sensitivity analysis proceed?
   (f) At what stages should validation and sensitivity analysis proceed, for example, every time models are combined? Every 6 months?
5. There will be an enormous library of physiological normals: equations, parameters, and so on.
   (a) How will one deal with variations from normal? Presumably this will be partly with sensitivity analysis.
   (b) How much variation from normal will be allowed? In other words, what are the limits on normal: upper and lower?
   (c) Does one anticipate large variations in normal, from parameter to parameter and from equation to equation?
   (d) Will clinicians be in on this and other decision processes?
6. How will one deal with the effects of aging and responses to the environment, including exercise, hypoxia, altitude, and temperature?
7. The genome project is relatively simple, and mistakes can be relatively easily corrected.
   (a) Is this the case with the physiome project?
   (b) How will one deal with errors in the physiome project?
   (c) If there is one error in one submodel, how will one find that error?
   (d) How will one determine the effect of errors in a submodel that is a component of the overall model on that larger model: by sensitivity analysis?
8. Can anyone input data, equations, formulae and models into the project database?
   (a) If so, who "edits" the deposits into databases, or are they done just ad hoc?
   (b) If not, is there a gatekeeper?
   (c) Asked another way, what is the quality control and who is in charge of it?
   (d) Who is going to determine what finally winds up in the library/database?
   (e) Who will be in charge of the information/data?
9. Several computer languages are currently mentioned. Will there ultimately be a single language?
10. Is there a freely available Web journal for the project? (The Virtual Journal of the Virtual Human)
    (a) Perhaps the VJVH could be the filter for the information and data, as well as a means for developing standards (see next question).
11. Standards, if done properly, should help in several areas. Some standards are being developed (216).

   - (a) We gather that part of the significant difference among models arises from their being published in different journals.
   - (b) Do common standards exist for the project?
   - (c) Who is responsible for implementing and coordinating generic and specific standards?

12. A very important set of questions involves intellectual philanthropy versus intellectual property.
    - (a) Are there any tools, including developmental tools, that are freely available, that is, public domain?
    - (b) The NLM has made the Visible Human data public domain. How much of the physiome database will be public domain and how much private domain?
    - (c) Would dividing the database into multiple public and private sections inhibit the integrity, the wholeness and the usefulness of the database?
    - (d) How will one audit, in one huge model, the amalgamated data from government employees, universities, research institutes and industry?
    - (e) How will one determine credit, cash, brownie points, or whatever for data, equations, models, and so on?
    - (f) The same questions hold for developmental tools.
13. How will ongoing projects fit into the project? How will future projects fit in?
14. What data are required, for each step?
    - (a) Who decides what data are required, or do data just appear in the database because they were there?

Excellent reviews on the Physiome Project include (216–219). In addition try an IUPS Web site (220).

EPILOGUE

Over one-half of a century ago, two papers were published in the same year, in the same country. These two papers point the way to the future in physiological systems modeling, given all the enormous amount of models and data pouring in. The authors of one paper went on to win the Nobel Prize. The author of the other committed suicide within two years. Neither knew of the other's existence, and one paper probably went unnoticed, and was certainly misunderstood.

If nothing else, our article has illustrated the need to integrate mathematical biology with experimental biology. To achieve this integration, experimentalists must learn the language of mathematics and dynamical modeling and theorists must learn the language of biology. Hodgkin and Huxley's quantitative model of the nerve action potential and Alan Turing's work on pattern formation in activator–inhibitor systems (221) represent those two languages. These classic studies illustrate two ends of the spectrum in mathematical biology: the detailed-model approach and the minimal-model approach. When combined, they are highly synergistic in analyzing the mechanisms underlying the behavior of complex biological systems. Their effective integration will be essential for unraveling the physical basis of the mysteries of life. For more detail on this important concept, see the fascinating account by Weiss et al. (222).

ACKNOWLEDGMENTS

The authors are grateful to the American Physiological Society and to the American Society for Biological Chemistry and Molecular Biology for making available PDF files of all issues of their journals. The 1924 Haggard articles were required reading for the senior author, during his residency. Classic and macro physiology, much of which is to be found in the older literature have been emphasized.

The senior author is also grateful to all the mentors in his modeling life. There are so many that they can only be listed and thanked, in more or less chronological order. Avram Goldstein, Aldo Corbascio, Bram Noordergraaf, Isaac Starr, Bob Fleming, Jan Beneken, Karel Wesseling, The late Vince Rideout, Yas Fukui, James Martin, Ken Starko. The senior author is an anesthesiologist, and the bias occasionally shows.

GLOSSARY

CVS	Cardiovascular system
CNS	Central nervous system
ANS	Autonomic nervous system
MAP	Mean arterial pressure
CVP	Central venous pressure
CO	Cardiac output
CBF	Cerebral blood flow
CSF	Cerebrospinal fluid
ICP	Intracranial pressure
LV	Left ventricle
LA	Left atrium
RV	Right ventricle
PV	Pressure volume
SVR	Systemic vascular resistance
avu	Arteriovenous
AV	Atrioventricular
$P\text{O}_2$	Partial pressure of oxygen
PKPD	Pharmacokinetic/pharmacodynamic
Physiological PKPD	physiologically based pharmacokinetic/pharmacodynamic (PBPKPD!)
PBPM	Physiologically based pharmacological models

Some systems are so familiar that an abbreviation suffices: CVS for cardiovascular system and CNS for central nervous system. A glossary of abbreviations follows for your reference.

BIBLIOGRAPHY

1. Elenkov IJ, et al. The sympathetic nerve—an integrative interface between two supersystems: The brain and the immune system. Pharmacol Rev 2000;52:595–638.
2. Bassingthwaighte JB. A view of the physiome. Available at http://physiome.org/files/Petrodvoret.1997/abstracts/jbb.html. 1997.
3. Weinstein AM. Mathematical models of renal fluid and electrolyte transport: Acknowledging our uncertainty. Am J Physiol Renal Physiol 2003;284(5):F871–F884.
4. Werner J, Böhringer D, Hexamer M. Simulation and prediction of cardiotherapeutical phenomena from a pulsatile model coupled to the Guyton circulatory model. IEEE Trans Biomed Eng 2002;49(5):430–439.
5. Rideout VC. Mathematical and computer modeling of physiological systems. 1st ed. Biophysics and bioengineering. Noordergraaf A, editor. Englewood Cliffs (NJ): Prentice-Hall; 1991. p 261.
6. Baan J, Noordergraaf A, Raines J. Cardiovascular system dynamics. Cambridge (MA): The MIT Press; 1978. p 618.
7. Papper EM, Kitz RJ. editors. Uptake and distribution of anesthetic agents. McGraw-Hill: New York; 1963. p 321.
8. Bird R, Stewart W, Lightfoot E. Transport phenomena. New York: Wiley; 1960.
9. Fukui Y. A study of the human cardiovascular-respiratory system using hybrid computer modeling. in Department of Engineering. University of Wisconsin: Madison (WI); 1971.
10. Guyton AC, Polizo D, Armstrong GG. Mean circulatory filling pressure measured immediately after cessation of heart pumping. Am J Physiol 1954;179:262–267.
11. Guyton A, Lindsey A, Kaufmann B. Effect of mean circulatory filling pressure and other peripheral circulatory factors on cardiac output. Am J Physiol 1955;180:463–468.
12. Manning RD Jr, et al. Essential role of mean circulatory filling pressure in salt-induced hypertension. Am J Physiol Regulatory Integrative Comp Physiol 1979;5:R40–R47.
13. Guyton A, et al. Systems analysis of arterial pressure regulation and hypertension. Ann Biomed Eng 1972;1(2):254–281.
14. Werner J, Böhringer D, Hexamer M. Simulation and prediction of cardiotherapeutical phenomena from a pulsatile model coupled to the Guyton circulatory model. IEEE Trans Biomed Eng 2002 May;49(5):430–439.
15. Hardman J, Wills J, Aitkenhead A. Investigating hypoxemia during apnea: Validation of a set of physiological models. Anesth Analg 2000;90:614–618.
16. Hardman J, Wills J, Aitkenhead A. Factors determining the onset and course of hypoxemia during apnea: An investigation using physiological modelling. Anesth Analg 2000;90:619–624.
17. Hardman J, Bedforth N. Estimating venous admixture using a physiological simulator. Br J Anaesth 1999;82:346–349.
18. Hardman J, et al. A physiology simulator: Validation of its respiratory components and its ability to predict the patient's response to changes in mechanical ventilation. Br J Anaesth 1998;81:327–332.
19. Snow J. On narcotism by the inhalation of vapours: Part iv. Lond Med Gaz 1848;7:330–334.
20. Snow J. On narcotism by the inhalation of vapours: Part xv. Lond Med Gaz 1850;11:749–754.
21. Frantz R. Ueber das verhalten des aethers im thierischen organismus (cited by Kunkle, a. J., Handbuch der toxikologie, Jena, 1899, I, 434). 1895: Wurzburg.
22. Nicloux M. Les anethesiques generaux au point de vue chemicophysiologique. Bull Acad Med 1908;lx:297.
23. Haggard HW. The absorption, distribution, and elimination of ethyl ether, iii. The relation of the concentration of ether, or any similar volatile substance, in the central nervous system to the concentration in the arterial blood, and, the buffer action of the body. J Biol Chem 1924;59(3):771–781.
24. Severinghaus J. Role of lung factors. In: Papper EM, Kitz RJ, editors. Uptake and distribution of anesthetic agents. 1963. pp. 59–71.
25. Eger EI,Ii. A mathematical model of uptake and distribution. In: Papper EM, Kitz RJ, editors. Uptake and distribution of anesthetic agents. New York: McGraw-Hill 1963. pp. 72–87.
26. Price H. A dynamic concept of the distribution of thiopental in the human body. Anesthesiology 1960;21(1):40–45.
27. Rackow H, et al. Simultaneous uptake of N_2O and cyclopropane in man as a test of compartment model. J Appl Physiol 1965;20:611–620.
28. Ashman M, Blesser W, Epstein R. A nonlinear model for the uptake and distribution of halothane in man. Anesthesiology 1970;33:419–429.
29. Zwart A, Smith NT, Beneken JEW. Multiple model approach to uptake and distribution of halothane. Use of an analog computer. Comp Biol Med Res 1972;55:228–238.
30. Smith NT, Zwart A, Beneken JEW. An analog computer multiple model of the uptake and distribution of halothane. Proc San Diego Biomed Symp 1972;11:235–241.
31. Smith NT, Zwart A, Beneken JEW. Effects of halothane-induced changes in skin, muscle, and renal blood flows on the uptake and distribution of halothane. Use of a multiple model. Proc 5th World Cong Anaesth 1972.
32. Fukui Y, Smith NT. A hybrid computer multiple model for the uptake and distribution of halothane. I. The basic model. Proc San Diego Biomed Symp 1974.
33. Fukui Y, Smith NT. A hybrid computer multiple model for the uptake and distribution of halothane. Ii. Spontaneous vs. Controlled ventilation, and the effects of CO_2. Proc San Diego Biomed Symp 1974.
34. Fukui Y, Smith NT. Interaction among ventilation, the circulation, and the uptake and distribution of halothane. Use of a hybrid computer model I. The basic model. Anesthesiology 1981;54:107–118.
35. Fukui Y, Smith NT. Interaction among ventilation, the circulation, and the uptake and distribution of halothane. Use of a hybrid computer model ii. Spontaneous vs. Controlled ventilation, and the effects of CO_2. Anesthesiology 1981;54:119–124.
36. Schwid HA, Wakeland C, Smith NT. A simulator for general anesthesia. Anesthesiology 1986;65:A475.
37. Smith NT, Starko K. Body simulation enhancements, including chemicalreaction simulation of cyanide therapy. Anaes Analg 2004;98(5s):S38.
38. Smith NT, Starko K. Physiologic and chemical simulation of cyanide and sarin toxicity and therapy. Stud Health Technol Inform 2005;111:492–497.
39. Smith NT, Starko K. The physiology and pharmacology of growing old, as shown in body simulation. Medicine Meets Virtual Reality. The Magical Next Becomes the Medical Now. Long Beach (CA): IOS Press; 2005.
40. Smith NT, Starko K. http://www.advsim.com/biomedical/body_manual/index.htm.
41. Oliver RE, Jones AF, Rowland MA. whole-body physiologically based pharmacokinetic model incorporating dispersion concepts: Short and long time characteristics. J Pharmacokin Pharmacodyna 2001;28(1):27–55.
42. Levitt DG, Quest PK. A general physiologically based pharmacokinetic model. Introduction and application to propranolol. BMC Clin Pharmacol 2002;2(5).
43. Levitt DG, Quest PK. Measurement of intestinal absorption and first pass metabolism—application to human ethanol pharmacokinetics. BMC Clin Pharmacol 2002;2(4).
44. Levitt DG, Quest PK. Volatile solutes—application to enflurane, nitrous oxide, halothane, methoxyflurane

and toluene pharmacokinetics. BMC Anesthesiology 2002;2 (5).

45. Levitt DG: http:/www.pkquest.com/.
46. Shafer S, Stanski DR. Stanford PK/PD software server. 2005.
47. Ursino M, Magosso E. Acute cardiovascular response to isocapnic hypoxia. I. A mathematical model. Am J Physiol Heart Circ Physiol 2000;279:H149–H165.
48. Groebe K. Precapillary servo control of blood pressure and postcapillary adjustment of flow to tissue metabolic status. A new paradigm for local perfusion regulation. Circulation 1996;94:1876–1885.
49. Krejci V, et al. Continuous measurements of microcirculatory blood flow in gastrointestinal organs during acute haemorrhage. Br J Anaesth 2000;84:468–475.
50. Hart BJ, et al. Right ventricular oxygen supply/demand balance in exercising dogs. Am J Physiol Heart Circ Physiol 2001;281:H823–H830.
51. Paneraiy RB. Assessment of cerebral pressure autoregulation in humans—a review of measurement methods. Physiol Meas 1998;19:305–338.
52. Lu K, et al. Cerebral autoregulation and gas exchange studied using a human cardiopulmonary model. Am J Physiol Heart Circ Physiol 2004;286:H584–H601.
53. Panerai RB, Dawson SL, Potter JF. Linear and nonlinear analysis of human dynamic cerebral autoregulation. Am J Physiol 1999;277:H1089–H1099.
54. Gao E, et al. Mathematical considerations for modeling cerebral blood flow autoregulation to systemic arterial pressure. Am J Physiol 1998;274:H1023–H1031.
55. Zhang R, et al. Transfer function analysis of dynamic cerebral autoregulation in humans. Am J Physiol 1998;274:H233–H241.
56. Hughson RL, et al. Critical analysis of cerebrovascular autoregulation during repeated head-up tilt. Stroke 2001;32:2403–2408.
57. Czosnyka M, et al. Contribution of mathematical modelling to the interpretation of bedside tests of cerebrovascular autoregulation. J Neurol Neurosurg Psychiat 1997;63:721–731.
58. Ursino M, Lodi CA. A simple mathematical model of the interaction between intracranial pressure and cerebral hemodynamics. J Appl Physiol 1997;82(4):1256–1269.
59. Hyder F, Shulman RG, Rothman DL. A model for the regulation of cerebral oxygen delivery. J Appl Physio 1998;85:554–564.
60. Kirkham SK, Craine RE, Birch AA. A new mathematical model of dynamic cerebral autoregulation based on a flow dependent feedback mechanism. Physiol Meas 2001;22:461–473.
61. Cornelissen AJM, et al. Myogenic reactivity and resistance distribution in the coronary arterial tree: A model study. Am J Physiol Heart Circ Physiol 2000;278:H1490–H1499.
62. Broten TP, Feigl E. Role of myocardial oxygen and carbon dioxide in coronary autoregulation. Am J Physiol 1992; 262:Hl231–H1237.
63. Cornelissen AJM, et al. Balance between myogenic, flow-dependent, and metabolic flow control in the coronary arterial tree: A model study. Am J Physiol Heart Circ Physiol 2002;282:H2224–H2237.
64. Vergroesen I, et al. Quantification of O_2 consumption and arterial pressure as determinants of coronary flow. Am J Physiol 1987; 252.
65. Guiota C, et al. Model-based assessment of pressure and flow-dependent coronary responses following abrupt pressure drops. Computers Biol Med 2000;30:111–126.
66. Jayaweera AR, et al. Role of capillaries in determining CBF reserve: New insights using myocardial contrast echocardiography. Am J Physiol 1999;277:H2363–H2372.
67. Geven MCF, et al. A physiologically representative in vitro model of the coronary circulation. Physiol Meas 2004;25:891–904.
68. Beyar R, Sideman S. Time-dependent coronary blood flow distribution in the left ventricular wall. Am J Physiol 1987; 252.
69. Pollack GH, Reddy RV, Noordergraaf A. Input impedance wave travel and reflections in the pulmonary arterial tree: Studies using an electrical analog. IEEE Trans Biomed Eng 1968;15:151–164.
70. Jager GN, Westerhof N, Noordergraaf A. Oscillatory flow impedance in electrical analog of arterial system. Circ Res 1965;16:121–133.
71. Beneken JEW, Beneken JEW. Some computer models in cardiovascular research. In: Bergel H, editor. Cardiovascular fluid dynamics. London: Academic Press; 1972. p 173–223.
72. Beneken JEW, DeWit B. A physical approach to hemodynamic aspects of the human cardiovascular system. In: Reeve E, Guyton A, editors. Physical. Basis of circulatory transport. Philadelphia: Saunders; 1967.
73. Rothe CF, Gersting JM. Cardiovascular interactions: An interactive tutorial and mathematical model. Adv Physiol Educ 2002;26:98–109.
74. Chung DC, et al. A dynamic model of ventricular interaction and pericardial influence. Am J Physiol 1997;272:H2942–H2962.
75. Lu K, et al. A human cardiopulmonary system model applied to the analysis of the Valsalva maneuver. Am J Physiol Heart Circ Physiol 2001;281:H2661–H2679.
76. Smith NT, Starko K. Anesthesia circuit. In: Atlee J, editor. Complications in anesthesiology. 1998.
77. Moore JA, et al. Accuracy of computational hemodynamics in complex arterial geometries reconstructed from magnetic resonance imaging. Ann Biomed Eng 1999;27(1):32–41.
78. Huang W, et al. Comparison of theory and experiment in pulsatile flow in cat lung. Ann Biomed Eng 1998;26:812–820.
79. De Gaetano A, Cremona G. Direct estimation of general pulmonary vascular models from occlusion experiments. Cardiovas Eng 2004;4(1).
80. Karamanoglu M, et al. Functional origin of reflected pressure waves in a multibranched model of the human arterial system. Am J Physiol 1994;267:H1681–H1688.
81. Karamanoglu M, Feneley MP. Late systolic pressure augmentation: Role of left ventricular out- flow patterns. Am J Physiol 1999;277:H481–H487.
82. Sun Y, et al. Mathematical model that characterizes transmitral and pulmonary venous flow velocity patterns. Am J Physiol 1995;268:H476–H489.
83. Yellin EL, et al. Mechanisms of mitral valve motion during diastole. Am J Physiol 1981;241:H389–H400.
84. Thomas JD, et al. Physical and physiological determinants of transmitral velocity: Numerical analysis. Am J Physiol 1991;260:H1718–H1730.
85. Salem JE, et al. Mechanistic model of myocardial energy metabolism under normal and ischemic conditions. Ann Biomed Eng 2002;30:202–216.
86. Olufsen M, et al. Numerical simulation and experimental validation of blood flow in arteries with structured-tree outflow conditions. Ann Biomed Eng 2000;28(28):1281–1299.
87. Wootton DM, Ku DN. Fluid mechanics of vascular systems, dis- eases, and thrombosis. Annu Rev Biomed Eng 1999;1:299–329.
88. Vasquez EC, et al. Neural reflex regulation of arterial pressure in pathophysiological conditions: Interplay among the baroreflex, the cardiopulmonary reflexes and the chemoreflex. Braz J Med Biol Res 1997;30:521–532.
89. Marshall JM. Peripheral chemoreceptors and cardiovascular regulation. Physiol Rev 1994;74:543–594.
90. Lanfranchi PA, Somers VK. Arterial baroreflex function and cardiovascular variability: Interactions and implications. Am J Physiol Regul Integr Comp Physiol 2002;283:R815–R826.
91. Wesseling KH, et al. Baromodulation as the cause of short term blood pressure variability? International Conference on

Applications of Physics to Medicine and Biology. Trieste, Italy: World Scientific Publishing Co; 1982.

92. Wesseling KH, Settels J. Baromodulation explains short-term blood pressure variability. In: Orlebeke J, Mulder G, Van Doornen L, editors. Psychophysiology of cardiovascular control: Models, methods and data. New York: Plenum Press; 1983. pp. 69–97.
93. Settels J, Wesseling KH. Explanation of short-term blood pressure responses needs baromodulation. Ann Int Conf IEEE Eng Med Biol Soc 1990;12(2):696–697.
94. Rose WC, Schwaber JS. Analysis of heart rate-based control of arterial blood pressure. Am J Physiol 1996;271:H812–H822.
95. Ursino M, Magosso E. Acute cardiovascular response to isocapnic hypoxia. I. A mathematical model. Am J Physiol Heart Circ Physiol 2000;279:H149–H165.
96. Ursino M, Magosso E. Acute cardiovascular response to isocapnic hypoxia. II. Model validation. Am J Physiol Heart Circ Physiol 2000;279:H166–H175.
97. Magosso E, Ursino M. A mathematical model of CO_2 effect on cardiovascular regulation. Am J Physiol Heart Circ Physiol 2001;281:H2036–H2052.
98. Melchior FM, Srinivasan RS, Charles JB. Mathematical modeling of human cardiovascular system for simulation of orthostatic response. Am J Physiol 1992;262:H1920–H1933.
99. Lerma C, et al. A mathematical analysis for the cardiovascular control adaptations in chronic renal failure. Art Org 2004;28(4):398–409.
100. Ursino M, Magosso E. A theoretical analysis of the carotid body chemoreceptor response to O_2 and CO_2 pressure changes. Resp. Physiol Neurobiol 2002;130:99–110.
101. Ursino M, Antonucci M, Belardinelli E. Role of active changes in venous capacity by the carotid baroreflex: Analysis with a mathematical model. Am J Physiol 1994;267:H253l–H2546.
102. Cohen MA, Taylor JA. Short-term cardiovascular oscillations in man: Measuring and modelling the physiologies. J. Physiol 2002;542(3):669–683.
103. Seydnejad SR, Kitney RI. Modeling of Mayer waves generation mechanisms determining the origin of the low- and very low frequency components of BPV and HRV. IEEE Eng Med Biol 2001;20:92–100.
104. Cavalcanti S, Belardinelli E. Modeling of cardiovascular variability using a differential delay equation. IEEE Trans Biomed Eng 1996;43(10):982–989.
105. Magosso E, Biavati V, Ursino M. Role of the baroreflex in cardiovascular instability: A modeling study. Cardiovas Eng 2001;1(2):101–115.
106. Aljuri N, Cohen RJ. Theoretical considerations in the dynamic closed-loop baroreflex and autoregulatory control of total peripheral resistance. Am J Physiol Heart Circ Physiol 2004;287(5):H2252–H2273.
107. Ben-Haim SA, et al. Periodicities of cardiac mechanics. Am J Physiol 1991;261:H424–H433.
108. Ursino M. Interaction between carotid baroregulation and the pulsating heart: A mathematical model. Am J Physiol 1998;275:H1733–H1747.
109. Hughson RL, et al. Searching for the vascular component of the arterial baroreflex. Cardiovasc Eng 2004;4:155–162.
110. O'Leary DD, et al. Spontaneous beat-by-beat fluctuations of total peripheral and cerebrovascular resistance in response to tilt. Am J Physiol Regul Integr Comp Physiol 2004;287(3):R670–R679.
111. Toska K, Eriksen M, Walloe L. Short-term control of cardiovascular function: Estimation of control parameters in healthy humans. Am J Physiol 1996;270:H651–H660.
112. Toska K, Eriksen M, Walloe L. Short-term cardiovascular responses to a step decrease in peripheral conductance in humans. Am J Physiol 1994;266:H199–H211.
113. Lee J-S. 1998 distinguished lecture: Biomechanics of the microcirculation, an integrative and therapeutic perspective. Ann Biomed Eng 2000;28:1–13.
114. Carr RT, Lacoin M. Nonlinear dynamics of microvascular blood flow. Ann Biomed Eng 2000;28:641–652.
115. Schmid-Schonbein GW. Biomechanics of microcirculatory blood perfusion. Annu Rev Biomed Eng 1999;1:73–102.
116. Pries AR, Secomb TW. Microcirculatory network structures and models. Ann Biomed Eng 2000;28:916–921.
117. Beard DA, Bassingthwaighte JB. Modeling advection and diffusion of oxygen in complex vascular networks. Ann Biomed Eng 2001;29:298–310.
118. Clark ME, Kufahl RH. Simulation of the cerebral macrocirculation. In: Baan J, editors. Cardiovascular system dynamics. Cambridge (MA): The MIT Press; 1978. pp. 380–390.
119. Lakin WD, et al. A whole-body mathematical model for intracranial pressure dynamics. J Math Biol 2003;46:347–383.
120. Olufsen M, Tran H, Ottesen J. Modeling cerebral blood flow control during posture change from sitting to standing. Cardiov Eng: An Int J 2004;4(1).
121. Sato J, et al. Differences in the dynamic cerebrovascular response between stepwise up tilt and down tilt in humans. Am J Physiol Heart Circ Physiol 2001;281:H774–H783.
122. Zhu L. Theoretical evaluation of contributions of heat conduction and countercurrent heat exchange in selective brain cooling in humans. Ann Biomed Eng 2000;28:269–277.
123. Ursino M, Di Giammarco P, Belardinelli E. A mathematical model of cerebral blood flow chemical regulation—part I: Diffusion processes. IEEE Trans Biomed Eng 1989;36(2): 183—191.
124. Ursino M, DiGiammarco P, Belardinelli E. A mathematical model of cerebral blood flow chemical regulation—part I: Diffusion processes. IEEE Trans Biomed Eng 1989;36(2): 1922–2201.
125. Ursino M, Magosso E. Role of tissue hypoxia in cerebrovascular regulation: A mathematical modeling study. Ann Biomed Eng 2001;29:563–574.
126. Wakeland W, et al. Modeling intracranial fluid flows and volumes during traumatic brain injury to better understand pressure dynamics. IEEE Eng Med Biol 2003;23:402–405.
127. Lodi CA, Ursino M. Hemodynamic effect of cerebral vasospasm in humans: A modeling study. Ann Biomed Eng 1999;27:257–273.
128. Pasley RL, Leffler CW, Daley ML. Modeling modulation of intracranial pressure by variation of cerebral venous resistance induced by ventilation. Ann Biomed Eng 2003;31:1238–1245.
129. Ursino M, Lodi CA. A simple mathematical model of the interaction between intracranial pressure and cerebral hemodynamics. J Appl Physiol 1997;82(4):1256–1269.
130. Ursino M. A mathematical study of human intracranial hydrodynamics. Part 1–the cerebrospinal fluid pulse pressure. Ann Biomed Eng 1988;16(4):379–401.
131. Ursino M. A mathematical study of human intracranial hydrodynamics. Part 2–simulation of clinical tests. Ann Biomed Eng 1988;16(4):403–416.
132. Ursino M, Lodi CA. Interaction among autoregulation, CO_2 reactivity, and intracranial pressure: A mathematical model. Am J Physiol 1988;274:H1715–H1728.
133. Loewe S, et al. Modeling cerebral autoregulation and CO_2 reactivity in patients with severe head injury. Am J Physiol 1998;274:H1729–H1741.
134. Ursino M, et al. Cerebral hemodynamics during arterial and CO_2 pressure changes: In vivo prediction by a mathematical model. Am J Physiol Heart Circ Physiol 2000;279:H2439–H2455.
135. Ursino M, Iezzi M, Stochetti N. Intracranial pressure dynamics in patients with acute brain damage: A critical

analysis with the aid of a mathematical model. IEEE Trans Biomed Eng 1995;42(6):529–540.

136. Sharan M, et al. An analysis of hypoxia in sheep brain using a mathematical model. Ann Biomed Eng 1998;26:48–59.
137. Cammarota JP Jr , Onaral B. State transitions in physiologic systems: A complexity model for loss of consciousness. IEEE Trans Biomed Eng 1998;45(8): 1017–1023.
138. Smith NP. A computational study of the interaction between coronary blood flow and myocardial mechanics. Physiol Meas 2004;25:863–877.
139. Smith NP, Pullan AJ, Hunter PJ. Generation of an anatomically based geometric coronary model. Ann Biomed Eng 2000;28:14–25.
140. Grotberg JB. Respiratory fluid mechanics and transport processes. Annu Rev Biomed Eng 2001;3:421–457.
141. Defares JG, Derksen HE, Duyff JW. Cerebral blood flow in the regulation of respiration. Acta Physiol Pharmacol Neerl 1960;9:327–360.
142. Defares JG, Principles of feedback control and their application to the respiratory control system. Handbook of physiology. Respiration. Washington (DC): American Physiology Society. 1964. pp. 649–680.
143. Grodins FS, Buell J, Bart AJ. Mathematical analysis and digital simulation of the respiratory control system. J Appl Physiol 1967;22(2):260–276.
144. Chiari L, Avanzolini G, Ursino M. A comprehensive simulator of the human respiratory system: Validation with experimental and simulated data. Ann Biomed Eng 1997; 25:985–999.
145. Hyuan U, Suki B, Lutchen KR. Sensitivity analysis for evaluating nonlinear models of lung mechanics. Ann Biomed Eng 1998;26:230–241.
146. Avanzolini G, et al. Role of the mechanical properties of tracheobronchial airways in determining the respiratory resistance time course. Ann Biomed Eng 2001;29:575–586.
147. Ursino M, Magosso E, Avanzolini G. An integrated model of the human ventilatory control system: The response to hypercapnia. Clin Physiol 2001;21(4):447–464.
148. Grodins FS, Yamashiro SM. Respiratory function of the lung and its control. New York: Macmillan; 1978.
149. Batzel JJ. Modeling and stability analysis of the human respiratory control system, in Department of Mathematics. Raleigh (NC): North Carolina State University; 1998. p 195.
150. Khoo MC, Gottschalk A, Pack AI. Sleep-induced periodic breathing and apnea: A theoretical study. J Appl Physiol 1991;70(5):2014–2024.
151. Kamm RD. Airway wall mechanics. Annu Rev Biomed Eng 1999;1:47–72.
152. Tehrani FT. Mathematical analysis and computer simulation of the respiratory system in the newborn infant. IEEE Trans Biomed Eng 1993;40:475–481.
153. Joyce CJ, Williams AB. Kinetics of absorption atelectasis during anesthesia: A mathematical model. J Appl Physiol 1999;86:1116–1125.
154. Dempsey JA, Forster HV. Mediation of ventilatory adaptations. Physiol Rev 1982;62(1):262–346.
155. Schafer JA. Interaction of modeling and experimental approaches to understanding renal salt and water balance. Ann Biomed Eng 2000;28:1002–1009.
156. Russell JM. Sodium-potassium-chloride cotransport. Physiol Rev 2000;80(1):211–276.
157. Dibona GF. Neural control of the kidney: Functionally specific renal sympathetic nerve fibers. Am J Physiol Reg Integ Comp Physiol 2000;279:R1517–R1524.
158. Kellen MR, Bassingthwaighte JB. An integrative model of coupled water and solute exchange in the heart. Am J Physiol 2003;285:H1303–H1316.
159. Makroglou A, Li J, Kuang Y. Mathematical models and software tools for the glucose-insulin regulatory system and diabetes: An overview. 2005.
160. Tibell A, Binder C, Madsbad S. Insulin secretion rates estimated by two mathematical methods in pancreas-kidney transplant recipients. Am J Physiol 1998;274:E716–E725.
161. Ibbini MS, Masadeh MA, Bani MM. Amer, A semiclosed-loop optimal control system for blood glucose level in diabetics. J Med Eng Technol 2004;28(5):189–196.
162. Parker RS, Doyle I, Francis J, Peppas NA. A model-based algorithm for blood glucose control in type 1 diabetic patients. IEEE Trans Biomed Eng 1999;46(2).
163. Bequette B. A critical assessment of algorithms and challenges in the development of a closed-loop artificial pancreas. Diabetes Technol Ther 2005; Feb; 7(1):28–47, 2005;7(1):28–47.
164. Berman N, et al. A mathematical model of oscillatory insulin secretion. Am J Physiol 1993;264:R839–R851.
165. Straub SG, Sharp GWG. Hypothesis: One rate-limiting step controls the magnitude of both phases of glucose-stimulated insulin secretion. Am J Physiol Cell Physiol 2004;287:C565–C571.
166. Lenbury Y, Ruktamatakul S, Amornsamarnkul S. Modeling insulin kinetics: Responses to a single oral glucose administration or ambulatory-fed conditions. BioSystems 2001; 59:15–25.
167. Chay TR, Keizer J. Minimal model for membrane oscillations in the pancreatic beta-cell. Biophys J 1983;42:181–190.
168. Sherman A, Rinzel J, Keizer J. Emergence of organized bursting in clusters of pancreatic beta-cells by channel sharing. Biophys J 1988;54:411–425.
169. Keizer J. Electrical activity and insulin release in pancreatic beta cells. Math Biosci 1988;90:127–138.
170. Keizer J, Magnus G. ATP-sensitive potassium channel and bursting in the pancreatic beta cell. A theoretical study Biophys J 1989;89:229–242.
171. Sherman A, Keizer J, Rinzel J. Domain model for Ca^{2+}-inactivation of Ca^{2+} channels at low channel density. Biophys J 1990;56:985–995.
172. Keizer J, Young GWD. Effect of voltage-gated plasma membrane Ca^{2+} fluxes on ip3-linked Ca^{2+} oscillations. Cell Calcium 1993;14:397–410.
173. Sherman A. Contributions of modeling to understanding stimulus-secretion coupling in pancreatic beta-cells. Am J Physiol 1996;271:E362–E372.
174. Tolic I, Mosekilde E, Sturis J. Modeling the insulin-glucose feedback system: The significance of pulsatile insulin secretion. J Theor Biol 2000;207:361–375.
175. Tolic I, Mosekilde E, Sturis J. Modelling the insulin-glucose feedback system. http://dev.cellml.org/examples/repository/.
176. Nickerson D, Lloyd CM. http://www.cellml.org/examples/repository/.
177. Shannahoff-Khalsa DS, et al. Low-frequency ultradian insulin rhythms are coupled to cardiovascular, autonomic, and neuroendocrine rhythms. Am J Physiol 1997;272:R962–R968.
178. Erzen FC, et al. GlucoSim: A web-based educational simulation package for glucose-insulin levels in the human body. Available at http://216.47.139.198/glucosim/index.html. 2005.
179. Hodgkin AL, Huxley AF. A quantitative description of membrane current and its application to conduction and excitation in nerve. J Physiol (London) 1952;117:500–544.
180. Varghese A. Membrane models, in The biomedical engineering handbook. In: Bronzino JD, editor. Boca Raton (FL): CRC Press LLC; 2000.
181. Barr RC. Basic electrophysiology. 2nd ed. In: Bronzino JD, editor. The biomedical engineering handbook. Boca Raton (FL): CRC Press LLC; 2000. p 18.

182. Cuellar A, et al. CellML specification. Available at http://www.cellml.org/public/specification/index.html. 2005.
183. Kuznetz LH. A two-dimensional transient mathematical model of human thermoregulation. Am J Physiol 1979;237(6):266–277.
184. Downey D, Seagrave RC. Mathematical modeling of the human body during water replacement and dehydration: Body water changes. Ann Biomed Eng 2000;28:278–290.
185. Fiala D, Lomas KJ, Stohrer M. A computer model of human thermoregulation for a wide range of environmental conditions: The passive system. J Appl Physiol 1999;87:1957–1972.
186. Fiala D, Lomas KJ, Stohrer M. Computer prediction of human thermoregulatory and temperature responses to a wide range of environmental conditions. Int J Biometeorol 2001;45:143–159.
187. Boregowda SC. Thermodynamic modeling and analysis of human stress responses Mechanical Engineering. Old Dominion University; 1998. p 175.
188. Gardner GG, Martin CJ. The mathematical modelling of thermal responses of normal subjects and burned patients. Physiol Meas 1994;15:381–400.
189. Quinn ML, et al. The case for designer animals: (use of simulation to reduce animal studies). Anesthesiology 1987;67(3):A215.
190. Martin JF, et al. A new cardiovascular model for real-time applications. Trans-Soc Computer Simulation 1986;3:31–65.
191. Slate JB. Model-based design of a controller for infusing sodium nitroprusside during postsurgical hypertension. Electrical Engineering. Madison (WI): University of Wisconsin; 1980.
192. Wesseling KH, A baroreflex paradox solution. Utrecht: TNO; 1982. pp. 152–164.
193. Kjaergaard S, et al. Non-invasive estimation of shunt and ventilation-perfusion mismatch. Intensive Care Med 2003;29:727–734.
194. Mesic S, et al. Computer-controlled mechanical simulation of the artificially ventilated human respiratory system. IEEE Trans Biomed Eng 2003;50(6):731–743.
195. Smith NT, Wesseling KH, De Wit B. Evaluation of two prototype devices producing noninvasive, pulsatile calibrated blood pressure from a finger. J Clin Monit 1985;1(1):17–29.
196. Wesseling KH, et al. A simple device for the continuous measurement of cardiac output. Adv Cardiovasc Phys 1983;5:16–52.
197. Wesseling KH, et al. Continuous monitoring of cardiac output. Medicamundi 1976;21:78–90.
198. Jansen J, et al. Continuous cardiac output computed from arterial pressure in the operating room. Br J Anaesth 2001;87(2):212–222.
199. Hirschl MM, et al. Noninvasive assessment of cardiac output in critically ill patients by analysis of the finger blood pressure waveform. Crit Care Med 1997;25(11).
200. Wesseling KH, et al. Computation of aortic flow from pressure in humans using a nonlinear, three-element model. J Appl Physiol 1993;74(5):2566–2573.
201. Jansen J, et al. A comparison of cardiac output derived from the arterial pressure wave against thermodilution in cardiac surgery patients. Br J Anaesth 2001;87(3):212–222.
202. Gizdulich P, Prentza A, Wesseling KH. Models of brachial to finger pulse wave distortion and pressure decrement. Cardiovasc Res 1997;33:698–705.
203. Thoman W, et al. A computer model of intracranial dynamics integrated to a full-scale patient simulator. Computers Biomed Res 1998;31:32–46.
204. Thoman W, et al. Autoregulation in a simulator-based educational model of intracranial physiology. J Clin Monit 1999;15:481–491.
205. Van Meurs W, Nikkelen E, Good M. Pharmacokinetic-pharmacodynamic model for educational simulations. IEEE Trans Biomed Eng 1998;45(5):582–590.
206. Euliano T, et al. Modeling obstetric cardiovascular physiology on a full-scale patient simulator. J Clin Monit 1997;13: 293–297.
207. Garfield J, Paskin S, Philip J. An evaluation of the effectiveness of a computer simulation of anaesthetic uptake and distribution as a teaching tool. Medi Educ 1989;23:457–462.
208. Bassingthwaighte JB. Strategies for the physiome project. Ann Biomed Eng 2000;28:1043–1058.
209. Anon., The physiome project. http://nsr.bioeng.washington/PLN. 2005.
210. Anon., The heart of the matter, in The Economist. Available at http://www.economist.com/science/tq/PrinterFriendly.cfm?Story_ID=885127. 2005. p 6.
211. Sideman S. Preface: Cardiac engineering—deciphering the cardiome, in Cardiac engineering: From genes and cells to structure and function. In: Sideman S, Beyar R, editors. New York: New York Academy of Sciences; 2004.
212. Popel AS, et al. The microcirculation physiome project. Ann Biomed Eng 1998;26:911–913.
213. Tawhai MH, Burrowes KS. Developing integrative computational models of pulmonary structure. The Anatom Rec (P B: New A) 2003;275B:207–218.
214. Tawhai MH, Ben-Tal A. Multiscale modeling for the lung physiome. Cardiovas Eng 2004;4(1).
215. Lonie A. The kidney simulation project. 2005.
216. Crampin EJ, et al. Computational physiology and the physiome project. Exp Physiol 2004;89(1):1–26.
217. Hunter P, Robbins P, Noble D. The IUPS human physiome project. Pflugers Arch - Eur J Physiol 2002;445:1–9.
218. Crampin EJ, et al. Computational physiology and the physiome project. Exp Physiol 2003;89(1):1–26.
219. Hunter PJ, Borg TK. Integration from proteins to organs-the physiome project. Nat Rev: Mol Cell Biol 2003;4:237–243.
220. Anon., The IUPS physiome project. Available at http://www.physiome.org.nz/anatml/pages/specification.html. 2005.
221. Turing A. The chemical basis of morphogenesis. Philos Trans R Soc (London), Ser A 1952;237:37–72.
222. Weiss JN, Qu Z, Garfinkel X. Understanding biological complexity: Lessons from the past. FASEB J 2003;17.

See also ELECTROCARDIOGRAPHY, COMPUTERS IN; ELECTROPHYSIOLOGY; PULMONARY PHYSIOLOGY; RADIOTHERAPY TREATMENT PLANNING, OPTIMIZATION OF.

PICTURE ARCHIVING AND COMMUNICATION SYSTEMS

KATHERINE ANDRIOLE P.
Harvard Medical School
Boston, Massachusetts

INTRODUCTION AND HISTORICAL OVERVIEW

A picture archiving and communication system (PACS) is a medical image management system, or a collection of electronic technologies that enable the digital or filmless imaging department. In a PACS, images are acquired, stored, and transmitted or distributed digitally, as well as interpreted digitally using what is known as soft copy display.

In the analog world, even images that were created by inherently digital modalities are printed to film for display on a light box or alternator, and for image archival as the legal record where films are stored in large rooms of filed film jackets. Imaging examinations on film must be transported from one location to another by foot for viewing by radiologists and referring clinicians. Films are retrieved and re-archived manually by film library personnel. Using a PACS, images are acquired as digital computer files, stored on computer disks or other digital media, transmitted across computer networks, and viewed and manipulated on computer workstations.

The benefits of PACS are numerous and include rapid and remote data distribution within and between health-care enterprises. Digital archival is more permanent than film with regard to media degradation as well as the problem of lost films. A PACS gives multiple users in distinct locations simultaneous access to the same imaging examinations. Also, the digital nature of the data allows for image manipulation and processing, which may lead to enhanced visualization of radiological features and improved interpretation of imaging studies. The potential impact of using a PACS can be more expedient care, more efficient workflow, and more cost-effective and higher quality care.

PACS and filmless radiology are a better way to practice imaging in medicine today. The number of images per study has grown beyond what is feasible for viewing on film. Still, today, only about 20% of health-care enterprises have implemented PACS.

Picture archiving and communication systems have come about via a convergence of technological and economic factors. The facilitating technologies responsible include a dramatic improvement in computing power, the advancement of network capabilities and storage devices, the development of imaging standards, and systems integration. In the 1990s, PACS applications challenged computer hardware. Today, PACS applications are only a subset of what computers can do.

Early PACS

In 1979, the earliest paper proposing the concept of a PACS was published by Heinz U. Lemke, Ph.D., entitled “Applications of Picture Processing, Image Analysis and Computer Graphics Techniques to Cranial CT Scans” (1). In the early 1970s, M. Paul Capp, M.D., Sol Nudelman, Ph.D., and colleagues at the University of Arizona Health Sciences Center organized a digital imaging group that developed the first digital subtraction angiography (DSA) device that was the precursor to clinical digital imaging. They introduced the notion of a “photoelectronic radiology department” and depicted a system block diagram of the demonstration facility they had built (2).

Samuel J. Dwyer, III, Ph.D. predicted the cost of managing digital diagnostic images in a radiology department (3) and, along with Andre Duerinckx, M.D., Ph.D., organized a landmark conference at which the acronym PACS was coined (4). This meeting, sponsored by the International Society for Photo-Optical Engineering (SPIE), titled “The First International Conference and Workshop on Picture Archiving and Communications Systems (PACS) for Medical Applications,” was held in Newport Beach, CA, January 18–21, 1982, and continues today as the Annual SPIE International Symposium on Medical Imaging. Two panel discussions “Equipment Manufacturers’ View on PACS” and “The Medical Community’s View on PACS” that took place at the first PACS conference were captured in the proceedings (4). Talk occurred of linking imaging modalities into a single digital imaging network and the recognition that, in order for this linking to be practical, standards would be required. Steven C. Horii, M.D., participated in those beginning discussions and has been instrumental in bringing about the creation and implementation of a standard for digital medical imaging, now known as the Digital Imaging and Communications in Medicine or DICOM Standard.

A number of PACS pioneers have contributed to the advancement of digital medical imaging to its current status through efforts in research and development, design, implementation, testing, analysis, standards creation, and education of the technical and medical communities. In 1982–1983, Dwyer oversaw the building of what is often considered the first PACS. In 1983, the first of numerous papers was presented by H. K. Bernie Huang, D.Sc., FRCR, detailing the PACS efforts at the University of California at Los Angeles, which culminated years later in a clinically operational filmless radiology department (5). G. James Blaine, D.SC., and R. Gilbert Jost, M.D., at the Washington University School of Medicine, focused their efforts on the development of utilities enabling PACS research and development (6). In the mid- to late-1980s, several researchers described their prototype PACS hardware and software efforts (7–10).

Similar activities were taking place in Europe and Asia. Hruby opened a completely digital radiology department in the Danube Hospital in Vienna in 1990, setting the tone for the future (11). Several academic radiology departments in the United States began working with major vendor partners to further the technology and its clinical implementation. Such academic–industry collaborations continue the advancement of PACS today.

Standards

The development of standards in medical imaging is one of the facilitating factors that has enabled PACS to mature and become more widely used. DICOM (Digital Imaging and Communications in Medicine) (12) along with several other integration standards, has been one of the most important accomplishments for PACS. DICOM is a standard that was created to promote an open architecture for imaging systems, allowing interoperability between systems for the transfer of medical images and associated information. As an exchange protocol, it was designed to bridge differing hardware devices and software applications.

With the increasing use of computers in clinical applications, and with the introduction of digital subtraction angiography and computed tomography (CT) in the 1970s, followed by other digital diagnostic imaging modalities, the American College of Radiology (ACR) and the

National Electrical Manufacturers Association (NEMA) recognized the emerging need for a standard method for transferring images and associated information between devices manufactured by the various vendors (12). These devices produced a variety of digital image formats. The push by the radiological community for a standard format across imaging devices of different models and makes began in 1982. ACR and NEMA formed a joint committee in 1983 to develop a standard to promote communication of digital image information, regardless of device manufacturer. It was felt that this committee would facilitate the development and expansion of picture archiving and communication systems that could also interface with other hospital information systems and allow the creation of diagnostic information databases that could be interrogated by a wide variety of geographically distributed devices.

The ACR-NEMA Standard version 1.0 was published in 1985. Two revisions followed, one in October 1986 and the second in January 1988 as version 2.0. It included version 1.0, the published revisions, and additional revisions. ACR-NEMA 2.0 consisted of a file header followed by the image data. The file header contained information relevant to the image, such as matrix size or number of rows and columns, pixel size, gray-scale bit depth, and so on, as well as information about the imaging device and technique (i.e., Brand X CT scanner, acquired with contrast). Patient demographic data, such as name and date of birth, were also included in the image header. The ACR-NEMA 2.0 standard specified exactly where in the header each bit of information was to be stored, such that the standard required image information could be read by any device, simply by going to the designated location in the header. Version 2.0 also included new material to provide command support for display devices, to introduce a new hierarchy scheme to identify an image, and to add data elements for increased specificity when describing an image. These standards publications specified a hardware interface, a minimum set of software commands, and a consistent set of data formats. Data included patient demographic information as well as imaging parameters. This standard unified the format of imaging data but functioned only as a point-to-point procedure, not including a networking communications protocol until later versions.

In 1994, at the Radiological Society of North America (RSNA) Meeting, a variety of imaging vendors participated in an impressive demonstration of the new and evolving imaging standard (ACR-NEMA 3.0). Participants attached their devices to a common network and transmitted their images to one another. In addition to the standard image format of ACR-NEMA 2.0, the DICOM standard included a network communications protocol or a common language for sending and receiving images and relevant data over a network.

Today, this standard, which is currently designated Digital Imaging and Communications in Medicine (DICOM), embodies a number of major enhancements to previous versions of the ACR-NEMA Standard, the first that is applicable to a networked environment. The original ACR-NEMA Standard included a well-defined format for the image data but worked only in point-to-point configurations. In order to function in a networked environment, a Network Interface Unit (NIU) was required. Operation in a networked environment is supported today using the industry standard networking protocol TCP-IP (transfer communication protocol - Internet protocol). Thus, in addition to the format of the data being exchanged between medical imaging devices, the DICOM Standard also specifies how the devices themselves should communicate using simple commands such as Find, Get, Move, Send, and Store. These commands operate on objects such as images and text, which are formatted in terms of groups and elements. The hierarchy of data is of the form patient, study, series or sequence, and image.

DICOM specifies, through the concept of service classes, the semantics of commands and associated data, and it also specifies levels of conformance. The DICOM standard language structure is built on information objects (IO), application entities (AE), and service class users (SCU) and providers (SCP). Information objects include, for example, the image types, such as CT, MRI, and CR. The application entities include the devices, such as a scanner, workstation, or printer. The service classes (SC*) define an operation on the information object via service object pairs (SOP) of IO and SCU and SCP. The types of operations performed by an SCU-SCP on an IO include storage, query-retrieve, verification, print, study content notification, study content notification and patient, and study and results management. An example DICOM-formatted message is written in terms of a tag (consisting of a group and an element) followed by the length of the tag, followed by the value: 0008,0020-8-20050402 represents group 8, element 20, which corresponds to the study date given as an 8 character field. DICOM is a continuously evolving standard with significant updates yearly.

The information technologies most familiar to radiology departments are PACS and Radiology Information Systems (RIS). PACS or image management systems typically perform the functions of image acquisition, distribution, display, and archival. Often, separate systems exist for primary interpretation in the radiology department and for use enterprise-wide by nonradiologist clinicians. The RIS typically maintains radiology-specific data, such as imaging examination orders, reports, and billing information.

Although implementation of either one or both of these systems can improve workflow and reduce operating costs, the elimination of film and paper from the practice of radiology is not easily realized without integrating the functions performed by several other information technologies. These systems include hospital information systems (HIS), Computerized Physician Order-Entry (CPOE) Systems, Report Generation Systems, Decision Support Systems, and Case-Based Teaching Files. Together, these systems make up the electronic medical record (EMR).

Several of these systems are more widely known to the health-care enterprise outside of diagnostic imaging departments. They, none-the-less, contain data essential to high quality low cost radiological practice. The value these systems can bring to medical imaging in the clinical enterprise is high, but they must be seamlessly integrated. Including several features, such as single sign-on, electronic master patient index, and context sensitivity, can help make these information systems tools and technologies most useful to radiology.

An effort known as Integrating the Healthcare Enterprise (IHE) provides a framework using existing standards, such as DICOM and Health Level 7 (HL7), to facilitate intercommunications among these disparate systems and to optimize information efficiency. The IHE is a joint effort of the Radiological Society of North America (RSNA) and the Healthcare Information and Management Systems Society (HIMSS) begun in 1998 to more clearly define how existing standards should be used to resolve common information system communication tasks in radiology.

The IHE technical framework defines a common information model and a common vocabulary for systems to use in communicating medical information. It specifies exactly how the DICOM and HL7 standards are to be used by the various information systems to perform a set of well-defined transactions that accomplish a particular task. The original seven tasks facilitated by the IHE include scheduled workflow, patient information reconciliation, consistent presentation of images, presentation of grouped procedures, access to radiology information, key image notes, and simple image and numeric reports. Profiles continue to be added yearly to the framework, enhancing its value in integrating information systems in a healthcare environment.

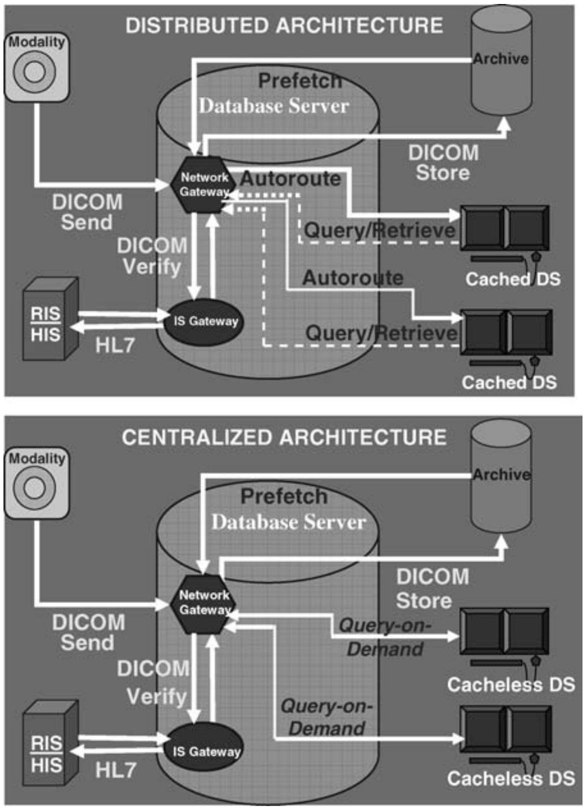


Figure 1. **(a)** Distributed versus **(b)** centralized PACS architectures.

Architectures

Two basic architectures are used in PACS today, distributed or cached and centralized or cacheless depicted in Figs. 1a and 1b, respectively. Data is acquired into the PACS in the same manner for both architectures, from the imaging modality via a DICOM sent to a network gateway. Demographic data is verified by interfacing to the radiology or hospital information system (RIS-HIS) through an IS gateway. Studies are permanently archived by a DICOM store to an electronic archive device.

In a distributed system, images and other relevant data are automatically routed to the workstation(s) where the studies are expected to be viewed and cached or stored on the local display station disk. The best-distributed PACS also prefetch relevant prior examinations from the long-term archive and automatically route them to the pertinent display for immediate access to comparison images. Studies not automatically routed to a workstation can be queried for and retrieved on request.

In a centralized system, images remain on a large central server and are only sent to the workstation when they are called up for display. In this query-on-demand architecture, data is retrieved instantaneously into memory for viewing and manipulation. Images are never stored on the local display station disk, thus centralized systems are also known as cacheless.

A centralized architecture is easier to implement and maintain and uses a simple on-demand data access model, but also has a single point-of-failure in the central server component. A cacheless system is also bandwith-limited, requiring a fast network connection from display stations to the central server. The distributed architecture requires more complex workflow logic such as autorouting and prefetching of data to implement. However, distributed PACS may have more functionality and may be more easily scalable than centralized systems. Early PACS were predominately distributed systems. With the increasing availability of high bandwidth networks and large amounts of inexpensive storage media, most PACS today follow the centralized architecture. Web-based PACS typical of today's systems are a specialized subset of centralized architectures. Future PACS may evolve to be a combination of both distributed architectures and centralized architectures, encompassing the best of each design.

If a PACS operates with a cached architecture in which data is automatically distributed to and stored at the display station, then the online storage capabilities should include space for maintaining all the pertinent examinations for a given episode of current care, (i.e., three days for outpatients and six days for inpatients). Space for prefetched relevant historical examinations should also be included in the anticipated storage requirements.

If the PACS operates as a cacheless centralized system, then it is important to have adequate capacity to store a patient's clinical encounter on the server. In this case, it is important to have large amounts of online RAID (redundant array of independent disks—see section on RAID below) at the central server instead of large amounts of local storage at each display station. RAID capacity should also encompass relevant prior examinations prefetched to the server.

KEY COMPONENTS AND ESSENTIAL FEATURES

Image Acquisition

Digital acquisition of data from the various imaging modalities for input to a PACS is the first step to eliminating film in medical imaging. Essential features for successful clinical implementation include conformance with the DICOM standard, radiology information system – hospital information system (RIS-HIS) interfacing, and workflow integration Quality assurance (QA) and quality control (QC) and troubleshooting problems occurring specifically at image acquisition are also critical as these problems affect the integrity of data in the archive.

Integration with PACS. Image acquisition is the first point of data entry into a PACS,system and, as such, errors generated here can propagate throughout the system, adversely affecting clinical operations. General predictors for successful incorporation of image acquisition devices into a digital imaging department include the following: ease of device integration into the established daily workflow routine of the clinical environment, high reliability and fault-tolerance of the device, simplicity and intuitiveness of the user interface, and device speed (13). The integration of modalities with PACS and information systems using the DICOM modality worklist feature (see below) can reduce the number of patient demographic errors and the number of cases that are inappropriately or unspecified and therefore not archiveable, which also ensures the correctness of permanently archived information.

DICOM. Imaging modality conformance with the DICOM standard is critical. DICOM consists of a standard image format as well as a network communications protocol. Compliance with this standard enables an open architecture for imaging systems, bridging hardware and software entities, allowing interoperability for the transfer of medical images and associated information between disparate systems.

The DICOM standard is used, for example, to negotiate a transaction between a compliant imaging modality and a compliant PACS workstation. The scanner notifies the workstation, in a language both understand, that it has an image study to send to it. The workstation replies to the modality when it is ready to receive the data. The data is sent in a format known to all, the workstation acknowledges receipt of the image, and then the devices end their negotiation. Data is formatted in terms of groups and elements. Group 8, for example, pertains to image identification parameters (such as study, series, and image number) and Group 10 includes patient demographics (such as patient name, medical record number, and date of birth).

Prior to DICOM, the acquisition of digital image data and relevant information was extremely difficult, often requiring separate hardware devices and software programs for different vendors' products, and even for different models of devices made by the same manufacturer because each vendor used their own proprietary data format and communication's protocol. Most of the major manufacturers of imaging devices currently comply with the DICOM standard, thus greatly facilitating an open systems architecture consisting of multivendor systems. For many legacy devices purchased prior to the establishment of DICOM, an upgrade path to compliance can be performed. For those few devices that do not yet meet the standard, interface boxes consisting of hardware equipment and software programs that convert the image data from the manufacturer's proprietary format to the standard form are available.

RIS-HIS Interfacing for Data Verification. Equally essential, particularly at acquisition, is integrating the radiology information system (RIS) or hospital information system (HIS) with the PACS, which greatly facilitates input of patient demographics (name, date, time, medical record number (MRN) to uniquely identify a patient, accession number (AccNum) to uniquely identify an imaging examination, exam type, imaging parameters, etc.) and enables automatic PACS data verification, correlation, and error correction with the data recorded in the RIS-HIS. Most imaging modalities are now tightly coupled with the RIS, providing automatic downloading of demographic information from the RIS via barcode readers or directly to the scanner console (via modality worklist capability) and, hence, to the DICOM header. This copling eliminates the highly error-prone manual entry of data at acquisition.

HL7 is the RIS-HIS standard, and compliance with HL7 is desirable. RIS-HIS databases are typically patient-centric, enabling query and retrieval of information by the patient and study, series, or image data hierarchy. Integration of RIS-HIS data with the PACS adds intelligence to the system, helping to move data around the system based on "*how, what* data should be delivered *where* and *when*", automating the functions performed traditionally by the film librarian.

Modality Worklist. Many vendors now provide the capability to download RIS-HIS schedules and worklists directly to the imaging modality, such as most computed tomography (CT), magnetic resonance imaging (MRI), digital fluoroscopy (DF), and ultrasound (US) scanners. In these circumstances, the imaging technologist need only choose the appropriate patient's name from a list on the scanner console monitor (i.e., by pointing to it on a touchscreen pad), and the information contained within the RIS-HIS database will be downloaded into the PACS header and associated with the image data for that patient examination.

The general DICOM model for acquisition of image and relevant data from the imaging modality involves the modality device acting as a SCU, which provides the data, and storing it to a SCP, which provides the service: devices such as a PACS acquisition gateway or an image display workstation. In the modality worklist function, however, the image device receives the pertinent patient demographics and image study information from a worklist server, such as a PACS- RIS- or RIS-HIS-interfaced device.

Two modes exist for accomplishing the RIS-HIS data transfer to the imaging modality. The first involves data, being transferred automatically to the modality based on the occurrence of an event trigger, such as an examination being scheduled or a patient having arrived. The second method involves a query from the modality to the RIS-HIS or some other information system that holds relevant data, such as an electronic order-entry system or even some PACS databases, which may be initiated by entry of some identifier at the modality, such as bar coding of the study accession number or the patient medical record number from the scheduling card. This method then initiates a request for the associated RIS-HIS information (patient name, date of birth) to be sent from the worklist server on demand.

The benefits of the DICOM modality worklist cannot be overstated. Incorrectly (manually) entered patient demographic data, such as all the permutations of patient name (i.e., James Jones, J Jones, Jones J) can result in mislabeled image files and incomplete study information and, as such, is crucial to maintaining the integrity of the PACS database. Furthermore, the improvements in departmental workflow efficiency and device usability are greatly facilitated by modality worklist capabilities. For those few vendors not offering DICOM modality worklist for their imaging devices, several interface or broker boxes are available that interconnect PACS to RIS-HIS databases translating DICOM to HL7 and vice versa. Figure 2 diagrams an example of how RIS, HIS, and PACS systems might interact upon scheduling an examination for image acquisition into a PACS (14).

Acquisition of the Native Digital Cross-Sectional Modalities. Image acquisition from the inherently digital modalities, such as CT, MRI, and US, should be a direct digital DICOM capture. Direct digital interfaces allow capture and transmission of image data from the modality at the full spatial resolution and full bit depth of gray scale inherent to the modality, whereas the currently outdated analog (video) frame grabbers digitize the video signal voltage output going to an image display, such as a scanner console monitor. In the frame-grabbing method, as in printing an image to film, the image quality is limited by the process to just 8 bits (or 256 gray values), whereas most modalities have the capability to acquire in 12, 16, or even 32 bits for color data. Capture of only 8 bits may not allow viewing in all the appropriate clinical windows and levels or contrast and brightness settings and is, therefore, not optimal.

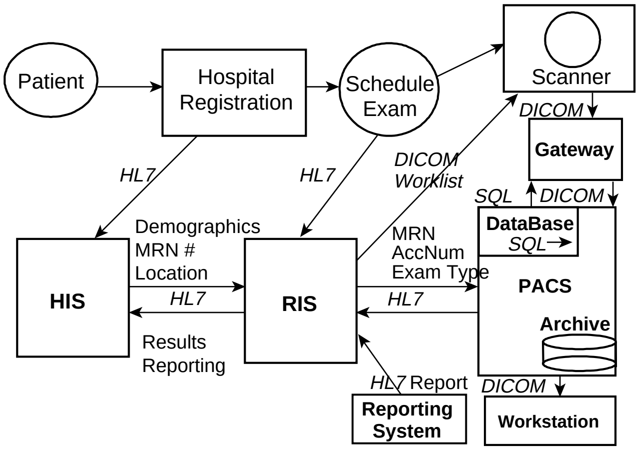


Figure 2. Diagram of how RIS, HIS, and PACS systems might interact on scheduling an examination for image acquisition into a PACS.

Table 1. The Commonly PACS-Interfaced Cross-Sectional Modalities and their Inherent File Sizes

Modality	Image Matrix Size	Grayscale Bit Depth
Computed Tomography (CT)	512 × 512 pixels	12 – 16 bits
Digital Angiography & (RA)	512 × 512 pixels or	8 – 12 bits
Digital Fluoroscopy (DF)	1024 × 1024 pixels or 2048 × 2048 pixels	
Magnetic Resonance Imaging (MRI)	256 × 256 pixels	12 – 16 bits
Nuclear Medicine Images (NUC)	64 × 64 pixels or	8 – 32 bits
	128 × 128 pixels or 256 × 256 pixels	
Ultrasound (US)	64 × 64 pixels or 128 × 128 pixels	16 – 32 bits

For example, when viewing a CT of the chest, one may wish to view in lung window and level settings and in mediastinal and bone windows and levels. Direct capture of the digital data will allow the viewer to dynamically window and level through each of these settings on-the-fly (in real time) at the softcopy display station. However, to view all appropriate window and level settings on film, several copies of the study would have to be printed, one at each window and level setting. If one performs the analog acquisition or frame grabbing of the digital data, the viewer can only window and level through the 8 bits captured, which may not be sufficient. Thus, direct capture of digital data from the inherently digital modalities is the preferred method of acquisition. Table 1 lists the cross-sectional modalities commonly interfaced to PACS along with their inherent file sizes and bit depths.

Acquisition of Projection Radiography. Methods for digital image acquisition of the conventional projection X ray include computed radiography (CR) scanners or imaging with photostimulable or storage phosphors, digitization of existing analog film, and digital radiography (DR) devices. Digital acquisition of images already on film can be accomplished using a variety of image digitization devices or film scanners, including the no longer used analog video cameras with analog-to-digital converters (ADC), digital cameras, charge-coupled devices (CCD), and laser scanners. Both CR and DR are replacement methods for capturing conventional screen-film projection radiography.

Film Digitizers. Film digitizers will still be necessary even in the all digital or filmless imaging department, so that film images from outside referrals lacking digital capabilities can be acquired into the system and viewed digitally. Film digitizers convert the continuous optical density values on film into a digital image by sampling at discrete evenly spaced locations and quantizing the transmitted light from a scan of the film into digital numbers. Several types of film digitizers exist today, with some used more frequently than others in PACS and teleradiology applications.

The analog video camera with ADC, or camera on a stick, was used in low cost, entry-level teleradiology applications but is no longer used in PACS applications today because of its manual operation. The analog video camera requires an illumination source and careful attention to lens settings, focus, f-stop, and so on. In addition, it has a maximum resolution of 1024 by 1024 by 8 bits (256 grays), thus limiting the range of window and level, or contrast and brightness values, the resulting digital image can be displayed in. Digital cameras produce a digital signal output directly from the camera at a maximum resolution of 2048 by 2048 by 12 bits (4096 grays) but are still infrequently used in PACS due to their high cost.

More commonly used are film scanners such as the CCD and laser scanners sometimes called flat-bed scanners. CCD scanners use a row of photocells and uniform bright light illumination to capture the image. A lens focuses the transmitted light from the collimated, diffuse light source onto a linear CCD detector, and the signal is collected and converted to a digital electronic signal via an ADC. CCD scanners have a maximum resolution of 4096 by 4096 by 8–12 bits, but they have a narrow film optical density range to which they can respond. CCD scanners have been used in high-end teleradiology or entry-level in-house film distribution systems, such as image transmission to the intensive care units (ICUs).

The laser scanner or laser film digitizer uses either a helium-neon (HeNe) gas laser or a solid-state diode laser source. The laser beam is focused by lenses and directed by mirror deflection components, and the light transmitted through the film is collected by a light guide, its intensity detected by a photomultiplier tube, converted to a proportional electronic signal, and digitized in an ADC. Laser scanners use a fine laser beam of generally variable or adjustable spot sizes down to 50 μm (producing an image sharpness of approximately 10 line pairs per millimeter). They have a maximum spatial resolution of 4096 by 5120 and a grayscale resolution of 12 bits, and can accommodate the full optical density range of film. They are semi- or fully-automatic in operation and are currently the scanner of choice for PACS applications even though they are often more expensive than CCD scanners.

Computed Radiography (CR). Computed Radiography (CR) refers to projection X-ray imaging using photostimulable or storage phosphors as the detector. In this modality, X rays incident upon a photostimulable phosphor-based image sensor or imaging plate produce a latent image that is stored in the imaging plate until stimulated to luminesce by laser light. This released light energy can be captured and converted to a digital electronic signal for transmission of images to display and archival devices. Unlike conventional screen-film radiography in which the film functions as the imaging sensor, or recording medium, as well as the display device and storage media, CR eliminates film from the image recording step, resulting in a separation of image capture from image display and image storage. This separation of functions potentiates optimization of each of these steps individually. In addition, CR can capitalize on features common to all digital images, namely, electronic transmission, manipulation, display, and storage of radiographs (15).

Technological advances in CR over time have made this modality widely accepted in digital departments. Hardware and software improvements have occurred in the photostimulable phosphor plate, in image reading-scanning devices, and in image processing algorithms. Overall reduced cost of CR devices, as well as a reduction in the cost and increased utility of image display devices, have contributed to the increased acceptance of CR as a viable digital counterpart to conventional screen-film projection radiography.

Review of the Fundamentals

Process Description. ACR system consists of a screen or plate of a stimulable phosphor material that is usually contained in a cassette and is exposed in a manner similar to the traditional screen-film cassette. The photostimulable phosphor in the imaging plate (IP) absorbs X rays that have passed through the patient, "recording" the X-ray image. Like the conventional intensifying screen, CR plates produce light in response to an X ray, at the time of exposure. However, storage phosphor plates have the additional property of being capable of storing some of the absorbed X-ray energy as a latent image. Plates are typically made of an europium-doped barium-fluoro-halide-halide crystallized matrix. Electrons from the dopant ion become trapped just below the conduction band when exposed to X rays. Irradiating the imaging plate at some time after the X ray exposure with red or near-infrared laser light liberates the electrons into the conduction band, stimulating the phosphor to release some of its stored energy in the form of green, blue, or ultraviolet light—the phenomenon of photostimulable luminescence. The intensity of light emitted is proportional to the amount of X ray absorbed by the storage phosphor (16).

The readout process uses a precision laser spot-scanning mechanism in which the laser beam traverses the imaging plate surface in a raster pattern. The stimulated light emitted from the IP is collected and converted into an electrical signal, with optics coupled to a photomultiplier tube (PMT). The PMT converts the collected light from the IP into an electrical signal, which is then amplified, sampled to produce discrete pixels of the digital image, and sent through an ADC to quantize the value of each pixel (i.e., a value between 0 and 1023 for a 10 bit ADC or between 0 and 4095 for a 12 bit ADC).

Not all of the stored energy in the IP is released during the readout process. Thus, to prepare the imaging plate for a new exposure, the IP is briefly flooded with high intensity (typically fluorescent) light. This erasure step ensures removal of any residual latent image.

A diagram of the process steps involved in a CR system is shown in Fig. 3. In principle, CR inserts a digital computer between the imaging plate receptor (photostimulable phosphor screen) and the output film. This digital

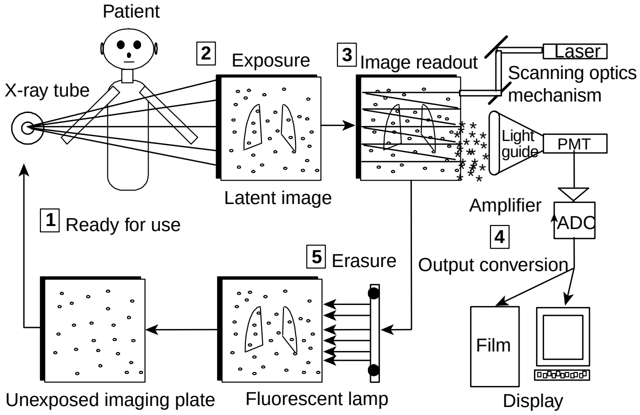


Figure 3. The image production steps involved in CR. The imaging plate is exposed to X rays, read out by a laser scanning mechanism, and erased for reuse. A light guide collects the photostimulated luminescence and feeds it to a photomultiplier tube (PMT) that converts the light signal to an electrical signal. Amplification, logarithmic conversion, and analog-to-digital conversion produce the final digital signal that can be displayed on a cathode ray tube monitor or sent to a laser printer for image reproduction on film.

processor can perform a number of image processing tasks including compensating for exposure errors, applying appropriate contrast characteristics, enhancing image detail, and storing and distributing image information in digital form.

System Characteristics. One of the most important differences between CR and screen-film systems is in exposure latitude. The response of a digital imaging system relates the incident X ray exposure to the resulting pixel value output. System sensitivity is the lowest exposure that will produce a useful pixel value, and the dynamic range is the ratio of the exposures of the highest and lowest useful pixel values (17). Storage phosphor systems have extremely wide exposure latitude. The wide latitude of storage phosphor systems, and the effectively linear detector characteristic curve, allows for a wider range of exposure information to be captured in a single image than is possible with any screen-film system. In addition, the wide dynamic range of CR allows it to be used under a broad range of exposure conditions without the need for changing the basic detector, also making CR an ideal choice for applications in which exposures are highly variable or difficult to control, as in portable or bedside radiography. Through image processing, CR systems can usually create a diagnostic image out of under- or over-exposures via appropriate look-up table correction. In the screen-film environment, such under- or over-exposures might have necessitated retakes and additional exposure to the patient.

Dose requirements of a medical imaging system depend on the system's ability to detect and convert the incoming signal into a usable output signal. It is important to stress that CR systems are not inherently lower dose systems than screen-film. In fact, several studies have demonstrated a higher required exposure for CR to achieve equivalent optical density on screen-film (18,19). However, the wider latitude of storage phosphor systems makes them much more forgiving of under- or over-exposure. As in any digital radiography system, when dose is decreased, the noise due to quantum mottle increases (20). Reader tolerance of this noise tends to be the limiting factor on the lowest acceptable dose.

In some clinical situations, the radiologist may feel comfortable in lowering the exposure technique factor to reduce dose to the patient, such as in pediatric extremity X-ray exams. In others, such as imaging the chest of the newborn, one may wish to increase exposure to reduce the more visible mottle (at lower doses) to avoid mistaking the noise over the lungs as an indication of pulmonary interstitial emphysema, for example. CR systems are signal-to-noise-limited (SNR-limited), whereas screen-film systems are contrast-limited.

Image Quality. DQE: Objective descriptors of digital image quality include detective quantum efficiency (DQE), which is a measure of the fidelity with which a resultant digital image represents the transmitted X-ray fluence pattern (i.e., how efficiently a system converts the X-ray input signal into a useful output image), and includes a measure of the noise added (17). Also taken into account are the input/output characteristics of the system and the resolution response of unsharpness or blur added during the image capture process. The linear, wide-latitude input/output characteristic of CR systems relative to screen-film leads to a wider DQE latitude for CR, which implies that CR has the ability to convert incoming X-ray quanta into "useful" output over a much wider range of exposures than can be accommodated with screen-film systems (20).

Spatial Resolution: The spatial resolution response or sharpness of an image capture process can be expressed in terms of its modulation transfer function (MTF), which, in practice, is determined by taking the Fourier Transform of the line spread function (LSF) and relates input subject contrast to imaged subject contrast as a function of spatial frequency (17). The ideal image receptor adds no blur or broadening to the input LSF, resulting in an MTF response of one at all spatial frequencies. A real image receptor adds blur, typically resulting in a loss of MTF at higher spatial frequencies.

The main factors limiting the spatial resolution in CR, similar to screen-film systems, is X-ray scattering within the phosphor layer. However, it is the scattering of the stimulating beam in CR, rather than the emitted light as in screen-film, that determines system sharpness (20,21). Broadening of the laser light spot within the IP phosphor layer spreads with the depth of the plate. Thus, the spatial resolution response of CR is largely dependent on the initial laser beam diameter and on the thickness of the IP detector. The reproducible spatial frequency of CR is also limited by the sampling used in the digital readout process. The spatial resolution of CR is less than that of screen-film, with CR ranging from 2.5 to 5 line pairs per millimeter (lp/mm) using a 200 μm laser spot size and a digital matrix size of approximately 2,000 by 2,500 pixels versus the 5–10 lp/mm or higher spatial resolution of screen-film.

Finer spatial resolution can technically be achieved today with the ability to tune laser spot sizes down to

50 μm or less. However, the image must be sampled more finely (approximately 4,000 by 5,000 pixels) to achieve 10 lp/mm. Thus, a tradeoff exists between the spatial resolution that can technically be achieved and the file size to practically transmit and store. Most general CR examinations are acquired using a 200 μm laser spot size and a sampling of 2 k by 2.5 k pixels. For examinations requiring very fine detail resolution, such as in mammography, images are acquired with a 50 μm laser spot size and sampled at 4 k by 5 k pixels.

Contrast Resolution: The contrast or gray-scale resolution for CR is much greater than that for screen-film. Note that because overall image quality resolution is a combination of spatial and gray-scale resolution, the superior contrast resolution of CR can often compensate for its lack of inherent spatial resolution. By manipulating the image contrast and brightness, or window and level values, respectively, small features often become more readily apparent in the image, which is analogous to "bright-lighting" or "hot-lighting" a bone film, for example, when looking for a small fracture. The overall impression is that the spatial resolution of the image has been improved when, in fact, it has not changed—only the contrast resolution has been manipulated. More work needs to be done to determine the most appropriate window and level settings with which to initially display a CR image. Lacking the optimum default settings, it is often useful to "dynamically" view CR softcopy images with a variety of window and level settings.

Noise: The types of noise affecting CR images include X-ray dose-dependent noise and fixed noise (independent of X-ray dose). The dose-dependent noise components can be classified into X-ray quantum noise, or mottle, and light photon noise (21). The quantum mottle inherent in the input X-ray beam is the limiting noise factor, and it develops in the process of absorption by the imaging plate, with noise being inversely proportional to the detector X-ray dose absorption. Light photon noise occurs in the process of photoelectric transmission of the photostimulable luminescence light at the surface of the PMT.

Fixed-noise sources in CR systems include IP structural noise (the predominant factor), noise in the electronics chain, laser power fluctuations, quantization noise in the analog-to-digital conversion process, and so on (20,21). IP structural noise develops from the nonuniformity of phosphor particle distribution, with finer particles providing noise improvement. Note that for CR systems, it is the noise sources that limit the DQE system latitude, whereas in conventional X-ray systems, the DQE latitude is limited by the narrower exposure response of screen-film.

Comparison with Screen-Film: The extremely large latitude of CR systems makes CR more forgiving in difficult imaging situations, such as portable examinations, and enables decreased retake rates for improper exposure technique, as compared with screen-film. The superior contrast resolution of CR can compensate in many cases for its lesser spatial resolution. Cost savings and improved radiology departmental workflow can be realized with CR and the elimination of film for projection radiographs.

Available CR Systems.

Historical Perspective. Most of the progress in storage phosphor imaging has been made since World War II (22). In 1975, Eastman Kodak Company (Rochester, NY) patented an apparatus using infrared-stimulable phosphors or thermoluminescent materials to store an image (23). In 1980, Fuji Photo Film (Tokyo, Japan) patented a process in which photostimulable phosphors were used to record and reproduce an image by absorbing radiation and then releasing the stored energy as light when stimulated by a helium-neon laser (24). The emitted phosphor luminescence was detected by a PMT, and the electronic signal produced reconstructed the image.

Fuji was the first to commercialize a storage phosphor-based CR system in 1983 (as the FCR 101) and published the first technical paper (in Radiology) describing CR for acquiring clinical digital X-ray images (25). The central processing type second-generation scanners (FCR 201) were marketed in 1985 (21). Third-generation Fuji systems marketed in 1989 included distributed processing (FCR 7000) and stand-alone (AC-1) types (21). Fuji systems in the FCR 9000 series are improved, higher speed, higher performance third-generation scanners. Current Fuji systems include upright chest units, CR detectors in wall and table buckeyes, multiplate autoloaders, and more compact stand-alone units.

In 1992, Kodak installed its first commercial storage phosphor reader (Model 3110) (16). Later models include autoloader devices. In 1994, Agfa-Gevaert N.V. (Belgium) debuted its own CR system design (the ADC 70) (26). In 1997, Agfa showed its ADC Compact with greatly reduced footprint. Agfa also introduced a low cost, entry-level single-plate reader (the ADC Solo) in 1998, appropriate for distributed CR environments such as clinics, trauma centers, and ICUs. In 1998, Lumisys presented its low cost, desktop CR unit (the ACR 2000) with manual-feed, single-plate reading. Numerous desktop units have been introduced including the Orex CR. Konica Corp. debuted its own device (XPress) in 2002 and, later, the Regius upright unit, both of which have relatively fast scan times (at 40 and 16 s cycle times, respectively).

Many companies have been involved in CR research and development, including N.A. Philips Corp.; E.I. DuPont de Nemours & Co.; 3M Co.; Hitachi, Ltd.; Siemens AG; Toshiba Corp.; General Electric Corp.; Kasei Optonix, Ltd.; Mitsubishi Chemical Industries, Ltd.; Nichia Corp.; GTE Products Co.; and DigiRad Corp. (20).

Technological Advances. Major improvements in the overall CR system design and performance characteristics include a reduction in the physical size of the reading/scanning units, increased plate-reading capacity per unit time, and better image quality. These advances have been achieved through a combination of changes in the imaging plates themselves, in the image reader or scanning devices, and in the application of image processing algorithms to affect image output.

The newer imaging plates developed for the latest CR devices have higher image quality (increased sharpness) and improved fading and residual image characteristics. Higher image quality has resulted from several modifications in the imaging plate phosphor and layer thickness. Smaller phosphor grain size in the IP (down to approximately 4 μm) diminishes fixed noise of the imaging plate, whereas increased packing density of phosphor particles counteracts a concomitant decrease in photostimulable luminescence (21). A thinner protective layer is used in the plates tending to reduce X-ray quantum noise and, in and of itself, would improve the spatial resolution response characteristics of the plates as a result of diminished beam scattering. However, in the newest IPs, the quantity of phosphor coated onto the plate is increased for durability purposes, resulting in the same response characteristic of previous imaging plates (27).

An historical review of CR scanning units chronicles improved compactness and increased processing speed. The first Fuji unit (FCR 101) from 1983 required roughly 6 m^2 of floor space to house the reader and could only process about 45 plates per hour, whereas today's Fuji models as well as other vendor's devices occupy less than 1 m^2 and can process over 110 plates per hour, which represents a decrease in apparatus size by a factor of approximately one-sixth and an increase in processing capacity of roughly 2.5 times. Desktop models reduce the physical device footprint even further.

CR imaging plate sizes, pixel resolutions, and their associated digital file sizes are roughly the same across manufacturers for the various cassette sizes offered. For example, the 14″ by 17″ (or 35 cm by 43 cm metric equivalent) plates are read with a sampling rate of 5–5.81 pixels per mm, at a digital image matrix size of roughly 2 k by 2 k pixels (1760 by 2140 pixels for Fuji (21) and 2048 by 2508 pixels for Agfa and Kodak (16). Images are typically quantized to 12 bits (for 4096 gray levels). Thus, total image file sizes range from roughly 8 megabytes (MB) to 11.5 MB. The smaller plates are scanned at the same laser spot size (100 μm), and the digitization rate does not change; therefore, the pixel size is smaller (16). The 10″ by 12″ (24 cm by 30 cm) plates are typically read at a sampling rate of 6.7–9 pixels per millimeter (mm) and the 8″ by 10″ (18 cm by 24 cm) plates are read at 10 pixels per mm (16,21).

Cassetteless CR devices have been introduced in which the detector is incorporated into a chest unit, wall, or table buckey to speed throughput and facilitate workflow much like DR devices do. Dual-sided signal collection capability is available by Fuji, increasing overall signal-to-noise. Agfa has shown a product in development (ScanHead CR) that stimulates and reads out the imaging plate line-by-line, as opposed to the point-by-point scanning that occurs in most CR devices today. Increased speed (5 s scan time) and higher DQE have been demonstrated. In addition, needle phosphors have been explored as a possible replacement to powder phosphors, having shown improved spatial resolution and DQE.

Image Processing Algorithms. Image processing is performed to optimize the radiograph for output display. Each manufacturer has a set of proprietary algorithms that can be applied to the image for printing on laser film or display initially only on their own proprietary workstations. Prior to the DICOM standard, only the raw data could be directly acquired digitally. Therefore, to attain the same image appearance on other display stations, the appropriate image processing algorithms (if known) had to be implemented somewhere along the chain from acquisition to display. Now image processing parameters can be passed in the DICOM header and algorithms applied to CR images displayed on generic workstations. Typically, however, advanced real-time manipulation of images can only be done on each manufacturers' specific processing station. In general, the digital image processing applied to CR consists of a recognition or analysis phase, followed by contrast enhancement or frequency processing. Note that the same general types of image processing applied to CR can also be applied to DR images.

Image Segmentation. In the image recognition stage, the region of exposure is detected (i.e., the collimation edges are detected), a histogram analysis of the pixel gray values in the image is performed to assess the actual exposure to the plate, and the appropriate look-up table specific to the region of anatomy imaged and chosen by the X-ray technologist at the time of patient demographic information input is selected. Proper recognition of the exposed region of interest is extremely important as it affects future processing applied to the image data. For example, if the bright-white area of the image caused by collimation at the time of exposure is not detected properly, its very high gray values will be taken into account during histogram analysis, increasing the "window" of values to be accommodated by a given display device (softcopy or hardcopy). The effect would be to decrease the overall contrast in the image.

Some segmentation algorithms, in addition to detection of collimation edges in the image, enable users to blacken the region outside these edges in the final image if so desired (16,28), which tends to improve image contrast appearance by removing this bright-white background in images of small body parts or pediatric patients. The photo in Fig. 4B demonstrates this feature of "blackened surround," as applied to the image in Fig. 4A.

Contrast Enhancement. Conventional contrast enhancement, also called gradation processing, tone scaling, and latitude reduction, is performed next. This processing amounts to choosing the best characteristic curve (usually a nonlinear transformation of X-ray exposure to image density) to apply to the image data. These algorithms are quite flexible and can be tuned to satisfy a particular user's preferences for a given "look" of the image (29). Look-up tables are specific to the region of anatomy imaged. Figure 5 shows an example of the default adult chest look-up table (Fig. 5a) applied to an image and the same image with high contrast processing (Fig. 5b). A reverse-contrast scale or "black bone" technique, in which what was originally black in the image becomes white and what was originally white in the image becomes black, is sometimes felt to be beneficial for identifying and locating tubes and lines. An example is shown in Fig. 6 where the contrast reversal algorithm has been applied to the image in Fig. 6a, resulting in the image in Fig. 6b.

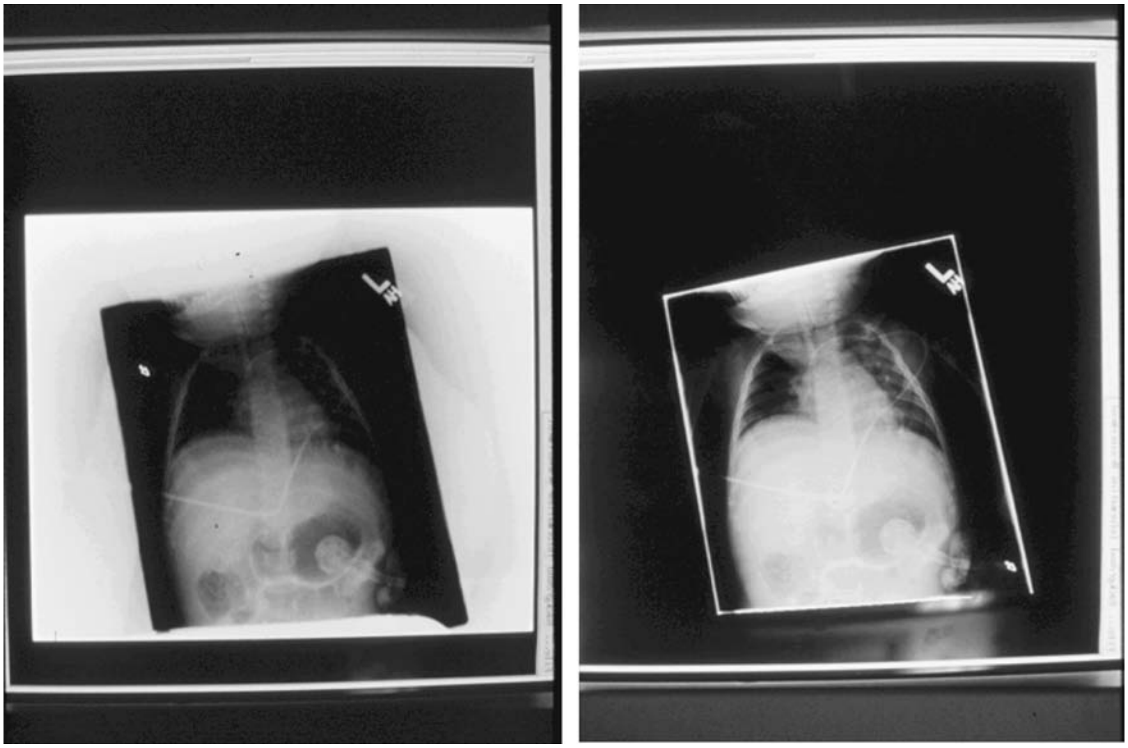


Figure 4. Example image segmentation algorithm detection of (white) collimation edges of exposure region in image **A**, with "blackened surround" applied in image **B**. Note the improved overall contrast in the image in **B**.

Spatial Frequency Processing. The next type of image processing usually performed is spatial frequency processing, sometimes called edge enhancement. These algorithms adjust the frequency response characteristics of the CR systems essentially implementing a high- or band-pass filter operation to enhance the high spatial frequency content contained in edge information. Unfortunately, noise also contains high spatial frequency information and can be exacerbated by edge enhancement techniques. To lessen this problem, a nonlinear unsharp masking technique is typically implemented serving to suppress noise via a smoothing process. Unsharp masking is an averaging technique that, via summation, tends to blur the image. When this result is subtracted from the original image data, the effect is one of noise suppression. Specific spatial frequencies can be preferentially selected and emphasized by changing the mask size and weighting parameters. For example, low spatial frequency information in the image can be augmented by using a relatively large mask, whearas high spatial frequency or edge information can be enhanced by using a small mask size (16).

Dynamic Range Control. An advanced algorithm by Fuji, for selective compression or emphasis of low density

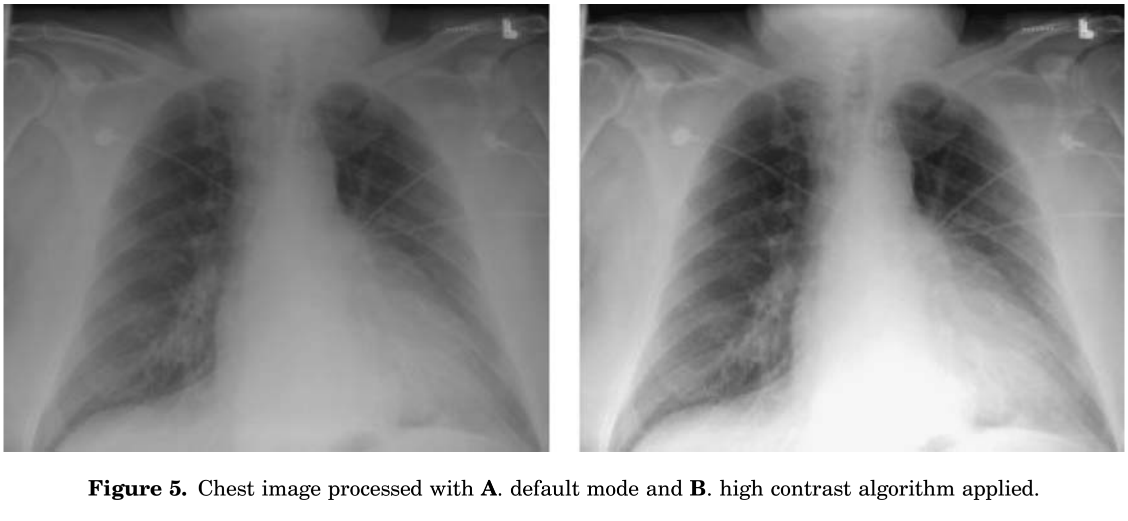


Figure 5. Chest image processed with **A**. default mode and **B**. high contrast algorithm applied.

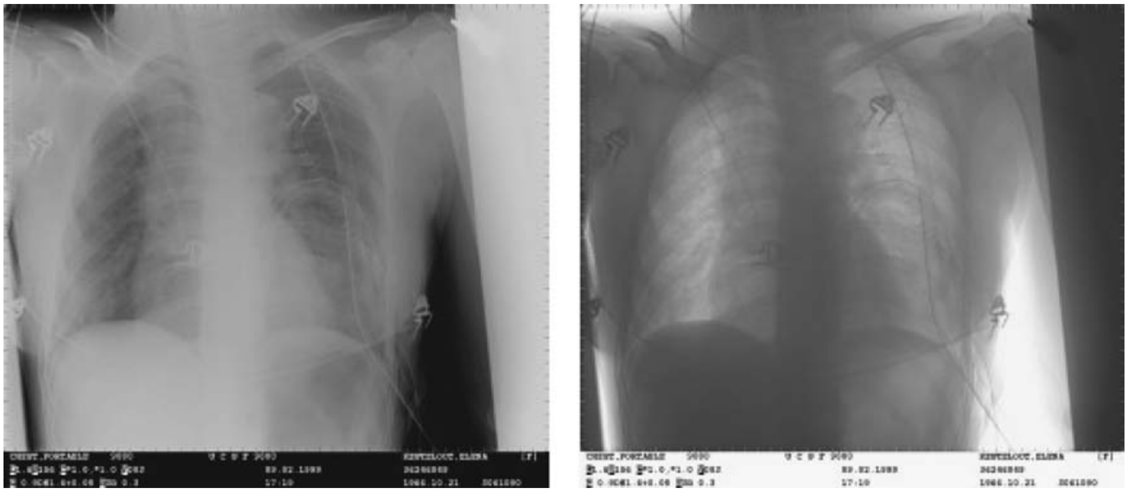


Figure 6. Chest image processed with **A**. default mode and **B**. blackbone or contrast reversal algorithm applied.

regions in an image, independent of contrast and spatial frequency is known as dynamic range control (DRC) processing (30). The algorithm consists of performing an unsharp mask for suppression of high spatial frequency information, then application of a specific look-up table mapping to selected regions (i.e., low density areas). This mask is then added back to the original data with the overall result being improved contrast in poorly penetrated regions, without loss of high frequency and contrast emphasis. In a clinical evaluation of the algorithm for processing of adult portable chest exams, DRC was found to be preferred by five thoracic radiologists in a side-by-side comparison, providing improved visibility of mediastinal details and enhanced subdiaphragmatic regions (31).

Multiscale Image Contrast Amplification. Multiscale image contrast amplification (MUSICA) is a very flexible advanced image processing algorithm developed by Agfa (26,32). MUSICA is a local contrast enhancement technique based on the principle of detail amplitude or strength and the notion that image features can be striking or subtle, large in size or small. MUSICA processing is independent of the size or diameter of the object with the feature to be enhanced. The method is carried out by decomposing the original image into a set of detail images, where each detail image represents an image feature of a specific scale. This set of detail images or basis functions completely describes the original image. Each detail image representation and the image background are contrast equalized separately; some details can be enhanced and others attenuated as desired. All the separate detail images are recombined into a single image, and the result is diminished differences in contrast between features regardless of size, such that all image features become more visible.

Image Artifacts. The appearance and causes of image artifacts that can occur with CR systems should be recognized and corrected. Artifacts can develop from a variety of sources, including those related to the imaging plates themselves, to image readers, and to image processing. Several types of artifacts potentially encountered with CR have been minimized with the latest technology improvements but may still be seen in older systems.

Lead backing added to the aluminum-framed, carbon-fiber cassettes has eliminated the so-called light-bulb effect, darkened outer portions of a film due to backscattered radiation (33). High sensitivity of the CR plates renders them extremely susceptible to scattered radiation or inadvertent exposure, thus routine erasure of all CR plates on the day of use is recommended as is the storing of imaging plates on end, rather than stacking of cassettes one on top of the other (34). The occurrence of persistent latent images after high exposures or after prolonged intervals between plate erasure and reuse (33,35) has been lessened by the improved efficiency of the two-stage erasure procedure used in the latest CR systems (34). Improved recognition of the collimation pattern employed for a given image allows varied (including off-angle) collimation fields and in turn, improves histogram analysis and subsequent processing of the imaged region (34), although these algorithms can fail in some instances. Plate cracking, from wear-and-tear, can create troublesome artifacts as depicted in Volpe (34).

Inadvertent double exposures can occur with the present CR systems, potentially masking low density findings, such as regions of parenchymal consolidation, or leading to errors in interpreting line positions. Such artifacts are more difficult to detect than with screen-film systems because of CR's linear frequency processing response, optimizing image intensity over a wide range of exposures (i.e., due to its wide dynamic range). Figure 7 shows an example double-exposure artifact, and additional examples are included in Volpe (34). Laser scanning artifacts can still occur with current CR readers and are seen as a linear artifact across the image, caused by dust on the light source (34). Proper and frequent cleaning of the laser and light guide apparatus as well as the imaging plates themselves can prevent such artifacts.

The ability of CR to produce clinically diagnostic images over a wide range of exposures is dependent on

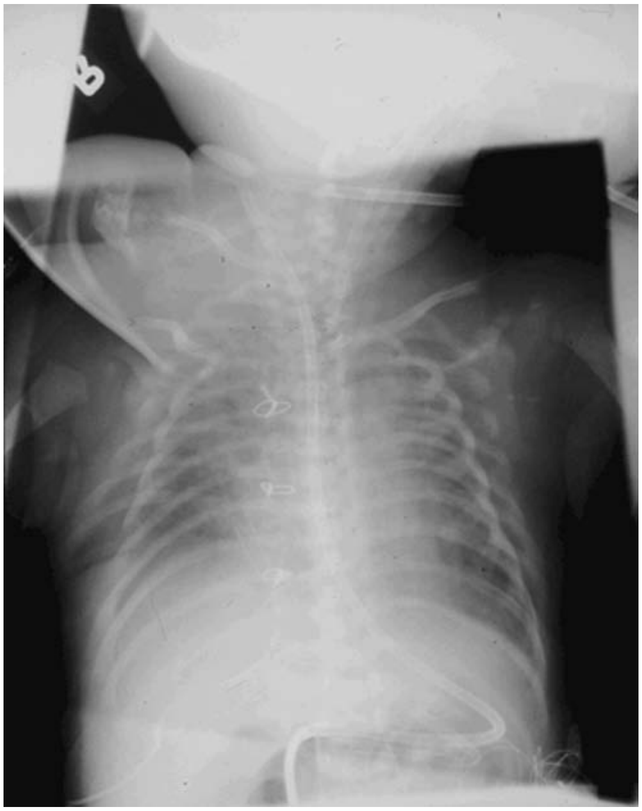


Figure 7. Example inadvertent double exposure.

the effectiveness of the image analysis algorithms applied to each dataset. The specific processing parameters used are based on standards tuned to the anatomic region under examination. Incorrect selection of diagnostic specifier or inappropriate anatomic region can result in an image of unacceptable quality. Understanding the causes of some of these CR imaging artifacts described here, as well as maintaining formal, routine quality assurance procedures, can help to recognize, correct for, and avoid future difficulties.

Summary of CR. CR can be used for the digital image acquisition of projection radiography examinations into a PACS. As a result of its wide exposure latitude and relative forgiveness of exposure technique, CR can improve the quality of images in difficult imaging situations, such as in portable or bedside examinations of critically ill or hospitalized patients. As such, CR systems have been successfully used in the ICU setting, in the emergency room (ER) or trauma center, as well as in the operating room (OR). CR can also be cost-effective for a high volume clinic setting, or in a low volume site as input to a teleradiology service, and have successfully reduced retake rates for portable and other examinations.

Technological advances in CR hardware and software have contributed to the increased acceptance of CR as a counterpart to conventional screen-film projection radiography, making the use of this modality for clinical purposes more widespread. CR is compatible with existing X-ray equipment, yet separates out the functions of image acquisition or capture, image display, and image archival versus traditional screen-film, in which film serves as the image detector, display, and storage medium. This separation in image capture, display, and storage functions by CR enables optimization of each of these steps individually. Potential expected benefits are improved diagnostic capability (via the wide dynamic range of CR and the ability to manipulate the exam through image processing) and enhanced radiology department productivity (via networking capabilities for transmission of images to remotely located digital softcopy displays and for storage and retrieval of the digital data).

Digital Radiography (DR). In addition to CR devices for digital image acquisition of projection X rays are the maturing direct digital detectors falling under the general heading of digital radiography (DR). Note that digital mammography is typically done using DR devices, although CR acquired at much higher sampling matrices has also been tested.

Unlike conventional screen-film radiography in which the film functions as the imaging sensor or recording medium as well as the display and storage media, DR, like CR, eliminates film from the image recording step, resulting in a separation of image capture from image display and image storage. This separation of functions potentiates optimization of each of these steps individually. In addition, DR, like CR, can capitalize on features common to digital or filmless imaging, namely the ability to acquire, transmit, display, manipulate, and archive data electronically, overcoming some of the limitations of conventional screen-film radiography. Digital imaging benefits include remote access to images and clinical information by multiple users simultaneously, permanent storage and subsequent retrieval of image data, expedient information delivery to those who need it, and efficient cost-effective workflow with elimination of film from the equation.

Review of the Fundamentals.

Process Description. Indirect versus Direct Conversion: DR refers to devices for direct digital acquisition of projection radiographs in which the digitization of the X-ray signal takes place within the detector. Compare this method with CR, which uses a photostimulable phosphor imaging plate detector in a cassette design that must be processed in a CR reader following X-ray exposure, for conversion to a digital image. DR devices, also called flat-panel detectors, include two types, indirect conversion devices in which light is first generated using a scintillator or phosphor and then detected by a CCD or a thin-film-transistor (TFT) array in conjunction with photodiodes; and DDR devices, which consist of a top electrode, dielectric layer, selenium X-ray photoconductor, and thin-film pixel array (36). Figure 8 shows a comparison of the direct and indirect energy conversion steps in the production of a digital X-ray image. DDR devices offer direct energy conversion of X ray for immediate readout without the intermediate light-conversion step.

The basis of DR devices is the large area thin-film-transistor (TFT) active matrix array, or flat panel, in which each pixel consists of a signal collection area or charge

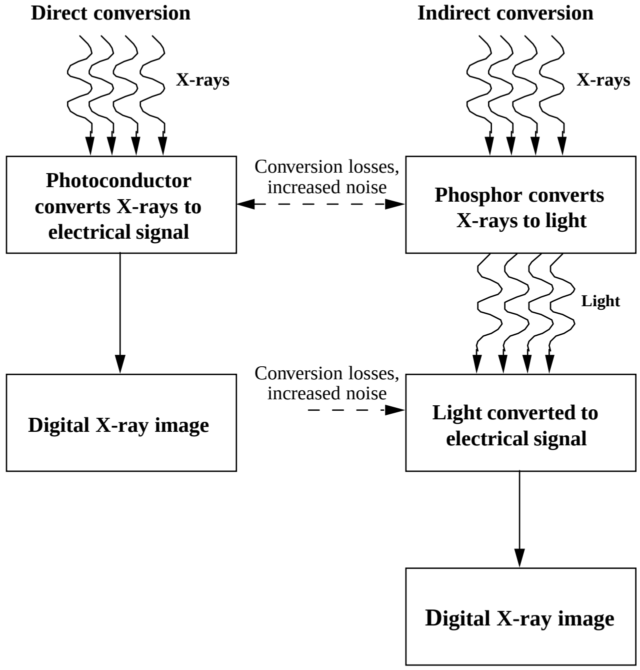

Direct conversion
Indirect conversion
X-rays
X-rays
Photoconductor converts X-rays to electrical signal
Conversion losses, increased noise
Phosphor converts X-rays to light
Light
Conversion losses, increased noise
Digital X-ray image
Light converted to electrical signal
Digital X-ray image

Figure 8. The image production steps involved in direct and indirect digital radiography detectors.

collection electrode, a storage capacitor, and an amorphous silicon field-effect transistor (FET) switch that allows the active readout of the charge stored in the capacitor (36). Arrays of individual detector areas are addressed by orthogonally arranged gate switches and data lines to read the signal generated by the absorption of X rays in the detector. The TFT arrays are used in conjunction with a direct X-ray photoconductor layer or an indirect X-ray-sensitive phosphor-coated light-sensitive detector or photodiode array.

An example DDR device, diagramed in cross section in Fig. 9 (36), uses a multilayer detector in a cassette design, in which the X-ray energy is converted directly to electron-hole pairs in an amorphous selenium (Se) photoconductive conversion layer. Charge pairs are separated in a bias field such that the holes are collected in the storage capacitors and the electrons drift toward the Se-dielectric interface. At the end of exposure, the image resides in the pixel matrix in the form of charges, with the charge proportional to the absorbed radiation. At the end of the readout, the charges are erased to prepare for another detection cycle.

An example indirect DR device uses an X-ray-sensitive phosphor coating on top of a light-sensitive flat-panel amorphous silicon (Am-Si) detector TFT array. The X rays are first converted to light and then to a proportional charge in the photodiode [typically a cesium iodide (CsI) scintillator], which is then stored in the TFT array where the image signal is recorded.

System Characteristics. DR detectors have high efficiency, low noise and good spatial resolution, wide latitude, and all the benefits of digital or filmless imaging. Similarly, DR has a very wide dynamic range of quantization to thousands of gray levels. These devices are becoming more widely used clinically and are available in table buckeys as well as chest units. DR units have superior workflow and increased patient throughput due to the elimination of cassette handling (37).

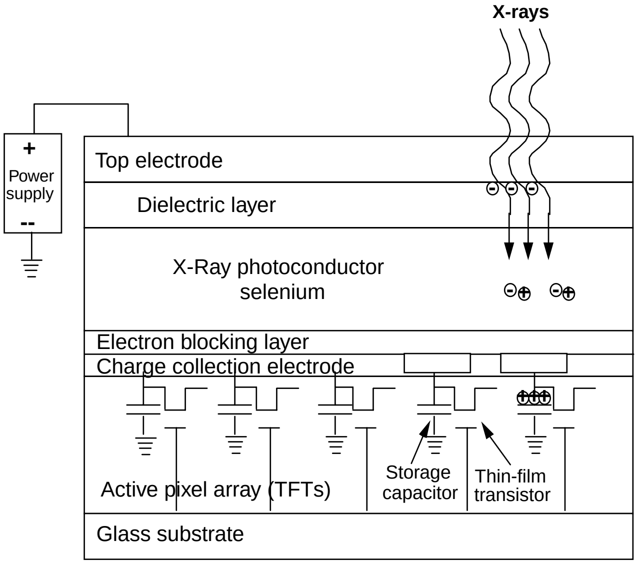


Figure 9. Cross-sectional view of an example direct digital radiography (DDR) detector panel.

The short imaging cycle time of DR may lend itself to combined static radiographic and dynamic fluoroscopic uses in future applications, which is true especially for the indirect devices. The direct Se detector, for example, has a ghosting problem due to charge trapping, which introduces a lag time at the end of each cycle, lengthening the time between readiness for the next exposure.

The cost of sensor production is still high such that the overall price of devices has not dropped appreciably. DR is sometimes referred to as a one-room-at-a-time technology because the detectors are built into the room and matched to the X-ray source. Detector fragility and poor portability makes DR difficult to use in the bedside X-ray environment, but some portable devices are now being introduced.

Image Quality. DR image quality is comparable with that of CR. However, DR devices have higher DQEs than CR, capturing roughly 80% absorption of the X-ray energy at optimum exposures. Thus, DR is a higher efficiency, low noise detector, converting much of the incoming X-ray signal into useful output. Several recent studies have demonstrated high image quality at lower radiation dose to the patient. The ability to lower exposure would be a significant advantage for DR. A factor limiting DR efficiency involves the packing fraction, or active detector area to dead space taken up by the data readout devices (transistors, data lines, capacitors, etc.). As the physical size of the data readout components is currently fixed, the smaller the pixel size, the smaller the packing fraction, with a larger proportion of dead area overwhelming the active area, in some cases reducing the active area to 30% or less (36). The overall effect is a reduction in geometric and quantum efficiency.

The spatial resolution of DR is comparable with CR, which is still less than that for analog X ray. Typical matrix sizes are on the order of 2000 to 2500 pixels by 2000 to 2500 pixels. The pixel size of the TFT array detector is the limiting factor for spatial resolution, with the direct Se detector yielding a better inherent spatial resolution than indirect detectors, which can lead to better signal modulation and superior contrast.

DR design presents a delicate tradeoff between detector efficiency, inversely proportional to pixel size, and spatial resolution, affected directly by pixel size. Typically, DR devices are specified for higher detection efficiency at a cost of less spatial resolution than screen-film, with compensation by a wide dynamic range or high contrast resolution. Design complexities requiring further development include wiring configurations to minimize dead space and maximize the detector packing fraction, fast and robust signal readout methods, and better error-correction matrices for more accurate signal readout.

Comparison of CR and DR. Table 2 lists the advantages of CR and DR, including all the benefits of digital images that can be electronically processed, manipulated, distributed, displayed, and archived. The superior contrast resolution of the digital modalities can compensate, in many cases, for the lesser spatial resolution as compared with screen-film. Both CR and DR can be used for the digital image acquisition of projection radiography examinations into a PACS.

As for any digital image acquisition device, CR or DR would be the first point of entry into a PACS. Errors may propagate from here, with the quality of the PACS output being directly dependent on the quality of the signal in. In addition to image quality, essential features for successful clinical implementation of CR or DR systems for a PACS include the following. DICOM conformance of the modality is essential and includes compliance with the image data and header format, as well as the DICOM communication protocol. Equally critical is interfacing to the RIS-HIS. Integration of the CR/DR system with the RIS-HIS can reduce human errors on patient demographic information input and improve efficiency. Ease of integration of the device into the daily workflow routine, and simplicity and robustness of the user interface are very important. Reliability, fault-tolerance, and capabilities for error tracking are also major issues to consider, as are device speed and performance.

As a result of CR's convenient workflow and portability, as well as its wide exposure latitude and relative forgiveness of exposure technique, CR can improve the quality of images in difficult imaging situations, such as in portable or bedside examinations of critically ill or hospitalized patients, and enable decreased retake rates for improper exposure technique. As such, CR systems have been successfully used in the ICU setting, in the ER or trauma center, as well as in the OR. CR can also be cost effective for a high volume clinic setting, or in a low volume site as input to a teleradiology service. Cost savings and improved radiology departmental workflow can be realized with CR and the elimination of film (37).

Table 2. Summary of Advantages of CR and DR Systems

- Produce digital images capable of being electronically processed, manipulated, distributed, displayed, and archived.
- Large latitude systems allowing excellent visualization of both soft tissue and bone in the same exposure image.
- Superior contrast resolution can compensate for lack of spatial resolution.
- Decreased retake rates.
- Potential cost savings if film is eliminated.
- Improved radiology department workflow with elimination of film handling routines.

Table 3. Summary of Future Trends in Image Acquisition

Image Matrix Size	↑
Image Quality	↑↑
Spatial Resolution	↑
# Image Slices	↑↑↑
Size of Imaging Examinations	↑↑↑
Size of Devices	↓↓
Portability of Devices	↑
Cost of Devices	↓
% of Image Devices that are Digital	↑↑↑
% of Image Acquisition that is Digital (Elimination of Film)	↑↑

Technological advances in CR hardware and software have contributed to the increased acceptance of CR as the current counterpart to conventional screen-film projection radiography, making the use of this modality for clinical purposes more widespread. CR is compatible with existing X-ray equipment, yet separates out the functions of image acquisition or capture, image display, and image archival versus traditional screen-film, in which film serves as the image detector, display, and storage medium. This separation in image capture, display, and storage functions by CR enables optimization of each of these steps individually. Potential expected benefits are improved diagnostic capability (via the wide dynamic range of CR and the ability to manipulate the data through image processing) and enhanced radiology department productivity (via networking capabilities for transmission of images to remotely located digital softcopy displays and for storage and retrieval of the digital data).

DR devices have more efficient detectors, offering direct energy conversion of X ray for immediate readout. The higher DQE may enable DR to produce high quality images at a lower radiation dose to the patient. These detectors have low noise and good spatial resolution, wide latitude, and all the benefits of digital or filmless imaging. However, cost is still high because detector production is difficult and expensive, and DR is a one-room-at-a-time detector. DR may be cost-effective in high volume settings with constant high patient throughput (37).

However, meeting the cost competitiveness of screen-film systems is difficult unless film printing is eliminated from the cost equation. DR may be preferable for imaging examinations requiring very high quality, such as in mammography, upright chest exams and bone work. DR devices integrated into table and wall buckeys are now making these devices highly efficient for emergency department trauma cases.

Future improvements in image processing algorithms, with a better understanding of optimum display settings for soft copy viewing, have the potential to greatly facilitate and standardize softcopy reading of digital projection radiographs, and further the acceptance of CR and DR in the clinical arena. It is likely that CR and DR devices will coexist for some time.

Future Trends in Image Acquisition. Although the types of imaging modalities will probably not change all that much in the next several years, the anticipated future trends in image acquisition for digital radiology and PACS include changes in the image dataset sizes, changes in the imaging devices themselves, and improvement in image processing for softcopy display of digital images.

Image Data Sets. No new types of imaging modalities are foreseen for the near future. However, it is anticipated, and has to a certain extent already begun, that the image datasets acquired from the existing modalities will increase in overall study file size, in some cases dramatically. For example, many radiology departments have begun installing multiple detector array or multislice CT scanners that tend to generate a greater number of individual images than do the single detector array scanners because the slice thickness in helical acquisition (~ 0.75 mm) versus the single detector arrays (~ 7–10 mm), and the clinical imaging protocols used, as well as the increasing clinical utility of 3D image display representations.

Image matrix sizes for the digital projection radiography devices (CR and DR) have gone up from roughly from one and two thousand square matrices to four by five thousand pixels squared for mammography applications. The increased sampling was done to improve the spatial resolution. Most laser film digitizers can now vary their spot sizes from 200 μm down to 50 μm, greatly improving the inherent spatial resolution of the resulting images of the scanned analog film, with a concomitant increase in file size.

The bit depth representation of gray-scale pixel values has also increased from 8 bits to 10, 12, and 16 bits, and color images are stored as 32 bit or 4 byte per pixel data files. Further, the addition of post-processing results or slice reconstructions, and cinegraphic sequences to the image dataset, while improving the overall quality of the image, may greatly increase the amount of data to be acquired into a PACS.

Devices. While image datasets and file sizes are getting larger, the imaging devices themselves will continue to get smaller in physical footprint, which has been seen most dramatically with the CR devices, going from requiring roughly 36 m^2 of floor space and special electrical power and cooling, to desktop devices that can be placed in most any location. CT and MRI devices are also becoming smaller in size, more portable, and more robust. Hopefully, these devices will continue to become less expensive. Terahertz imaging currently used in aerospace applications may become developed for uses in imaging humans for medical purposes. These devices acquire images at 0.25 and 0.3 THz, creating a binary (two-color) picture to contrast between materials with different transmission and reflection properties. The main advantage of a terahertz imager is that it does not emit any radiation and it is a passive camera, capturing pictures of the natural terahertz rays emitted by almost all objects.

Image Processing. An important area of increased attention continues to be image processing capabilities for softcopy image display. Future processing techniques will most likely go above and beyond the simple window and level (or contrast and brightness) manipulation techniques. These post-processing algorithms are currently available and tunable at the imaging modality, or accompanying modality acquisition workstation, but may, in time, be manipulable in real-time at the display station. Stand-alone 3D workstations are becoming more common. Efforts to embed advanced processing and visualization in the PACS workstation will ultimately allow real-time processing to be performed by the radiologist or even the referring clinician.

Image compression is currently being debated, but may, in time, be available at the modality to reduce image transmission time and archival space. Some techniques, such as the wavelet transform, may become more widely used not only as a compression technique, but also for image enhancement at the imaging devices.

In time, it is anticipated that the percentage of all imaging devices used by health-care enterprises that are digital in nature will increase greatly. Further, the percentage of digital image acquisition from the devices that are capable should increase, decreasing the amount of film used as an acquisition, display, and archival medium.

Medical Image Archival

Digital image archives were once thought of as costly inefficient impediments to moving toward PACS and digital imaging departments (38). However, current trends in archival technology have shown the cost of digital storage media decreasing steadily with capacity increasing, whereas analog devices such as paper and film continue to increase in overall cost (39). Improvements in storage devices along with the use of intelligent software have removed digital archives as a major stumbling block to implementing PACS. The following tutorial on electronic archival technologies for medical images includes a discussion of available digital media, PACS system architectures, and storage management strategies.

Digital image archival can be more efficient than the manual data storage of the traditional film file room. A study of image examination retrieval from a PACS versus a film-based system showed statistically significant reduction in times for the digital method, in many cases down from hours to minutes (40). The improved retrieval times with PACS were particularly striking for studies between six months and one year old, and for studies greater than one year (40).

An informal survey of 75 radiologists operating in a traditional film-based radiology department found that 70% experienced delayed access to films, which caused them and their staff money in terms of decreased efficiency

(41). Rarely did this delayed access to films result in repeated or unnecessary studies, or result in longer hospital stays. However, inaccessible or lost films did result in time spent, often by the radiologist or clinician, looking for films.

Digital archives are generally less people-intensive, eliminating the physical handling of films, and are, therefore, less expensive and less subject to the errors in filing and lost films that often plague film stores. Electronic archives can improve the security of stored image data and related records, assuring no loss of exam data while offering simultaneous case availability to many.

Digital archives must have an intelligent patient-centric system database interface to enable easy retrieval of imaging examinations. They should conform to the DICOM standard format and communications protocol by being able to accept and return DICOM format files. Many archive systems reformat the data once inside the storage device to a more efficient schema appropriate for the specific archive architecture.

Medical image data files are large compared with text-based clinical data, and are growing in size as new digital applications prove clinically useful. A single view chest X ray, for example, requires approximately 10 MB of storage space. With the expanding availability of multi-detector CT scanners and increasing use of magnetic resonance angiography examinations, thousand-slice studies are not uncommon. Imaging activity continues to increase significantly as it becomes a key diagnostic triage event, with most diagnostic imaging departments showing an increase in overall volume of cases. A typical 500 bed health-care enterprise performing approximately 200,000 examinations, for example, can generate on the order of 5–6 terabytes (TB) of data per year (42).

Compression can be used to reduce both image transmission time and storage requirements. Note that compression that can be achieved via hardware or software also occurs clinically (i.e., not all images of a study are filmed). Lossless (or bit-preserving) compression at 2:1 is done by most PACS archive systems already. Lossy or non-bit-preserving compression, by definition, does not provide an exact bit-for-bit replica of the original image data on decompression. However, studies have shown that numerically lossy compression can produce visually lossless images at compression ratios of 5:1 to 30:1 depending on modality (43–45). Compression at these levels can achieve much greater space savings and appear to be of adequate image quality for image comparison and review of prior studies. For perspective, without compression only 50 two-view digital projection X-ray examinations at approximately 10 MB per image can be stored on a 1 gigabyte (GB) disk. With compression at 25:1, approximately 1250 examinations can be stored on a 1 GB disk.

Digital Archival Media. The digital storage device media used in PACS today include computer hard drives or magnetic disks, RAID disks (redundant array of inexpensive disks), optical disks (OD), magneto-optical disks (MOD), and tape. Newer technologies such as digital video disks (DVD) and ultra-density optical (UDO) disks are being introduced. The properties and attributes of each storage device, including capacity, performance, and relative cost, are compared and summarized below. Figure 10 graphs the capacity versus performance characteristics of the various digital archive media widely used in PACS. Table 4 summarizes the relative cost of digital archive media per capacity or number of GB that can be stored per dollar, in addition to typical capacities and retrieval times.

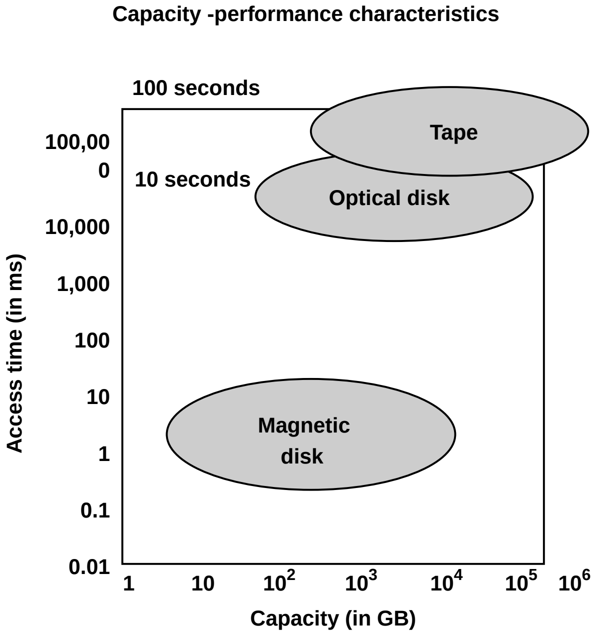


Figure 10. Graph of capacity versus performance characteristics of digital archive media.

Magnetic Disk (MD). The standard computer hard drive or magnetic disk (MD), also known as a direct access storage device (DASD), is the fastest medium from which to retrieve data. Retrieval times are on the order of 1 to 50 milliseconds (ms). However, MDs have the lowest capacity, typically hundreds of MB to hundreds of GB, and the highest cost per amount of data storage of all the archival media used in PACS today, although prices are continuing to decrease rapidly. As a result of the relative high cost in the past, MDs have historically been used for online local storage at the display workstation where fast access was required, yet large capacity was not cost-effective. Today, it is becoming cost-effective to use spinning media for all stages of storage – the trend toward everything-on-line.

Redundant Array of Inexpensive Disks (RAID). RAID devices consist of multiple MDs with high performance and larger capacity (now TB worth per device). These devices can offer redundancy, lessening the concerns with a single point of failure, and have hot-swapable components that can be replaced as needed without bringing the entire archive system down.

RAID has traditionally been used for "near line," intermediate short-term storage, or clinical-operational storage cache to minimize the number of transactions hitting the deep or long-term archive. It is becoming cheap enough per

Table 4. Digital Archive Media Capacity, Retrieval Times, and Relative Costs per GB of Storage

Archive Media Type	Storage Capacity	Performance Retrieval Times	Cost per Capacity (in Order of $ per GB)
Magnetic Disk	100s MB–10s GB	1 to 50 ms	$1.00/GB
Optical Disk	1 – 10s GB for TB devices	s to min	$0.40/GB
Tape	10s – 100s GB for 10s TB devices	10s s to min	$0.25/GB
RAID	10s–100s GB for 10s TB devices	100–300 ms	$10.00/GB
DVD	GB for TB devices	s	$2.50/GB
UDO	30 GB for 10s – 100s TB devices	s to min	$2.00/GB

capacity to consider using RAID in larger configurations for high performance longer-term storage. In these configurations, a higher percentage of studies, perhaps accounting for several years or more, can remain available online for immediate access.

Optical/Magneto-optical Disk (OD/MOD). Optical disks and magneto-optical disks are of the removable spinning storage media class typically stored in an automated media movement device or jukebox giving them total device storage amounts equal to hundreds of times the media capacity. ODs and MODs have higher capacity than MDs, typically GB to tens of GB yielding hundreds of GB to tens of TB total device storage. They are lower cost per capacity than RAID, on the order of a half dollar per GB of storage. Optical disks are a slower medium than RAID, on the order of seconds to minutes for data retrieval in batch and random seek modes.

ODs are also known as WORM or write once, read many disks with data permanently written onto the disks. MODs are erasable reusable platters and are able to hold more data per unit than ODs. As a result of the slower performance and lower cost per capacity, ODs and MODs have traditionally been used for long-term permanent PACS storage.

Tape. Tape is also a removable storage medium typically kept in a jukebox or tape library. The magnetic tape type most often used for PACS is digital linear tape (DLT). It has very high capacity, tens to hundreds of GB for many TB per tape library, and low cost on the order of a quarter dollar or less per GB of storage. It is, however, a slower medium than MOD, for example, in random retrieval times because of its sequential nature. Tape performance is competitive with MOD for retrievals of large files, however, using very high batch read-write rates. Even random retrievals of very large files (on the order of 50 MB) can be transferred faster with DLT than with MODs. Tape has historically and is currently being used for disaster backup as well as for long-term permanent storage.

Newer Technologies

Digital Video Disk (DVD). Newer technologies, such as DVDs, appear promising but have failed to move significantly into the medical arena due to their high cost and slow performance. DVDs use dual-sided storage, thus achieving greater amounts of storage (GB) than MODs fairly inexpensively. However, the cost of the drives still remain quite high and the lack of a universal standard read-write format currently limits the use of DVDs for PACS, although the DICOM standard is currently addressing this issue.

Ultra-Density Optical (UDO). Recently released ultra-density optical (UDO) disks use a blue laser recording technology to achieve much greater data storage densities, on the order of 30 GB capacity per disk, predicted to double within six months. UDO is a WORM disk technology with a 50 year lifespan and a current cost of approximately $2 per GB. Although just a first-generation device release in the medical arena (other fields including the defense industry have used UDOs), it may prove to be a useful technology for PACS.

A summary of capacity, performance, and relative cost of the types of digital archive media available today for PACS is given in Table 4. Figure 10 graphs the capacity versus performance characteristics of the MD, OD, and tape. Note that tape and OD are relatively similar in their tradeoff between capacity and performance.

Archival Strategies

Data Migration. Note that medical images have a life cycle in which, early on, quick access to the data is critical and is often needed by several people in many different locations simultaneously. After a patient has been treated and discharged, however, that same imaging study may rarely need to be accessed again, and if it is, taking minutes or even hours to retrieve it may be acceptable. This pattern of use suggests that hierarchical or staged archival strategies can be implemented for optimum cost-effective use of storage technologies, particularly for the older distributed PACS architectures.

The stages or terms of storage include online or local storage, short- or intermediate-term near-line storage, long-term or offline storage, and disaster recovery or backup storage. Online storage contains information that must be available to the user immediately at the display station and, therefore, requires high-speed access. As this performance is costly, online storage is usually reserved for clinically critical data needed during a single episode of current care (i.e., three days for outpatient clinical encounters and six days on average for a typical inpatient stay). The medium best meeting online local storage needs is the standard computer magnetic disk.

Short-term or near-line storage is used to provide relevant prior or historical imaging studies for comparison during a patient's return visit. This method does not require immediate access, particularly if the data can be automatically prefetched with advanced notice of scheduled appointments. Note that most patients who do not return for continuing care within 18 months of the acute visit are unlikely to return at all. In other words, a large percentage of imaging examinations performed will never be re-reviewed after the original clinical episode, so slower access may be acceptable. As such, RAID devices work well as short-term or near-line storage, although RAID devices have nearly the high performance of a single magnetic disk but are more costly because of the controller and redundancy capabilities built in.

Long-term or permanent storage provides availability to data with advance notice for retrieval, especially when long-term storage is offline or on-the-shelf. Removable storage media devices such as OD, MOD, or tape jukeboxes are typically used for long-term storage due to their high capacity per cost characteristics. Long-term archives must cover an institution's entire medico-legal storage requirements, which vary from state to state (i.e., 5 years for general adult studies, 21 years for pediatric studies, and life for mammograms and images with orthopedic appliances for Massachusetts).

The requirements for fast retrieval of images initially followed by slower retrieval later, if at all, suggests that different types of storage devices could be used over time to archive images with cost savings. As fast retrieval times grow less important, images could be migrated to less costly, higher capacity, slower storage devices, as diagramed in Fig. 11. Software is used to handle the movement of data from one medium to another, and the strategy makes the actual physical storage device transparent to the end user. Such a strategy is known as a hierarchical storage management (HSM) scheme.

Hierarchical Storage Management and Compression. Note that data compression can be used to maximize the amount of online or near-line storage available to a PACS. Although the full resolution image data can be viewed originally for primary diagnosis, a losslessly compressed version can be sent off site to an inexpensive tape backup archive. The original data can also be wavelet lossy compressed and stored on a large RAID device for maximum cost-effective online storage and retrieval of images for review and comparison (44). This scheme is depicted in Fig. 12.

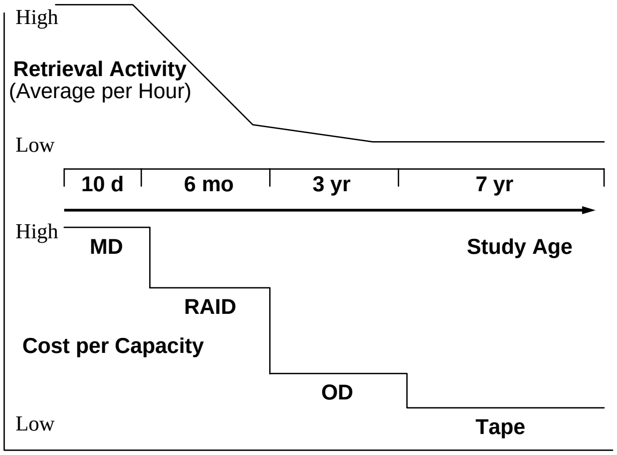


Figure 11. Migration strategy and retrieval requirements versus cost over time.

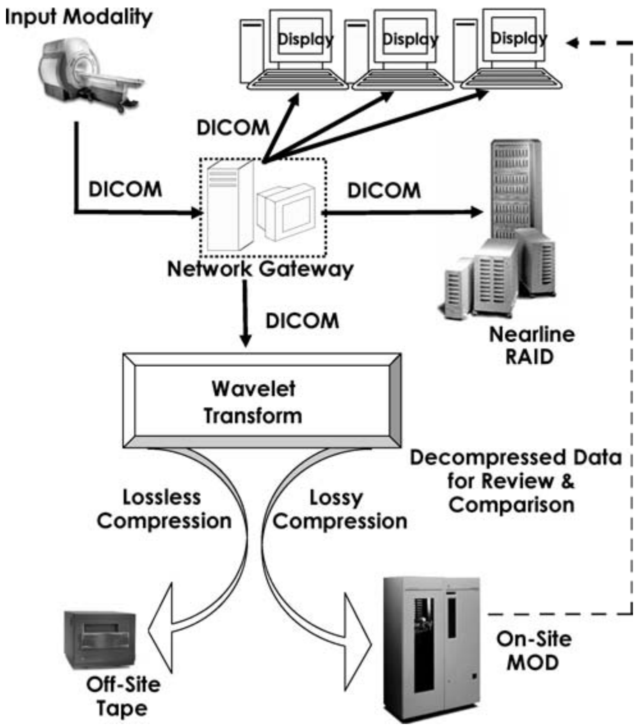


Figure 12. HSM scheme. Image data is viewed at its original content for primary diagnosis, losslessly compressed for the offsite legal record on tape, and wavelet lossy compressed for the onsite near-line storage on MOD for review and historic comparison.

An HSM scheme using short-term archival of uncompressed DICOM data for primary diagnosis, in an onsite RAID, coupled with a very deep long-term archive of diagnostic quality wavelet compressed data in an onsite optical jukebox, cost effectively maximizes online storage for immediate image retrieval. Diagnostically lossy compressed data (at ratios of 25:1 for CR, 10:1 for CT, and 5:1 for MCI) grows the onsite jukebox by 10 times row, depending on the mix of cases, making 20 or more years available online (44), which effectively maintains the entire legal record worth of original plus two relevant prior examinations all online. Note that no large-scale facility has implemented this schema, preferring to accommodate their needs by purchasing more hardware. Also it is unclear what the medico-legal ramifications would be in using lossy compressed images for historical comparison.

A hierarchical storage management scheme such as this provides a solution for maximum intermediate storage and retrieval through the use of onsite lossy compression and offsite tape backup of losslessly compressed data for the legal record and disaster recovery of data. The use of compression in this HSM scheme provides a cost-effective, high performance archive system. This HSM can be tailored to a given health-care enterprise's need to provide clinically and economically beneficial digital archival of medical images.

Other Scalable Solutions: EOL, NAS, SAN, CAS.

Everything-On-Line (EOL). With the dramatic decline in the cost and increase in capacity of RAID devices, it may become feasible to have all studies with their relevant prior examinations accessible online, which is particularly important for centralized or cacheless PACS architectures. On the other hand, imaging volume and study sizes continue to increase and may continue to overburden archive technologies. Thus, perhaps the intelligent use of the hardware and software technologies currently available through data migration schema is a sound strategy.

Networked-Attached Storage (NAS). Networked-Attached Storage (NAS) involves automated storage on a direct-access but separate local drive. NAS uses the existing local area network (LAN) to connect storage devices and the systems requiring data. A NAS server is optimized to perform file-sharing functions without the application processing overhead of typical network file servers, which enables files to be served rapidly to its clients. Performance is affected by the LAN capabilities and system configuration.

Storage Access Networks (SAN). Storage Access Networks (SAN) are dedicated networks that link storage devices and servers, creating an independent directly accessible pool of storage. SANs typically use fiber channel (FC) technology for high speed serial interconnections, usually over optical fiber. This network can provide simultaneous access from one or many servers to one or many storage devices and eliminates potential loading on the LAN.

Content-Addressed Storage (CAS). In Content-Addressed Storage (CAS) systems, data is stored and retrieved based on unique content ID keys. As medical image data is "fixed content" in that its information needs to be stored but cannot (typically) be altered in any way, CAS may prove useful. CAS associates a digital fingerprint, ID, or logical address to a stored element of data providing content security and integrity. The object-oriented nature of CAS could be exploited to improve database searchability.

Application Service Provider (ASP). An Application Service Provider (ASP) approach to medical data archival may be practical for some small entities. This strategy, in which an outside vendor provides services for storage using their hardware and software, has been around for several years. The advantages include less capital requirements for onsite hardware, technology obsolescence protection, maintenance and migration shifted to the ASP vendor, and offsite data backup. Disadvantages include potential vulnerability in performance and long-term viability of the ASP vendor, security issues, and potential high cost of service, particularly for large-volume sites.

Computer Networking

Computer networks enable communication of information between two or more physically distinct devices. They provide a path by which end user radiologists and clinicians sitting at one geographic location, for example, can access radiological images and diagnostic reports from a computer at another location. A private locally owned and controlled network (i.e., within a building or hospital) is called a LAN, whereas a network used outside of a local area is known as a wide area network or WAN. A WAN uses an external service provider and usually has lower bandwidth services than LANs. Intranet communication refers to communication across a private limited-access LAN. Internet communication is across public shared-access WANs.

Signals are transmitted via either bound media such as over cables or unbound broadcast media. Analog communications systems encode information into a continuous wave form of voltage signals, whereas digital systems encode the data into two discrete states or bits, either "0" or "1". The bits are packaged to form bytes, words, packets, blocks, and files based on a specified communication protocol. These communications standards give detailed specifications of the media, the physical connections between devices, the signal levels and timings, the packaging of the signals, and the software needed for the transport of data (46).

Serial data transmission sends digital signals one bit at a time over a single wire; the single bit stream is reassembled at the receiving end of transmission into meaningful byte-word-packet-block-file data (46). Parallel data transmission uses multiple wires to transmit bits simultaneously and, as such, provides increased transmission speeds. Synchronous communication is used in applications that require maximum speed and is carried out between two nodes that share a common clock. Asynchronous communication relies on start and stop signals to identify the beginning and end of data packets (46). An example of this technology is asynchronous transfer mode (ATM) technology.

Hardware. Important networking infrastructure considerations include bandwidth, or how much data can be transferred per period of time; latency, or how long the trip takes; topology or network segmentation, which describes the path data takes; and reliability and redundancy. Table 5 lists different types of network bandwidths or speeds available, in bits per second (bps), along with example transmission times for a single 10 MB CR digital projection radiograph. Note that over a telephone modem it would take approximately 24 minutes at best to transmit a single 10 MB image, whereas over Fast Ethernet it would

Table 5. Network Bandwidths and Example Image Transmission Times

	Maximum Bandwith (bps)	Min Transmission Time for 10 MB CR
Modem	56 kbps	23.8 min
T1 Line	1.54 Mbps	52 s
Ethernet	10 Mbps	8 s
Fast Ethernet	100 Mbps	0.8 s
ATM	155 Mbps	0.52 s
Gigabit Net	1 Gbps	0.08 s

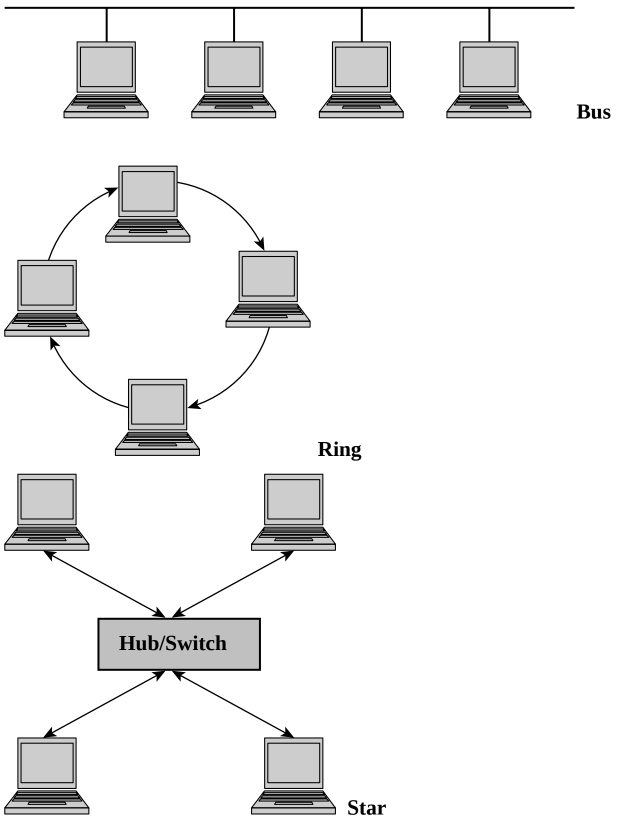


Figure 13. Network topologies typically used in PACS: bus, ring, and star.

be only a fraction of a second. LANs can consist of low speed Ethernet, medium speed Fast Ethernet, or fast speed Gigabit infrastructure. WANs can consist of a range of digital service speeds from slow telephone modem to medium speed T1 lines to fast ATM. Transmission speeds track with cost.

The most common network topologies used in PACS include bus, ring, and star configurations. These are diagramed in Fig. 13. Bus topologies commonly use Ethernet and have the advantage of network simplicity, but the disadvantage of upper-level bottlenecking and difficult to identify channel failure. The ring topology uses fiber distributed data interface (FDDI) or high speed ATM SONET (synchronous optical NETwork) ring technology. Ring topologies offer simplicity and no bottleneck, but, in a single ring, if the channel between two nodes fails, then the network is down. The star or hub topology uses high speed Ethernet or ATM switching and offers network simplicity but a bottleneck as well as a single point of failure at the hub or switch.

The physical media or cabling that makes up PACS networks varies from telephone wires to unshielded twisted pair (UTP) copper cabling, also referred to as CAT5 or CAT3, depending on the quality of the wiring, to coax cable (also known as thinnet or 10Base5), and fiber optic cabling. Fiber optic cabling can transmit more data over longer distances than conventional cabling by using light or lasers instead of electrical signals, but are of relatively high cost. The network interface card or NIC connects a computer to the physical media or cabling of the network. A unique address, the Media Access Control or MAC address is derived from the NIC. This address is used to identify each individual computer on a network.

A hub or multiport repeater connects multiple computers on a network. A bridge isolates network traffic and connects two or more networks together. The bridge listens to all network traffic and keeps track of where individual computers are. A properly located bridge can take a large congested network segment and reduce the number of data collisions, improving performance.

A switch or router can scale a small-bandwidth network to a large bandwidth. Switches tend to be protocol-independent, whereas routers are protocol-dependent. Routers or relays are used in large networks because they can limit the broadcasts necessary to find devices and can more efficiently use the bandwidth. Routers, sometimes referred to as gateways, while beneficial in large complicated environments, can unfortunately slow traffic down because it has to examine data packets in order to make routing decisions. Bridges work much more quickly because they have fewer decisions to make. Switches have revolutionized the networking industry. They look at only as much of the data packet as bridges look at and are, in a sense, bridges with many interfaces. Switching incorporated with routing helps make network bottlenecks easier to remove. Some switches are used in the core of a network whereas others are used to replace hubs. Figure 14 shows example network switch and router devices.

Figure 15 diagrams an example PACS network, the one used at the University of California at San Francisco Medical Center for transmission of radiological images and related data around the hospital LAN and the health-care center WAN. The image acquisition devices, including the various modalities, such as CT and magnetic resonance scanners, are at the top of the diagram. As image acquisition devices are only capable of connecting to a network with Ethernet speeds, a switch or router is used to take scanner outputs over 10 Mbps in and transmit that data to the PACS servers using faster speeds of 100 Mbps. Images are sent to the display stations using the fastest network available. The circle in the upper-right corner of the diagram represents the UCSF Radiology WAN over which images and information are sent to other health-care facilities over 155 Mbps ATM.

Figure 14. Example network switches and routers.

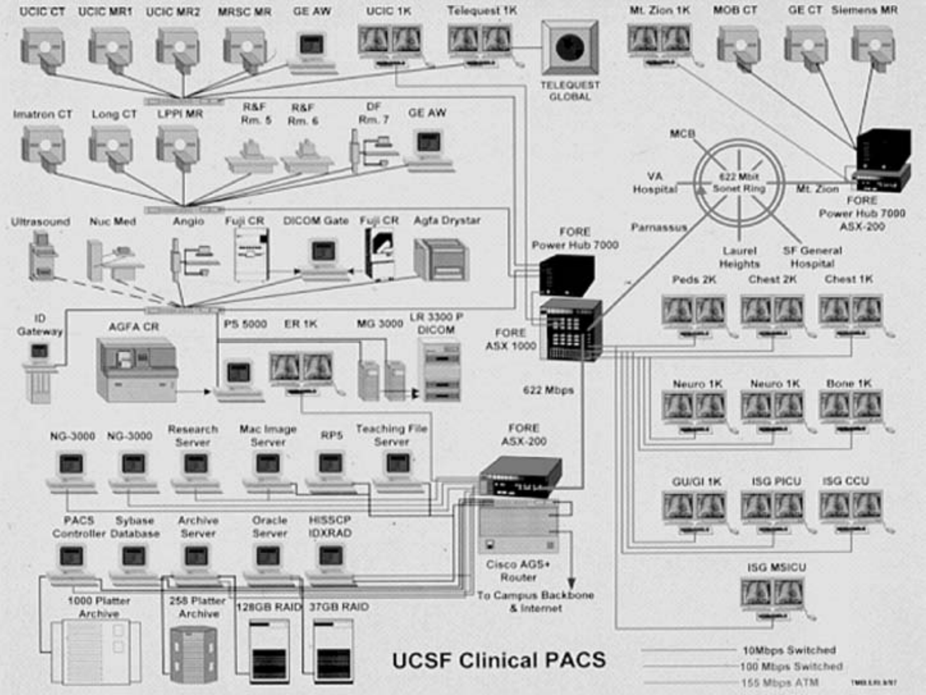


Figure 15. Example PACS network used at the University of California at San Francisco Medical Center for radiological images and related data.

Networking Software. The International Standards Organization (ISO) developed the Open Systems Interconnect (OSI) model as a framework to facilitate interoperation of computer networks from the application layer (i.e., the image viewer) all the way down to the physical layer (i.e., the wires). The ISO/OSI communication protocol stack is shown in Fig. 16. It consists of seven layers (46). Each layer in the stack is interested only in the exchange of information between the layer directly above or directly below, and each layer has different and well-defined tasks.

The top or seventh layer of the ISO/OSI stack is the Application Layer, which provides services to users. The Application Layer knows the data it wants to transmit and which machine it wants to communicate with. The sixth layer is the Presentation Layer, which takes care of data transformation such as encryption, compression, or reformatting. Layer five is the Session Layer, which controls applications running on different workstations, which is followed in the stack by the fourth or Transport Layer, which transfers data between end points and is handled here with error recovery. Layer three is the Network Layer, which establishes, maintains and terminates network connections. The second layer is the Data Link Layer, which handles network access, collision detection, token passing, and so on, and network control of logical links such as sending and receiving data messages or packets. The bottom layer or layer one is the Physical Layer corresponding to the hardware layer or the cable itself.

Also diagramed in Fig. 16 are the stacks for TCP/IP (Transmission Control Protocol/Internet Protocol) widely

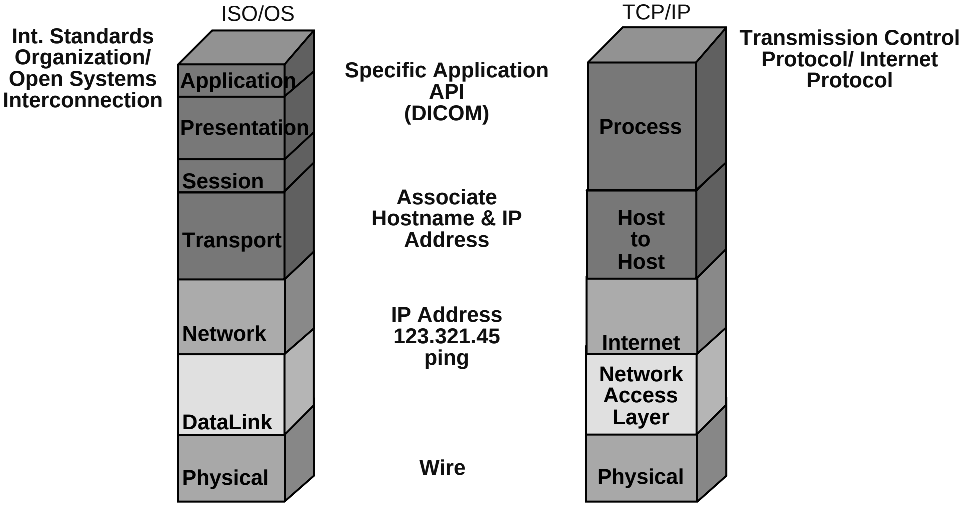

ISO/OS
Int. Standards Organization/ Open Systems Interconnection
Application
Presentation
Session
Transport
Network
DataLink
Physical
Specific Application API (DICOM)
Associate Hostname & IP Address
IP Address 123.321.45 ping
Wire
TCP/IP
Transmission Control Protocol/ Internet Protocol
Process
Host to Host
Internet
Network Access Layer
Physical

Figure 16. Communication protocol stacks for ISO/OSI and TCP/IP.

used in PACS applications. TCP/IP has four layers but is shown here as five layers, with the lowest level split out into two layers for the network access and physical layers. The top layer is called the Process Layer, and it corresponds to the Application and Presentation Layers in the ISO/OSI model. Such an application in the PACS world might be the DICOM communications protocol application. The layer below is the Host-to-Host or Transport Layer, followed by the Internet Layer and then the Network Access Layer, which encompasses Ethernet, token ring, for example, and the Physical or media Layer. To determine if two devices in a PACS network are communicating, the physical connection is tested. Using the unique IP address of the devices, a “ping” command will validate whether the one computer can reach the other over some network path. The TCP/IP hostname of the computer relates the unique IP address of the device to its DICOM AE (application entity) title so that computers can communicate using this protocol.

Figure 17 demonstrates the path messages take from one subnet, through a router, to another subnet. The process starts at the top of the stack in host A, where a DICOM port is opened. The information travels down the stack through the TCP Host-to-Host Layer to the Internet (IP) Layer and out the Network Access Layer across the Physical Layer. Messages then pass from the Physical Layer up the stack to the Network Access Layer, then the Internet Layer, to the Host-to-Host or Transport Layer, and, finally, a port is opened in the top Processor Application Layer.

Security and Redundancy. Advances in networking and communications devices have greatly facilitated the transfer of information between computers all over the world. For the secure transmission of private information such as clinical images and patient data, additional technologies are put into place to make the Internet a safe medium. The Health Insurance Portability and Accountability Act (HIPAA) of 1996 required the Department of Health and Human Services to establish national standards for electronic health-care transactions and national identifiers for providers, health plans, and employers. It also addresses the security and privacy of health data, such that adopting these standards will improve the efficiency and effectiveness of the nation’s health-care system by encouraging the widespread use of electronic data interchange in health care.

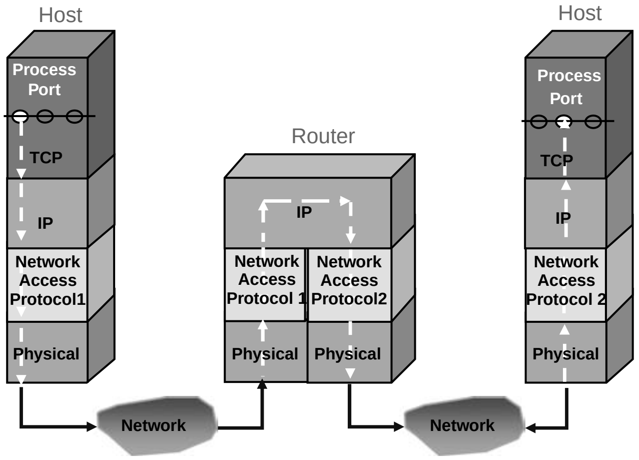


Figure 17. Connecting subnets via a router and DICOM.

Firewalls are protocol-dependent devices often used to secure a network by filtering traffic based on rules and policies. Intrusion detection systems (IDS) are another form of security device that uses policy-based monitoring, event logging, and alarms and alerting messaging to protect a network. Virtual Private Networks (VPNs) protect data by encrypting the information at the source device and decrypting it at the destination device. VPN clients are often used to securely access a hospital network for a location outside its LAN. A path can be created through the firewall or directly to a specific server enabling transmission of data.

PACS networks have become mission-critical devices in the transmission of digital medical images in and among health-care enterprises. As such, the networks must be highly available and have fault-tolerance mechanisms built in. Requirements for high availability networks include having redundant technology with automatic failover when devices are down. Redundant media and multiple paths should exist for the same information to get from one place to another. Redundant power for the devices involved is generally considered routine, as is proactive monitoring and problem mitigation with automated fault-detection processes in place.

Medical Image Display

Hardware – Monitors: CRT versus LCD. The choice of diagnostic display monitors was relatively straightforward for early adopters of PACS. Hardware was of a single type—cathode ray tube (CRT) technology—and was usually oriented in portrait mode emulating the shape of film. Monitors had high brightness, typically 200 to 300 candelas per square meter (cd/m^2), relative to other computer and television monitors, and high refresh rates of greater than 72 Hz to reduce flicker visible to the human eye. The devices themselves were physically large, heavy, and expensive. They generated noticeable quantities of heat while consuming relatively high amounts of power, and their display quality degraded quickly in time, requiring frequent monitor replacement.

Early medical-grade monitors were available in two spatial resolutions (high and low) reflecting their pixel matrix sizes (2 k or 2048 columns by 2500 rows and 1 k or 1024 columns by 1280 columns, respectively). Medium resolution 1.5 k monitors of 1500 columns by 1500 rows were later added to the mix. As a result of the exponentially higher cost of 2 k monitors as compared with 1 k monitors, radiology departments typically had a combination of a few strategically placed high resolution displays and many low or medium resolution displays. The American College of Radiology (ACR) recommended that 2 k monitors be used for primary diagnosis of digital projection radiographs because a single image could be displayed per monitor in its full inherent acquired spatial resolution. The cross-sectional modalities with slice matrix sizes of 512 by 512 for CT and 256 by 256 for MRI were considered adequately displayed on 1 k monitors. As display application software and graphical user interfaces (GUIs) improved, many radiologists became comfortable reading from 1 k monitors even for projection radiography, as long as the images were acquired at their full spatial and contrast resolutions and the GUIs allowed for easy manipulation, magnification, and comparison of images.

Today, a richer array of hardware technologies exist for the purposes of displaying digital medical images. Unfortunately, no formally defined standards or specification guidelines currently exist to clarify choices of monitors for users. The different display devices available today and the specifications to consider when purchasing monitors for use in radiological imaging are described. An explanation of the monitor types including CRTs, active-matrix liquid crystal displays (AM-LCDs), and plasma technologies is given along with a discussion of spatial resolution capabilities and requirements, contrast resolution and monitor luminance, the orientation or shape and number of displays necessary, and a comparison of color versus monochrome or gray-scale monitors. Device calibration and quality assurance practices are also addressed.

Two technology types of hardware displays are currently used in medical imaging, the half-century-old mature cathode ray tubes and what the popular literature refers to as flat-panel technology, of which several types exist (47). Of the two broad categories of flat-panel displays, one filters reflected light or light from a source behind the filter, whereas the second type creates light by exciting a phosphor. Note that the term flat panel is not meant to refer to the face of the monitor as some CRTs have a flat face (48). Rather, it refers to the thin-film transistor array panel that addresses each pixel.

CRTs produce light by exciting a phosphor-luminescent coating with a focused electron beam. Light is generated in an emissive structure, where it diffuses in a controlled manner forming the displayed image. The highest spatial resolution CRT monitors available have a display area of 2048 by 2560 or roughly 5 million (mega) pixels (Mpixels). They come in low- and high-bright versions of 50–60 FL and 100 FL or greater, respectively. High and low resolution (2 k and 1 k) monitors typically come in the portrait mode with a 9:16 or 3:4 aspect ratio emulating the shape of film. Most medium resolution CRTs (1.5 k) are square or in the landscape mode with an aspect ratio of 16:9 or 4:3. Choice of portrait versus landscape monitor shape is a function of personal user preference, with no technical issues bearing on the issue.

The flat-panel display type predominantly used in medical imaging is the active-matrix liquid crystal display (AM-LCD). LCDs use a transistor-driven matrix of organic liquid crystals that filter reflected light. LCDs use a light-modulating as opposed to a light-emitting mechanism for creating the display image. Polarization controls the light intensity such that the maximum intensity is perpendicular to the LCD panel. Consequently, this technology suffers from marked variations in luminance and contrast depending on viewing angle (47), which is the off-axis viewing or angle-of-regard problem in which images can appear quite different if viewed from different angles or heights above and below the center axes of the screen. Newer LCD designs have a more uniform luminance and contrast profile within a larger viewing angle cone (some as high as 170°). Users should inquire about the horizontal and vertical viewing angle capabilities and, better yet, ask the vendor for a demonstration monitor for clinician testing. LCDs typically have the capability to display in portrait and landscape modes.

Plasma display panels (PDPs) are currently being developed largely for high definition television (HDTV) viewing with 37 inch or larger screens. A current passed through ionized gas (Ne-Xe) contained between a pair of glass layers causes emission of ultraviolet light that excites visible light-emitting phosphors to produce the display image. PDPs are very expensive and have roughly the same number of addressable pixels as 17 inch LCDs, can be hung on a wall, have a wide viewing angle with no loss of quality, and have high brightness but relatively slow response time (48). They are not used currently in medical imaging because of their high cost, slow refresh rates, ghosting artifacts, and contrast limitations. Other types of displays, such as field-emissive display (FEDs) and organic light-emitting diodes (OLEDs), are undergoing heavy developmental efforts but are not yet viable for medical imaging display purposes (48).

Although the spatial resolution terminology used for both CRTs and LCDs is based on the device pixel matrix dimensions — 1 k to 2 k for CRTs and 3, 5, and 9 Mpixels for LCDs — not all monitors are created equal. For example, 1 k and 2 k CRT monitors tend to have standard diagonals so that the larger pixel matrix size connotes smaller pixel size and, hence, better spatial resolution capabilities, and all 1 k monitors have had equivalent spatial resolution, as did all 2 k monitors, which is not the case for LCD displays. For example, 3 Mpixel monitors come with different sized diagonals, that is, different physical sizes such that the physically bigger monitor actually has larger pixel size and hence poorer spatial resolution. Users need to understand what the pixel or spot size is, because it directly reflects spatial resolution and perception of fine detail, and not necessarily choose the largest screen. Pixel size can be determined from the physical screen size, typically given as a display area diagonal in inches and total pixel count or horizontal and vertical matrix resolution. Often, vendors give the device pixel density or pixel pitch spacing and, to confuse the issue, it is often given in millimeters. As for comparison between CRT and LCD monitors, the 1 k CRTs at 1024 by 1280 correspond to 1 Mpixel monitors, 1500 by 1500 correspond to 2 Mpixel monitors, and 1760 by 1760 correspond to 3 Mpixel displays. The 2 k or 2048 by 2500 CRTs correspond to the 5 Mpixel LCD. The recently introduced 9 Mpixel LCD display has 200 pixels per inch on a 22 inch diagonal screen.

The brightness of a monitor or its luminance affects perceived contrast or the number of discernable gray levels. Studies have shown that diagnostic accuracy increases as monitor luminance increases. To gain a perspective on luminance values, the typical lightbox or alternator used to display film is on the order of 400 to 600 FL (1360 to 2040 cd/m^2), whereas the standard PC color monitor is roughly 20 to 40 FL (68 to 136 cd/m^2). An LCD color monitor has 65 to 75 FL (221 to 255 cd/m^2) monitor luminance, whereas the low-bright medical-grade CRT monitors have 50 to 60 FL (170 to 204 cd/m^2) and the high-bright CRTs have 100 FL (340 cd/m^2) or greater monitor luminance. Among the device specifications reflecting monitor brightness and affecting contrast resolution are the monitor and display card bit depth (typically 8 bits for 256 potential gray values) and the monitor dynamic range or contrast ratio reflecting the maximum discernable luminance over the minimum, with typical values of 600:1 or greater.

In comparing CRT versus LCD display technologies, the advantages of LCD over CRT monitors include better stability for longer device lifetime. The change in brightness of standard LCD monitors has been measured at less than 0.5% per month (47). LCDs are not prone to the geometric distortion typical of CRTs, they tend to consume less power, and this have reduced sensitivity and reflection artifacts from ambient room lighting. Disadvantages of LCD monitors versus CRTs include the afore-mentioned off-axis viewing or angle-of-regard distortion of LCDs, backlight instabilities, liquid crystal fluctuations with temperature, and manufacturing defects creating dead or nonresponsive pixel areas.

Receiver Operating Characteristic (ROC) studies are currently the best methodology available to compare monitor quality and associate it with reader performance, that is diagnostic accuracy, sensitivity, and specificity. Numerous clinical studies have been performed, most showing no significant difference between diagnostic performance on CRTs and LCDs. Recent studies representative of CRT versus LCD comparison for radiological diagnosis include one that examined brain CTs for identifying early infarction (49) and the other looked at CRs of the chest for the evaluation of interstitial lung disease (50). The CT ROC study showed no statistically significant differences in diagnostic performance between a 21 inch monochrome CRT monitor with a pixel matrix of 1280 by 1600 and a brightness of 175 FL versus an 18 inch color LCD monitor with a pixel matrix of 1024 by 1280 and a luminance of 55 FL, when 10 radiologists were asked to rate the presence or absence of disease on a 5 point scale. Similarly, an ROC study comparing the efficacy of a 5 Mpixel CRT display versus a 3 Mpixel LCD for the evaluation of interstitial lung disease in digital chest radiography showed no statistically significant change in observer performance sensitivity between the two types of monitors.

Several studies have investigated the comparison of color versus monochrome (technically achromatic) or gray-scale monitors, and a clear consensus does not seem to exist, which is an important issue because color monitors tend to have decreased luminance, contrast, and spatial resolution capabilities than monochrome monitors, and the human visual system has decreased spatial resolution perception in the color channels, but greater dynamic range (500 just-noticeable-differences (JND) versus 60 to 90 JNDs in gray-scale). On the other hand, high performance monochrome monitors are expensive and have a relatively short lifetime of approximately 3 years, and color is becoming increasingly useful in diagnostic imaging with the emergence of 3D display renderings. Although a study comparing monochromatic versus color CRT monitors found no statistically significant differences in display of CR chest images for the detection of subtle pulmonary disease, they did find higher sensitivity rates for specialty chest radiologists on the monochromatic monitor, perhaps due to the lower maximum luminance levels of the color displays (51). Another study comparing pulmonary nodule detection on P45 and P104 monochrome and color 1600 by 1200 pixel monitors found significantly greater false-positive and false-negative responses with the color monitors as well as longer search times (52). So, for primary diagnosis of projection radiographs in particular, monochrome monitors may still be the way to go. Note, however, that users prefer color LCDs when compared with color CRTs. This fact may be related to the Gaussian spot pixel and emissive structure of CRTs and the use of black matrix (shadow mask or aperture grille), which separates the red-green-blue phosphor dots that form an arrangement of color dots or stripes for luminance and chromatic contrast (47). Grille misalignment can degrade color purity and contrast.

Early PACS adopters equipped their radiology reading rooms with the highest quality display monitors, some 2 k, others 1 k, but all high brightness. The software applications were more complex than those targeted for the nonradiologist enterprise user. It was common to provide an intermediate application for use by image-intensive specialists such as orthopedists, neurosurgeons, and radiation oncologists as well as in image-intensive areas such as emergency departments and ICUs. Lesser quality monitors with stripped-down software capabilities were used by enterprise image users. It is interesting to note that as display hardware and software continue to evolve, display application software moves toward melding into one flexible easily configurable GUI, and one monitor type may, in time, meet most needs.

Monitor calibration and QA practices are important to maintaining high performing medical displays. The DICOM 14 Gray-scale Standard Display Function (GSDF) and the AAPM (American Association of Physicists in Medicine) Task Group 18 recommend that monitors be calibrated to a perceptually linearized display function as this is matched to the perceptual capabilities of the human visual system. Monitors forced to follow these standards produce more uniform images with optimum contrast. CRTs are less stable than LCD monitors requiring luminance calibration and matching to be conducted monthly, physically measuring light levels with a photometer. Many LCD displays have embedded luminance meters for automated QA measurements, Although some studies have also recommended doing external luminance measures but less frequently. LCDs must still be manually inspected for nonresponsive pixels. Routine manual viewing of test patterns, such as the SMPTE (Society of Motion Picture and Television Engineers) Test Pattern, are usually sufficient for evaluating overall monitor performance, low contrast, and fine detail detection.

Software Functionality. How many individual monitors does a user need per display workstation — 1, 2, 4, or 8?

Many feel that for primary diagnosis, dual-headed configurations are most efficient for comparison of current and prior relevant studies, particularly for projection radiographs. Note that a good GUI design can reduce the need for multiple monitors. The ability to page through and move images around the screen, the ability to instantaneously switch between tile and stack or cine modes of display, and the ability to view multiple studies on one monitor as well as side-by-side comparison of studies are critical to reducing the amount of hardware and physical display space required. In most cases, the two-monitor setup is sufficient for primary diagnosis and image intensive use with perhaps a third (color) monitor for worklist creation and access to other relevant medical data. The most common configuration for enterprise users is the single-headed or one-monitor display.

First and foremost, a software GUI must be intuitive and easy to use. The software application must be responsive, robust, and reliable. Most display workstations have GUIs to perform two basic functions. The first is to deliver a patient list or worklist of imaging examinations to be read, for example, "today's unread emergency department CTs." The worklist environment allows users to interrogate the entire PACS database for a subset of cases with which they wish to work. Typically, the database can be searched for by a specific patient name or other identifier, for individual imaging modalities, over specified time frames, by imaging or patient location within the hospital or health-care enterprise, and so on. The second basic function workstations perform is study display and image manipulation.

Automated hanging protocols based on examination type, the existence of prior historical examinations, and so on can greatly enhance radiology interpretation workflow. For example, if the current imaging study requiring interpretation is a two-view chest projection radiograph, then the default display might be to place the posterior-anterior (PA) view on the left display monitor and the lateral view on the right. If the patient has had a prior chest X ray, then the current PA should be placed on the left monitor with the lateral view behind and the prior PA on the right monitor with its corresponding lateral view waiting behind. If the current study is an MR of the brain, then automatically hang each sequence as a separate stack and place multiple (i.e., four-on-one) stacks per monitor so that they can be cined through simultaneously.

Basic display manipulation tools include the ability to dynamically change the window and level or contrast and brightness of the displayed image, the ability to magnify a portion of the image or zoom and pan through the entire image, and monitor configuration and image navigation tools such as paging, cine, and linked stack modes. Image mensuration capabilities, including linear, angle, and region-of-interest measurements, are also typical. Some advanced tools include the ability to track on a corresponding perpendicular slice, in MR studies, for example, where the cursor is on the perpendicular view for 3D referencing. Figure 18 depicts several example softcopy image displays on PACS workstations.

Well-designed PACS are tightly integrated with other information systems such as the hospital or radiology information system, enabling access to other relevant data about the patient or imaging examination from within the display application. Sometimes diagnostic radiology report generation systems, such as speech recognition, are embedded in the diagnostic workstation.

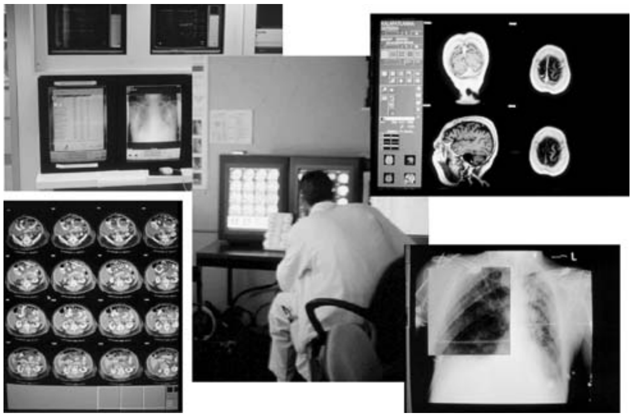


Figure 18. Example soft copy image display on PACS workstations.

It is anticipated that the number of specialty display applications available on display stations will continue to increase as more features are added. Examples of the results of several advanced processing techniques are shown in Fig. 19. Some systems also provide algorithms for image processing such as image sharpening, or edge enhancement, and image smoothing. Maximum intensity projection (MIP) displays and multiplanar reformats (MPR) are also appearing on PACS workstations. Real-time 3D reconstruction of image data at the PACS display is beginning to be seen. Multimodality image data fusion such as CT and positron emission tomography (PET) images to depict the functional maps overlaid on the anatomical data maps will also likely be viewable.

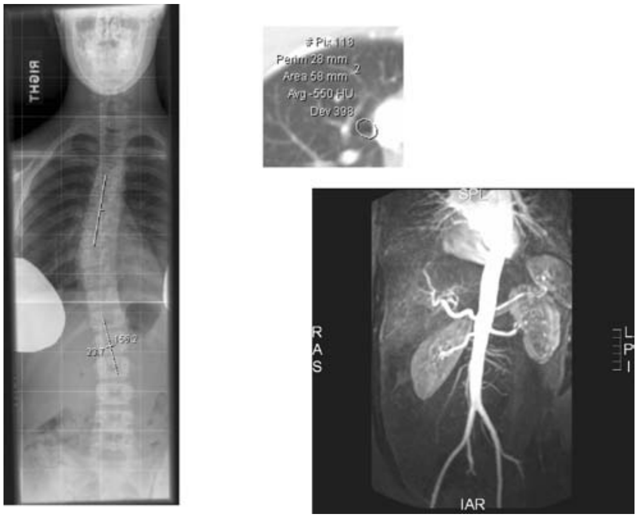


Figure 19. Example advanced processing techniques: calculation of scoliosis Cobb angle, region-of-interest measurements, 3D magnetic resonance angiogram.

CURRENT TRENDS AND FUTURE PROSPECTS

Medical imaging is increasingly the key triage event in a patient's encounter with the health-care system. The ability to eliminate the use of film in radiological imaging is a reality today. In addition to the economic advantages of using PACS for digital medical imaging, rapid access to all clinical information on patients, including imaging studies, anytime and anywhere with security, enhances the quality of patient care.

Hurdles still exist today. PACS are not just a radiological tool. Images are required enterprise-wide by many different types of clinicians and other health-care providers. Images are required for viewing in emergency departments, ICUs, surgical ORs, outpatient clinics and referring physicians' offices as well as for teaching conferences, and at home for viewing by patients and clinical providers. Unless PACS can also deliver images to everyone in the health-care enterprise who requires them, film cannot be turned off, and facilities will have to operate in a painful mixed environment.

Web-based enterprise PACS applications do exist and continue to improve in their performance. Note that web-based PACS for within the radiology department are also becoming more common. Several requirements in the enterprise environment make transition to the all-digital, totally filmless medical imaging facility more difficult than just within the radiology department. The web-PACS enterprise application user interface or GUI must be very intuitive—like the rental car model. Most licensed drivers are able to get into a rental car, orient themselves to where the lights and windshield wipers are, and then go ahead and drive the car. They do not need to know how a car works in order to drive it, and the user interface is very standard across most all cars. The car works robustly and reliably and the user does not need to read a manual before they can drive the car. Unfortunately, the state-of-the-art in computer GUIs is not quite as intuitive, but much progress has been and continues to be made in this area, making GUIs self-training and bullet-proof. Human-computer interfacing and interaction along with GUI design are currently active areas of research and discovery.

Web-PACS applications are often required to operate in mixed-platform environments to accommodate PC, Macintosh, and Unix boxes, which is sometimes problematic. Applications must be improved to be able to handle bottlenecks in the system at both the input to the database and the output to multiple users accessing the system simultaneously. The best applications are purely web-based; that is, they are not dependent on having to download Active-X components to every physical machine that uses the system.

In summarizing the key components and essential features for clinical implementation of a PACS, at acquisition, all image acquisition devices or imaging modalities should be required to conform to the DICOM standard image format and communications protocol. Devices and PACS in general operate best when well interfaced to other clinical information systems such as the HIS-RIS, report generation systems such as speech recognition applications, computerized physician order-entry (CPOE), and decision support systems. The inherently digital imaging modalities should be acquired into a PACS using direct DICOM capture. Film digitizers such as laser scanners can be used to acquire imaging examinations existing only on film into a PACS if required. Acquisition of digital projection radiographs, such as the conventional chest X ray, can be achieved using CR or photostimulable phosphor devices or digital radiography devices, which directly convert images to digital at the time of X ray exposure. CR and DR devices are likely to coexist for some time.

Archival media and devices will continue to advance, with databases becoming more patient-centric and seamlessly searchable. Computer networking will also improve in not only the hardware devices, but also the network management software strategies. Display GUIs must continue to become more intuitive and robust, and the monitors themselves will trend toward LCD devices as opposed to CRTs.

Predictors for success of the introduction of new technologies into the clinical arena include ease of integration into the existing workflow or change management activities to optimize new workflow with new devices. Systems must be reliable, simple to use, and have optimum performance so that processes can be completed more efficiently than in the analog film-based world. Systems must be flexible and configurable on a per-user basis and they must include fault-tolerance, redundancy, and error-tracking capabilities.

In the future, radiology can help to drive the changes that will greatly impact all of health care. Medical image management systems are maturing outside of the radiology department, providing hospital enterprise-wide access to clinical images and information via the Intranet as well as web access from outside the enterprise via a secure Internet connection. The web will change the way people communicate and perform their duties. Web-based PACS applications will become the norm offering ubiquitous distribution of and access to clinical data. Computer workstations will become less costly, more reliable, and have more intuitive GUIs. All relevant medical information systems will become more tightly integrated with each other, sharing and maintaining user, patient, image, and application sensitivity such that multiple distinct applications perform virtually as one.

Future PACS are likely to be on the less expensive off-the-shelf PC platform using industry standards. PACS display stations are likely to be configured with fewer numbers of monitors—two within radiology and image-intensive specialty areas and one out in the enterprise. Small and medium sized community hospitals, private practices, outpatient centers in rural areas, and some indigent care facilities will begin realizing the benefits of PACS and digital medical imaging through better access to high quality diagnostic imaging services.

PACS functionality will be incorporated into handheld devices for some applications, and wireless transmission will mature in the clinical arena. Improved integration with other information technologies into the total electronic medical record (EMR), including automated speech recognition systems, will enable a more efficient filmless environment as well as a paperless workflow.

Advanced image processing utility and translation from the research environment to clinical applications will increase. 3D displays, the use of color and video will increase as will the incorporation of computer-aided detection and decision support through outcomes research and evidence-based medicine will become more prevalent. Multimodality functional and molecular imaging will mature clinically, and value-added applications for specialties outside of diagnostic imaging will increase. Virtual reality imaging presentations and image-guided surgery applications are likely to become more commonly used clinically.

It is likely that the radiological interpretation process will need to transform in order to handle the information and image data overload currently plaguing medical imaging. This image interpretation paradigm shift will be required in order to evaluate, manage, and exploit the massive amounts of data acquired in a more timely, efficient, and accurate manner. Discovery and development of this new paradigm will require research into technological, environmental, and human factors. Interdisciplinary research into several broad areas will be necessary to make progress and ultimately to improve the quality and safety of patient care with respect to medical imaging. These areas are likely to include studies in human perception, image processing and computer-aided detection, visualization, navigation and usability of devices, databases and integration, and evaluation and validation of methods and performance. The result of this transformation will affect several key processes in radiology, including image interpretation, communication of imaging results, workflow and efficiency within health-care enterprises, diagnostic accuracy and a reduction in medical errors, and, ultimately, the overall quality of patient care (53).

Twentieth century medical imaging was film-based in which images were interpreted on analog viewboxes, film was stored as the legal imaging record. Film had to be manually disturbed from one location to another and could be accessed in only one physical location at a time. Twenty-first century medical imaging will be characterized by digital image acquisition, softcopy computer interpretation, digital image archives, and electronic distribution. It is anticipated that the use of PACS and other information technology tools will enable the filmless, paperless, errorless era of imaging in medicine.

BIBLIOGRAPHY

1. Lemke HU, Stiehl HS, Scharnweber H, Jackel D. Applications of picture processing, image analysis and computer graphics techniques to cranial CT scans. Proceedings of the Sixth Conference on Computer Applications in Radiology and Computer Aided Analysis of Radiological Images; Los Alamitos, CA: IEEE Computer Society Press; 1979. 341–354.
2. Capp ML, et al. Photoelectronic radiology department. Proc SPIE–Int Soc Opt Eng 1981;314:2–8.
3. Dwyer III, SJ, et al. Cost of managing digital diagnostic images. Radiology 1982;144:313.
4. Duerinckx A., editor. Picture archiving and communication systems (PACS) for medical applications. First International Conference and Workshop, Proc SPIE; 1982; 318, Parts 1 and 2.
5. Huang HK, et al. Digital radiology at the University of California, Los Angeles: A feasibility study. Proc SPIE 1983; 418:259–265.
6. Blaine GJ, Hill RL, Cox JR, Jost RG. PACS workbench at mallinckrodt Institute of Radiology (MIR). Proc SPIE 1983; 418:80–86.
7. Seshadri SB, et al. Prototype medical image management system (MIMS) at the University of Pennsylvania: Software design considerations. Proc SPIE 1987;767:793–800.
8. Kim Y, Fahy JB, DeSoto LA, Haynor DR, Loop JW. Development of a PC-based radiological imaging workstation. Proc SPIE 1988;914:1257–1264.
9. Horii SC, et al. Environmental designs for reading from imaging workstations: Ergonomic and architectural features. Proc SPIE 1989;1091:172–178.
10. Arenson RL, Chakraborty DP, Seshadri SB, Kundel HL. The digital imaging workstation. Radiology 1990;176:303–315.
11. Hruby W, Maltsidis A. A view to the past of the future – A decade of digital revolution at the Danube Hospital. In: Hruby W, editor. Digital (R)evolution in Radiology. Vienna: Springer Publishers; 2000.
12. DICOM. 2004. Online. Available at http://medical.nema.org/ .http://medical.nema.org/dicom/2004/.
13. Andriole KP. Anatomy of picture archiving and communication systems: nuts and bolts – image acquisition: getting digital images for imaging modalities. Dig Imag 1999;12(2) Suppl 1: 216–217.
14. Andriole KP, Avrin DE, Yin L, Gould RG, Arenson RL. PACS databases and enrichment of the folder manager concept. Dig Imag 2000;13(1):3–12.
15. Andriole KP. Computed radiography overview. In: Seibert JA, Filipow LJ, Andriole KP, editors. Practical Digital Imaging and PACS. Medisson, WI: Medical Physics Publishing; 1999. p 135–155.
16. Bogucki TM, Trauernicht DP, Kocher TE. Characteristics of a storage phosphor system for medical imaging. Kodak Health Sciences Technical and Scientific Monograph. No. 6, New York: Eastman Kodak Co.; July 1995.
17. Barnes GT. Digital X-ray image capture with image intensifier and storage phosphor plates: Imaging principles, performance and limitations. Proceedings of the AAPM 1993 Summer School: Digital Imaging; Charlottesville, VA: University of Virginia; Monograph 22: 23–48.
18. Wilson AJ, West OC. Single-exposure conventional and computed radiography: The hybrid cassette revisited. Invest Radiol 1993;28(5):409–412.
19. Andriole KP, Gooding CA, Gould RG, Huang HK. Analysis of a high-resolution computed radiography imaging plate versus conventional screen-film radiography for neonatal intensive care unit applications. SPIE Phys Med Imag 1994;2163: 80–97.
20. Kodak. Digital radiography using storage phosphors. Kodak Health Sciences Technical and Scientific Monograph. New York: Eastman Kodak Co.; April 1992.
21. Matsuda T, Arakawa S, Kohda K, Torii S, Nakajima N. Fuji Computed Radiography Technical Review. No. 2. Tokyo: Fuji Photo Film Co., Ltd.; 1993.
22. Berg GE, Kaiser HF. The X-ray storage properties of the infrared storage phosphor and application to radiography. Appl Phys 1947;18:343–347.
23. Luckey G. Apparatus and methods for producing images corresponding to patterns of high energy radiation. U.S. Patent 3,859,527. June 7, 1975. Revision No. 31847. March 12, 1985.
24. Kotera N, Eguchi S, Miyahara J, Matsumoto S, Kato H. Method and apparatus for recording and reproducing a radiation image. U.S. Patent 4,236,078. 1980.

25. Sonoda M, Takano M, Miyahara J, Kato H. Computed radiography utilizing scanning laser stimulated luminescence. Radiology 1983;148:833–838.
26. Agfa. The highest productivity in computed radiography. Agfa-Gevaert N.V. Report. Belgium: AGFA; 1994.
27. Ogawa E, Arakawa S, Ishida M, Kato H. Quantitative analysis of imaging performance for computed radiography systems. SPIE Phys Med Imag 1995;2432:421–431.
28. Kodak. Optimizing CR images with image processing: Segmentation, tone scaling, edge enhancement. Kodak Health Sciences Technical and Scientific Manuscript. New York: Eastman Kodak; March 1994.
29. Gringold EL, Tucker DM, Barnes GT. Computed radiography: User-programmable features and capabilities. Dig Imag 1994;7(3):113–122.
30. Ishida M. Fuji computed radiography technical review, No. 1. Tokyo: Fuji Photo Film Co., Ltd.; 1993.
31. Storto ML, Andriole KP, Kee ST, Webb WR, Gamsu G. Portable chest imaging: clinical evaluation of a new processing algorithm in digital radiography. 81st Scientific Assembly and Annual Meeting of the Radiological Society of North America. Chicago, IL: November 26 – December 1, 1995.
32. Vuylsteke P, Dewaele P, Schoeters E. Optimizing Radiography Imaging Performance. Proceedings of the 1997 AAPM Summer School; 1997; 107–151.
33. Solomon SL, Jost RG, Glazer HS, Sagel SS, Anderson DJ, Molina PL. Artifacts in Computed Radiography. AJR 1991;157: 181–185.
34. Volpe JP, Storto ML, Andriole KP, Gamsu G. Artifacts in chest radiography with a third-generation computed radiography system. AJR 1996;166:653–657.
35. Oestman JW, Prokop M, Schaefer CM, Galanski M. Hardware and software artifacts in storage phosphor radiography. RadioGraphics 1991;11:795–805.
36. Lee DL, Cheung LK, Jeromin LS. A new digital detector for projection radiography. Proc SPIE Phys Med Imag 1995;2432: 237–249.
37. Andriole KP. Productivity and cost assessment of CR, DR and screen-film for outpatient chest examinations. Dig Imag 2003;15(3):161–169.
38. Pratt HM, et al. Incremental cost of department-wide implementation of a PACS and computed radiography. Radiology 1998;206:245–252.
39. Chunn T. Tape storage for imaging. Imag World 1996;5(8): 1–3.
40. Horii S, et al. A Comparison of case-retrieval times: Film versus PACS. Dig Imag 1992;5(3):138–143. htpp://www.diagnosticimaging.com.
41. Eisenman, et al.. Diagnost Imag 1996;9:27.
42. Siegel E, Shannon R. Understanding Compression. Great Fall, VA: SCAR Publications; 1997; 11–15.
43. Erickson BJ, Manduca A, Palisson P, Persons KR, Earnest F 4th, Savcenko V, Hangiandreou NJ. Wavelet compression of medicl images. Radiology. 1998;206(3):599–607.
44. Avrin DE, et al. Hierarchical storage management scheme for cost-effective on-line archival using lossy compression. Dig Imag 2001;14(1):18–23.
45. Savcenko V, Erickson BJ, Palisson PM, Persons KR, Manduca A, Hartman TE, Harms GF, Brown LR. Detection of subtle abnormalities on chest radiographs after irreversible compression. Radiology 1998;206(3):609–616.
46. Huang HK. PACS and Imaging Informatics: Basic Principles and Applications. Hoboken, NJ: John Wiley & Sons, Inc.; 2004.
47. Badano A. Principles of cathode-ray tube and liquid crystal display devices. In: Samei E. Flynn MJ, editors. Advances in Digital Radiography: Categorical Course in Diagnostic Radiology Physics. Oak Brook, IL: Syllabus, RSNA; 2003. pp. 91–102.
48. Leachtenauer JC. Electronic Image Display: Equipment Selection and Operation. Bellingham, WA: SPIE Press; 2004.
49. Partan G, et al. Diagnostic performance of liquid crystal and cathode-ray-tube monitors in brain computed tomography. Eur Radiol 2003;13:2397–2401.
50. Langer S, et al. Comparing the efficacy of 5-MP CRT versus 3-MP LCD in the evaluation of interstitial lung disease. Dig Imag, Online publication date, June 29, 2004.
51. Iwano S, et al. Detection of subtle pulmonary disease on CR chest images: Monochromatic CRT monitor vs color CRT monitor. Eur Radiol 2001;11:59–64.
52. Krupinski E, Roehrig H. Pulmonary nodule detection and visual search: P45 and P104 monochrome versus color monitor displays. Academ Radiol 2002;9:638–645.
53. Andriole KP, et al. Addressing the coming radiology crisis: Transforming the radiological interpretation process. Dig Imag Online October 2004.

Further Reading

Andriole KP, Gould RG, Avrin DE, Bazzill TM, Yin L, Arenson RL. Continuing quality improvement procedures for a clinical PACS. *Dig Imag* 1998;11(31):111–114.

Honeyman JC, et al. PACS quality control and automatic problem notifier. *SPIE Med Imag 1997: PACS Design and Evaluation* 1997;3035:396–404.

Honeyman JC, Staab EV. Operational concerns with PACS implementations. *Appl Radiol* 1997; August: 13–16.

Seibert JA. Photostimulable phosphor system acceptance testing. In: Seibert JA, Barnes GT, Gould RG, editors. Specification, Acceptance Testing and Quality Control of Diagnostic X-ray Imaging Equipment. Medical Physics Monograph no. 20. Woodbury, NY: AAPM; 1994. 771–800.

Willis CE, Leckie RG, Carter J, Williamson MP, Scotti SD, Norton G. Objective measures of quality assurance in a computed radiography-based radiology department. *Proc SPIE* 1995; 2432:588–599.

See also Photography, medical; radiology information systems; ultrasonic imaging.

PIEZOELECTRIC SENSORS

Yanbin Li
Department of Biological and
Agricultural Engineering,
University of Arkansas
Fayetteville, AR

Xiao-Li Su
BioDetection Instruments LLC
Fayetteville, AR

INTRODUCTION

Piezoelectric sensors are generally referred to as analytical devices for detection of chemical or biological agents with piezoelectric quartz crystals (PQCs) as transducers. The origin of piezoelectric sensors can be traced back to 1880 when Jacques and Pierre Curie discovered normal and converse piezoelectric effects (1). In the former, the application of a mechanic stress to the surface of quartz and some other crystals induces an electric potential across the crystal; in the latter, conversely, the application of a voltage across the crystal results in an internal mechanical strain. The converse piezoelectric effect is the basis of all piezoelectric sensors.

PQCs have been widely used in oscillators and filter circuits for high-precision frequency control. The use of PQCs as a mass sensor is based on the work of Sauerbrey (2), which established a linear relationship between the decrease in resonant frequency and the increase in surface mass loading of a PQC. Because of its high sensitivity for mass detection, a piezoelectric sensor is usually called quartz crystal microbalance (QCM). QCM was originally and is still used to measure thickness of coatings in vacuum and air. The first example for application of QCM to analytical chemistry was reported by King in 1964 (3). He coated PQCs with various materials and used them as a sorption detector in gas chromatography to detect and measure the composition of vapors and gases. The applications of piezoelectric sensors were limited to the determination of environmental and other gas species for a long time. The liquid-phase measurements began in the 1980s when new oscillator technology emerged and advanced to make PQCs oscillate in solution as stably as in gas. Then, numerous piezoelectric chemical sensors and biosensor have been reported.

Piezoelectric sensors, characterized by their simplicity, low cost, high mass-detection sensitivity, and versatility, have found increasing applications in biomedical analyses. The objective of this article is to briefly present the theory, equipment, and applications of piezoelectric sensors in the biomedical area. Interested readers are referred to some key review articles (4–9) for more theoretical and technical details of piezoelectric sensors.

THEORY

Typically, a piezoelectric sensor is fabricated by modifying a PQC with a layer of sensing material, chemical or biological, that has specific affinity to a target analyte. The specific binding between the sensing material and the target analyte causes a change in the resonant frequency of PQC that is proportional to the amount of target analyte adsorbed or bound on the sensor surface, which can be correlated to the concentration of target analyte in the original sample.

A typical PQC consists of a quartz crystal wafer and two excitation electrodes plated on opposite sides of the crystal. The wafer is cut from a natural or synthetic crystal of quartz. The electromechanical coupling and stresses resulting from an applied electric field depend on the crystal symmetry, cut angle, and electrode configuration (4). Different modes of electromechanical coupling lead to different types of acoustic waves, including thickness shear mode (TSM), surface acoustic wave (SAW), shear horizontal (SH) SAW, SH acoustic plate mode (APM), and flexural plate wave (FPW).

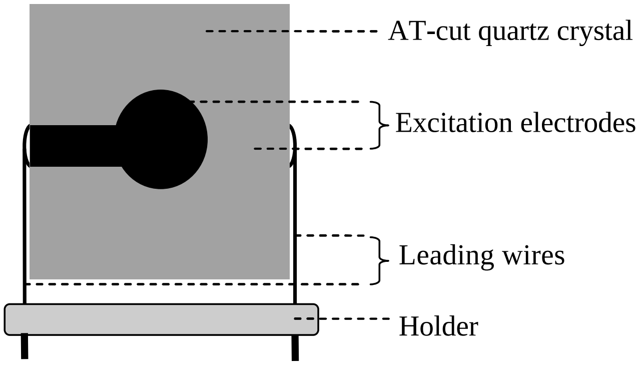


Figure 1. Schematic of an AT-cut piezoelectric quartz crystal.

The TSM device, widely referred to as QCM, is the simplest and most popular piezoelectric sensor. Hence, we will focus on TSM piezoelectric sensors in this article. A TSM sensor typically uses AT-cut quartz as a piezoelectric substrate, which has a minimal temperature coefficient and is obtained by cutting quartz crystals at approximately 35° from the z-axis of the crystal. Figure 1 shows the schematic of an AT-cut PQC. When an alternating electric field is applied across an AT-cut quartz crystal through the excitation electrodes, the crystal produces a shear vibration in the x-axis direction parallel to the electric field and propagation of a transverse shear wave through the crystal in the thickness direction. The resonant frequency of the vibration is determined by the properties of the crystal and the contacting medium.

Mass Response

Most piezoelectric sensors are used as mass sensors that are based on the Sauerbrey equation (2):

$$\Delta F = \frac{-2F_0^2}{A\sqrt{\mu_q \rho_q}} \Delta M \tag{1}$$

where ΔF is the frequency change measured (Hz), F_0 is the resonant frequency of the fundamental mode of the crystal (Hz), A is the area of the gold disk coated onto the crystal (cm^2), ρ_q is the density of the quartz crystal (2.648 g·cm^{-3}), μ_q is the shear modulus of quartz (2.947×10^{11} g·cm^{-1}·s^{-2}), and ΔM is the mass change (g). The Sauerbrey equation is applicable only to a thin ($\sim$1 μm) and elastic film coupled to the crystal surface, where the mass loading can be up to 0.05% of the crystal mass.

According to the Sauerbrey equation, the magnitude of frequency decrease corresponding to a mass increase is proportional to ${F_0}^2$; i.e. the higher the fundamental resonant frequency, the higher the mass sensitivity. Typical AT-cut PQCs have $F_0 = 5 \sim 30$ MHz with a frequency resolution of $\sim$0.1 Hz and a mass sensitivity of $0.056 \sim 2.04$ Hz·cm^2·ng^{-1}. Thinner crystal wafers have higher F_0 and thus have higher mass sensitivity, but they are also more fragile and thus are more difficult to manufacture and handle. For one of the most commonly used AT-cut PQCs, $F_0 = 9$ MHz, $A = 0.2$ cm^2, and the detectable mass is 1.09 ng per Hz, which is approximately 100 times higher than that of an electronic fine-balance with a sensitivity of 0.1 μg. As AT-cut piezoelectric sensors can sensitively detect mass change at nanogram levels, they are frequently referred to as quartz crystal microbalances (QCMs) or nanobalances.

However, as a mass sensor, the QCM does not have selectivity. To make a selective QCM chemical sensor or biosensor, the QCM must be coated with a film of chemical or biological recognition material that is selective to a target analyte.

Viscosity-Density Effect

When a QCM is employed in liquid phase, in addition to the mass change, it also responds to other factors such as liquid density and viscosity, surface energy, viscoelasticity, roughness, and surface charge density. For a QCM with only one side in contact with a Newtonian liquid, ΔF is linearly proportional to the squared root of the product of viscosity (η_L) and density (ρ_L) of the liquid (10):

$$\Delta F = -F_0^{3/2}\sqrt{\rho_L\eta_L/\pi\mu_q\rho_q} \tag{2}$$

For a QCM with simultaneous mass and liquid loading, ΔF can be expressed as (11)

$$\Delta F = -\frac{2F_0^2}{\sqrt{\mu_q\rho_q}}(\Delta M/A + \sqrt{\rho_L\eta_L/4\pi F_0}) \tag{3}$$

In equation 3, the first term is equivalent to the Sauerbrey equation, and the second term is equivalent to Kanazawa equation. Equation 3 indicates the additive nature of mass and viscosity-density effects in changing the resonant frequency. It also shows that it is impossible to distinguish the mass effect from the viscosity-density effect when only the resonant frequency is monitored.

Some piezoelectric sensors are based on the density-viscosity change rather than on the elastic pure mass change. For example, a piezoelectric sensor was used to detect *E. coli* based on the gelation of Tachypleus amebocyte lysate (TAL) (12), and the detection range is $2.7 \times 10^4 - 2.7 \times 10^8$ cells·mL^{-1}. Gee et al. (13) used a piezoelectric senor for measuring microbial polymer production and growth of an environmental isolate obtained from river sediment contaminated with petroleum hydrocarbons. The increasing amount of produced polymer corresponded to an increase in the viscosity of the liquid, which was directly measurable as the fluid contacts the surface of the quartz crystal in the sensor system. These methods, although they lack specificity, are advantageous in that coating on the PQC surface is unnecessary.

Equivalent Circuit Analysis

Equivalent circuit analysis can provide more detailed information about the surface/interface changes of a piezoelectric sensor (11,14–17). A PQC can be represented by a Butterworth–Van Dyke (BVD) model (Fig. 2), which is composed of a static capacitance (C_0) in parallel with a motional branch containing a motional inductance (L_m), a motional capacitance (C_m), and a motional resistance (R_m) in series. Each parameter has its distinct physical meaning: C_0 reflects the dielectric properties between the electrodes located on opposite sides of the insulating quartz crystal; C_m represents the energy stored during oscillation, which corresponds to the mechanical elasticity of the vibrating body; L_m is related to the displaced mass; and R_m is the energy dissipation during oscillation, which is closely related to viscoelasticity of the deposited films and viscosity-density of the adjacent liquid.

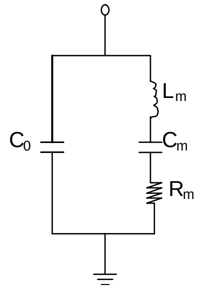


Figure 2. Butterworth–Van Dyke model of a piezoelectric quartz crystal.

The BVD model can be described by the following admittance equations:

$$Y(\omega) = G(\omega) + jB(\omega) \tag{4}$$

$$G(\omega) = \frac{R_m}{R_m^2 + (\omega L_m - 1/\omega C_m)^2} \tag{5}$$

$$B(\omega) = -\frac{(\omega L_m - 1/\omega C_m)}{R_m^2 + (\omega L_m - 1/\omega C_m)^2} + \omega C_0 \tag{6}$$

where Y is admittance, i.e., the reciprocal of impedance. Y is a complex quantity, its real part G is conductance, and its imaginary part B is susceptance. Both G and B are a function of the scanning frequency f ($\omega = 2\pi f$) and the four equivalent circuit parameters. These parameters are determined by physical properties of the quartz crystal, perturbing mass layer and contacting liquid, and can be obtained by fitting the measured impedance/admittance data to the BVD model using the admittance equations. Figure 3 shows typical admittance spectra of an unperturbed 8 MHz AT-cut PQC in air. The fitted results of F_0, R_m, L_m, C_m, and C_0 were 7.99 MHz, 9.6 Ω, 17.9 mH, 22.2 fF, and 8.2 pF (including parasitic capacitance in the test fixture) for the quartz crystal (18).

High-frequency admittance/impedance analysis has been extensively used in surface/interface studies. A simpler way to provide insights into the viscoelastic properties of a bound surface mass is to simultaneously monitor F_0 and R_m or F_0 and the dissipation factor D using a quartz crystal analyzer that is much less expensive than the impedance analyzer. This method has been applied to study the behavior of adherent cells in response to chemical, biological, or physical changes in the environment.

For a QCM in contact with liquid, the change of motional resistance was first derived by Muramatsu et al. (14) as follows:

$$\Delta R = (2\pi F_0\rho_L\eta_L)^{1/2}A/k^2 \tag{7}$$

where k is the electromechanical coupling factor.

Simultaneous measurements of ΔF and ΔR can differentiate an elastic mass effect from the viscosity-induced effect. ΔR is a good measure of the viscoelastic change. For an elastic mass change, ΔR will be zero and ΔF will be linearly proportional to the mass change in accordance with the Sauerbrey equation. For a QCM with only one side in contact with a Newtonian liquid, both ΔF and ΔR are linearly proportional to the squared root of the product

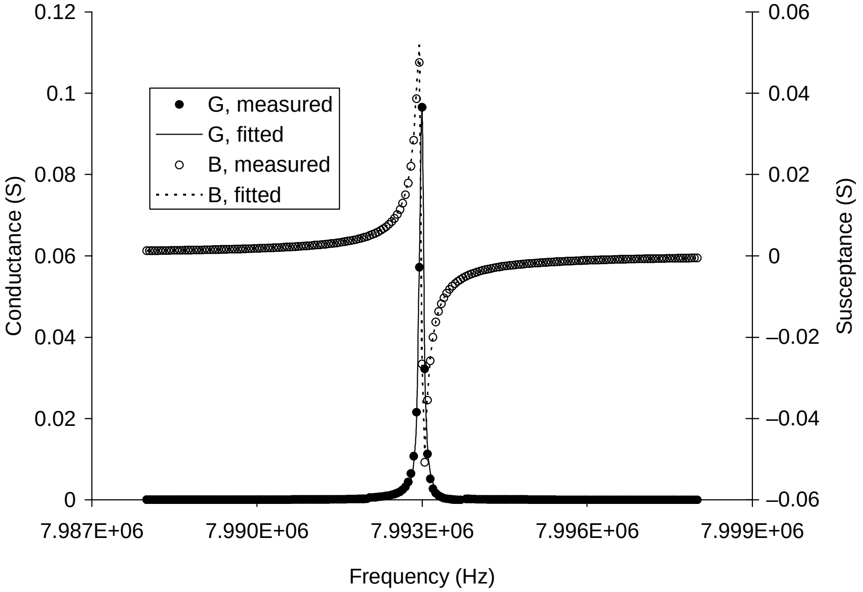


Figure 3. Typical conductance and susceptance spectra of an unperturbed 8 MHz AT-cut PQC in air.

of viscosity and density of the liquid. Therefore, a pure viscosity-density change will result in a linear $\Delta F \sim \Delta R$ plot. In the presence of a viscoelastic change, the $\Delta R \sim \Delta F$ plot will lie between the pure viscosity-density effect line and the elastic mass effect line or even above the former (8,18).

EQUIPMENT AND EXPERIMENTS

Traditionally, a piezoelectric sensor (resonator) is driven by a homemade oscillator, and the oscillation frequency is measured by a frequency counter that is connected to a recorder or computer for data collection. The Pierce oscillator with a fundamental mode AT-cut resonator is the most popular oscillator design type and operates with the piezoelectric sensor as an inductive element. Now such oscillators are commercially available. One of the suppliers is International Crystal Manufacturing (Oklahoma City, OK), which produces a standard (clock) oscillator for gas-phase QCMs and lever oscillator for liquid-phase QCMs.

Highly sophisticated, automatic, and microprocessor-controlled piezoelectric detectors or QCM instruments are commercially available from several manufacturers (19). The main commercial systems include QCA-917- and 922 quartz crystal analyzers (Princeton Applied Research, Oak Ridge, TN), EQCM 400 series electrochemical quartz crystal microbalances (CH Instruments, Austin, TX), EQCN-700 and -900 electrochemical quartz crystal nanobalances (Elchema, Potsdam, NY), the PZ-1000 immuno-biosensor system and PZ-105 gas phase piezoelectric detector (Universal Sensors, Metairie, LA), RQCM research QCM (Maxtek, Santa Fe Springs, CA), and Mark series cryogenic and thermally controlled QCMs (QCM Research, Lake Forest, CA). These systems are designed to reliably measure mass change up to ~100 μg with a resolution of ~1 ng·cm^{-2}. Most of them are programmed and controlled by easy-to-use Windows-based software (Microsoft Corporation, Redmond, WA). The QCA 922, designed for EQCM with a potentiostat or stand-alone operation, can simultaneously measure resonant frequency and resistance of QCM. The RQCM can measure crystal frequency and crystal resistance for up to three crystals simultaneously. Moreover, high-frequency impedance/admittance analyzers such as E4991A (Agilent Technologies, Palo Alto, CA) can be used to obtain the impedance/admittance spectra of the quartz crystal and to acquire the equivalent circuit parameters by fitting the impedance/admittance data to the BVD model as described earlier.

Previously, for liquid-phase applications, the dip-and-dry approach is made in the measurement of piezoelectric sensors, in which the resonant frequency of the same sensor is measured in gas phase before and after the sample solution is dipped and dried. This approach does not need a special fixture to mount the crystal; however, it is unsuitable for automation and has poor reproducibility. By mounting the crystal to a dip or well cell, like those from Princeton Applied Research and International Quartz Manufacturing, and letting only one side of the crystal exposed to the test solution, it is possible to monitor the frequency change in solution in real time.

Flow-through QCM biosensors have drawn increasing attention due to its ease for automation. For example, Su and Li (20) developed a flow-through QCM immunosensor system for automatic detection of *Salmonella typhimurium*. The QCM immunosensor was fabricated by immobilizing anti-*Salmonella* antibodies on the surface of an 8 MHz AT-cut quartz crystal with Protein A method, and then installed into a 70 μL flow-through detection cell. The flow cell was composed of acrylic, with upper and lower pieces held together by two screws with O-rings. One face of the sensor was exposed to the 70 μL chamber that was connected to a peristaltic pump and multiposition switching valve. The flow cell was designed to reduce the potential of air bubbles remaining on the crystal after filling from the dry state and to allow air bubbles in the liquid phase to pass out without sticking to the crystal. The oscillation frequency of the QCM sensor was monitored in real time by a Model 400 EQCM system controlled by a laptop PC

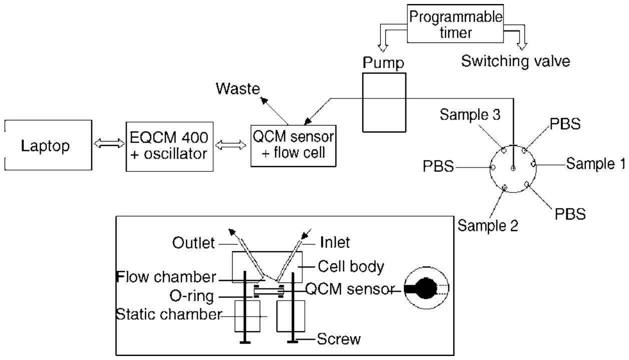


Figure 4. Components of an automatic QCM immunosensor system (top) and the flow cell (bottom).

under the Windows environment. Both the pump and the valve were controlled by a DVSP programmable timer. A schematic diagram of the whole QCM sensor system is illustrated in Fig. 4. The operation of this QCM sensor system was totally automated. As shown in Fig. 5, negative frequency shifts exist between every two neighboring phosphate buffered saline solution (PBS) baselines, which were attributed to the specific adsorption of target bacteria onto the biosensor surface.

APPLICATIONS

Piezoelectric sensors, as simple yet powerful tools, have been extensively employed in detection of chemical and biological agents as well as in the study of chemical, electrochemical, and biological interfacial processes. An online search from SciFinder Scholar (Chemical Abstracts) with the keyword "quartz crystal microbalance" resulted in 4722 references, and 2203 of them are journal articles published during 2000–2005. These studies were conducted to develop (1) antibody/antigen-based biosensors (immunosensors) for detecting biomacromolecules, viruses, cells, as well as small molecules; (2) DNA/RNA probe-based biosensors (genosensors) for *in situ* detection of nucleic acid hybridization; (3) biosensors based on immobilized enzymes, proteins, receptors, lipids, or whole cells; and (4) chemical sensors based on inorganic or organic films for measurements of organic vapors, metal ions, and drugs. Also piezoelectric sensors reported in these studies were used for (1) studies of adsorption of biomolecules and living cells by bare QCM or QCM with functionalized surfaces; and (2) QCM/EQCM investigation/analyses of interfacial phenomena and processes, including self-assembles monolayers (SAMs), films formed using the layer-by-layer assembly technique, molecularly imprinted polymers, biopolymer films, micellar systems, ion transfer at and ion exchange in thin polymer films, doping reactions of conducting polymers, electrodedeposition of metals, and dissolution of metal films.

In the following sections, immunosensors and genosensors, which are most relevant and important to biomedicine, are chosen to discuss the applications of piezoelectric sensors. Some review articles (4–8,21,22) are available for more comprehensive information about applications of piezoelectric sensors.

Immunosensors

One important feature of piezoelectric sensors is that they can be designed as label-free immunosensors. The

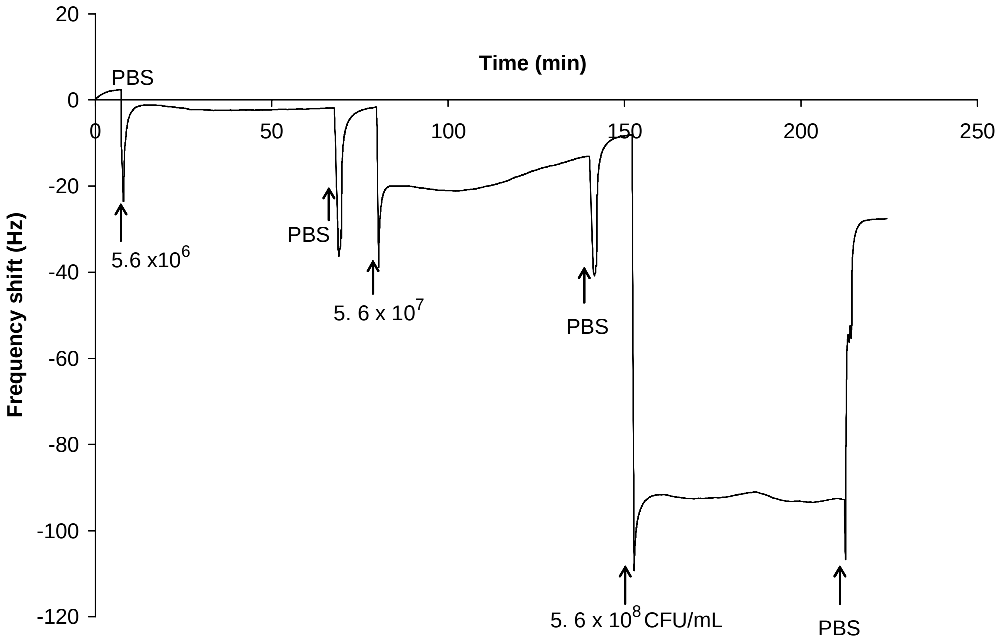


Figure 5. Typical output of an automatic QCM immunosensor system for detection of *S. typhimurium*.

immunosensors taking advantage of antibody-antigen affinity reaction are among the most promising biosensors due to their high specificity and versatility. Conventional immunosensors generally involve the formation of a sandwich immuno-complex consisting of an immobilized primary antibody, captured target analyte, and labeled secondary antibody followed by an optical or electrochemical measurement to detect the label directly or indirectly. Piezoelectric immunosensors do not need a labeled antibody and are thus much simpler and easier in operation than the sandwich immunosensors.

The first piezoelectric immunosensors is reported by Shons et al. (23), who modified a quartz crystal with bovine serum albumin (BSA) and used it to detect anti-BSA antibodies. Since then, numerous piezoelectric immunosensors have been reported for the detection of various analytes from small molecules to biological macromolecules, whole viruses, and cells. In brief, a piezoelectric immunosensor is fabricated by immobilizing a specific antibody/antigen on the surface of an AT-cut PQC. When the immunosensing surface is exposed to a sample solution, a binding reaction occurs between the immobilized antibody/antigen and its complementary part (target analyte). The binding event can be monitored *in situ* by QCM based on the change of surface mass loading and/or other properties such as viscoelasiticity, and thus, the target species is quantitatively detected. Figure 6 illustrates the stepwise assembly and the principle of piezoelectric immunosensor for direct detection of the binding of target analyte and immobilized antibody.

Microbial detection is probably the most common area in which piezoelectric immunosensors are applied. Current practice for effective treatment of infectious diseases without abuse of antibiotics relies on rapid identification of specific pathogens in clinical diagnostics. Nevertheless, conventional methods for microbial detection are inadequate due to being tedious and laborious. Although traditional culture methods hypothetically allow the detection of a single cell, they are extremely time-consuming, typically requiring at least 24 hours and multistep tests to confirm the analysis. Even current rapid methods such as enzyme-linked immunosorbent assay (ELISA) and polymerase chain reaction (PCR) still take several hours to generate only tentative results and require skilled personnel.

Numerous piezoelectric immunosensors have been reported for rapid and specific detection of pathogenic bacteria as alternatives to the conventional methods since the pioneer work of Muramatsu et al. (24), which involved the determination of *Candida albicans* with an AT-cut PQC coated with a specific antibody. Piezoelectric immunosensors have been developed for detection of different bacteria including *S. typhimurium*, *S. paratyphi*, *E. coli*, *E. coli* K12, *Chlamydia trachomatis*, *Yersisinia pestis*, *Candida albicans*, and *Shigella dysenteriae* (25–27). The lower limits of detection typically ranged between 10^5 and 10^7 cells mL^{-1} along with a detection time of tens of minutes to several hours.

QCM has been used to detect various infectious viruses. In the study by König and Gratzel (28), Herpes simplex types 1 and 2, Varicella-zoster virus, Cytomegalovirus, and Epstein–Barr virus were detected using a reusable QCM immunosensor with a detection range from 10^4 to 10^9 cells. They reported that a similar QCM immunosensor could detect Rotavirus and Adenovirus with a linear detection range from 10^6 to 10^{10} cells (25) as well as hepatitis A and B viruses (29). Kosslinger et al. (30) demonstrated the feasibility of detecting HIV viruses using a QCM sensor. Antibodies specific to the HIV were absorbed on the crystal surface, and a serum sample could be detected in 10 min in a flow QCM system.

A piezoelectric immunosensor was developed for the detection of cortisol in a range of 36–3628 ppb (31). Cortisol antibodies were covalently bound onto the Au electrode of a 10 MHz crystal with a water-insoluble polymer and thyroxine antibodies.

Piezoelectric immunosensors have also been frequently reported for determination of biological macromolecules. For example, Kurosawa et al. (32) constructed a high-affinity

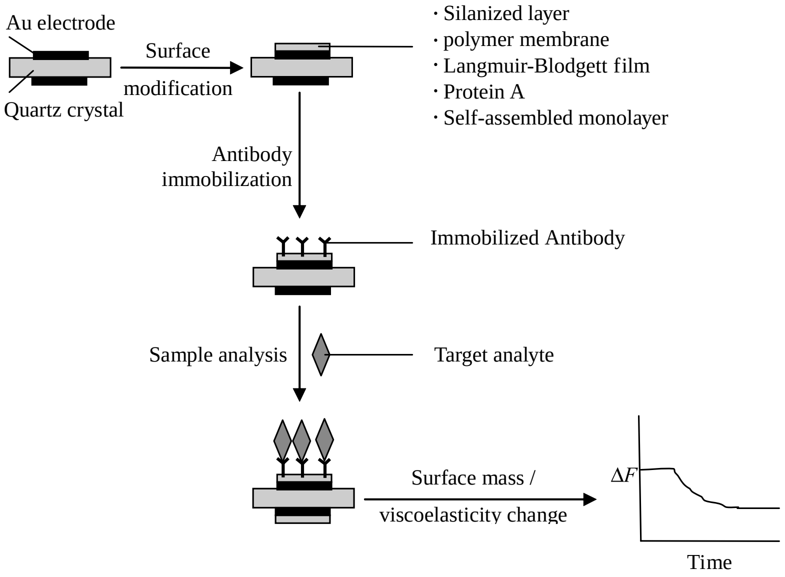


Figure 6. Mechanism of a piezoelectric immunosensor for direct detection of the binding of target analyte and immobilized antibody.

piezoelectric immunosensor using anti-C-reactive protein (CRP) antibody and its fragments for CRP detection. When anti-CRP F(ab')2-IgG antibody was immobilized on the PQC, the detection limit and the linearity of CRP calibration curve were achieved at concentrations from 0.001 to 100 $\mu g \cdot dL^{-1}$ even in serum samples.

Antibody immobilization is vital in successful development of a piezoelectric immunosensor. The current immobilization methods are mainly based on a silanized layer, polymer membrane, Langmuir–Blodgett film, Protein A, and SAM. Protein A, due to its natural affinity toward the Fc region of IgG molecules, has been commonly used to orient antibodies for immunoassays. The orient immobilization has the advantage that it does not block the active sites of the antibodies for binding of target antigens. The procedure for antibody immobilization based on Protein A is simple. Briefly, the Au surface of PQC is coated first with Protein A, and then the antibody is bound to the immobilized Protein A directly. The SAM technique offers one of the simplest ways to provide a reproducible, ultra-thin, and well-ordered layer suitable for further modification with antibodies, which has the potential of improving detection sensitivity, speed, and reproducibility.

Analytical applications of piezoelectric immunosensors based on the direct binding between immobilized antibodies/antigens and target analytes are attractive, owing to the versatility and simplicity of the method. However, the sensitivity of theses approaches is relatively low due to the relatively small numbers of analyte entities that can specifically bind to the limited number of antibody/antigen sites on the surface. Some amplification techniques have been investigated for the sensitivity of piezoelectric immunosensors. Ebersole and Ward (33) reported an amplified mass immunosorbent assay with a QCM for the detection of adenosine 5'-phosphosulfate (APS). The enzymatic amplification led to significant enhancement of the detection sensitivity; levels of approximately 5 $ng \cdot mL^{-1}$ (10^{-14} M) APS reductase could be detected, whereas the direct binding of APS reductase at even more elevated concentrations could not be measured. A sensitive QCM immunosensor was developed by incorporating the Au nanoparticle-amplified sandwiched immunoassay and silver enhancement reaction (34). Au nanoparticle-promoted silver (I) reduction and silver metal deposition resulted in about a two-orders-of-magnitude improvement in human IgG quantification. Su and Li (18) described a piezoelectric immunosensor for the detection of *S. typhimurium* with simultaneous measurements of the changes in resonant frequency and motional resistance (ΔF and ΔR). In the direct detection of *S. typhimurium*, ΔF and ΔR were proportional to the cell concentration in the range of 10^5 to 10^8 and 10^6 to 10^8 $cells \cdot mL^{-1}$, respectively. Using anti-*Salmonella* magnetic microbeads as a separator/concentrator for sample pretreatment as well as a marker for signal amplification, the detection limit was lowered to 10^2 $cells \cdot mL^{-1}$ based on the ΔR measurements.

Genosensors

Piezoelectric genosensors are fabricated by immobilizing a single-stranded (ss) DNA/RNA probe on the PQC surface. Specific hybridization between the immobilized DNA/RNA probe and its complementary strand in sample causes a change in the resonant frequency of the QCM. Various methods have been used for the immobilization of DNA probes onto the QCM surface. Among these, the SAM method is most commonly used because it offers an ordered, stable, and convenient immobilization. Thiolated oligonucleotides can directly form a SAM on the gold surface of the QCM electrode via the Au-thiolate bond. Figure 7 is an illustration of the piezoelectric genosensor for DNA hybridization detection.

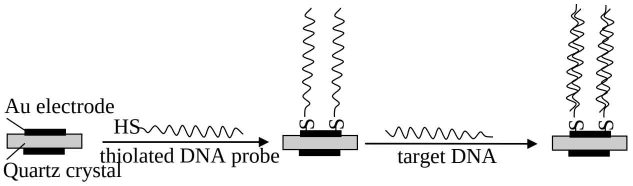


Figure 7. Illustration of piezoelectric genosensor for *in situ* detection of DNA hybridization.

DNA/RNA probes have been applied to detect pathogenic microorganisms in clinical samples. Bacterial and viral pathogens are detectable because of their unique nucleic acid sequences. The DNA/RNA probing process is usually preceded by PCR amplification as target nucleic acid may be present in a sample in very small quantities. Most PCR formats are followed by the detection of amplicons using gel electrophoresis or the membrane-based hybridization method. The former lacks of sensitivity; the latter is more specific, but it requires multistep processing and is thus time-consuming.

Over the past decade, many attempts have been made to develop biosensors for sensitive and reliable detection of DNA hybridization. Fawcett et al. (35) were the first to develop a piezoelectric DNA sensor. Piezoelectric genosensors, due to their simplicity and cost effectiveness, have been recently applied to detect gene mutation, genetically modified organisms, and bacterial pathogens. Su et al. (36) described a piezoelectric DNA sensor for detection of point mutation and insertion mutations in DNA. The method involved the immobilization of ssDNA probes on QCM, the hybridization of target DNA to form homoduplex or heteroduplex DNA, and finally the application of MutS for the mutation recognition. By measuring the MutS binding signal, DNA containing a T:G mismatch or unpaired base was discriminated against perfectly matched DNA at a target concentration ranging from 1 nM to 5 μM.

The sensitivity of piezoelectric genosensors may be improved through optimizations of probe immobilization or by means of signal amplification using anti-dsDNA antibodies, liposomes, enzymes, or nanoparticles. A nanoparticle is an effective marker for mass amplification as it has relatively greater mass in comparison with a DNA molecule. Amplified with nanoparticles, the limit of DNA detection by QCM can be lowered for several orders to as low as 10^{-16} M (37). Mao et al. (38) developed a piezoelectric genosensor for the detection of *E. coli* O157:H7 with nanoparticle amplification, in which thiolated ssDNA probes specific to *E. coli* O157:H7 *eaeA* gene were immobilized onto the QCM surface through self-assembly and streptavidin conjugated Fe_3O_4 nanoparticles were used as "mass enhancers" to amplify the detection signal. As low as 10^{-12} *M* synthesized oligonucleotides and 2.67×10^2 cells·mL^{-1} of *E. coli* O157:H7 could be detected by the piezoelectric genosensor.

CONCLUSIONS

Piezoelectric sensors, as sensitive mass sensors or QCMs, have been applied to detect and measure a broad variety of biomedical analytes in both gas and liquid phases based on the adsorption/desorption of target analyte(s) on/from the sensor surface, in which the selectivity is controlled by the sensing material. Piezoelectric sensors are more than pure mass sensors as they are also capable of detecting subtle changes in the solution–surface interface that can be due to density-viscosity changes in the solution, viscoelastic changes in the bound interfacial material, and changes in the surface free energy. The most attractive advantage of QCMs is that they are suitable for label-free detection and flow-through, in real-time detection. However, using the direct detection approach, the sensitivity of QCM is inadequate for some applications in biomedical analysis such as detection of low levels of pathogens and other biological agents in clinical samples. The lack of sensitivity may be addressed by employing amplification techniques as introduced earlier, by using piezoelectric films and bulk silicon micromaching techniques to manufacture high-frequency QCMs (21), or by designing devices of other acoustic wave modes such as SAW, APM, and FPW.

BIBLIOGRAPHY

1. Curie J, Curie P. An oscillating quartz crystal mass detector. Comp Rend 1880;91:294–297.
2. Sauerbrey GZ. Use of quartz vibration for weighing thin films on a microbalance. J Phys 1959;155:206–212.
3. King WH Jr. Piezoelectric sorption detector. Anal Chem 1964;36:1735–1739.
4. Guilbault GG, Jordan JM. Analytical uses of piezoelectric crystals: A review. CRC Crit Rev Anal Chem 1988;19:1–28.
5. Ward MD, Buttry DA. In situ interfacial mass detection with piezoelectric transducers. Science 1990;249:1000–1007.
6. Buttry DA, Ward MD. Measurement of interfacial process at electrode surfaces with the electrochemical quartz crystal microbalance. Chem Rev 1992;92:1355–1379.
7. Janshoff A, Galla H-J, Steinem C. Piezoelectric mass-sensing devices as biosensors—An alternative to optical biosensors? Angew Chem Int Ed 2000;39:4004–4032.
8. Marx KA. Quartz crystal microbalance: a useful tool for studying thin polymer films and complex biomolecular systems at the solution-surface interface. Biomacromolecules 2003;4: 1099–1120.
9. Buck RP, Lindner E, Kutner W, Inzelt AG. Piezoelectric chemical sensors. Pure Appl Chem 2004;76:1139–1160.
10. Kanazawa KK, Gordon JG. The oscillation frequency of a quartz resonator in contact with a liquid. Anal Chim Acta 1985;175:99–105.
11. Martin SJ, Granstaff VE, Frye GC. Characterization of a quartz crystal microbalance with simultaneous mass and liquid loading. Anal Chem 1991;63:2272–2281.
12. Qu X, Bao LL, Su X-L, Wei W. Rapid detection of *Escherichia coli* form with a bulk acoustic wave sensor based on the gelation of *Tachypleus* amebocyte Lyste. Talanta 1998;47: 285–290.
13. Gee WA, Ritalahti KM, Hunt WD, Loffler FE. QCM viscometer for bioremediation and microbial activity monitoring. IEEE Sens J 2003;3:304–309.
14. Muramatsu H, Tamiya E, Karbue I. Computation of equivalent circuit parameters of quartz crystals in contact with liquids and study of liquid properties. Anal Chem 1988;60: 2142–2146.
15. Zhou T, Nie L, Yao S. On equivalent circuits of piezoelectric quartz crystals in a liquid and liquid properties, Part I, Theoretical derivation of the equivalent circuit and effects of density and viscosity of liquids. J Electroanal Chem Interf Electrochem 1990;293:1–18.
16. Nöel MAM, Topart PA. High-frequency impedance analysis of quartz crystal microbalance, 1. General considerations. Anal Chem 1994;66:484–491.
17. Xie Q, Wang J, Zhou A, Zhang Y, Liu H, Xu Z, Yuan Y, Deng M, Yao S. A study of depletion layer effects on equivalent circuit parameters using an electrochemical quartz crystal impedance system. Anal Chem 1999;71:4649–4656.
18. Su X-L, Li Y. A QCM immunosensor for *Salmonella* detection with simultaneous measurements of resonant frequency and motional resistance. Biosens Bioelectron. In press.
19. O'Sullivan CK, Guilbault GG. Commercial quartz crystal micobalances theory and applications. Biosens Bioelectron 1999;14:663–670.
20. Su X-L, Li Y. An automatic quartz crystal microbalance immunosensor system for *Salmonella* detection. ASAE Paper No. 047043. St. Joseph, MI: The American Society of Agricultural Engineers; 2004.
21. Martin SJ, Frye GC, Spates JJ, Butler MA. Gas sensing with acoustic devices. Proc-IEEE Ultrasonics Symp 1996;1:423–434.
22. Vaughan RD, Geary E, Pravda M, Guilbault GG. Piezoelectric immunosensors for environmental monitoring. Int J Environ Anal Chem 2003;83:555–571.
23. Shons A, Dorman F, Najarian J. The piezoelectric quartz immunosensor. J Biomed Mater Res 1972;6:565–570.
24. Muramatsu H, Kajiwara K, Tamiya E, Karube I. Piezoelectric immunosensor for detection of *Candida albicans* microbes. Anal Chim Acta 1986;188:257–261.
25. Konig B, Gratzel M. Detection of viruses and bacteria with piezoelectric immunosensor. Anal Lett 1993;26:1567–1575.
26. Ivnitski D, Abel-Hamid I, Atanasov P, Wilkins E. Biosensors for detection of pathogenic bacteria. Biosens Bioelectron 1999;14:599–624.
27. Deisingh AK, Thompson M. Detection of infectious and toxigenic bacteria. Analyst 2002;127:567–581.
28. König B, Gratzel M. A novel immunosensor for Herpes virus. Anal Chem 1994;66:341–348.
29. König B, Gratzel M. Long term stability and improved reusability of piezoelectric immunosensor for human erythrocytes. Anal Chim Acta 1993;280:37–42.
30. Kösslinger C, Crost S, Aberl F. A quartz crystal microbalance for measurements in liquids. Biosens Bioelectron 1992;7: 397–410.
31. Attili BS, Suleiman AA. Peizoelectric immunosensor for detection of cortisol. Anal Lett 1995;28:2149–2159.
32. Kurosawa S, Nakamura M, Park JW, Aizawa H, Yamada K, Hirata M. Evaluation of a high-affinity QCM immunosensor using antibody fragmentation and 2-methacryloyloxyethyl phosphorylcholine (MPC) polymer. Biosens Bioelectron 2004;20:1134–1139.

33. Ebersole RC, Ward MD. Amplified mass immunosorbent assay with a quartz crystal microbalance. J Am Chem Soc 1988;110: 8623–8628.
34. Su X, Li SFY, O'Shea SJ. Au nanoparticle- and silver-enhancement reaction-amplified microgravimetric biosensor. Chem Commun 2001; 755–756.
35. Fawcett NC, Evans JA, Chen LC, Drozda KA, Flowers N. A quartz crystal detector for DNA. Anal Lett 1988;21:1099–1110.
36. Su X, Robelek R, Wu Y, Wang G, Knoll W. Detection of point mutation and insertion mutations in DNA using a quartz crystal microbalance and MutS, a mismatch binding protein. Anal Chem 2004;76:489–494.
37. Liu T, Tang J, Jiang L. The enhancement effect of gold nanoparticles as a surface modifier on DNA sensor sensitivity. Biochem Biophys Res Commun 2004;313:3–7.
38. Mao X, Yang L, Su X-L, Li Y. A nanoparticle-based quartz crystal microbalance DNA sensor for the detection of *Escherichia coli* O157:H7. Biosens Bioelectron. In press.

See also COCHLEAR PROSTHESES.

PLETHYSMOGRAPHY. See IMPEDANCE PLETHYSMOGRAPHY.

PNEUMATIC ANTISHOCK GARMENT. See SHOCK, TREATMENT OF.

PNEUMOTACHOMETERS

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

INTRODUCTION

From the beginning of time, breathing has been important since it has been the most common indicator of life. The Bible indicates that God infused life into man by "blowing into his nostrils the breath of life" (1). Not surprisingly then, the way in which we breath is an important indicator of our health. Consequently, the medical profession has tried to learn our physical condition by focusing on the behavior of the respiratory system. Two main mechanisms take place in the breathing process: (1) Movement of gases from the nose and mouth to the alveoli, and (2) CO_2 and O_2 gas exchange at the alveoli. An interesting historical perspective, standardization of pulmonary function tests (PFTs) and the associated equipment received a great push from the mobile PFT trucks that were part of a campaign to eliminate lung cancer in the United States in the 1950s. Even today, PFTs are often used as a preliminary screen for lung cancer.

This section focuses on the equipment used to make flow and volume measurements in the first category: pneumotachometers, also known as respirometers, spirometers, or simply flowmeters.

There are two clinical areas where flow measurement devices are used. These are (1) the field of spirometry (2–4), dealing with the actual performance of the respiratory system as reflected by the volumes that the lung can realize, and the speed in which these volumes can be moved in and out of the lungs. Indicators such as tidal volume (TV or V_t) and vital capacity (VC) provide a glimpse of the range of motion of the lungs. Similarly, parameters, such as FEV_1 (forced expiratory volume in 1 s) and $FEF_{25-75\%}$ (forced expiratory flow at the mid-portion of forced vital capacity). Notice that these parameters are the result of: (a) the patient's ability to cooperate, (b) condition of the diaphragm, the respiratory system's main workhorse, (c) range of motion of the lungs, and (d) the mechanical components associated with the respiratory pathways (size of the conducting airways). (2) The field of parameter estimation (5–12), dealing specifically with the noninvasive measurement of components descriptive of the mechanical characteristics of the respiratory apparatus. These parameters include resistance, compliance, and inertance. One advantage of this approach is its independence from the patient's ability to cooperate. This approach is particularly useful in unconscious, and very young or very old patient populations. However, this approach requires a much higher level of computation.

Before proceeding with a presentation on pneumotachometer, some definitions and conventions are appropriate.

Open and Closed Systems

In many pulmonary function tests, the test procedure requires that the subject inhales maximally, then place the pneumotachometer in then mouth, and then exhales as fast as possible. This type of set up is referred to as an "open" system since no exhaled air is rebreathed. In other types of pulmonary function testing, the patient exchanges air back and forth with a reservoir. This later system is referred to as a "closed" system.

Variable Used for Flow

The most common variable used to describe flow is $\dot{V}$. The "dot" comes from the mathematical notation of differentiation with respect to time, originally developed by Sir Isaac Newton, or

$$\dot{V} = \frac{dV}{dt}$$

which implies that flow, $\dot{V}$, is the time rate or change of lung volume, V.

Polarity

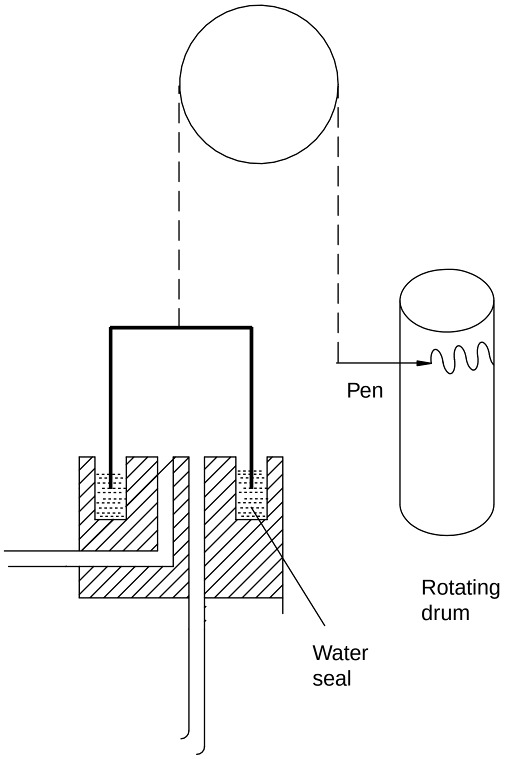
Spirometer tests consider expiratory flow as positive while parameter estimation procedures look at inspiration as positive. Perhaps it is reflective that most spirometer tests are performed during expiration while in parameter estimation procedures, inspiration is the primary arena.

DEVICES FOR MEASURING RESPIRATORY FLOW

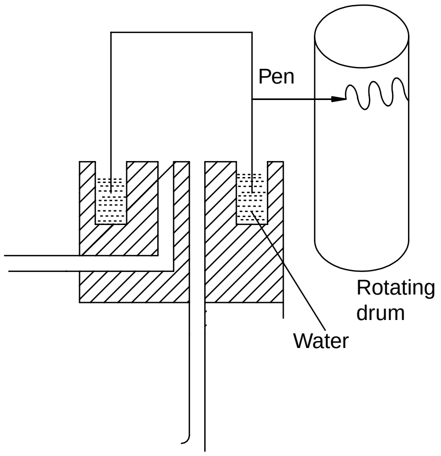
This section presents devices that have and continue to be used for measuring respiratory flow and volume. Even though some of them are not used as often, they are part of the toolbox that clinicians and researchers have used in getting a handle in the respiratory system.

Volume Displacement Type

This type of device, often called volumetric type, captures the volume of gas coming out of the subject's mouth with an expanding reservoir whose degree of expansion can be recorded either mechanically or electronically. Flow can be obtained by differentiating the changing volume. This type of instrument is still used in many pulmonary function facilities and exercise physiology laboratories, since they offer the highest accuracy available. There two types that use a water seal: (1) the counterbalanced bell and (2) the lightweight bell over a water seal (often called the Stead–Wells spirometer after the individuals who requested the device. These are shown in Fig. 1. Even though these devices have served the medical community well, their bulkiness and expense have motivated biomedical equipment developers to design smaller, more portable nonvolumetric units, even though they are not as accurate as the volumetric type. Three other variations of the volume displacement type exist: the wedge type, the bellows, and the rolling diaphragm, as shown in Fig. 2 (13).

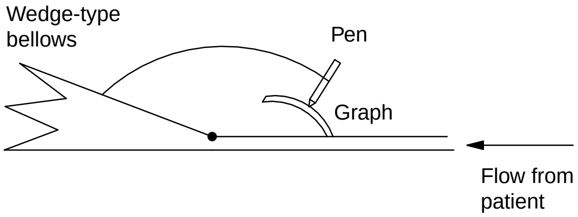


(a) Counterbalanced bell

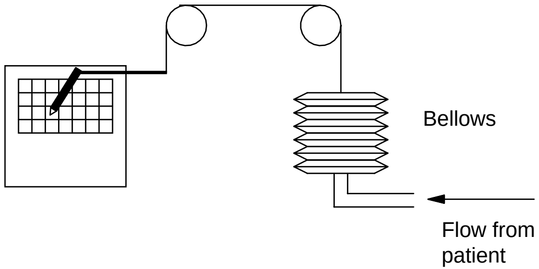


(b) Lightweight bell (stead-wells)

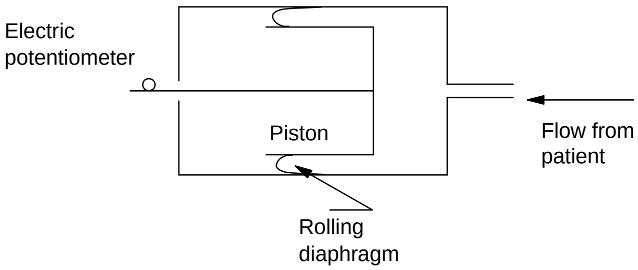
Figure 1. Water seal spirometer.



(a) Wedge type



(b) Bellows type



(c) Rolling diaphragm type

Figure 2. Non-water seal spirometers: (a) wedge type, (b) the bellows, and the (c) rolling diaphragm.

Other Applications of Volume Displacement Type. The water seal spirometers have also been used to monitor breathing over longer periods of time. Simply closing the circuit creates some problems since O_2 in the air is being consumed while and the mixture becomes progressively CO_2 rich. To solve this problem, the bell is originally filled with O_2, and a CO_2 scrubber (Baralyme, a trade name for generic BaO) is inserted into the circuit. After a few seconds, additional oxygen is inserted into the circuit. The setup and the resulting the waveform are shown in Fig. 3. This device has found applications to evaluate the metabolic rate (proportional to O_2 consumption).

This device has also been used to measure compliance, the elasticity of the respiratory system (RS). It is carried out by adding a series of weights on top of the spirometer bell, which increases the overall system pressure (14,15). The corresponding change in system volume

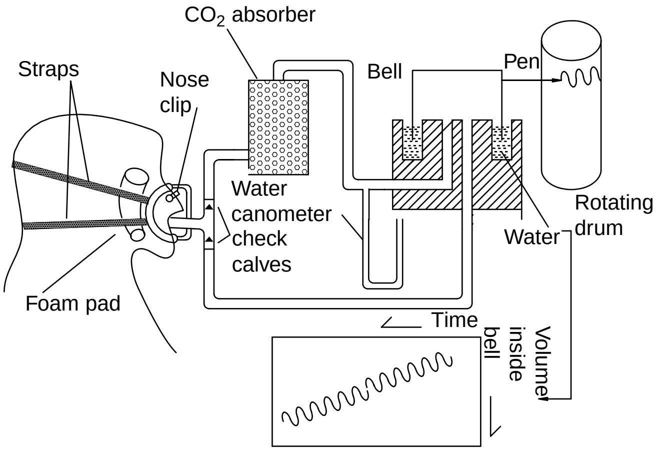


Figure 3. Closed-circuit spirometer.

detected at the bell, which corresponds to the increase in lung volume, is measured. The resulting respiratory compliance is given by

$$C_{\mathrm{RS}} = \frac{\Delta V}{\Delta P}$$

This approach has been automated by Crawford (16). Notice that the resulting compliance, the compliance of the respiratory system, C_{RS}, captures both the compliance of the lungs, C_{L}, and the compliance of the chest wall C_{CW}:

$$\frac{1}{C_{\mathrm{RS}}} = \frac{1}{C_{\mathrm{L}}} + \frac{1}{C_{\mathrm{CW}}}$$

and the lung compliance is made up of the compliance of the left and right lobes.

$$C_{\mathrm{L}} = C_{\mathrm{L-left}} + C_{\mathrm{L-right}}$$

The following type of pneumotachometers measure flow directly, instead of measuring change in volume. These devices electronically integrate the flow signal to obtain volume.

FLUID RESISTANCE TYPE OF PNEUMOTACHOMETERS

Fleish Pneumotachometers

Another device that has been and continues to be used extensively is the Fleish (or Fleisch) pneumotachometer (17). It consists of a fluid resistive device through which the air passes. The pressure drop across a fluid resistive element created by the respiratory flow is applied to a pressure transducer. Since there is a linear correlation between the measured pressure drop and flow, flow can be determined from the pressure drop. This approach has also been the primary workhorse of pulmonary instrumentation. There are two areas of concern with this type of device: (1) the linearity of the correlation between pressure drop and volumetric flow, and (2) the potential condensation of vapor droplets in the resistive element. Both of these factors can affect the device's accuracy. The American Thoracic Society (ATS), the medical section of the American Lung Association (18), has published standards regarding the required accuracy of the equipment (5% in some tests, while 3% in others) They have also established the conditions in which the results should be reported: BTPS. [BTPS stands for body conditions: normal body temperature (37 °C) ambient pressure saturated with water vapor].

Methods for Obtaining a Linear Flow-Pressure Drop Relationship. The design of the fluid resistive element in the Fleish pneumotachometer produces laminar flow. Three general approaches have been implemented to obtain this linearity. The first one uses capillary tubing placed in parallel (bundle); The second method uses a coiled metal strip with capillary tubing-like corrugations; The third one uses a porous medium, for example, a screen or paper similar to what is used in a vacuum cleaner bag.

In the capillary version, if the flow is laminar, the resulting pressure drop is given by

$$\Delta p = \frac{128\,\mu L}{N\pi D^4}\dot{V}$$

where Δp is the pressure drop, L is the length of the capillaries, N is the number of capillaries, D is the diameter of the capillaries, and μ is the absolute viscosity. The above equation can be expressed as

$$\Delta p = R\dot{V}$$

where R is the linear fluid resistive coefficient. Even though effort is taken to make the flow laminar, there are always some turbulent components. This turbulent behavior occurs primarily at the entrance and exit of the capillary tube bundle.

If the pressure drop is created with a square-edge orifice, the flow most likely will be turbulent, and the pressure drop is given by

$$\Delta p = \frac{\rho}{2C_{\mathrm{D}}A^2}\dot{V}^2$$

where ρ is the density, C_{D} is the discharge coefficient, and A is the area of the orifice. Since this function is truly an odd function [$f(-x) = -f(x)$], the even function above [$f(-x) = f(x)$] is modified by means of the absolute value sign is

$$\Delta p = \frac{\rho}{2C_{\mathrm{D}}A^2}|\dot{V}|\dot{V}$$

The above equation can be expressed as

$$\Delta p = k|\dot{V}|\dot{V}$$

where k is the nonlinear (quadratic) fluid resistive coefficient. In actual practice, most fluid resistors can be described as a combination of laminar and turbulent components:

$$\Delta p = R\dot{V} + k|\dot{V}|\dot{V}$$

where R and k are the linear and nonlinear (quadratic) fluid resistive coefficients. The expression for flow in terms of pressure drop is

$$\dot{V} = \frac{-R + \sqrt{R^2 + 4k\Delta p}}{2k}$$

The reader might suggest to themselves: Why not simply use a square-edge orifice and linearize the resulting signal (i.e., take the square root of the resulting pressure to obtain flow)? In some cases this is done, however, it presents some challenges. If the dynamic range of the flow (maximum flow to minimum flow) is 10–1, the resulting range of the corresponding pressure drops would be 100–1, a pressure signal that would be either too small (and consequently being difficult to measure) or too large (and consequently offering some detectable resistance to the patient).

The other concern with any fluid resistance type flowmeter is condensation. If the fluid resistive element of the flowmeter is at a lower temperature than the air exhaled from the subject (37 °C, saturated), there is a likelihood that the water vapor present in the air would condense on the fluid resistive element, changing the flow–pressure drop relationship. To avoid this problem, most Fleish-like pneumotachometers use a heating element to keep the fluid resistive element hotter than the flow. Despite these drawbacks, fluid resistance type pneumotachometers continue to be one of the most popular methods of measuring respiratory flow.

Osborne Pneumotachometer

The last section introduced problems associated with a square edge orifice for producing the necessary pressure drop. An innovative idea that overcomes these problems is the variable area Osborne flowmeter (19,20). In this flowmeter, the resistive element consists of a thin disk with a flap cut into it, as shown in Fig. 4. As the flow attempts to pass through the space between the flap and the disk, the flap bends, increasing the effective area of the orifice. As a result, the square-type relationship between flow and pressure no longer occurs; instead a linear one is obtained. The specific pressure–flow relationship is captured electronically at the time of calibration and later used to obtain flow from the measured pressure drop.

NONFLUID RESISTANCE TYPE OF PNEUMOTACHOMETERS

Respirometers

Another type of flow and volume indicator is called the Wright respirometer (21,22). It operates on the principle that the moving gas imparts movement on a rotating vane. In the newer units, the motion of the vane is sensed electronically. This type of device is often classified as turbine type.

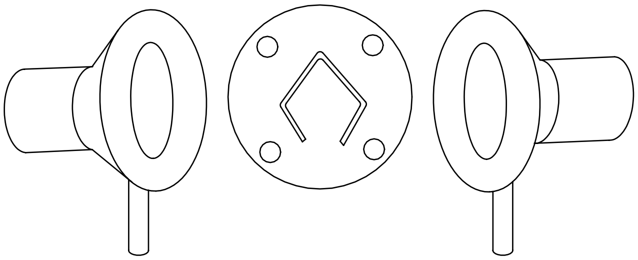


Figure 4. Flow element of the Osborne flowmeter (Adapted from Ref. (20).

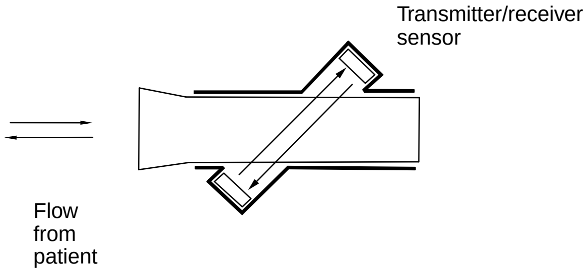


Figure 5. Ultrasound acoustic flowmeter. (Adapted from easy one literature, ndd medical technologies).

Another similar device made by the Wright company (13) records flow instead of expired volume. As flow passes through the device, it simultaneously deflects a vane that as it moves it opens additional passageways through which the flow exits. This mechanical unit is used for measuring peak flow.

Ultrasound–Acoustic Pneumotachometers

This flowmeter uses the Doppler phenomenon, which states that sound travels faster if the medium is also moving. It is very attractive since it provides a property-independent measurement of flow, that is, the measurement is independent of gas composition, pressure, temperature, and humidity. It does not compensate for changes to the air as it enters the respiratory system. (See Calibration section). Until recently, this type of flowmeter was too expensive for regular clinical use. One implementation (23) utilizes a sound pulse along a path that intersects the respiratory flow at an angle, as shown in Fig. 5. A pair of transmitters sends and receives sound in an alternating fashion. The sound pulse gets to the receiver faster if the sound wave is traveling in the same direction as the measured flow. On the other hand, the sound pulse that opposes the direction the respiratory flow takes longer. The difference between these two transit times is used for calculating flow.

Thermal Units

Another class of flowmeter uses thermistors (24,25) to measure flow. Thermistors are electrical resistors made of a material whose resistance decreases with temperature. As the flow passes by the thermistor bead, it attempts to decrease its temperature, which translates to an increase in resistance. An electronic circuit supplies the necessary current to maintain the thermistor at a constant temperature. As a result, the change in current is proportional to gas flow. Often these units are referred to as the hot-wire anemometer type.

Vortex Shedding Flowmeter

Whenever a flow field passes a structure (bluff body), it produces eddies (turbulence) past the structure (26,27). For a particular flow range, the frequency of the shedding is proportional to flow. This principle has been used to measure flow. The flowmeter made by Bourns Medical Systems (28) uses an ultrasonic transducer–receiver pair

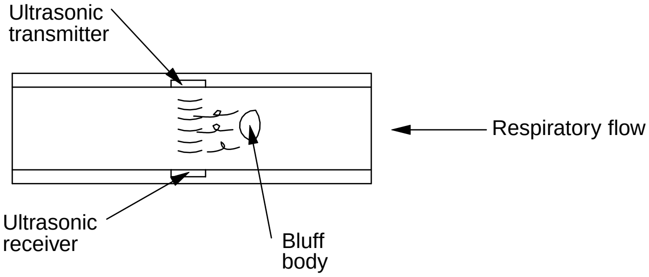


Figure 6. Ohio vortex (Bluff Body) respiration monitor. (Adapted from Ref. 13).

that detects the vortex created by the flow, and converts it to an electric signal. The manufacturer of the device claims that gas composition, temperature, and humidity will not affect the measurement process.

Flutter Flowmeter

Whenever flow passes a flexible structure, there is a possibility of a fluid–structure interaction, which makes the structure vibrate. This phenomenon is often called flutter. In aircraft, this is a major concern because wings that experience this phenomenon can break off; consequently aircraft designers have learned to avoid such a condition. On the other hand, developers of respiratory flowmeters have taken advantage of this principle to measure flow. The Ohio Vortex Respiration Monitor shown in Fig. 6 utilizes a light beam–photoelectric eye to capture the resulting vibration and convert it to flow.

Lift Force Gas Flow Sensor

Airfoils, (the shape of the wing in an aircraft), produce lift when subjected to a flow field. Svedin (29–31) is using this concept to measure respiratory gasses in medical applications. The sensor consists of two plates with polysilicon strain gages connected to a Wheatstone bridge. The plates deflect in response to the resulting lift force, in a direction normal to the flow field. One of the claims made by the developers of this device is the sensor's relative insensibility to inertial forces.

Fluidic Oscillator

Fluidics is a technology that was invented in 1959 and has been used in analogue and digital applications (32,33). It is very similar to pneumatics, but few or no moving parts are used. The devices can also be operated using liquids. The device that gave birth to the technology is the fluid amplifier, a device that with no moving parts can amplify a pressure differential. Figure 7 shows a cross-section of an amplifier, made by staking several perforated stainless steel sheets. Several amplifiers can be staged to achieve gains close to one-half of a million. The most successful fluidics device is the windshield water spray found in most cars. The windshield cleaning fluid or water enters into a cavity that makes the exiting jets oscillate, as shown in Fig. 8. Many applications are possible with this novel technology (34). It must be clarified that prior to the advent of MEMS and microchannels, the word "fluidics" was used only in dealing specifically with this type of devices. Today, any small channel carrying a fluid is called a fluidic device or a microfluidic channel.

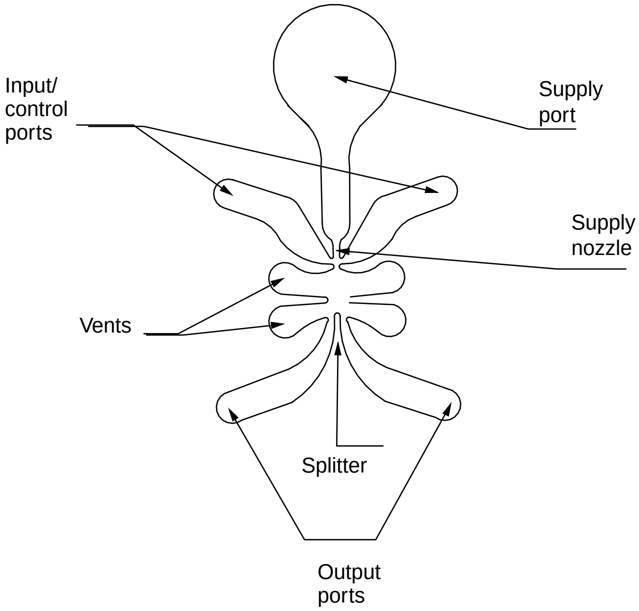


Figure 7. Fluidic fluid amplifier.

The fluidic oscillator flowmeter (unidirectional) offers much promise. It is constructed by configuring a fluid amplifier with feedback: the outputs are connected to the inputs, as shown in Fig. 9. This produces an oscillation whose frequency is proportional to flow. Even though this type of flowmeter has not been specifically used to measure respiratory flow, its performance has been demonstrated successfully as a flow sensor in gasoline delivery systems (35). It aspirates the displaced air in the fuel tank through small holes in the delivery nozzle. It withdraws a volumetric flow equal to the gasoline volumetric flow delivered to the tank. The purpose of course is to minimize the discharge of unburned hydrocarbons into the atmosphere.

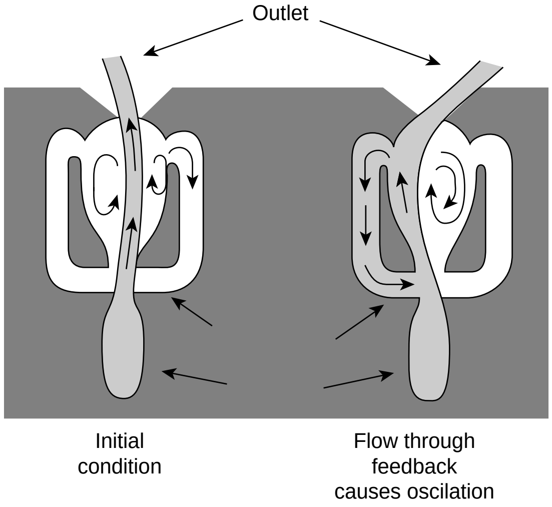


Figure 8. Fluidic windshield water spray.

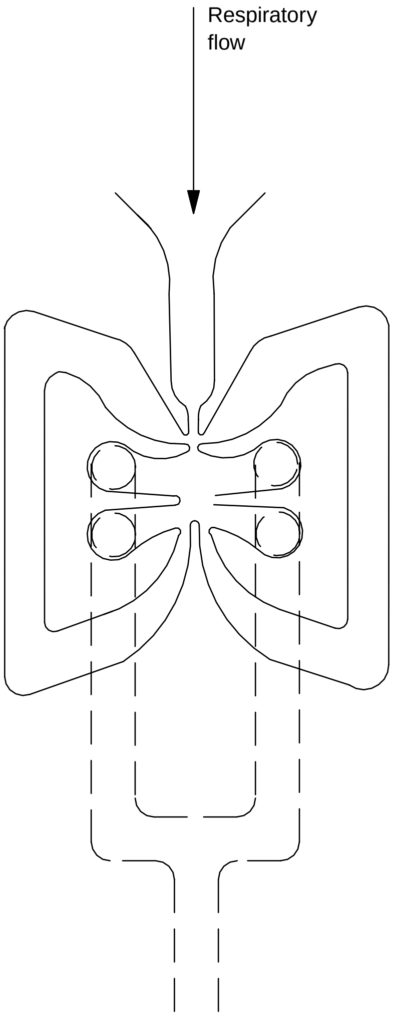


Figure 9. Amplifier with positive feedback: fluidic oscillator.

COMMERCIALLY AVAILABLE SPIROMETERS

For a list of most of the commercially available spirometers, see the online source of information by ADVANCE for Managers of Respiratory Care (36) and American Association of Respiratory Care (37).

CALIBRATION

Calibration of Volume Displacement Type Spirometers

This type of volumetric device (Figs. 1 and 2) rarely loses calibration since the geometry is fixed, especially in the type where the measurements of the volume changes are recorded by means of a pen writing on a revolving graph. Even in systems that have electronic position sensors, due to the robustness of these components, recalibration is seldom necessary. However, recalibration should be done on a regular basis to insure the absence of artifacts such as leaks, rough spots, and so on.

The most common method for performing this procedure is the calibrated syringe: a large piston cylinder device that allows the piston to move within two stops and delivering a fixed volume (typically 3L). The calibration process consists of pulling the piston out to the outer stop, connecting the syringe to the spirometer, and gently pushing the piston to the other stop. The resulting volume indication should be that of the calibrated syringe volume.

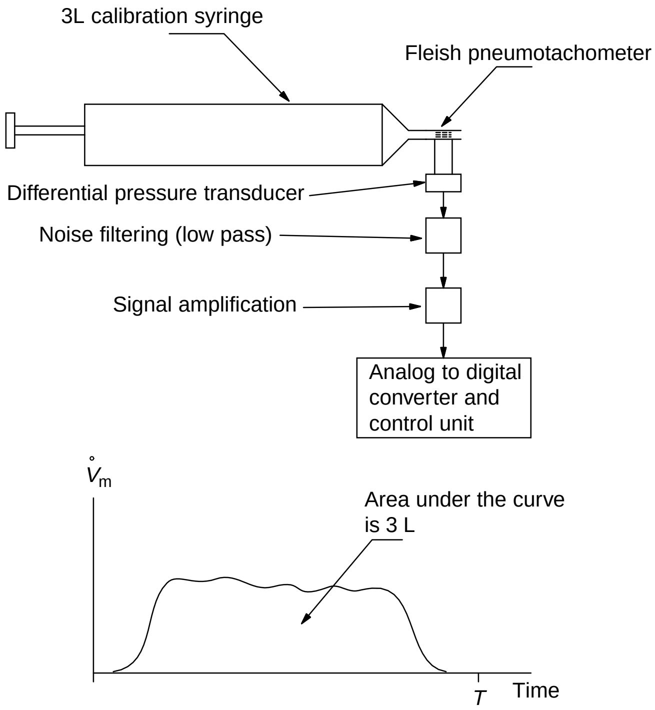

3L calibration syringe
Fleish pneumotachometer
Differential pressure transducer
Noise filtering (low pass)
Signal amplification
Analog to digital converter and control unit
$\mathring{V}_m$
Area under the curve is 3 L
T
Time

Figure 10. Syringe calibration procedure.

Calibration of Spirometers other than the Volume Displacement Type

There are three approaches for calibrating pneumotachometer that do not utilize volume displacement. They are (1) syringe, (2) wet test gas meter; and (3) comparison with a standard calibration flowmeter.

Syringe Approach. In performing this calibration procedure, it is assumed that the pneumotachometer is integrated into a data acquisition–calibration system. To perform the calibration, a fixed volume is passed through the pneumotachometer using a calibrated syringe, typically 3L. The data acquisition system records the resulting flow signal, $\dot{V}_M$, as shown schematically in Fig. 10. Then it integrates it (i.e., finds the area under the curve) to obtain the measured volume:

$$V_M = \int_0^T \dot{V}_M \, dt$$

Then it determines the correction factor, K, so that:

$$KV_M = 3$$

Afterward it uses K to adjust the measured flow:

$$\dot{V}_{\mathrm{TRUE}} = K\dot{V}_M$$

This approach assumes that the relationship between pressure drop and flow are linear. The user should consult the user's manual, since some of the pneumotachometers automatically adjust for BTPS.

Wet Test Gas Meter. This device is similar to the gas meter found in homes and industrial sites (38). It consists of a series of constant volume chambers that are filled by the incoming flow. They are attached to a rotating structure that enables the chambers to go through the following sequence: (1) fill the chamber with the incoming air, (2) create a water seal at the bottom of the chamber, (3) move the sealed chamber to the exhaust side, and (4) release the volume in the chamber to the outside. As a result, the incoming flow imparts a rotation that is observable from the outside by means of rotating hands. Consequently, there is a 1:1 correlation between the volume that passes through the device and the number of rotations indicated by the hand. To perform a calibration test, a known flow of gas is passed through the Wet Test Gas Meter and the flow recorded:

$$\dot{V} = \frac{(\text{number of revolutions})(\text{volume/revolution})}{\text{time required for above revolutions}}$$

Immediately afterward, the same flow is passed through the pneumotachometer under calibration. The constant flow is produced by applying a source of high pressure to a needle valve. Since the pressure resistance created by the Wet Test Gas Meter or the pneumotachometer is very small (a couple of inches of water), the flow is determined by the upstream pressure, Ps and the size of the orifice in the needle valve. The resulting output signal, *e*, is recorded, as shown in Fig. 11. The procedure is repeated for various needle valve settings to produce a curve similar to the one

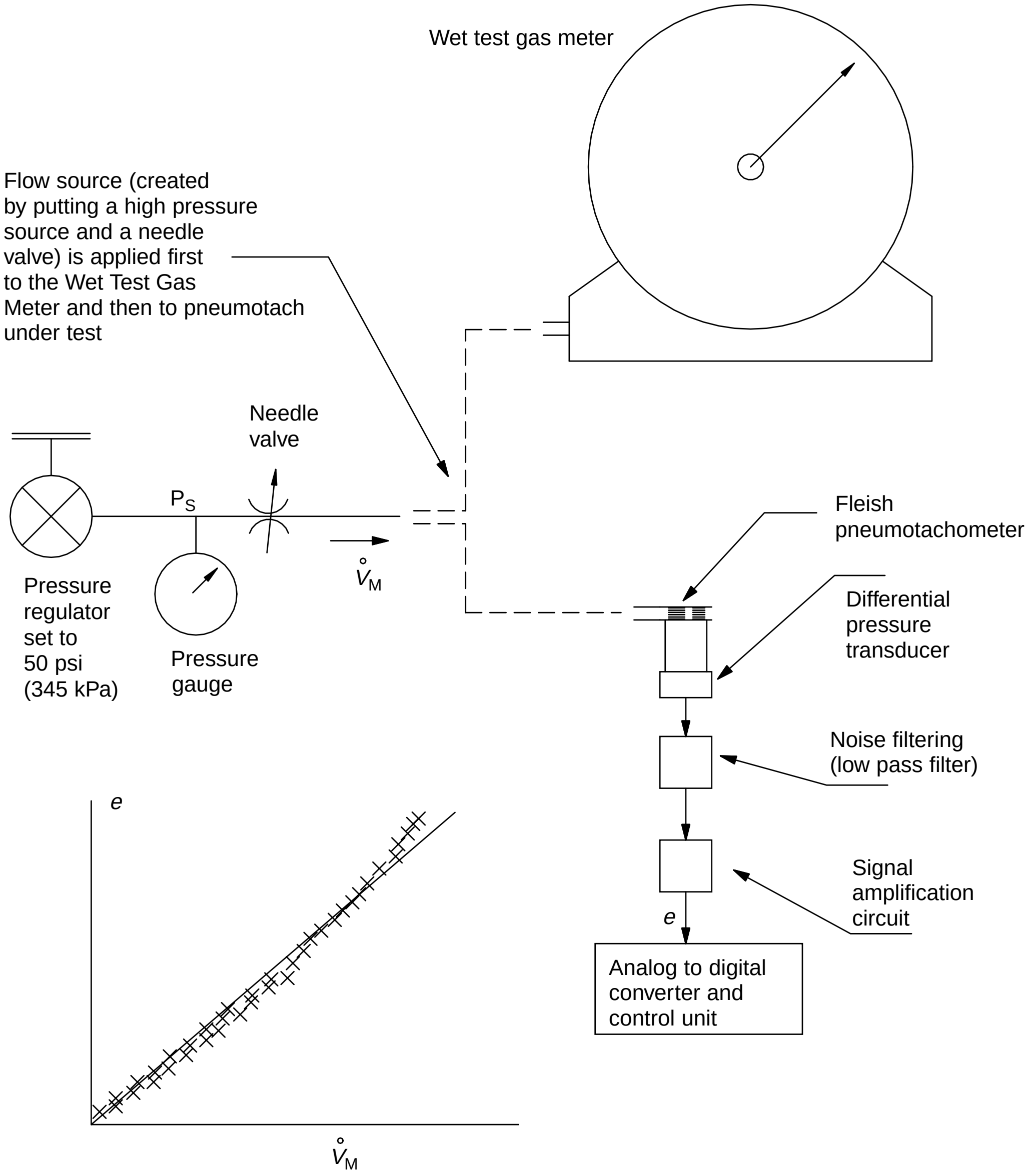

Wet test gas meter
Flow source (created by putting a high pressure source and a needle valve) is applied first to the Wet Test Gas Meter and then to pneumotach under test
Needle valve
P_S
$\mathring{V}_M$
Pressure regulator set to 50 psi (345 kPa)
Pressure gauge
Fleish pneumotachometer
Differential pressure transducer
Noise filtering (low pass filter)
Signal amplification circuit
e
Analog to digital converter and control unit
e
$\mathring{V}_M$

Figure 11. Calibration procedure using the Wet Test Gas Meter.

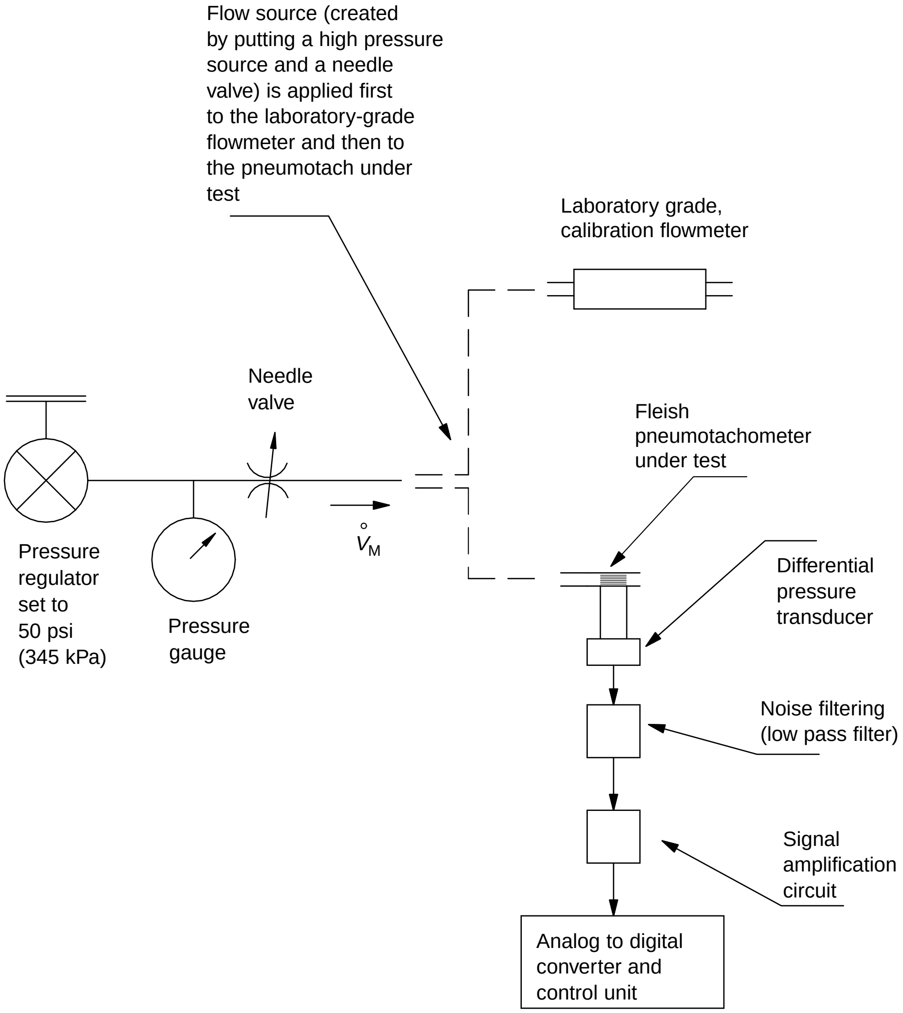

Flow source (created by putting a high pressure source and a needle valve) is applied first to the laboratory-grade flowmeter and then to the pneumotach under test
Laboratory grade, calibration flowmeter
Needle valve
Fleish pneumotachometer under test
Differential pressure transducer
Pressure regulator set to 50 psi (345 kPa)
Pressure gauge
$\overset{\circ}{V}_M$
Noise filtering (low pass filter)
Signal amplification circuit
Analog to digital converter and control unit

Figure 12. Calibration using a Laboratory grade Flowmeter.

shown in Fig. 11 that can then be used to obtain either a linear approximation or a polynomial fit. Afterward, the calculated flow is adjusted for conversion into standard conditions: BTPS [normal body temperature (36 °C) ambient pressure saturated with water vapor]. This type of calibration device has an accuracy of ±0.5%.

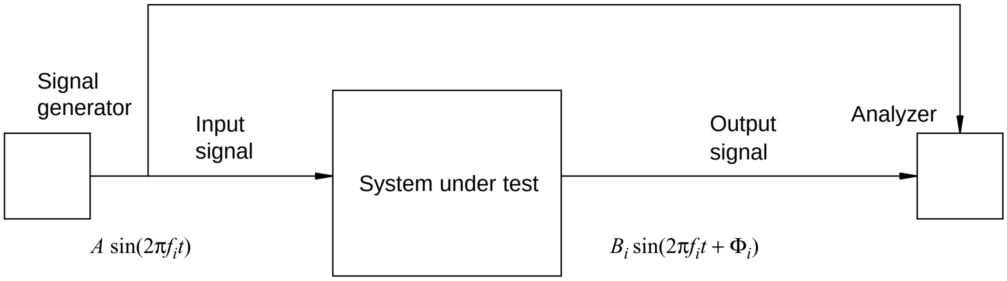
Calibration Using a Laboratory Grade Flowmeter. This method is possible when a laboratory-grade flowmeter is available. As done in the Wet Test Gas Meter, a series of flows are passed through both the laboratory grade flowmeter and the pneumotachometer under test, as shown in Fig. 12. The output of the pneumotachometer under test is made to agree with the flow indicated by the laboratory grade flowmeter.

Piston Prover. Flow meters can also be calibrated using a piston prover (39). This device consists of a precision bore glass tube with a piston, that moves due to the displacement of the gas whose flow rate is being measured. The piston, which is slightly smaller than the bore of the tube, has a groove filled with mercury to create a leak-less, low friction seal. The flow is determined by the time that it takes for the piston to travel between two reference positions, timed using light beams. This device has an accuracy of better than ±0.2%. NIST provides calibration services for flowmeters using this technique and apparatus.

Soap Bubble Technique. From time to time, researchers need to measure a particular flow, but do not have the specialized calibration equipment required. In cases like this, the soap bubble technique can be used. It is implemented by placing a soap film across the open end of a constant, cross-section tube. The other end of the tube is connected to a barb fitting that will easily receive a plastic tubing through which the flow to be measured is passing. Next, the soap film is moved toward the other end (the side with the fitting), by using a source of vacuum that is momentarily connected to the end with the fitting. Then, the tubing with the flow to be measured is connected to fitting, and the time that it takes for the soap film to travel between two reference positions is recorded. This information is then used to calculate flow. This technique resembles the piston prover, except that a soap film is used instead of a piston.

PNEUMOTACHOMETER DYNAMIC CHARACTERISTICS

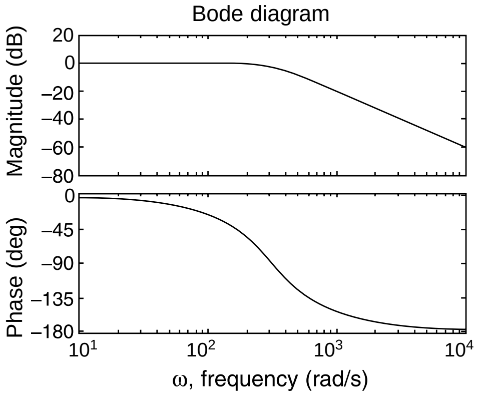
Another important characteristic of a pneumotachometer is its ability to measure rapidly changing flows. If one



(a) General schematic

test no	f_i	B_i	B_i/A	$20\log(B_i/A)$	ϕ_i
1					
2					
3					
⋮					
n					

(b) Experimental data



(c) Magnitude and phase Bode plot

Figure 13. Block diagram of procedure used for obtaining a frequency response test.

attempts to measure a fast changing flow signal with a flowmeter that does not have the adequate dynamic characteristics or "frequency response", one would obtain inaccurate results. One of the most common vehicles for capturing a device's frequency response is called the Bode Plot (40). This section addresses the general test procedure and presentation format used for obtaining a frequency response or Bode plot. Afterward, it presents the method used for performing a frequency response test on flowmeters.

One of the fundamental properties of a linear system is that if the system is excited with a sinusoidal input signal, after the transients die out, its output will also exhibit a sinusoidal behavior. A frequency response test consists of applying sinusoids of different frequencies and measuring the amplitude of the output and the phase difference between the input and output. Figure 13a shows a sinusoidal generator producing a sine wave with amplitude A and frequency f_i, being applied to a linear system. It also shows the corresponding output, another sinusoid of the same frequency, but with amplitude B_i and phase Φ_i. This procedure is repeated for different frequencies in order to obtain a table as shown Fig. 13b. Notice that the ratio of the output–input amplitude is calculated as well as 20 log of this ratio. This latter quantity has units called dB or deciBels in honor of the communication pioneer, Alexander Graham Bell. The Bode plot, which is actually made up of two plots, presents this information in graphical form. The magnitude plot shows the amplitude ratio in dB (i.e., 20 log B_i/A) on the vertical axis and the frequency [$\omega = 2\pi f$], in a logarithmic scale, on the horizontal axis. This is shown in Fig. 13c. Similarly, the Phase Bode Plot shows the phase against the various input frequencies. This is shown in Fig. 13d. The particular plot shown here is descriptive of a

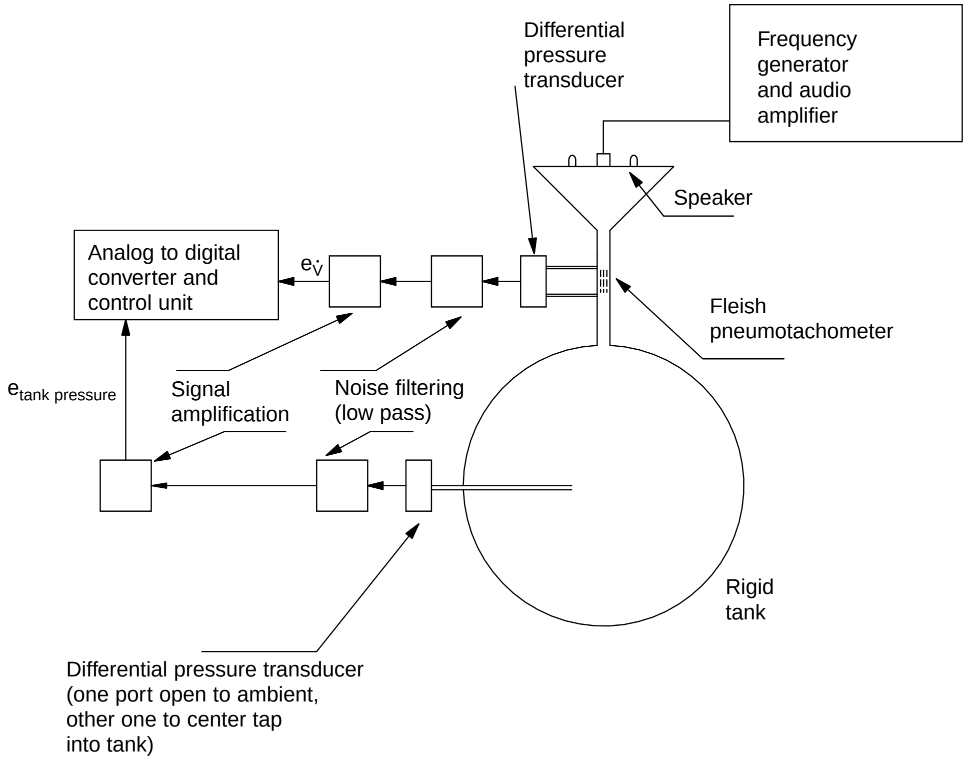

Differential pressure transducer
Frequency generator and audio amplifier
Speaker
Analog to digital converter and control unit
$e_{\dot{V}}$
Fleish pneumotachometer
$e_{\text{tank pressure}}$
Signal amplification
Noise filtering (low pass)
Rigid tank
Differential pressure transducer (one port open to ambient, other one to center tap into tank)

Figure 14. Test setup used for measuring frequency response of a pneumotachometer.

device that has low pass characteristics. This means that it passes signals of low frequencies at a constant gain (flat or horizontal portion), but attenuates ones that are of high frequencies (sloped portion of the graph). For any particular pneumotachometer application, it is desirable to use one that is flat in the range of frequencies that will be present in flow waveform that is to be measured. As an example of this, the ATS Standardization of Spirometry (18) stipulates, "Measuring PEF requires an instrument that has a frequency response that is flat (±5%) up to 12 Hz".

The reader may question the emphasis on sinusoidal inputs when in the real world, very few signals are sinusoidal. Representation in terms of sinusoids provides a general framework for dealing with any type of periodic signals, periodic function can be decomposed into a summation of sinusoids (Fourier Series).

Performing a frequency response test on a pneumotachometer is more challenging than for an electronic device where both the input and output are electronic signals that are easily measurable. Development of this experimental procedure for pneumotachometers has been refined by Jackson and Vinegar (41). The test set up can be represented as shown in Fig. 14. For purposes of this explanation, assume that the pneumotachometer being tested is of the Fleish type, a fluid resistance with a pressure transducer for measurement of the pressure drop. To perform the frequency response test, a small loudspeaker is used to generate the sinusoidal flow that is applied to the pneumotachometer under test. On the other side of the pneumotachometer, a large, rigid tank is connected. Notice that because the tank is filled with ambient air, a compressible fluid, flow enters and exits it. Of course, when the flow enters the tank, the pressure in it increases, and when the flow exits, the pressure decreases. The experimental procedure requires a second pressure transducer that measures the pressure inside the tank. This signal is used to calculate the flow in and out of the tank. Referring to Fig. 14, $e_{\dot{V}\,\text{measured}}$ is the pneumotachometer output signal (the output voltage of the pressure transducer connected to the pneumotachometer) and $e_{\text{tank pressure}}$ is the tank's pressure transducer output signal. The flow into the tank is given by

$$\dot{V}_{\text{true}} = K_1 \frac{d(e_{\text{tank pressure}})}{dt}$$

A frequency response test is performed by varying the frequency of the loudspeaker, collecting the data and plotting it in Bode format. The response is flat as long as the ratio of $e_{\dot{V}\text{measured}}/\dot{V}_{\text{true}}$ is constant.

CORRECTION FOR STANDARD CONDITIONS

Often the state in which the flow is measured is not reflective of the physiological conditions in which flow becomes meaningful. In other cases the calibration of the flowmeter takes place under a different set of conditions than what the flowmeter will experience in the clinical setting. In both cases, one needs to correct the resulting measurements to accurately capture the true physiological event happening. To illustrate the correction process, an example is provided here. Consider a pneumotachometer that is to be used for measurement of inspiratory flow. Further assume that the pneumotachometer has been calibrated with room air, using the standard 3L syringe. Consequently, there is no need to correct the flowmeter reading since it will measure properly the flow that passes through it during the test. That is, the conditions for calibration and clinical testing are the same. However,

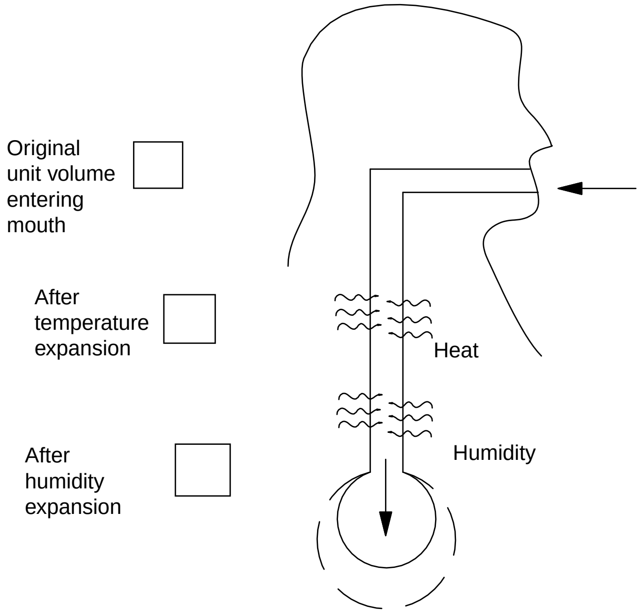

Original unit volume entering mouth
After temperature expansion
After humidity expansion
Heat
Humidity

Figure 15. Correction for volume expansion due to increase in temperature and humidity.

when the air enters the respiratory system, it experiences two changes: its temperature increases from room temperature to body temperature, and the humidity from 0% (assuming a very dry day) to 100% RH. Both of these processes need to be taken into consideration if it is desired to know the actual volume that is expanding the alveoli. In reality, both processes are happening simultaneously, but for purposes of this example, we will assume that the air is heated first and then it becomes saturated. Both of these transformations cause the air volume to expand. Figure 15 shows the expansion process.

The increase in temperature causes the air to expand (Dalton’s law of gasses). The correction factor for temperature is

$$\mathrm{CF}_{\mathrm{Temperature}} = \frac{T_{\mathrm{body}} + 273.16\,^{\circ}\mathrm{C}}{T_{\mathrm{inlet}} + 273.16\,^{\circ}\mathrm{C}}$$

Assuming that the inlet temperature is 21.1 °C (70 °F) and that the body temperature 37 °C (98.6 °F) this correction factor $\mathrm{CF}_{\mathrm{Temperature}}$ is 1.05.

The increase in humidity also creates an expansion. It is due to the effect of water vapor on the partial pressures. The total pressure of the dry gas is equal to the sum of the partial of its three main constituents N_2, O_2, and CO_2.

$$P_{\mathrm{total}} = P_{\mathrm{N_2}} + P_{\mathrm{O_2}} + P_{\mathrm{CO_2}}$$

However, as the air travels through the airways, it is humidified. We assume that the total pressure remains constant as the air moves through the airways. Now the total pressure becomes:

$$P_{\mathrm{total}} = P_{\mathrm{H_2O}} + P_{\mathrm{N_2}} + P_{\mathrm{O_2}} + P_{\mathrm{CO_2}}$$

The presence of the partial pressure of the water vapor ($P_{\mathrm{H_2O}}$) causes the sum of partial pressures of the last three constituents to decrease. This in turn causes the volume to expand. Thus the correction factor expression for humidity is

$$\mathrm{CF}_{\mathrm{HUMIDITY}} = \frac{p_{\mathrm{ambient}}}{p_{\mathrm{ambient}} - p_{\mathrm{sat}}}$$

Assume that the ambient barometric pressure is 29.92 mmHg (98.3 kPa) and at a body temperature of 36 °C, the water vapor pressure is 1.78 mmHg (6 kPa). Substituting these values into the previous equation produces a correction factor $\mathrm{CF}_{\mathrm{HUMIDITY}}$ of 1.06.

Combining both factors, the total correction factor becomes 1.12. This means that every milliliter that enters the mouth expands by 12% by the time that it reaches the alveoli. Consequently, measured flows need to be increased by 12%.

Comments about Complexities in the Measurement and Correction Process

More accurate (than the upper limits established by the ATS), bidirectional flow measurement is a challenging task when using pneumothachometers other than the volumetric type. This is true since the air volume varies due to density changes (resulting from temperature and composition differences of the incoming and outgoing air). The composition of inspired and expired air is shown in Table 1. The problem is made even more complicated if Fleish-type pneumotachometers are used, since now viscosity (the critical variable in the flow–pressure drop relationship) is influenced by both temperature and gas composition.

OVERVIEW OF DESIGN SPECIFICATIONS

This section presents some of the technical specifications of Fleish-type pneumotachometers, specifically those manufactured by Hans Rudolph a major manufacturer of respiratory components (42). These characteristics are shown in Table 2.

It is difficult to describe accuracy of each device in detail, since it depends not only on the hardware (i.e., resistive fluid sensor element) and the measurement process, but also on the correction factors that are applied (gas composition,

Table 1. Normal Gas Concentrations of Air[a]

Source of sample	Nitrogen (N_2),%	Oxygen (O_2),%	Carbon Dioxide (CO_2),%	Water Vapor (H_2O),%
Inspired air (dry)	78.65	20.9	0.04	0.5
Alveolar air (saturated)	75.45	13.2	5.2	6.2
Expired air (saturated)	74.8	15.3	3.7	6.2

[a] Adapted from Ref. (42).

Table 2. Characteristics of Commercially Available Pneumotachs[a,b]

Application	Flow Range ($L \cdot min^{-1}$)	Fluid Resistance ($mmH_2O/(L \cdot min^{-1})$)	Max. Pressure Drop at End of Range (mmH_2O)
Premature (38 week gestational)	0–10	1.0	10
Neonate (Birth to 1 month)	0–10	0.70	7
Infant (1–12 month)	0–35	0.20	7
Pediatrics	0–100	0.10	10
Pediatrics	0–160	0.10	16
Adults	0–400	0.04	16
Adults	0–800	0.02	16

[a] adapted from Ref. 43.

[b] Note: 1 psi = 1 lbf/in^2 = 6.89 kPa = 704 mmH_2O.

temperature, relative humidity, differences between atmospheric conditions at the time of calibration and those at the time of use, and corrections for quadratic-type pressure drop). As mentioned before, most manufacturers simple indicate that their products meet the ATSs standards.

INDIRECT TECHNIQUES FOR MEASURING FLOW

There are various methods that provide an estimate of respiratory flow, by means of an indirect measurement. These techniques find an application in cases where continuous connection to a pneumotach through a mouthpiece or mark is not feasible.

INDUCTIVE PLETHYSMOGRAPHY

Respiratory inductive plethysmography (RIP) has been used for many years for respiratory monitoring. Used in intensive care units worldwide for monitoring respiratory activity, primarily tidal volume. During inspiration, due to the bucket-handle effect of the ribs and the outward displacement of the abdomen, there is an increase in the cross-sectional area of the rib cage and abdomen, which translates into an increase in circumference (or more accurately, perimeter). This increase in perimeter is measured using elastic bands and correlated to a specific lung volume increase, or often used as a simple, relative measurement of lung volume expansion.

This technique used by the LifeShirt (44) is a vest-like, portable physiological monitoring system. In the LifeShirt, “two parallel, sinusoidal arrays of insulated wires embedded in elastic bands are woven into a flexible garmet. Extremely low voltage electrical current is passed through the wire creating an oscillating circuit. As the body chambers expand and contract, the electrical sensors generate different magnetic fields that are converted into proportional voltage changes over time (i.e., waveforms)”. For calibration, which is done immediately after putting on the vest, the user breathes into a fixed-volume, calibration bag. Then the associated tidal volume is correlated to the measured body’s expansion.

Respiratory Sounds

There is a correlation between respiratory flow and breath sounds. This method has been pursued for many years and continues to be an ongoing topic for research (45,46). Tracheal breath sounds are preferred by many investigators since they are louder than chest sounds and since there is more correlation of flow at the trachea and less filtering from the chest wall (47).

The U.S. Army (48) has been extremely active in developing physiological monitoring systems based on the sounds detected at the neck. Their motivation for this research is based on the premise that continuous monitoring of the soldier’s health “can provide exceptional improvement to survivability, mobility, and lethality” (49). This group has generated a considerable amount of methodology as well as easy-to-build sensors and equipment for this application (50,51).

Some development has been done to obtain an estimate of actual flow from respiratory sounds (instantaneous measurement of the magnitude and direction of respiratory flow) (52). This is possible since inspiration and expiration have different “sounds”, due to the asymmetrical nature of the respiratory passageways. This approach is implemented by initially, simultaneous measuring both tracheal sounds (using a small microphone attached to the neck) and actual flow (using a standard pneumotachometer). After recording data for 1–2 min, these two waveforms that are fed into a computer program that produces a correlation algorithm, which can be used subsequently to predict flow (magnitude and direction) from the tracheal sounds. The long-terms aim of this effort is detection of SIDS candidacy. The basic concept is that there might be a flow patterns/signature that could be construe as an early-predictor of SIDS. This test would be done during the first night that a newborn spends at the hospital. If a suspicious pattern is found, then a baby monitor could be sent home with the baby. It could also find application in the analysis the breathing patterns of athletes (in which a mouthpiece-tachometer is not feasible).

SUMMARY

Pneumotachometers have contributed and will continue to contribute significantly to our understanding of the respiratory system. They are an essential tool in the fight against respiratory diseases. In addition new applications are being developed that aim at the prevention of disease.

BIBLIOGRAPHY

1. The Bible, Chapter 2, verse 7.
2. Rupple G. Manual of Pulmonary Function Testing. 2nd ed. St. Louis: C.V. Mosby Company; 1979.
3. Fishman AP. Assessment of Pulmonary Function. New York: McGraw-Hill; 1980.
4. Nunn JF. Applied respiratory Physiology. 4th ed.
5. Pimmel R, Fullon JM. Characterizing Respiratory Mechanics with Excitation Techniques. Ann Biomed Eng 1982;9:475–488.
6. Bakos JH. Estimation of the Total Respiratory Parameters in Paralyzed and Free-breathing Rabbits by the Technique of Forced Oscillation, M.S. dissentation, Pennsylvania State University; 1979.
7. Tsai MJ, et al. Respiratory parameter estimation using forced oscillatory impedance data. J Appl Physiol 1977; 43(2):322–330.
8. Goldman MD. Clinical applications of forced oscillation. Pulmonary Pharmacol Therapeut 2001;14:341–350.
9. Schmid-Schoenbein GW, Fung YC. Forced perturbation of respiratory system: (A) The traditional model. Ann Biomed Eng 1978;6:194–211.
10. Schmid-Schoenbein GW, Fung YC. Forced perturbation of respiratory system: (B) A continuum mechanics analysis. Ann Biomed Eng 1978;6:367–398.
11. Macia NF, Dorson WJ, Higgins WT Jr. Lung-diaphragm Model of the Respiratory System for Parameter Estimation Studies Presented at the 1997 International Mechanical Engineering Congress & Exposition, Dallas, TX, November 1997.
12. Macia NF. Noninvasive, Quick Obstruction Estimation Method for the Measurement of Parameters in the Nonlinear Lung Model, Ph.D. dissertation, Arizona State University, August 1988.
13. McPherson SP. Respiratory Therapy Equipment. 2nd ed. St. Louis: C.V. Mosby Co.; 1981.
14. Cherniack RM, Brown E. A simple method for measuring respiratory compliance: Normal values for males. J Appl Physiol 1965;20(1):
15. Merth IT, Quanjer PH. Respiratory system compliance assessed by the multiple occlusion and weighted spirometer method in non-intubated healthy newborns. Pediatr-Pulmonol 1990;8(4):273–279.
16. Crawford S. Automated System for Measurement of Total Respiratory Compliance, Applied Project Report, Arizona State University East, 2003 (paper also under preparation).
17. Fleish A. Der Pneumotachograph-ein Apparat zur Geshwindigkertsregistrierung der Atemluft Pflugers. Arch Gas Physiol 1925;209:713.
18. The American Thoracic Society, Standardization of Spirometry, 1994 Update. Standardization of Spirometry, 1994 Update. Am J Respir Crit Care Med 1995;152:1107–1136.
19. Osborne JJ. Monitoring respiratory function. Crit Care Med 1974;2:217.
20. Osborne JJ. A flowmeter for respiratory monitoring. Crit Care Med 1978;6:349.
21. Sukes MK, NcNicol MW, Campbell EJM. Respiratory Failure. 2nd ed. Oxford (UK): Blackwell Scientific Publications; 1976.
22. Wright Respiratoy, Harris Calorific; Cleveland, Ohio.
23. ndd Medical technologies, 17 Progress Ave., Chelmsford, (MA), www.nddmed.com.
24. Sulton FD, Nett LM,Petty TL. A new ventilation monitor for the intensive care unit. Care Resp 1974;19:196.
25. Petty TL. Intensive and Rehabilitative Respiratory Care. 2nd ed. Lea and Febiger: Philadelphia; 1974.
26. McShane JL, Geil FG. Measuring flow. Res/Devel February 1975; 30.
27. Frederick G. Application of bluff body vortex shedding and fluidic circuit techniques to control of volumetric flow, M.S.E. dissentation; Arizona State University; 1974.
28. Bourns Medical Systems, Riverside (CA).
29. Svedin N, Kälvesten E, Stemme G. A new edge-detected lift force flow sensor presented at The 10th International Conference on Solid-State Sensors and Actuators (TRANSDUCERS'99) in Sendai, Japan; June 7–10, 1999.
30. Svedin N. A lift force flow sensor designed for acceleration insensitivity. Sensors Actuators 1998;68(1–3): 263–268.
31. Svedin N. A new silicon gas-flow sensor based on lift force. IEEE/ASME J Microelectromech Syst 1998;7(3):303–308.
32. Behrens CW. What is fluidics. Appl Manuf July 1968.
33. Fluidics flown on. The Engineer 28 June 1990.
34. Drzewiecki TM, Macia NF. Fluidic Technology: Adding Control, Computation and Sensing Capability to Microfluidics., Smart Sensors, Actuators, and MEMS conference at SPIE's International Symposium on Microtechnologies for the New Millennium, Gran Canaria, Spain; May 2003.
35. Thurston J. Personal communications, Honeywell, Tempe, (AZ).
36. ADVANCE for Managers of Respiratory Care, www.ADVANCEforMRC.com.
37. American Association of Respiratory Care, www.aarc.org and http://buyersguide.aarc.org/.
38. Varlen Instruments Inc., 2777 Washington Blvd, Bellwood, IL 60104, (800) 648-3954.
39. Wright JD, Matingsly GE. NIST Calibration Services for Gas Flow Meters. NIST special Publication 250–49.
40. Macia NF, Thaler GJ. Modeling and Control of Dynamic Systems. Thomson Delmar Learning; 2004.
41. Jackson AC, Vinegar A. A technique for measuring frequency responses of pressure, volume and flow transducers. J Appl Physio Respirat Environ Excercise Physiol 1979;47(2): 462–467.
42. Martini FH. Fundamentals of Anatomy and Physiology. 4th ed. New York: Prentice Hall; 1998.
43. Hans Rudolph, Inc., Kansas City, Mo., www.rudolphkc.com.
44. Vivo Metrics, Inc. Venturo, CA, www.vivometrics.com.
45. Graviely N. Breath Sounds Methodology. Boca Raton (FL): CRC Press; 1995.
46. Soufflet G, et al. Interaction between tracheal sound and flow rate: A comparison of some flow evaluations for lung sounds. IEEE Trans Biomed Eng 1990;37(4):
47. Mussell MJ, Nakazono Y, Miyamoto Y. Effect of air flow and flow transducer on tracheal breath sounds. Med Biol Eng Comput November 1990.
48. www.arl.army.mil/sed/acoustics.
49. Scanlon MV. Acoustic Sensor Array Element Extracts Physiology During Movement (internal document, available at the above, Army web site)
50. Scanlon MV. Acoustics Sensor for Health Status Monitoring. Proceeding of IRIS Acoustic and Seismic Sensing. 1998. Vol. II, 205–222. (Also available at the above, Army web site.)
51. Bass JD, Scanlon MV, Mills TK, Morgan JJ. Getting Two Birds with one Phone: An acoustic sensor for both speech recognition and medical monitoring. presented in poster format at the 138th Meeting of the Acoustical Society of America. November, 1999. (Also available at the above, Army web site.)
52. Gudala SG. Estimation of Air Flow from Tracheal Breath Sounds, Master of Technology Applied Project, Arizona State University East; May 2003.

See also PULMONARY PHYSIOLOGY; RESPIRATORY MECHANICS AND GAS EXCHANGE; VENTILATORY MONITORING.

POLYMERASE CHAIN REACTION

Michael L. Metzker
Baylor College of Medicine
Houston, Texas
Thomas C. Caskey
Cogene Biotech Ventures
Houston, Texas

INTRODUCTION

Few techniques rival the impact that the polymerase chain reaction (PCR) has made in the age of molecular biology. Cloning and deoxyribonucleic acid (DNA) sequencing are other such techniques that have become embedded into everyday life on the molecular biologist's bench. Over 60 books alone (not to mention the tens of thousands of research articles) have been devoted to the strategies, methods and applications of PCR for the identification, detection and diagnosis of genetic and infectious diseases. Rightfully so, the inventor of PCR, Kary B. Mullis, was awarded the Nobel Prize in Chemistry for his discovery of the technique in 1993. However, PCR has not been without controversy. In 1989, DuPont challenged the validity of the Cetus PCR patents in federal court and with the Office of Patents and Trade Marks, and by 1991 the Cetus patents were unanimously upheld and later sold to Hoffman La Roche for $300 million. More recently, in 1993, Promega has challenged the validity of the Hoffmann La Roche *Taq* DNA polymerase patent that is currently pending. In this article, we attempt to provide a comprehensive overview for the molecular biologist when applying PCR to his/her application of interest.

DNA POLYMERASE REACTION

The DNA replication is an inherent process for the generation and evolution of future progeny in all living organisms. At the heart of this process is the DNA polymerase that primarily synthesizes new strands of DNA in a $5'\rightarrow3'$ direction from a single-stranded template. Most native DNA polymerases, however, are polyfunctional and show $5'$-exonuclease and/or $3'$-exonuclease activities that are important for cellular DNA repair and proofreading functions. Numerous molecular biology applications have harnessed these activities, such as labelling DNA by nick translation and TaqMan assays (see below), and endrepair of sheared DNA fragments and improving DNA synthesis fidelities, respectively. The PCR is an elegant, but simple, technique for the *In vitro* amplification of target DNA utilizing DNA polymerase and two specific oligonucleotide or primer sequences flanking the region of interest. PCR is a cyclic process of double-strand separation of DNA by heat denaturation, specific hybridization or annealing of short oligonucleotide primers to singlestranded DNA, and synthesis by DNA polymerase (1,2). Each cycle doubles the region marked by the primer sequences. By sequential iteration of the process, PCR exponentially generates up to a billion of copies of the target within just a few hours (Fig. 1).

The specificity of PCR is highly dependent on the careful design of unique primers with respect to the genome under investigation and the nucleotide composition of the primer sequences. Theoretically, a 16-mer (4^{16}) is of sufficient length to represent all unique primer sequences from a completely random genome size of 3 billion base pairs. In the real world, however, all genomes are not random and contain varying degrees of repetitive elements. For the human genome, Alus, LINEs (long interspersed DNA elements) and low complexity repeats are frequently observed and should be avoided in primer design when possible. There are a few simple rules for designing primer sequences that work well in PCR. In practice, PCR primers should be between 18 and 25 nucleotides long, have roughly an equal number of the four nucleotides, and show a $G + C$ composition of 50–60%. Commercially available oligonucleotide synthesizers that show phosphamidite coupling efficiencies $> 98\%$ mean that primers of this size can usually be used in PCR without purification. A variety of computer programs are available for selecting primer sequences from a target region. Many of these programs will reveal internal hairpin structures and self-annealing primer sequences, but manual inspection of the oligonucleotide is still necessary to maximize successful PCR amplifications.

The concentrations of the PCR cocktail ingredients are also important for product specificity, fidelity and yield. In addition to *Taq* DNA polymerase and primers, the PCR mixture contains the cofactor magnesium ion (Mg^{2+}), the four $2'$-deoxyribonucleoside-$5'$-triphosphates (dNTPs) and the buffer. In general, PCR reagent concentrations that are too high from standard conditions result in nonspecific products with high misincorporation errors, and those that are too low result in insufficient product. A typical 50 μL PCR cocktail that contains 0.4 $\mu mol{\cdot}L^{-1}$ of each primer, 200 $\mu mol{\cdot}L^{-1}$ of each dNTP, 1.5 $mmol{\cdot}L^{-1}$ $MgCl_2$, and 1.25 units *Taq* DNA polymerase in 10 $mmol{\cdot}L^{-1}$ tris-HCl, pH 8.3, 50 $mmol{\cdot}L^{-1}$ KCl buffer works well for most PCR applications. The optimal Mg^{2+} concentration, however, may need to be determined empirically for difficult target templates. The performance and fidelity of *Taq* DNA polymerase are sensitive to the free Mg^{2+} concentration (3), which ionically interacts with not only the dNTPs but also with the primers, the template DNA, ethylenediaminetetraacetic acid (EDTA), and other chelating agents. In most cases, the Mg^{2+} concentration will range from 1.0 to 4.0 $mmol{\cdot}L^{-1}$.

The number of cycles and the cycle temperature–length of time for template denaturation and primer annealing and extension are important parameters for high quality PCR results. The optimal number of cycles is dependent on the starting concentration or copy number of the target DNA and typically ranges from 25 to 35 cycles. Too many cycles will significantly increase the amount of nonspecific PCR products. For low copy number targets, such as the integrated provirus of *Human immunodeficiency virus type 1* (HIV-1) from human genomic DNA, two rounds of PCR are employed first using an outer primer pair set followed by an internal (nested) primer pair set flanking the region of interest to yield positive and specific PCR products. Brief, but effective denaturation conditions, that

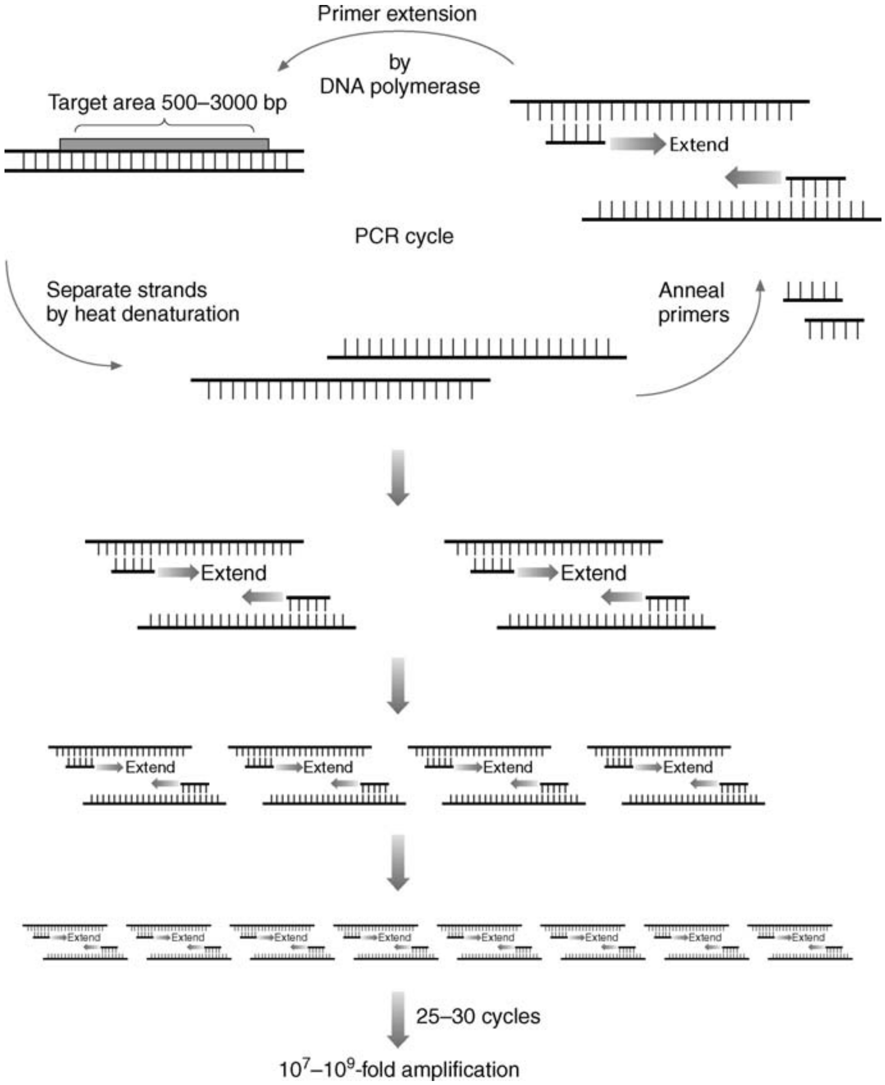


Figure 1. The PCR amplification cycle.

is 94–97 °C for 15–30 s, are necessary as *Taq* DNA polymerase has a half-life of only 40 min at 95 °C. Annealing conditions, on the other hand, are dependent on the concentration, base composition and the length of the oligonucleotide and typically range between 55 and 68 °C for 30–60 s. The length of the amplified target is directly proportional to the primer extension length of time. Primer extension is performed between 68 and 72 °C and, as a rule of thumb, is ~ 60 s for every 1 kb.

Crude extracts from blood, cerebral spinal fluid, urine, buccal smears, bacterial colonies, yeast spores, and so on are routinely used as sources of DNA for PCR templates. Due to the high sensitivity of PCR, rapid isolation protocols, such as heat and detergent disruptions, and enzymatic digestion of biological samples have been frequently used. Caution should be invoked when using crude extracts as starting materials for PCR amplifications because a number of impurities are known to inhibit *Taq* DNA polymerase. These include red blood cell components, sodium dodecyl sulfate (SDS), high salts, EDTA and too much DNA. Since only a few hundred target molecules are needed for successful PCRs, in most cases, these impurities can be effectively removed by simply diluting the starting material. Each sample should then be tested with control primers that specifically amplify a known target to determine the integrity of the crude extract. Alternatively, the isolation of the desired organism, such as HIV-1, human *Hepatitis A virus*, influenza virus, cytomegalovirus, and so on, or the isolation of specific cell fractions, such as peripheral blood mononuclear cells, can significantly increase the sensitivity and specificity of the PCR amplifications.

SENSITIVITY AND CONTAMINATION OF POLYMERASE CHAIN REACTION

Contamination is the dark side of the PCR force. The exquisite sensitivity of PCR can result in contamination from even a single molecule of foreign or exogenous DNA (4,5). To minimize false positives, standard operating procedures have been described, including the physical isolation of PCR reagents from the preparation of DNA templates and PCR products, using autoclaved solutions, premixing and aliquoting reagents, the use of disposable gloves, avoiding splashes, the use of positive displacement pipettes, adding DNA last, and carefully choosing positive and negative controls (6). Contamination is likely to surface for DNA samples that are difficult to amplify because of sequence content, or due to poor primer design and chemical impurities in DNA extractions. This is especially true for low copy number targets or degraded samples, as greater numbers of amplification cycles are generally required to achieve the desired product. In these cases, residual amounts of exogenous DNAs can compete and override the amplification process, resulting in spurious data. The best approach to challenge dubious results is to repeat the experiment with scrupulous care to details and controls. Biological samples collected at a single time point should be divided into multiple aliquots such that independent DNA extractions and PCR experiments can be performed to verify and validate initial results. Data should be discarded if inconsistent positive and negative PCR results occur upon repetition of the experiment. While negative controls can rule out reagent contamination, sporadic contamination can go unchecked. The probability of repeating spurious contamination in a consistent manner is extremely low.

There are three sources of contaminating DNA: (*1*) carryover contamination from previously amplified PCR products; (*2*) cross-contamination between multiple source materials; and (*3*) plasmid contamination from a recombinant clone that contains the target sequence. Of the three, carryover contamination is considered to be the major source of contamination because of the relative abundance of amplified target sequences. The substitution of dUTP for dTTP in the PCR cocktail has been routinely used as a method of preventing carryover contamination. Pretreatment of subsequent PCR mixtures prior to thermal cycling with uracil DNA glycosylase results in the removal of dU from any carryover PCR product, but does not affect the template DNA or dUTP. The dU removal creates an abasic site that is heat labile and degrades during thermal cycling, thus preventing carryover amplification. Moreover, ultraviolet (UV) light can reduce work surface and reagent contamination. Cross-contamination between samples is more difficult to diagnose, and suspicious results should be repeated from independent DNA extracts and PCR experiments for samples in question. Plasmid contamination, on the other hand, can be identified by sequence analysis and comparison to all laboratory plasmid sequences.

POLYMERASE CHAIN REACTION INTRODUCES MUTATIONS

The power and ease of PCR, however, were not fully appreciated until the introduction of the thermostable DNA polymerase isolated from *Thermus aquaticus* (*Taq*) (7) and automated instrumentation in 1988. It was here that PCR could be run in fully closed and automated systems. Fresh Klenow DNA polymerase did not have to be added at each cycle and PCR could be performed at higher annealing and extension temperatures, which increased the specificity and yields of the reactions while minimizing the risks of contamination. A hot start PCR further enhances specificity by preventing the formation of nonspecific products that arise during the initial steps of thermal cycling in PCR.

Taq DNA polymerase has been shown to incorporate nucleotides incorrectly at a frequency of 1 in 9000 bases by a mutation reversion assay (8). From sequence analysis of cloned PCR products, a slightly higher error frequency was determined (1 in 4000–5000 bp) for *Taq* DNA polymerase (9). The fidelity of DNA synthesis for *Taq* DNA polymerase, however, can vary significantly with changes in free Mg^{2+} concentration, changes in the pH of the buffer, or an imbalance in the four dNTP concentrations. Polymerase misincorporation errors are minimized when the four dNTPs are equimolar and between 50 and 200 $\mu mol{\cdot}L^{-1}$ (9). Since *Taq* DNA polymerase lacks a 3′-exonuclease activity, misincorporated bases typically cause chain termination of DNA synthesis that are not propagated in subsequent PCR cycles. In a worst-case scenario, a mutation occurring during the first round of PCR from a single target molecule and propagated thereafter would exist at a frequency of 25% in the final PCR product. Since hundreds of target copies are routinely used as starting DNA in PCR and most misincorporations terminate DNA synthesis, the observed error frequency is $\ll 25\%$.

Cloning of full-length genes from PCR products, however, has been problematic because PCR-induced mutations can cause amino acid substitutions in the wildtype sequence. Thus, significant effort must be employed in the complete sequencing of multiple PCR clones to identify mutation-free clones or ones that contain synonymous substitutions that do not change the protein coding sequence. Accordingly, thermostable DNA polymerases that contain a 3′-exonuclease (3′-exo) activity for proofreading of misincorporated bases have been recently introduced and include DNA polymerases isolated from *Pyrococcus furiosus (Pfu), Thermococcus litoralis* (Vent), *Pyrococcus* species GB-D (Deep Vent) and *Pyrococcus woesei* (*Pwo*). The error frequencies of these DNA polymerases are two- and sixfold < *Taq* DNA polymerase (10), but these polymerases are difficult for routine use, as the 3′-exonuclease activity can easily degrade the single-stranded PCR primers. 3′-Exo DNA polymerases, however, have been successfully used in long PCR in combination with *Taq* DNA polymerase and show an approximately twofold lower error frequency than *Taq* DNA polymerase alone (10).

POLYMERASE CHAIN REACTION LENGTH LIMITATIONS

For most applications, standard PCR conditions can reliably amplify target sizes up to 3–4 kb from a variety of source materials. Target sizes > 5 kb, however, have been described in the literature using standard PCR conditions, but generally yield low quantities of PCR product. The PCR size limitation can be attributed to the misincorporation of nucleotides that occurred 1 in 4000–5000 bp that ultimately reduced the efficiency of amplifying longer target regions. A breakthrough in long PCR came through the combined use of two thermostable DNA polymerases, one of which contains a 3′-exonuclease activity (11,12). The principle for long PCR is that the *Taq* DNA polymerase performs the high fidelity DNA synthesis part of PCR, coupled with the proofreading activity of *Pfu*, Vent or *Pwo* DNA polymerases. Once the nucleotide error is corrected, *Taq* DNA polymerase can then complete the synthesis of long PCR templates. From empirical studies, only a trace amount of the 3′-exo DNA polymerase, roughly 1% to that of *Taq* DNA polymerase or another DNA polymerase isolated from *Thermus thermophilis (Tth)*, is needed to perform long PCRs > 20 kb. Other important factors for long PCR are the isolation of high quality, high molecular weight DNA and protection against template damage, such as depurination during thermal cycling. The use of the cosolvents glycerol and dimethyl sulfoxide (DMSO) have been shown to protect against DNA damage by efficiently lowering the denaturation temperature by several degrees centigrade. The rule of thumb for primer extensions still applies for long PCRs (60 $\text{s}\cdot\text{kb}^{-1}$), although for targets > 20 kb, times extension should not exceed 22 min cycle^{-1}. The complexity and size of the genome under investigation can also affect the size of long PCR products. For example, PCR product lengths of 42 kb have been described for the amplification of λ bacteriophage DNA (11,12), compared with a 22 kb PCR product obtained from the human β-globin gene cluster (12).

CREATION OF NOVEL RECOMBINANT MOLECULES BY POLYMERASE CHAIN REACTION

Polymerase chain reaction can amplify both single- and double-stranded DNA templates as well as complementary DNA (cDNA) from the reverse transcription of messenger ribonucleic acid (mRNA) templates. Because of the flexibility of automated DNA synthesis, *In vitro* mutagenesis experiments can easily be performed by PCR. Recombinant PCR products can be created via the primer sequences by tolerated mismatches between the primer and the template DNA or by 5′-add-on sequences. Primer mediated mutagenesis can accommodate any nucleotide substitutions and small insertions or deletions of genetic material. The desired genetic alteration can be moved to any position within the target region by use of two overlapping PCR products with similar mutagenized ends (Fig. 2, left). This is accomplished by denaturing and reannealing the two overlapping PCR products to form heteroduplexes that have 3′-recessed ends. Following the extension of the 3′-recessed ends by *Taq* DNA polymerase, the full-length recombinant product is reamplified with the outer primers only to enrich selectively the full-length recombinant PCR product. 5′-Add-on adapters can also be used to join two unrelated DNA sequences, such as the splicing of an exogenous promoter sequence with a gene of interest (Fig. 2, right). The promoter–gene sequences are joined at the

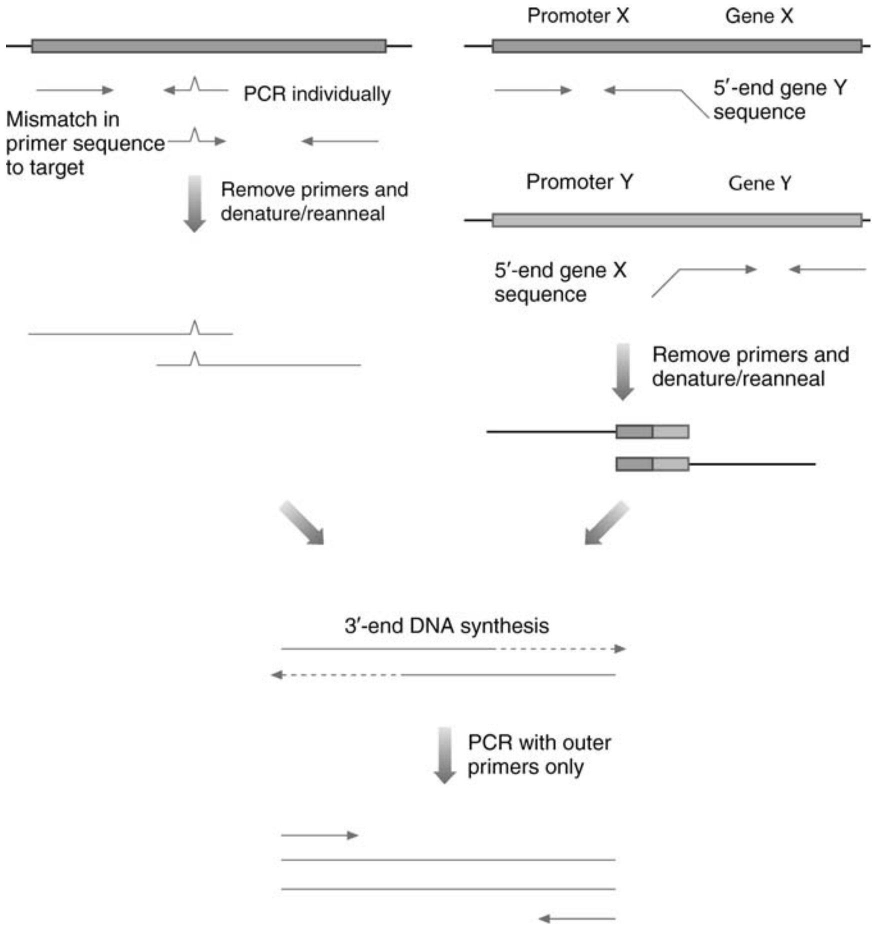

Promoter X
Gene X
PCR individually
Mismatch in primer sequence to target
5′-end gene Y sequence
Remove primers and denature/reanneal
Promoter Y
Gene Y
5′-end gene X sequence
Remove primers and denature/reanneal
3′-end DNA synthesis
PCR with outer primers only

Figure 2. Creation of mutagenized or recombinant PCR products via primer mismatches (left) or 5′-add-on sequences (right).

desired junction by 5′-add-on gene specific and 5′-add-on promoter-specific adapters that can PCR amplify the promoter and the gene targets, respectively. Heteroduplexes can then be formed, as described above, from the two overlapping PCR products, which are then selectively amplified with outer primers to generate the desired full-length recombinant PCR product.

POLYMERASE CHAIN REACTION AS A DETECTION SYSTEM

Polymerase chain reaction is a powerful tool for the detection of human polymorphic variation that has been associated with hereditary diseases. Many PCR techniques have been described that can discriminate between wild-type and mutant alleles, but in this section only a few of the most frequently used techniques are discussed. Of these, DNA sequencing of PCR products is the most widely used and most sensitive method for the detection of both novel and known polymorphic differences between individuals. Complementing scanning technologies, however, have been developed for the rapid detection of allelic differences because of the high costs associated with DNA sequencing and the ability to process large numbers of samples. Singlestrand conformation polymorphism (SSCP) has been commonly used as a technique for the identification of genetic polymorphisms. Following PCR, the product is heat denatured and subjected to native or nondenaturing gel electrophoresis. Allelic differences between samples are detected as mobility band shifts by radioactive and nonradioactive labeling procedures. PCR–SSCP, however, is limited in fragment size to $\sim$ 200 bp because the accuracy in discriminating between different alleles diminishes significantly with an increase in the fragment length.

Multiplex PCR allows for the simultaneous amplification of multiple target regions and has been particularly useful for the detection of exon deletion(s) in X-linked disorders, such as Duchenne muscular dystrophy (13) and Lesch–Nyhan syndrome (14). The multiplex PCR products are resolved by gel electrophoresis and are visualized by ethidium bromide staining. The absence of specific PCR product(s) is diagnostic of exon deletion(s) in affected males, and half-dosage PCR products are diagnostic of carrier mothers (15). Moreover, up to 46 primer pairs have been simultaneously amplified by multiplex PCR with excellent success (90%) for the large-scale identification of human single nucleotide polymorphisms (SNPs) by hybridization to high density DNA chip arrays.

Genetic polymorphisms can also be identified by immobilizing the PCR product on to a nylon membrane in a dot blot format and probing by hybridization with an allele-specific oligonucleotide (ASO) that contains a 5′-biotin group. The ASO hybridization is detected by adding streptavidin-horseradish peroxidase, which binds to the biotinylated probe, followed by a colorimeter assay. The colorimeter ASO assay has been applied to the genotyping of human leucocyte antigen (HLA)–DQA alleles and the detection of β-thalassaemia mutations. More recently, multiplex PCR and colorimeter ASO methodologies have been combined in a reverse fashion, in which ASOs are immobilized on to nylon membrane strips and probed against biotinylated gene-specific multiplex PCR reactions. Allele-specific PCR products are detected by hybridization and conversion of a colourless substrate to a blue precipitate for the simultaneous genotyping of HLA–DQA 1, low density lipoprotein receptor, glycophorin A, hemoglobin G gammaglobin, D7S8, and groupspecific component.

Lastly, *In situ* PCR enables the amplification of target sequences from sections of formalin-fixed, paraffin-embedded tissue specimens to determine the levels of gene expression in specific cell types that otherwise could not be detected by conventional *In situ* hybridization. The PCR is performed directly on glass slides by overlaying the PCR mixture on to the specimen, sealing the slides to prevent evaporation, and temperature cycling using a thermal sensor or modified thermal cycler that holds glass slides.

DEGENERATE POLYMERASE CHAIN REACTION

Degenerate PCR is a powerful strategy for obtaining novel full-length cDNA sequences from limited amino acid sequence information (16). The PCR primer sequences are derived from the reverse translation of 6–9 amino acid codons, which will result in varying levels of degeneracy except for methionine and tryptophan residues. Careful attention should be exercised in the design of degenerate primers because increasing the primer complexity (i.e., using codons that show more than twofold degeneracy) will typically result in an increase in nonspecific PCR products. One approach in reducing the complexity of the degenerate primer is the use of codon bias for the particular organism from which the gene will be cloned. Alternatively, the alignment of orthologous gene sequences from other species can greatly improve the specificity of cloning the gene of interest by revealing evolutionarily conserved domains. Once the optimal primer sequence is determined, the mixture of oligonucleotides can be simultaneously synthesized and will represent all possible amino acid combinations of the degenerate sequence. The specificity of PCR should then selectively amplify the correct primer sequences to generate a gene or gene family-specific probe from which the full-length cDNA can be obtained. Degenerate PCR has been successfully used in the screening of novel gene family members such as G-protein-coupled receptors, nuclear steroid receptors and protein tyrosine kinases.

ANCIENT DNA

Phylogenetics is the study of evolutionary relationships between specimens that are inferred from contemporaneous sequences. The ability to obtain DNA sequences from specimens or even fossils that are millions of years old could equip the phylogeneticist with a powerful means of directly testing an a priori hypothesis. Following death of the tissue or organism, however, DNA is rapidly degraded by, presumably, nuclease activities and hydrolytic processes, resulting in short fragment sizes that are generally no longer than 100–150 bp. Moreover, this old DNA is largely modified by oxidative processes and by intermolecular crosslinks that render it unsuitable for cloning by standard molecular biology procedures. Short PCRs, however, have been successfully performed from DNA samples isolated from archival and ancient specimens (17).

Museums hold vast collections of archived hospital files of patient specimens and of different species that have been collected over the last century. In a recent study, phylogenetic analyses of DNA sequences were performed from reverse transcriptase PCR (RT-PCR) of formalin-fixed, paraffin-embedded tissue specimens obtained from U.S. servicemen killed in the 1918 Spanish influenza pandemic. Viral sequences from three different gene regions were consistent with a novel H1N1 *Influenza A virus* that was most closely related to influenza strains that infect humans and swine, but not wild waterfowl, considered to be the natural reservoir for the influenza virus (18). Moreover, PCR and DNA sequencing have been performed on DNA extractions of archaeological findings, such as amplifying mitochondrial DNA sequences from a 7000 year old human brain, amplifying both mitochondrial and nuclear DNA sequences from bone specimens from a 14,000 year old saber-toothed cat, and amplifying chloroplast DNA sequences from fossil leaf samples from a 17 million-year-old Miocene *Magnolia* species.

QUANTITATIVE POLYMERASE CHAIN REACTION

Quantitative PCR (QPCR) has been widely used for detecting and diagnosing genetic deletions, for studying gene expression and for estimating the viral load of HIV-1. While DNA quantitation by multiplex PCR has been previously described (15), the quantitation of RNA has been wide reaching for the latter two areas. For many applications, estimating the relative amount of PCR product is sufficient to describe a biological observation. The absolute quantitation of RNA molecules, however, has been more difficult than for DNA because of the difficulty of generating accurate controls. Internal standards derived from synthetic RNA or cRNA have been designed to contain the same primer sequences as the target but yield a different-sized PCR product that can be easily separated by gel electrophoresis. cRNAs are not only coamplified with target sequences, but also coreverse transcribed to account for the variable efficiencies of cDNA syntheses. Moreover, QPCR is typically performed in the exponential or log phase of the amplification process (typically 14–22 cycles) to obtain accurate quantitative results. The absolute amount of target mRNA can be quantitated by serial dilutions of the target/internal control mixture and by extrapolating against the standard curve.

Both the variable range of initial target amounts and the presence of various inhibitors can, however, adversely affect the kinetics and efficiencies of PCR. Alternatively, a strategy based on a quantitative competitive (QC) approach has been used to minimize the effects of these variables. Known quantities of the competitor template, which contains the same primer sequences as the target but differs in size, are introduced into replicate PCRs containing identical quantities of the target. The point at which the intensities of the PCR products derived from the target sequence and the competitor template are equivalent is used to estimate the amount of target sequence in the original sample (19).

Recently, real-time QPCR and QCPCR (20) using a 5′-nuclease fluorogenic or TaqMan assay (21) has been developed to measure accurately the starting amounts of target sequences. Unlike gel electrophoresis, real-time QPCR has the unique advantage of being a closed-tube system, which can significantly reduce carryover contamination. Using this technique, one can easily monitor and quantitate the accumulation of PCR products during log phase amplification. The TaqMan assay utilizes dual reporter and quencher fluorescent dyes that are attached to a nonextendible probe sequence. During the extension phase of PCR, the 5′-nuclease activity of *Taq* DNA polymerase cleaves the hybridized fluorogenic probe, which releases the reporter signal and is measured during each cycle. In addition to real-time QPCR, TaqMan assays have broad utility for the identification of SNPs.

RELATED NUCLEIC ACID AMPLIFICATION PROCEDURES

Other *in vitro* systems can amplify nucleic acid targets such as the transcription-based amplification system (TAS) (6), its more recent version called the self-sustained sequence replication (3SR) (22) and the ligation-dependent Qβ-replication assay (23). These methods are best suited for the detection and semiquantitation of RNA target sequences. The strategy for TAS and 3SR is a continuous series of reverse transcription and transcription reactions that mimic retroviral replication by amplifying specific RNA sequences via cDNA intermediates. The primers contain 5′- add-on sequences for T7, T3, or SP6 promoters that are incorporated into the cDNA intermediates. The rapid kinetics of transcription-based amplifications is an attractive feature of these systems, which can amplify up to 10^7 molecules in 60 min. Short amplify products, however, which are due to incomplete transcription of the target region and incomplete RNase H digestion of the RNA–DNA hybrids, can be problematic in the TAS and 3SR assays.

Unlike PCR, TAS, or 3SR assays, the ligation-dependent Qβ-replication assay results in the amplification of probe, not target, sequences. This assay utilizes a single hybridization to the target sequence, which is embedded within, and divided between, a pair of adjacently positioned midvariant (MDV-1) RNA probes. MDV-1 RNA is the naturally occurring template for the bacteriophage Qβ RNA replicase. Following the isolation of the probe–target hybrids, ligation of the binary probes creates a full-length amplifiable MDV-1 RNA reporter. When Qβ replicase is added, newly synthesized MDV-1 RNA molecules are amplified from ligated binary probes that originally hybridized to the target sequence (23). Similar to TAS and 3SR, the Qβ-replication assay shows rapid kinetics, generating up to 10^9 molecules in 30 min, and all three methods have been successfully used for the detection and quantitation of HIV-1 RNA molecules.

LIGATION CHAIN REACTION

The ligase chain reaction (LCR) can also amplify short DNA regions of interest by iterative cycles of denaturation and annealing/ligation steps (24). The LCR utilizes four primers: two adjacent ones that specifically hybridize to one strand of target DNA and a complementary set of adjacent primers that hybridize to the opposite strand. LCR primers must contain a 5′-end phosphate group, such that thermostable ligase (24) can join the 3′-end hydroxyl group of the upstream primer to the 5′-end phosphate group of the downstream primer. Successful ligations of adjacent primers can subsequently act as the LCR template, resulting in an exponential amplification of the target region. The LCR is well suited for the detection of SNPs because a single-nucleotide mismatch at the 3′ end of the upstream primer will not ligate and amplify, thus discriminating it from the correct base. Although LCR is generally not quantitative, linear amplifications using one set of adjacent primers, called the ligase detection reaction, can be quantitative. Coupled to PCR, linear ligation assays can also be used as a mutation detection system for the identification of SNPs using both wild-type-specific and mutant-specific primers in separate reactions. The oligonucleotide ligase assay was first reported to detect SNPs from both cloned and clinical materials using a 5′-end biotin group attached to the upstream primer and a nonisotopic label attached to the downstream primer (25). Allele-specific hybridizations and ligations can be separated by immobilization to a streptavidin-coated solid support and directly imaged under appropriate conditions without the need for gel electrophoretic analysis.

SUMMARY

Some of the general concepts and practices of PCR have been reviewed here. Not only has PCR made a major and significant impact on basic and clinical research, but it has also been well accepted and utilized in forensic science. For any scientific methodology to be accepted in the courts as evidence, it must satisfy four criteria: that the method (*1*) be subject to empirical testing, (*2*) be subject to peer review and publication, (*3*) has a known error rate, and (*4*) is generally accepted in the scientific community. The application of PCR has been admitted in the U.S. courts as evidence in criminal cases for the analysis of human DNA sequences, and in January 1997 as evidence for the phylogenetic analysis of HIV DNA sequences (26). Clearly, the scope of applications for PCR seems endless and it is truly a remarkable technique that has been widely used in molecular biology.

BIBLIOGRAPHY

1. Saiki RK, et al. Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anaemia. Science 1985;230:1350–1354.
2. Mullis KB, Faloona FA. Specific synthesis of DNA *In vitro* via a polymerase-catalysed chain reaction. Methods Enzymol 1987;155:335–351.
3. Eckert KA, Kunkel TA. High fidelity DNA synthesis by the *Thermus aquaticus* DNA polymerase. Nucleic Acids Res 1990;18:3739–3744.
4. Gibbs RA, Chamberlain JS. The polymerase chain reaction: a meeting report. Genes Dev 1989;3:1095–1098.
5. Kwoh DY, et al. Transcription-based amplification system and detection of amplified human immunodeficiency virus type 1 with a bead-based sandwich hybridization format. Proc Nat Acad Sci U.S.A. 1989;86:1173–1177.
6. Kwok S, Higuchi R. Avoiding false positives with PCR. Nature (London) 1989;339:237–238.
7. Saiki RK, et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 1988;239:487–491.
8. Tindall KR, Kunkel TA. Fidelity of DNA synthesis by the *Thermus aquaticus* DNA polymerase. Biochemistry 1988;27:6008–6013.
9. Innis MA, Myambo KB, Gelfand DH, Brow MAD. DNA sequencing with *Thermus aquaticus* DNA polymerase and direct sequencing of polymerase chain reaction-amplified. Proc Nat Acad Sci U.S.A. 1988;85:9436–9440.
10. Cline J, Braman JC, Hogrefe HH. PCR fidelity of *Pfu* DNA polymerase and other thermostable DNA polymerases. Nucleic Acids Res 1996;24:3546–3551.
11. Barnes WM. PCR amplification of up to 35-kb DNA with high fidelity and high yield from λ bacteriophage templates. Proc Nat Acad Sci U.S.A. 1994;91:2216–2220.
12. Cheng S, Fockler C, Barnes WM, Higuchi R. Effective amplification of long targets from cloned inserts and human genomic DNA. Proc Nat Acad Sci U.S.A. 1994;91:5695–5699.
13. Chamberlain JS, et al. Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification. Nucleic Acids Res 1988;23:11141–11156.
14. Gibbs RA et al. Multiplex DNA deletion detection and exon sequencing of the hypoxanthine phosphoribosyltransferase gene in Lesch–Nyhan families. Genomics 1990;7:235–244.
15. Metzker ML, Allain KM, Gibbs RA. Accurate determination of DNA in agarose gels using the novel algorithm GelScann(1.0). Computer App Biosci 1995;11:187–195.
16. Lee CC, et al. Generation of cDNA probes directed by amino acid sequence: cloning of urate oxidase. Science 1988;239:1288–1291.
17. Pääbo S. Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. Proc Nat Acad Sci U.S.A. 1989;86:1939–1943.
18. Taubenberger JK, et al. Initial genetic characterization of the 1918 ‘Spanish’ influenza virus. Science 1997; 275:1793–1796.
19. Gilliland G, Perrin S, Blanchard K, Bunn HF. Analysis of cytokine mRNA and DNA: detection and quantitation by competitive polymerase chain reaction. Proc Nat Acad Sci U.S.A. 1990;87:2725–2729.
20. Heid CA, Stevens J, Livak KJ, Williams PM. Real time quantitative PCR. Genome Res 1996;6:986–994.
21. Holland PM, Abramson RD, Watson R, Gelfand DH. Detection of specific polymerase chain reaction product by utilizing the 5′→3′ exonuclease activity of *Thermus aquaticus* DNA polymerase. Proc Nat Acad Sci U.S.A. 1991;88:7276–7280.
22. Guatelli JC, et al. Isothermal, *In vitro* amplification of nucleic acids by a multienzyme reaction modeled after retroviral replication. Proc Nat Acad Sci U.S.A. 1990;87:1874–1878.
23. Tyagi S, Landegren U, Tazi M, Lizardi PM, Kramer FR. Extremely sensitive, background-free gene detection using binary probes and Qβ-replicase. Proc Nat Acad Sci U.S.A. 1996;93:5395–5400.
24. Barany F. Genetic disease detection and DNA amplification using cloned thermostable ligase. Proc Nat Acad Sci U.S.A. 1991;88:189–193.
25. Landegren U, Kaiser R, Sanders J, Hood L. A ligase-mediated gene detection technique. Science 1988;241:1077–1080.
26. State of Louisiana v. Richard J Schmidt. Reasons for ruling of Louisiana State 15th Judicial District Court Judge Durwood

Conque. 15th Judicial District Court, Lafayette Parish, Louisiana. Criminal Docket 73313. 1997.

Further Reading

Erlich HA, Gelfand D, Sninsky JJ. Recent advances in the polymerase chain reaction. Science 1991; 252: 1643–1650.

Innis MA, Gelfand DH, Sninsky JJ, White TJ, editors. PCR Protocols: A Guide to Methods and Applications. Academic; San Diego: 1990.

Mullis KB, Ferré F, Gibbs RA, editors. The Polymerase Chain Reaction. Birkhäuser; Boston: 1994.

See also ANALYTICAL METHODS, AUTOMATED; DNA SEQUENCE; MICROARRAYS.

POLYMERIC MATERIALS

XIAOHUA LIU
The University of Michigan
Ann Arbor, Michigan

SHOBANA SHANMUGASUNDARAM
TREENA LIVINGSTON ARINZEH
New Jersey Institute of
Technology
Newark, New Jersey

INTRODUCTION

This article aims to provide basic and contemporary information on polymeric materials used in medical devices and instrumentation. The fundamental concepts and features of polymeric materials are introduced in the first section. In the second section, the major commodity polymers used in medicine are reviewed in terms of their basic chemical and physical properties. The main part of this article, however, is devoted to polymers in biomedical engineering applications, including tissue engineering and drug delivery systems.

Polymers are a very important class of materials. A polymer can be defined as a long-chain molecule that is composed of a large number of repeating units of identical structure. Some polymers, (e.g., proteins, cellulose, and starch) are found in Nature, while many others, including polyethylene, polystyrene, and polycarbonate, are produced only by synthetic routes. Hundreds of thousands of polymers have been synthesized since the birth of polymer science. Today, polymeric materials are used in nearly all areas of daily life.

Polymers can simply be divided into two distinct groups based on their thermal processing behavior: thermoplastics and thermosets. Thermoplastics are linear or branched polymers, and they soften or melt when heated, so that they can be molded and remolded by heating. This property allows for easy processing and recycling. In comparison, thermosets are three-dimensional (3D) network polymers, and cannot be remelted. Once these polymers are formed, reheating will cause the material to scorch.

In addition to classification based on processing characteristics, polymers may also be grouped based on the chemical structure of their backbone. Polymers with one identical repeating unit in their chains are called homopolymers. The term copolymer is often used to describe a polymer with two or more repeating units. The sequence of repeating units along the polymer chain can form different structures, and copolymers can be further classified as random copolymers, alternating copolymers, block copolymers, and graft copolymers. In random copolymers, the sequence distribution of the repeating units is random, while in alternating copolymers the repeating unit are arranged alternately along the polymer chain. A block copolymer is one in which identical repeating units are clustered in blocks along the chain. In graft copolymers, the blocks of one type of repeating unit are attached as side chains to the backbone chains.

Unlike simple pure compounds, most polymers are not composed of identical molecules. A typical synthetic polymer sample contains chains with a wide distribution of chain lengths. Therefore, polymer molecular weights are usually given as averages. The number average molecular weight (M_n), which is calculated from the mole fraction distribution of different sized molecules in a sample, and the weight average molecular weight (M_w), which is calculated from the weight fraction distribution of different sized molecules, are two commonly used values. The statistical nature of polymerization reaction makes it impossible to characterize a polymer by a single molecular weight. A measure of the breadth of the molecular weight distribution is given by the ratios of molecular weight averages. The most commonly used ratio is M_w/M_n. As the weight dispersion of molecules in a sample narrows, M_w approaches M_n, and in the unlikely case that all the polymer molecules have identical weights, the ratio M_w/M_n becomes unity. Most commercial polymers have the molecular weight distribution of 1.5–10. In general, increasing molecular weight corresponds to increasing physical properties and decreasing polymer processability.

In many cases, individual polymer chains are randomly coiled and interviewed with no molecular order or structure. Such a physical state is termed amorphous. Amorphous polymers exhibit two distinctly different types of mechanical behavior. Some, like poly(methyl methacrylate, PMMA) and polystyrene are hard, rigid, glassy plastics at room temperature, while others, like polybutadiene and poly(ethyl acrylate), are soft, flexible, rubbery materials at room temperature. There is a temperature, or range of temperatures, below which an amorphous polymer is in a glassy state, and above which it is rubbery. This temperature is called the glass transition temperature (T_g). The value of T_g for a specific polymer will depend on the structure of the polymer. Side groups attached to the polymer chain will generally hinder rotation in the polymer backbone, necessitating higher temperatures to give enough energy to enable rotation to occur.

For most polymers, the T_g constitutes their most important mechanical properties. At low temperatures ($< T_g$), an amorphous polymer is glass-like, with a value of Young's modulus in the range of 10^9–10^{10} Pa, and it will break or yield at strains greater than a few percent. When the temperature is $> T_g$, the polymer becomes rubber-like, with a modulus in the range of 10^5–10^6 Pa, and it may withstand large extensions with no permanent deformation. At even higher temperatures, the polymer may undergo permanent deformation under load and behave like a highly viscous liquid. In the T_g range, the polymer is neither glassy nor rubber-like. It has an intermediate modulus and has viscoelastic properties.

CHEMICAL AND PHYSICAL PROPERTIES OF MAJOR COMMODITY POLYMERS

This section reviews the major polymers used in medical applications, with a brief discussion of chemical as well as physical properties and their application. They are grouped as homopolymers or copolymers.

Homopolymers

Polyacrylates (e.g., PMMA) and poly(hydryoxyethyl methacrylate) (PHEMA), are used for hard and soft contact lenses because of their excellent physical, coloring properties, and ease in fabrication. The PMMA polymer is a hydrophobic, linear chain polymer that is glassy at room temperature. It has very good light transmittance, toughness, and stability, making it an excellent material for intraocular lenses and hard contact lenses. The PHEMA polymer is used for soft contact lenses. With the addition of a –CH_2OH group to the methyl methacrylate side group of the PMMA structure, the polymer becomes hydrophilic. Typically, PHEMA is cross-linked with elthylene glycol dimethylacrylate (EGDM) to prevent the polymer from dissolving when hydrated. When fully hydrated, PHEMA is a hydrogel with potential use in advanced technology applications (e.g., biomedical separation and biomedical devices).

Polyolefins, which include polyethylene (PE) and polypropylene (PP), are linear chain polymers. Polyethylene is a highly crystalline polymer that is used in its high density form for biomedical applications because low density forms cannot withstand sterilization temperatures. The high density form is used for drains and catheters. The ultrahigh molecular weight form (UHMWPE) is used in orthopedic implants for load-bearing surfaces in total hip and knee joints. The material has good toughness, creep properties, resistance to environmental attack, and relatively low cost. The PP is related to PE by the addition of a methyl group along the polymer chain. It has similar properties to PE (e.g., high rigidity, good chemical resistance, and tensile strength) and is used for many of the same applications. It also has a high flex life, which is superior to PE and is therefore used for finger joint prostheses.

Polytetrafluoroethylene (PTFE), commonly known as Teflon, is similar in structure to PE except that the hydrogen in PE is substituted with fluorine. This polymer has a high crystallinity ($> 94\%$), high density, low modulus of elasticity, and tensile strength. It is a very stable polymer and difficult to process. The material also has very low surface tension and friction coefficient. It is used for vascular graft applications due to the lack of adherence of blood components.

Poly(vinyl chloride) (PVC) is typically used for tubing for blood transfusions, feeding, and dialysis. Pure PVC is hard and brittle. However, for these applications, the addition of plasticizers makes it soft and flexible. Issues concerning these plasticizers exist because they can be extracted during long-term use, making PVC less flexible over time.

Poly(dimethyl siloxane) (PDMS), or silicone rubber, is a versatile material. Low molecular weight polymers have low viscosity and can be cross-linked to make a higher molecular weight rubber-like material. It has a silicon–oxygen backbone instead of a carbon backbone. The material is less temperature sensitive than other rubbers because of its lower T_g. It also has excellent flexibility and stability. The applications of PDMS are widespread (e.g., catheter and drainage tubing, insulation for pacemaker leads, a component in some vascular graft systems, prostheses for finger joints, blood vessels, breast implants, outer ears, chin and nose implants). Since its oxygen permeability is very high, PDMS is also used in membrane oxygenators.

Polyamides, commonly known as nylons, are linear-chain polymers containing –CONH– groups. With the presence of these groups, the chains attract strongly to one another by hydrogen bonding. Increasing numbers of –CONH– groups and a high degree of crystallinity improves physical properties (e.g., strength and fiber forming ability). They are used for surgical sutures.

Polycarbonates are tough, amorphous, clear materials produced by the polymerization of biphenol A and phosgene. It is used as lenses for eyeglasses and safety glasses, and housings for oxygenators and heart–lung bypass machines.

Copolymers

Poly(glycolide lactide) (PGL) are random copolymers used in resorbable surgical sutures. The PGL is polymerized by a ring-opening reaction of glycolide and lactide and is gradually resorbed in the body due to the ester linkages in the polymer backbone via hydrolysis.

A copolymer of tetrafluoroethylene and hexafluoropropylene (FEP) is used similarly to PTFE. The advantage of FEP is that it is easier to process than PTFE, but still retains excellent chemical inertness and a low coefficient of friction. The FEP has a crystalline melting temperature of 265 °C, whereas PTFE is 375 °C.

Polyurethanes are copolymers, which contain “hard” and “soft” blocks. The “hard” blocks are composed of a diisocyanate and a chain extender, with a T_g above room temperature, and has a glassy or semi-crystalline character. The “soft” blocks are typically polyether or polyester polyols with a T_g below room temperature. Thus, the material also has rubbery characteristics. Polyurethanes are tough elastomers with good fatigue and blood-containing properties. They are typically used for pacemaker lead insulation, vascular grafts, heart assist balloon pumps, and artificial heart bladders.

POLYMERS IN BIOMEDICAL ENGINEERING APPLICATIONS

Polymers Used in Tissue Engineering

Synthetic Polymers. The most widely used synthetic polymers for tissue engineering products, either under development or on the market, are poly(lactic acid) (PLA), poly(glylic acid) (PGA), and their copolymers PLGA. Both PLA and PGA are linear aliphatic polyesters formed by ring-opening polymerization with a metal catalyst. The PLA can also be obtained from the renewable agricultural source, corn and degrades in two phases: hydrolysis and metabolization. The PLA and PGA polymers have similar chemical structures except that the PLA has a methyl pendant group. Both degrade by simple hydrolysis of their ester linkages. The PGA can also be broken down by nonspecific esterases and carboxypeptidases. The degradation rate is dependent on initial molecular weight, exposed surface area, crystallinity, and, in the case of copolymers, the PLA/PGA ratio present. PGA is highly crystalline, having a high melting point, and a low solubility in organic solvents. It is also hydrophilic in nature, losing its mechanical strength over a period of 2–4 weeks in the body.

The PLGA was developed to achieve a wider range of possible applications for PGA. Due to the extra methyl group in lactic acid, PLA is more hydrophobic and has a slower rate of backbone hydrolysis than PGA. The PLA is also more soluble in organic solvents. The copolymer PLGA degradation depends on the exact ratio of PLA and PGA present in the polymer. The PLGA polymer is less crystalline and tends to degrade more rapidly than either PGA or PLA. Lactic acid is a chiral molecule that exists in two stereoisomeric forms that yield four morphologically distinct polymers. Both *d*-PLA and *l*-PLA are two stereoregular polymers *d,l*-PLA is the racemic polymer and *meso*-PLA can be obtained from *d,l*-lactide.The amorphous polymer is *d,l*-PLA and is used typically for drug delivery applications where it is important to have a homogenous dispersion of the active agents within a monobasic matrix. The *l*-PLA polymer is semi-crystalline and most commonly used because the degradation product of *l*(+)-lactic acid is the naturally occurring steroisomer of lactic acid. It is typically used for high mechanical strength and toughness applications (e.g., orthopaedics).

Some of the other synthetic polymers currently under investigation for tissue engineering applications are described briefly. Polycaprolactone (PCL) is a synthetic aliphatic polyester with a melting point (T_m) of 55–65 °C. Degradation of PCL is a slow process that occurs either by hydrolysis or enzymatic degradation *in vivo*. The slow degradation rate of PCL is particularly interesting for long-term implants and controlled release application. Poly(hydroxy butyrate) (PHB) and its copolymers are semi-crystalline thermoplastic polyester made from renewable natural sources. *In vivo*, PHB degrades into hydroxybutyric acid that is a normal constituent of human blood. The PHB homopolymer is highly crystalline and has a high degradation rate. Its biodegradation and biocompatibility properties have led to research on its prospective use as a material for coronary stents, wound dressings, and drug delivery. Poly(propylene fumarate) (PPF) is an unsaturated linear polyester formed by the copolymerization of fumaric acid and propylene glycol. These polymer networks degrade by hydrolysis of the ester linkage to water-soluble products, namely, propylene glycol, poly(acrylic acid-*co*-fumaric acid), and fumaric acid. Due to its unsaturated sites along the polymer backbone, which are labile and can be cross-linked *in situ*, PPF is currently being evaluated for filling skeletal defects of varying shapes and sizes. Polyphosphoesters (PPE) are biodegradable polymers with physicochemical properties that can be altered by the manipulation of either the backbone or the side-chain structure. This property of PPE makes them potential drug delivery vehicles for low molecular drugs, proteins, deoxyribonucleic acid DNA plasmids, and as tissue engineering scaffolds. Since the phosphoester bond in a PPE backbone is cleaved by water, the more readily water penetrates, with greater bond cleavage and faster degradation rate. The products of hydrolytic breakdown of PPE are phosphate, alcohol, and diol. Polyphosphazenes are inorganic polymers having a phosphorus–nitrogen alternating backbone and each phosphorus atom is attached to two organic or organometallic side groups. They degrade by hydrolysis into phosphate, amino acid, and ammonia. The potential application is in low molecular weight drug release and in formulation of proteins and peptides. Polyanhydrides are a class of degradable polymers synthesized from photopolymerizable multimethacrylate monomers. Many polyanhydrides degrade from the surface by hydrolysis of the anhydride linkages. The rate of hydrolysis is controlled by the polymer backbone chemistry. They are useful for controlled drug delivery as they degrade uniformly into nontoxic metabolites. Polyorthoesters (POE), another class of biodegradable and biocompatible polymers, can be designed to possess a surface-dominant erosion mechanism. Acidic byproducts autocatalyzed the degradation process resulting in increased degradation rates than nonacidic byproducts. The POE, which is susceptible to acid-catalysed hydrolysis, has attracted considerable interest for the controlled delivery of therapeutic agents within biodegradable matrices.

Natural Polymers. Collagen is a widely used natural polymer in tissue engineering. It is a structural protein, being a significant constituent of the natural extracellular matrix. It has a triple-helical molecular structure that arises from the repetitious amino acid (glycine, proline, and hydroxyproline) sequence. *In vivo*, collagen in healthy tissues is resistant to attack by most proteases except specialized enzymes called collagenases that degrade the collagen molecules. Collagen can be used alone or in combination with other extracellular matrix components (e.g., glycosaminoglycan and growth factors) to improve cell attachment and proliferation. It has been tested as a carrier material in tissue engineering applications.

Other natural polymers under investigation for tissue engineering applications are described briefly. Gelatin, denatured collagen, is obtained by the partial hydrolysis of collagen. It can form a specific triple-stranded helical structure. The rate of the formation of a helical structure depends on many factors (e.g., the presence of covalent cross-bonds, gelatin molecular weight, the presence of amino acids, and the gelatin concentration in the solution). Gelatin is used in pharmaceuticals, wound dressings, and bioadhesives due to its good cell viability and lack of antigenicity. It has some potential for use in tissue engineering applications. Silk is a fibrous protein characterized by a highly repetitive primary sequence of glycine and alanine that leads to significant homogeneity in secondary structure, β-sheets in the case of many of the silks. Silk is biodegradable due to its susceptibility to proteolytic enzymes. Silk studies *in vitro* have demonstrated that protease cocktails and chymotrypsin are capable of enzymatically degrading silk. The mechanical properties of silk provide an important set of material options in the fields of controlled release, biomaterials, and scaffolds for tissue engineering. Alginate is a straight-chain polysaccharide composed of two monomers, mannuronic acid and guluronic acid residues, in varying proportions. Alginate forms stable gels on contact with certain divalent cations, such as calcium, barium, and strontium. Alginate is widely used as an instant gel for bone tissue engineering. Chitosan, a copolymer of glucosamine and *N*-acetylglucosamine is a crystalline polysaccharide. It is synthesized by the deacetylation of chitin. Chitosan degrades mainly through lysozyme-mediated hydrolysis, with the degradation rate being inversely related to the degree of crystallinity. Chitosan has excellent potential as a structural base material for a variety of tissue engineering application, wound dressings, drug delivery systems, and space-filling implants. Hyaluronate, a glycosaminoglycan is a straight-chain polymer composed of glucuronic acid and acetylglucosamine. It contributes to tissue hydrodynamics, movement, and proliferation of cells *in vivo*. Hyaluronan is enzymatically degraded into monosaccharides. It has been used in the treatment of osteoarthritis, dermal implants, and prevention of postsurgical adhesions.

Polymeric Scaffold Fabrication Techniques (6–8). Scaffolds for tissue engineering, in general, are porous to maximize cell attachment, nutrient transport, and tissue growth. A variety of processing technologies have been developed to fabricate porous 3D polymeric scaffolds for tissue engineering. These techniques mainly include solvent casting and particulate leaching, gas-foaming processing, electrospinning technique, rapid prototyping, and thermally induced phase-separation technique, which are described below.

Solvent casting and particulate leaching is a simple, but commonly used method for fabricating scaffolds. This method involves mixing water soluble salt (e.g., sodium chloride, sodium citrate) particles into a biodegradable polymer solution. The mixture is then cast into the desired shape mold. After the solvent is removed by evaporation or lyophilization, the salt particles are leached out and leave a porous structure. This method has advantages of simple operation and adequate control of pore size and porosity by salt/polymer ratio and particle size of the added salt. However, the interconnectivity between pores inside the scaffold is often low, which seems to be problematic for cell seeding and culture.

Gas foaming is marked by the ability to form highly porous polymer scaffold foams without using organic solvents. In this approach, carbon dioxide is usually used as a foaming agent for the formation of polymer foam. This approach allows the incorporation of heat sensitive pharmaceuticals and biological agents. The disadvantage of this method is that it yields mostly a nonporous surface and closed-pore structure.

Electrospinning is a fabrication process for tissue engineering that use an electric field to control the formation and deposition of polymer fibers onto a target substrate. In electrospinning, a polymer solution or melt is injected with an electrical potential to create a charge imbalance. At a critical voltage, the charge imbalance begins to overcome the surface tension of the polymer source, and forms an electrically charged jet. The jet within the electric field is directed toward the ground target, during which time the solvent evaporates and fibers are formed. This electrospinning technique can fabricate fibrous polymer scaffolds composed of fiber diameters ranging from several microns down to several hundred nanometers.

Rapid prototyping is a technology based on the advanced development of computer science and manufacturing industry. The main advantage of these techniques is their ability to produce complex products rapidly from a computer-aided design (CAD) model. The limitation of this method is that the resolution is determined by the jet size, which makes it difficult to design and fabricate scaffolds with fine microstructure. The controlled thermally induced phase-separation process was first used for the preparation of porous polymer membranes. This technique was recently utilized to fabricate biodegradable 3D polymer scaffolds. In this approach, the polymer is first dissolved in a solvent (e.g., dioxane) at a high temperature. Liquid–liquid or solid–liquid phase separation is induced by lowering the solution temperature. Subsequent removal of the solidified solvent-rich phase by sublimation leaves a porous polymer scaffold. The pore morphology and microstructure of the porous scaffolds varies depending on the polymer, solvent, concentration of the polymer solution, and phase separation temperature. One advantage of this method is that scaffolds fabricated with the technique have higher mechanical strength than those of the same porosity made with the well-documented salt-leaching technique.

Polymers for Drug Delivery. Over the past decade, the use of polymeric materials for the administration of pharmaceuticals and as biomedical devices has dramatically increased (9–11). One important medical application of polymeric materials is in the area of drug delivery systems. There are a few polymer molecules having a drug function, however, in most cases when polymers are used in drug delivery systems, they serve as a carrier of drugs. Table 1 lists some of the important biodegradable and nonbiodegradable polymers used in drug delivery systems.

Table 1. Typical Biodegradable and Nonbiodegradable Polymers Used in Controlled Release Systems

Nonbiodegradable Polymers	Biodegradable Polymers
Polyacrylates	Polyglycolides
Polyurethanes	Polylactides
Polyethylenes	Polyanhydrides
Polysiloxanes	Polyorthoesters
	Polycaprolactones
	Poly(β-hydroxybutyrate)
	Polyphosphazenes
	Polysaccharides

Most of the above biodegradable and nonbiodegradable polymers have been discussed in the previous sections; therefore, they are not described further here.

Stimuli-Responsive Hydrogels for Drug Delivery. Hydrogels have been used as carriers for a variety of drug molecules (10). A hydrogel is a network of hydrophilic polymers that are cross-linked by either covalent or physical bonds. It distinguishes itself from other polymer networks in that it swells dramatically in the presence of abundant water. The physicochemical and mechanical properties can be easily controlled, and hydrogels can be made to respond to changes in external factors.

In recent years, temporal control of drug delivery has been of great interest to achieve improved drug therapies. Stimuli-responsive hydrogels exhibit sharp changes in behavior in response to an external stimulus (e.g., temperature, pH, solvents, salts, chemical or biochemical agents, and electrical field). The stimuli-responsive hydrogels have the ability to sense external environmental changes, judge the degree of external signal, and trigger the release of appropriate amounts of drug. Such properties have made it very useful for temporal control of drug delivery (12,13).

Temperature-Sensitive Hydrogels. Temperature is the most widely used stimulus in environmentally responsive polymer systems. Temperature-sensitive hydrogels can respond to the change of environmental temperature. The change of temperature is not only relatively easy to control, but also easily applicable both *in vitro* and *in vivo*. Poly(*N*-isopropylacrylamide) (PNIPA) is representative of the group of temperature-responsive polymers that have a lower critical solution temperature (LCST), defined as the critical temperature at which a polymer solution undergoes phase transition from a soluble to an insoluble state above the critical temperature. The PNIPA exhibits a sharp phase transition in water at $\sim 32\ ^\circ\mathrm{C}$, which can be shifted to body temperatures by the presence of hydrophilic monomers (e.g., acrylic acid). Reversely, the introduction of a hydrophobic constituent to PNIPA would lower the LCST of the resulting copolymer.

When PNIPA chains are chemically cross-linked by a cross-linker (e.g., *N,N′*-methylenebisacrylamide and ethylene glycol dimethacrylate), the PNIPA hydrogel is formed, which swells, but does not dissolve in water. The PNIPA hydrogel undergoes a sharp swelling-shrinking transition near the LCST, instead of sol–gel phase separation. The sharp volume decrease of the PNIPA hydrogel above the LCST results in the formation of a dense, shrunken layer on the hydrogel surface, which hinders water permeation from inside the gel into the environment. The PNIPA hydrogels have been studied to the delivery of antithrombotic agents (e.g., heparin), at the site of a blood clot, utilizing biological conditions to trigger drug release. Drug release from the PNIPA hydrogels at temperatures below LCST is governed by diffusion, while above this temperature drug release is stopped, due to the dense layer formation on the hydrogel surface.

Some types of block copolymers made of poly(ethylene oxide) (PEO) and poly(propylene oxide) (PPO) also possess an inverse temperature sensitive property. Because of their LCST at around body temperature, they have been widely used in the development of controlled drug delivery systems based on the sol–gel phase transition at the body temperature.

pH-Sensitive Hydrogels. Polymers with a large number of ionizable groups are called polyelectrolytes. The pH-sensitive hydrogels are cross-linked polyelectrolytes containing either acidic or basic pendent groups, which show sudden changes in their swelling behavior as a result of changing the external pH. The pendant groups in the pH-sensitive hydrogels can ionize in aqueous media of appropriate pH value. As the degree of ionization increases (via increasing or decreasing pH value in the aqueous media), the number of fixed charges on the polymer chains increases, resulting in increased electrostatic repulsions between the chains. As a result of the electrostatic repulsions, the uptake of water in the network is increased and thus the hydrogels have higher swelling ratios. The swelling of pH-sensitive hydrogels can also be controlled by ionic strength and copolymerizing neutral comonomers, which provide certain hydrophobicity to the polymer chain. The pH-sensitive hydrogels have been used to develop control release formulations for oral administration. For polycationic hydrogels, the swelling is minimal at neutral pH, thus minimizing drug release from the hydrogels. The drug is released in the stomach as hydrogels swell in the low pH environment. This property has been used to prevent release of foul-tasting drugs into the neutral pH environment of the mouth.

Sometimes, it is desirable that hydrogels with certain compositions can respond to more than one environmental stimulus (e.g., temperature and pH). Hydrogel copolymers of *N*-isopropylacrylamide and acrylic acid with appropriate compositions have been designed to sense small changes in blood stream pH and temperature to deliver antithrombotic agents (e.g., streptokinase or heparin) to the site of a blood clot.

Electrosensitive Hydrogels. The electrosensitive hydrogels, which are capable of reversible swelling and shrinking under a change in electric potential, are usually made of polyelectrolytes. The electric sensitivity of the polyelectrolyte hydrogels occurs in the presence of ions in solution. In the presence of an applied electric field, the ions (both co-ions and counterions) move to the positive or negative electrode, while the polyions of the hydrogels cannot move. This results in a change in the ion concentration-dependent osmotic pressure, and hydrogels either swell or shrink to reach its new equilibrium. The electrosensitive hydrogels exhibit reversible swelling–shrinking behavior in response to on–off switching of an electric stimulus. Thus, drug molecules within the polyelectrolyte hydrogels might be squeezed out from the electric-induced gel contraction along with the solvent flow.

Other Stimuli-Sensitive Hydrogels. Hydrogels that respond to specific molecules found in the body are especially useful for some drug delivery purposes. One such hydrogel is glucose-sensitive hydrogel, which has potential applications in the development of self-regulating insulin delivery systems.

BIBLIOGRAPHY

1. Bower DI. An Introduction to Polymer Physics. Cambridge: Cambridge University Press; 2002.
2. Fried JR. Polymer Science and Technology. 2nd ed. NJ: Pearson Education Inc.; 2003.
3. Lakes R, Park J. Biomaterials: an Introduction. 2nd ed., New York: Plenum; 1992.
4. Ratner B, Hoffman A, Schoen F, Lemons J. Biomaterials Science: An Introduction to Materials in Medicine. 2nd ed. Burlington, (MA): Academic Press; 2004.
5. Lanza R, Langer R, Vacanti J. Principles of Tissue Engineering. 2nd ed. Burlington, (MA): Academic Press; 2000.
6. Agrawal CM, Ray RB. Biodegradable Polymeric Scaffolds for Musculoskeletal Tissue Engineering Hoboken, (NJ): John Wiley & Sons, Inc.; 2001.
7. Liu X, Ma PX. Polymeric scaffolds for bone tissue engineering. Ann Biomed Eng 2004;32:477–486.
8. Smith LA, Ma PX. Nano-fibrous scaffolds for tissue engineering. Colloids and Surfaces B: Biointerf 2004;39:125–131.
9. Langer R. New methods of drug delivery. Science 1990;249:1527–1533.
10. Hoffman AS, Hydrogels for biomedical applications. Adv Drug Deliv Rev 2002;54:3–12.
11. Brannon-Peppas L. Polymers in controlled drug delivery. Med Plast Biomater 1997;4:34–44.
12. Qiu Y, Park K. Environment-sensitive hydrogels for drug delivery. Adv Drug Deliv Rev 2001;53:321–339.
13. Kikuchi A, Okano T. Pulsatile drug release control using hydrogels. Adv Drug Deliv Rev 2002;54:53–77.

See also BIOMATERIALS: POLYMERS; BIOMATERIALS, TESTING AND STRUCTURAL PROPERTIES OF.

POLYMERS. See BIOMATERIALS: POLYMERS.

PRODUCT LIABILITY. See CODES AND REGULATIONS: MEDICAL DEVICES.

PROSTHESES, VISUAL. See VISUAL PROSTHESES.

PROSTHESIS FIXATION, ORTHOPEDIC. See ORTHOPEDICS, PROSTHESIS FIXATION FOR.

POROUS MATERIALS FOR BIOLOGICAL APPLICATIONS

GRACE E. PARK
THOMAS J. WEBSTER
Purdue University
West Lafayette, Indiana

INTRODUCTION

Porous materials have received much attention in the scientific community because of their ability to interact with biological ions and molecules not only at their surfaces, but also throughout their bulk (1). Because of this intrigue, traditional applications of porous materials have involved catalysis, bioseparations, adsorption of select species, and ion exchange (1). As the tissue engineering field has emerged due to the continuous need for better implantable materials, porous materials have also found their niche in regenerative medicine. Specifically, porous materials have been employed as implants for various parts of the body (e.g., bone, cartilage, vasculature, central and peripheral nervous systems, bladder, and skin) either as stand-alone regenerative devices or as drug delivery vehicles to promote tissue growth. Problems associated with current implants and the need for better porous biomaterials in numerous anatomical locations are described below.

Most significantly, estimated annual U.S. healthcare costs related to tissue loss or to organ failure surpassed \$400 billion in 1997 (2). An estimated 11 million people in the United States. have received at least one medical implant device; specifically, orthopedic implants (including fracture, fixation, and artificial joint devices) constitute the majority of these and accounted for 51.3% of all implants in 1992 (3). Among joint-replacement procedures, hip and knee surgeries represented 90% of the total and in 1988 were performed 310,000 times in the United States alone (3). Implanting an orthopedic material can be a costly procedure involving considerable patient discomfort, both of which can increase if surgical revisions become necessary after an orthopedic or dental implant is rejected by the host tissue, is insufficiently integrated into juxtaposed bone, and/or fails under physiological loading conditions. Unfortunately, the average lifetime of an orthopedic implant is only 15 years due to many factors including the lack of osseointegration into surrounding bone. Current metallic implants are for the most part nonporous with subsequent poor surface properties to promote new bone ingrowth quickly.

The reason for such a high number of implanted orthopedic–musculoskeletal devices stems from numerous bone diseases. For example, approximately one out of seven Americans suffer from some form of arthritis, which is an inflammatory condition due to wear and tear in the joint (4). The cost of arthritis and rheumatic diseases reaches \$86.2 billion a year, according to a study by the Arthritis Foundation and the National Institutes of Arthritis and Musculoskeletal and Skin Diseases (NIAMS) (5). Although a very common disease, repairing damaged articular cartilage is challenging due to its limited ability to self-repair as a result of its avascularity. In fact, one of the most popular surgical techniques to repair cartilage is not through the use of a biomaterial, but rather is a surgical technique that further injures cartilage to induce scar tissue formation. This scar tissue is intended to serve as new cartilage, but since it does not match the mechanical properties of cartilage tissue, patients receiving such treatments usually suffer from additional complications and pain after only 5 years of this procedure.

The story is not any better for vascular diseases requiring biomaterial intervention. Specifically, the leading cause of death in the United States is vascular disease (including atherosclerosis), affecting $\sim$ 58 million people (6). Atherosclerosis, which is hardening of the arteries, is caused by accumulation of cholesterol, fatty molecules, and other substances inside the vessel wall as the lumen becomes gradually narrower. Consequently, complete blockage of the lumen may result, inhibiting the blood flow through that blood vessel. Treatments for these conditions require the use of a vascular graft, initially seeking autologous (or taken from an individual's own tissue) materials. For those patients receiving a synthetic vascular graft, success rates for vessels < 7 mm approaches only 25% after 5 years. Current biomaterials used as small diameter vascular grafts are usually nonporous and result in the eventual reaccumulation of undesirable substances that clog the vessel lumen to block blood flow.

Neurological problems also necessitate the use of biomaterials. For example, Parkinson's disease, Huntington's disease, Alzheimer's disease, and epilepsy prevail as common central nervous system (CNS) degenerative pathologies, especially targeting the aging population. While most of these diseases may cause a form of dementia (a mental deterioration), Alzheimer's disease involves the loss of nerve cells related to memory and mental functions, whereas Parkinson's and Huntington's disease affect the mind and body. Among these, > 1.5 million Americans have been affected by Parkinson's disease (6) and ~ 24.4 million people are diagnosed with Alzheimer's disease and stroke, costing $>$ \$174 billion annually (7). These diseases, however, account for only those affecting a portion of the CNS: the brain. Equally as troubling are spinal cord disfunctions. Spinal cord injuries can seriously cause damage to a person's quality of life, contributing to $\sim 200{,}000$ Americans with this disability and expenses of up to \$250,000 a year per individual as reported in 1996 (8). Various treatment methods, such as the use of pharmaceutical agents, electrical stimulation probes, and bridges or conduits to physically connect damaged regions of the spinal cord have been developed and improved. However, few clinically approved porous biomaterials are available for treating peripheral and central nerve damage. This is despite the fact that pores in biomaterials could be very useful for guiding nerve fibers through damaged tissues.

Bladder is another organ that could benefit from the use of porous biomaterials. Urinary cancer stands as one of the most common forms of bladder disease, which is the second most common malignancy of the genitourinary tract and the fourth leading cause of cancer among American men (9). Conventional treatment methods include the resection of the cancerous portion of the bladder wall in conjunction with intravesical immunochemotherapy (10). However, these treatments have been less than successful due to local and systemic toxicity of chemotherapy agents (11) and possible reoccurrence of the cancer (12–14). The best approach to resolve these problems is to completely remove the bladder wall, which clearly leads to the need for a replacement porous biomaterial with highly effective designs matching the material and mechanical properties of the native bladder tissue.

The above statistics highlight the current state of diseases in numerous organs and the potential effect porous biomaterials could have in treating these ailments. It is currently believed that porous biomaterials may be the solution to healing these damaged organs if designed appropriately. The next section will emphasize the features a successful porous biomaterial should have for regenerating tissues.

FEATURES OF THE NEXT GENERATION OF SUCCESSFUL POROUS BIOMATERIALS

An ideal porous scaffold for regenerating the tissues–organs mentioned in the previous section should have these characteristics (15,16): a highly porous three-dimensional (3D) interconnected network of pores for cell infiltration and subsequent tissue ingrowth; biodegradable or bioresorbable in a controllable manner with a rate that matches that of tissue growth; appropriate surface chemistry to promote desirable cell adhesion, proliferation, and differentiation; permeable for transporting sufficient amount of nutrients and metabolic waste to and from cells; mechanical properties that match that of the tissues surrounding the biomaterial in situ; and ease of processibility for various desired shapes and sizes to match specific tissue abnormality.

Several studies have confirmed that biomaterial pore size, interconnectivity, and permeability (among other properties) play a crucial role in tissue repair (17–19). Specifically, from the aforementioned list, in the following sections surface, mechanical, degradation, porosity, pore size, and pore interconnectivity properties important for the success of porous biomaterials are elaborated.

Surface Properties

Porous Biomaterial Surface Interactions With the Biological Milieu. Assuming that the porous biomaterial has a clean surface after synthesis (Fig. 1a), the surface will be contaminated with various substances in air (e.g., hydrocarbons, sulfur, and nitrogen compounds) immediately before implantation (Fig. 1a and b) (20). Sterilization and/or introducing coatings can remove or reduce the level of contaminants. The initial interaction between an implant and the biological milieu *In vivo* occurs with water molecules (Fig. 1d) as a mono- or bilayer forms on the surface depending on the porous biomaterial's surface hydrophilicity (or binding strength of water molecules to the implant surface) (21). Water layers form within nanoseconds as other ions contained in body fluids (e.g., Cl^- and Na^+) interact with the adsorbed water molecules depending on the porous biomaterial surface chemistry (Fig. 1e). It is also possible that water interactions containing ions can penetrate the bulk porous biomaterial. Subsequently, proteins adsorb to their surfaces via initial adsorption, and then possible protein conformational changes or denaturation occurs (Fig. 1f). Replacement of these initial proteins with other proteins contained in bodily fluids may occur when biomolecules with stronger binding affinities approach the surface at a later time. Final conformations of the adsorbed proteins may differ from what occurred initially (Fig. 1g). Cells then interact with or bind to the adsorbed proteins on the porous biomaterial surface (Fig. 1h). The type of cells attached to proteins adsorbed on material surfaces and their subsequent activities will determine the tissue formed on the surface (Fig. 1i).

Protein Interactions with Porous Biomaterials

Protein Structure. Clearly, as just mentioned, one of the key events that will determine porous biomaterial success or failure is initial protein adsorption. To further explore

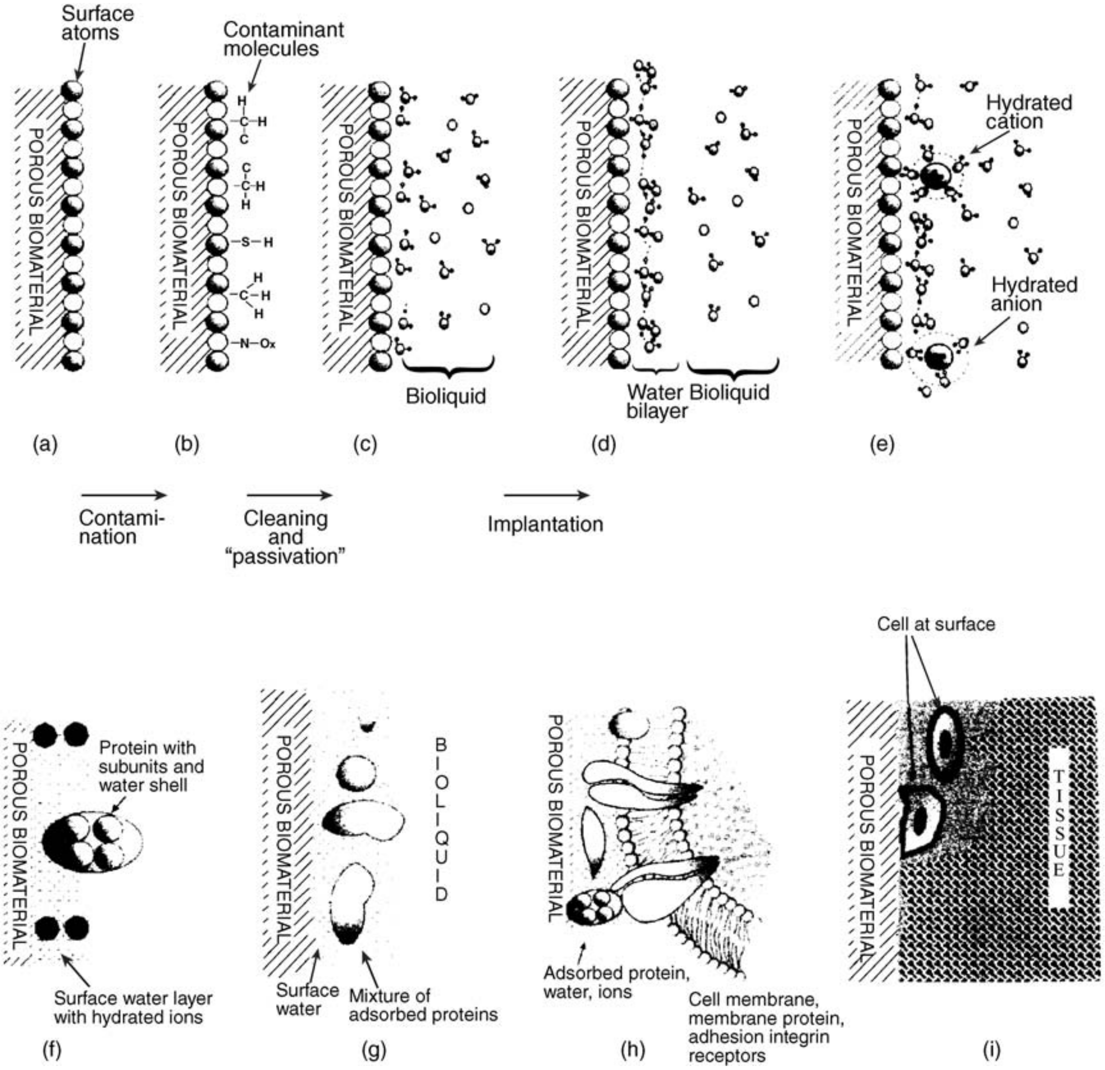

Surface atoms
Contaminant molecules
POROUS BIOMATERIAL
Bioliquid
Water bilayer
Bioliquid
Hydrated cation
Hydrated anion
(a)
(b)
(c)
(d)
(e)
Contamination
Cleaning and "passivation"
Implantation
Cell at surface
Protein with subunits and water shell
Surface water layer with hydrated ions
BIOLIQUID
Surface water
Mixture of adsorbed proteins
Adsorbed protein, water, ions
Cell membrane, membrane protein, adhesion integrin receptors
TISSUE
(f)
(g)
(h)
(i)

Figure 1. Schematic of the porous biomaterial–tissue interface. (Adapted from Ref. 20) (a) An initially clean porous biomaterial surface possesses surface atoms. (b) A porous biomaterial surface is contaminated with molecules from the ambient environment. (c) The surface is cleaned and passivated by saturation of dangling bonds. (d) A water bilayer forms immediately after implantation. (e) Hydrated ions (e.g., Na^+, and Cl^-, and Ca^{2+}) are incorporated into the water layer. (f) Proteins adsorb onto the surface depending on their concentration and size as well as properties of the porous biomaterial surface. (g) Various types of proteins adsorb to the surface at different conformations. (h) Cells bind to the proteins that adsorbed on the porous biomaterial surface. (i) Activity of the cells at the interface determines the type of tissue formed at that site.

this, first protein structure must be discussed. There are four levels of protein structure: primary, secondary, tertiary, and quaternary structures. It is important to understand how these different types of protein structures influence initial interactions with surfaces and consequently control cellular adhesion. The primary structure of a protein is its linear sequence of amino acids. Each amino acid is linked to another through peptide bonds. Some amino acids have side chains that are charged or are neutral. Those of particular importance in aqueous solutions exhibit polar characteristics. Other amino acids change their properties depending on the pH of the solution they reside in. Therefore, it should not be surprising that proteins exist with a wide range of properties as shown in Table 1. Table 1 describes the diverse nature of proteins in terms of size, shape, stability, and surface activity. To emphasize this diversity in protein properties, note the different interactions of albumin compared to fibrinogen on polyethylene (Table 1). Albumin is a cell nonadhesive protein while fibrinogen adsorption enhances a series of events leading to blood clot formation, a common problem of porous biomaterials in vascular applications.

Secondary protein structure consists of ordered structures in the protein chain. Two main secondary structures of proteins are the α-helix and β-pleated sheet. The degree of these structures may vary in a single protein and they are controlled by hydrogen-bonding mechanisms, which

Table 1. Diverse Properties of Proteins [a]

Protein	Function	Location	Size, kDa	Shape, nm	Stability	Surface Activity
Albumin	Carrier	Blood	65	4.2 × 14.1	Denatures at 60 °C	Low on polyethylene
Fibrinogen	Clotting	Blood	340	46.0 × 6. (trinodular string)	Denatures at 56 °C	High on polyethylene
IgG	Antibody	Blood	165	T-shaped		Low on polyethylene
Lysozyme	Bacterial lysis	Tear; hen egg	14.6	4.5 × 3.0 (globular)	$\Delta G_{\mathrm{n}} = -14\ \mathrm{kcl{\cdot}mol^{-1}}$	High on negatively charged surfaces
Hemoglobin	Oxygen carrier	Red blood cells	65	5.5 (spherical)	Normal form	Very high on polyethylene
Hemoglobin S	Oxygen carrier	Sickle red blood cells	65	5.5 (spherical)	Less than hemoglobin	Much higher air–water activity than hemoglobin
Myoglobin	Oxygen carrier	Muscle	16.7	4.5 × 3.5 × 2.5 spherical)	$\Delta G_{\mathrm{n}} = -12\ \mathrm{kcl{\cdot}mol^{-1}}$	
Collagen	Matrix factor	Tissue	285	300.0 × 1.5 (triple helical rod)	melts at 39 °C	
Bacteriorhodopsin	Membrane protein		26	3.0–4.0 long	$\Delta G_n = -8.8\ \mathrm{kcl{\cdot}mol^{-1}}$ denatures at 55 °C	High at cell membrane
Tryptophan Synthase alpha Subunit (wild type)	Enzyme		27		$\Delta G_n = -16.8\ \mathrm{kcl{\cdot}mol^{-1}}$	High air–water activity compared to ovalbumin
Tryptophasn Synthase Variant alpha Subunit	Enzyme		27			Much less active at air–water interface than wild type

[a] See Ref. 22.

are electrostatic attractions between oxygen of one chemical group and hydrogen of another chemical group.

Tertiary protein structures are the overall 3D shape of the protein that can be quite ordered or extremely complicated. The tertiary structure of proteins is a consequence of its primary structure as it depends on the spontaneous interactions between different amino acids and, under aqueous conditions, the spontaneous interactions between amino acids and water. There are four main interactions among residues of amino acids that contribute to the tertiary structure of proteins, each with different strengths: covalent, ionic, hydrogen, and van der Waals bonds. Of these interactions, covalent bonds are the strongest, ionic bonds are also strong (occurring between chemical groups with opposite charges), and van der Waals forces resulting from interactions between hydrophobic molecules are the weakest. However, the most influential bonds on protein tertiary structure are the weakest bonds: hydrogen bonds and van der Waals bonds. This is true since, compared to covalent and ionic bonds, these weaker bonds have many more opportunities for interacting in protein tertiary structure. In addition, because proteins exist in aqueous media, residues of amino acids must interact with water, which is a highly polar compound that forms strong hydrogen bonds. Therefore, the most stable structure of proteins in aqueous media is globular, having hydrophobic areas in the center and hydrophilic areas in the outer layer. Thus, although a generalization, it is possible that the adsorption of proteins to a porous biomaterial surface will be influenced by the presence of these hydrophilic amino acids on the outside of proteins in solution. However, when proteins come in contact with solid surfaces (e.g., porous biomaterials), protein structure will drastically change.

Only proteins that posses numerous subunits have quaternary structure. How these subunits interact will determine the quaternary structure of the protein. Interactions between amino acids on the exterior of the tertiary structure (mostly hydrophilic) will influence the quaternary structure, but certainly some hydrophobic interactions will also occur at the surface and impact quaternary structure.

Under certain extreme conditions (e.g., conditions that are outside of the physiological range or outside the range of 0–45 °C, pH 5–8, and in aqueous solutions of $\sim 0.15\ M$ ionic strength), proteins may loose their normal structure (23). In other words, under such conditions, the spherical or globular tertiary structure most soluble proteins assume in aqueous media will unfold or denature. The structure of denatured proteins has been described as a random coil structure similar to those found in synthetic polymers (23). Since the structure of the protein has changed from that of a hydrophilic–hydrophobic exterior–interior to a more random arrangement, often times denatured proteins loose their solubility, become less dense (folded protein structures have densities of $\sim 1.4\ \mathrm{g{\cdot}cm^{-3}}$), and loose their bioactivity (23). Although there have been many examples of protein denaturation in solution, in general, only few cases of full protein denaturation on porous biomaterial surfaces have been reported (23). That is, generally, proteins adsorbed at the solid–liquid interface are not fully denatured and retain some degree of structure necessary to mediate cell adhesion.

Protein Interactions Mediated by Surfaces. Soluble proteins present in biological fluids (e.g., blood plasma) are the type of proteins that are involved in immediate adsorption to surfaces (24). In contrast, insoluble proteins that comprise tissues (like collagen and elastin) are not normally free to diffuse to a solid surface; these proteins may, however, appear on solid surfaces of implantable devices due to synthesis and deposition by cells (23). As mentioned, in seconds to minutes, a monolayer of adsorbed protein will form on solid surfaces (23). The concentration of proteins adsorbed on a material surface is often 1000 times greater than in the bulk phase (23). Thus, extreme competition exists for protein adsorption due to a limited space available on the surface. Because of their diverse properties just described, proteins do not absorb indiscriminately to every material surface; that is, complimentary properties of the surface and of the protein as well as the relative bulk concentration of each protein determine the driving forces for adsorption (25,26). Moreover, this initial interaction is extremely important since some proteins are not free to rotate once adsorbed to material surfaces due to multiple bonding mechanisms. Thus, immediately upon adsorption, proteins are somewhat fixed in a preferred orientation or bioactivity to the bulk media that contains cells (23). Some porous biomaterial surface properties that have influenced protein adsorption events include chemistry (i.e., ceramic versus polymer), wettability (i.e., hydrophilicity compared to hydrophobicity), roughness, and charge as will be discussed later.

One of the major differences between a flat two-dimensional (2D) substrate surface and that of a 3D porous material is tortuosity. Clearly, protein interactions are much different on materials due to tortuosity. Specifically, a curved porous surface allows for greater surface area, enhanced interactions between adjacent electrons of the atoms on the surface of the pores, increased localization of point charges, and the potential for greater surface energy due to a larger juxtaposition of localized surface defects. Collectively, all of these differences between a nonporous and porous biomaterial provide for a much more complex environment for interactions between proteins and pore surfaces. It is the challenge of the porous biomaterial community to understand this challenge and thus design scaffolds that control select protein interactions.

Protein-Mediated Cell Adhesion. Interactions of proteins (both their adsorption and orientation or conformation) on porous biomaterials mediate cell adhesion. These interactions lead to extreme consequences for the ultimate function of an implanted device (27,28). An example of the importance of protein orientation for the adhesion of cells is illustrated in Fig. 2. A typical cell is pictured in this figure with integrin receptors that bind to select amino acid sequences exposed once a protein adsorbs to a surface (Fig. 1h). It is the ability of the cell to recognize such exposed amino acids that will determine whether a cell adheres or not. For example, many investigators are designing porous biomaterials to be more cytocompatible. However, it is the adhesion of select cells that must be emphasized. That is, many attempts have been made to immobilize select cell adhesive epitopes in proteins (e.g., the amino acid sequence arginine-glycine-aspartic acid or RGD) onto polymeric tissue engineering scaffolds. But, once implanted into bone, not only do desirable osteoblasts adhere, but so do undesirable fibroblasts (cells that contribute to soft not bony tissue juxtaposition).

Not only will cell adhesion be influenced by the exposure of amino acids in adsorbed proteins, but so will subsequent cell functions (e.g., extracellular matrix deposition). This is true since for anchorage-dependent cells, adhesion is a

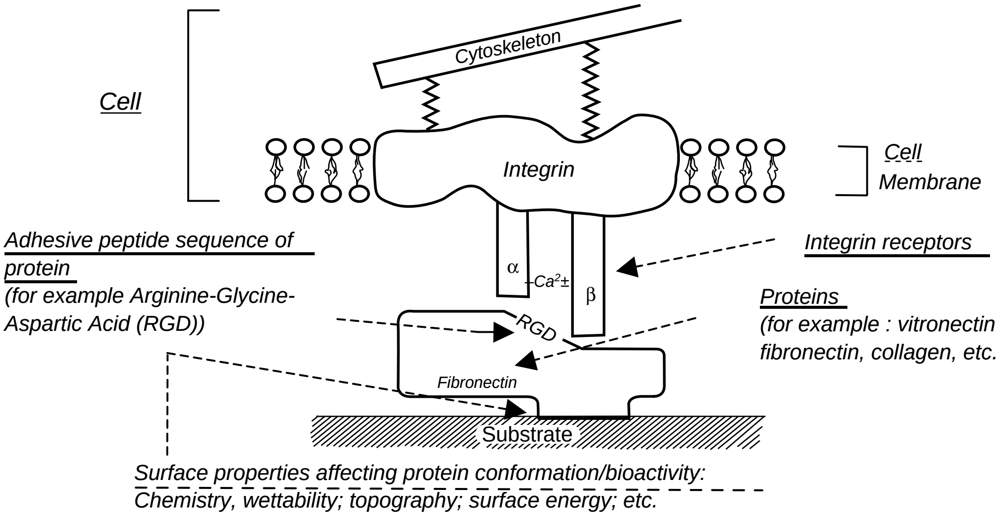

Cell
Cytoskeleton
Integrin
Cell Membrane
Adhesive peptide sequence of protein
(for example Arginine-Glycine-Aspartic Acid (RGD))
α
Ca2±
β
Integrin receptors
Proteins
(for example : vitronectin, fibronectin, collagen, etc.)
RGD
Fibronectin
Substrate
Surface properties affecting protein conformation/bioactivity:
Chemistry, wettability; topography; surface energy; etc.

Figure 2. Influence of protein conformation on cell integrin binding. Cell adhesion and its subsequent activity will be determined by the type of integrins that the cell uses to adhere to adsorbed proteins. (Adapted and redrawn from Ref. 35.)

crucial prerequisite for subsequent cell functions. Moreover, specific intracellular messages that control subsequent cell functions are transferred inside the cell depending on which integrin receptors are utilized by the cell to adhere to adsorbed proteins. For example, a recent study by Price et al. (29), demonstrated that new bone growth was promoted when osteoblasts adhere via heparin sulfate proteoglycan binding mechanisms (as opposed to RGD) to vitronectin adsorbed on porous ceramic scaffolds.

In this manner, it is clear that cells interact with their external environment through mechanical, electrical, and chemical signals transmitted through the cell membrane. As mentioned, cell adhesion is established through cell-binding regions of extracellular matrix proteins and respective cell-membrane-intercalated receptors (i.e., integrins) among other mechanisms. Integrins are a family of transmembrane heterodimeric glycoproteins that are receptors for specific epitopes of extracellular matrix proteins and for other cell-surface molecules (30). Integrins exist as a dimer complex composed of an α-subunit (120–180 kDa) noncovalently associated with a β-subunit (90–110 kDa) (31). Several of these integrins have been identified that are concentrated at loci, called focal adhesion sites, of close proximity between cells and extracellular matrices on substrates (31). Focal adhesion sites are points of aggregation of, and are physically associated with, intracellular cytoskeletal molecules that control, direct, and modulate cell function in response to extracellular signals (32).

However, integrin–protein interactions are not the only mechanisms by which cells adhere. Several articles suggested that *In vivo* (6) and *In vitro* (33,34) osteoblasts (bone-forming cells) attach to an implanted material through cell membrane heparin sulfate proteoglycan interactions with, for example, heparin-binding sites on fibronectin and collagen. Moreover, Nakamura and Ozawa (6) immunohistochemically detected heparin sulfate on the membranes of osteoblasts attached to bone matrix.

Whatever the method of cell attachment, protein orientation will alter from surface to surface, since neither proteins nor materials are homogeneous in properties or structure on the exterior. The existence of protein regions that are largely acidic–basic or hydrophilic–hydrophobic or have select amino acids exposed to the media will greatly influence how that protein adsorbs to a surface and, thus, its orientation. Similarly, ceramics, metals, polymers, and composites thereof have vastly different chemistries and atomic bonding mechanisms (i.e., ionic, metallic, and covalent) to influence protein interactions. The initial interactions between proteins important for cell functions and the design of better porous biomaterials is emphasized in the next section.

Design of Better Porous Biomaterial Surfaces. As mentioned, not only do properties of proteins determine the degree of their interactions with surfaces, but properties of the media and surface (specifically, wettability, surface energy, chemistry, roughness, etc.) also influence the degree of protein interactions (35). Clearly, altering surface properties to control such protein events for mediating cell function leading to tissue regeneration is at the heart of the research of many biomaterial scientists and engineers. Surface properties are so important because proteins have relatively large sizes and correspondingly large numbers of charged amino acid residues of different acidity/basicity well distributed on their exteriors. The polyelectrolytic property of proteins provides for exciting design criteria in surfaces to maximize or minimize specific protein interaction. Not surprisingly, at a neutral or slightly charged surface and at a pH in which the net charge on the protein is minimal, most proteins will exhibit maximum adsorption (23). For surfaces with a large net charge, initial protein interactions will be dominated by the degree of the opposite charge on the surface (23,35).

Consideration of the spatial organization of amino acids can be used in the design of surfaces to enhance protein interactions (36). As previously mentioned, for some proteins, hydrophilic and hydrophobic amino acids are present primarily on the exterior and interior, respectively. This spatial arrangement has a direct consequence on the initial interactions of these proteins with surfaces. For example, a surface that initiates interactions with the exterior hydrophilic amino acid residues in that type of a protein may promote its adsorption. In contrast, for the interior hydrophobic amino acid residues to interact with material surfaces, which may contain desirable cell adhesive epitopes (e.g., RGD), the soluble protein may have to unfold or loose tertiary structure. For this reason, one approach to increase the adsorption of a protein whose external amino acids are largely hydrophilic, would be to design a material surface which exhibits polar properties. The same can be said for any type of protein; that is, through an understanding of the amino acids that reside on the protein exterior when in the appropriate biological milieu, a complimentary surface can be designed. It is important to note, though, that this is a generalization as many proteins have a diverse collection of hydrophilic–hydrophobic amino acids externally that must be considered. In addition, as previously mentioned, proteins adsorb to surfaces in a competitive manner in which the adsorption of one protein will influence that of another.

Several studies have confirmed these speculations that properties (chemistry, charge, topography, etc.) of porous biomaterial surfaces dictate select interactions (type, concentration, and conformation–bioactivity, etc.) of proteins (24,37–40). It has been reported in the literature that changes in the type and concentration (up to 2100, 84, and 53% for albumin (40), fibronectin (41), and vitronectin (34), respectively) of protein adsorption on material surfaces depends on material surface properties, such as chemistry (i.e., polymer, metal, or ceramic), hydrophilicity–hydrophobicity, roughness, and surface energy. Consequently, since protein interactions can be controlled on porous biomaterial surfaces, so can cell adhesion. For example, a common porous biomaterial [poly(lactic-*co*-glycolic acid) or PLGA] has been modified to increase the adsorption of vitronectin and fibronectin through NaOH treatments (42–44). Since both vitronectin and fibronectin mediate osteoblast, vascular cell, and bladder cell adhesion, these NaOH treated PLGA scaffolds have found a home in numerous tissue engineering applications.

However, for the field of porous biomaterials to advance even further, instead of broadly speaking of protein adsorption on surfaces, researchers need to investigate and design succinct regions of surfaces to promote protein adsorption considering the complexities of their properties. Only when porous biomaterials are considered from the context of protein interactions necessary for desirable cell interactions, will better tissue engineering materials be formulated.

Mechanical Properties and Degradation Byproducts

Although porous biomaterial surface properties determine cell attachment, mechanical strength of the scaffold and the mechanical environment it provides plays an equally important role in enhancing subsequent cell functions leading to tissue growth (45). Mechanical forces felt by cell membrane molecules are interconnected to the cytoskeleton that can influence messenger ribonucleic acid (mRNA) and subsequent synthesis of intracellular proteins (all the way to the nucleus where gene expression can be changed). It is for these reasons that mechanical properties must also be carefully controlled in porous biomaterials. For example, a study of various mechanical stimuli placed on equine articular chondrocytes within nonwoven polyglycolic acid (PGA) mesh scaffolds indicated that when the stimuli were removed, after a period of 1 week, the mechanical integrity of the resulting tissue construct was lost (46). This result implies that the mechanical stimuli applied to cells within a porous biomaterial may influence the biomechanical functionality of the regenerated tissue.

Although most agree that the mechanical properties of a porous biomaterial should match those of the physiological tissue they are intending to replace, the specific parameters and values desirable in these studies vary. For bone tissue engineering, for example, Yaszemski et al. (47) stated that scaffolds should possess mechanical stiffness matching the low range values of trabecular bone (50–100 MPa), whereas Hutmacher's design principle (15,48) suggests matching the native tissue stiffness (10–1500 Mpa for trabecular bone (49)). Clearly, this wide range in mechanical values can provide for much different porous biomaterial efficacies and a consensus needs to be established.

Once deciding on the optimal mechanical properties needed in scaffold structures, there are numerous design parameters that can be exploited to match such values. For example, for a fibrous mesh, a decrease in fiber diameter increases mechanical strength due to an increase in fiber density (50). Obviously, increasing percent porosity and the diameters of individual pores can also be used to decrease the strength of scaffolds to match desired values. These properties not only influence inherent mechanical properties of scaffolds, but they can also be used to manipulate cell functions.

Specifically, Maroudas postulated that the scaffold surface rigidity or stiffness enhances cell adhesiveness and cell spreading (51). Pelham and Wang (52) have shown that focal adhesion contacts in cells and their migration on acrylamide gels are controlled by scaffold flexibility. They also suggested that tyrosine phosphorylation might be involved, activated by local tension at cell adhesion sites (53). Recently, Ohya et al. (54) studied the effects of hydrogel mechanical properties on cell adhesiveness and found that the higher the strength of the hydrogel formulation, the greater the capability to withstand cell traction forces, thereby resulting in greater cell spreading. These authors also noticed that cells preferred to adhere to stiffer regions within the hydrogel.

Common pore shapes in porous biomaterials include tube-like, spherical, and randomly spaced shapes. Differences in cell attachment, growth, migration, and matrix deposition by cells have all been observed depending on pore structure. Specifically, certain cell types prefer a select pore structure in accordance to their physiological matrix environment. For example, orthopedic tissue engineering scaffolds should have spherical pores with a high porosity to allow for immediate bone ingrowth, while maintaining the mechanical strength and integrity necessary due to their harsh mechanical environments *In vivo* (36). Porous biomaterial pore shapes are critically related to pore interconnectivity. Not only does pore interconnectivity in a porous biomaterial affect nutrient–waste diffusion, but it also influences cell growth. Bignon et al. (55) observed that the density of pore interconnections determines cellular colonization rates; meaning that the larger the macropores (within limits), the fewer pore interconnections that have to be transversed by the cells thus resulting in higher colonization rates. Of course, guided cell growth or migration is possible through deliberate pore shape and interconnectivity. For example, tube-like or fibrous pore shapes may promote neurite extension from neurons in specific directions. Studies have also shown that cells prefer discontinuities within a porous material in terms of growth and migration; clearly pores provide such discontinuities (56–58).

In addition, maintaining mechanical strength and structural integrity of porous biomaterials are crucial because scaffolds may be crushed when implanted or may degrade over time. Mechanical properties are especially important to characterize when they change over time. A thorough knowledge of the degradation process of the porous materials of interest (including degradation byproducts) should be mapped in order to control the mechanical stability and the degradation rate until the native tissue is formed at the site of implantation.

For porous biomaterials, a range of biodegradation choices exist, from nondegradable metals to degradable ceramics and polymers. Importantly as well, degradation rates of porous materials have in some cases been shown to be faster compared to solid block polymers (59,60) because acidic byproducts become trapped inside the bulk as they degrade, therefore causing an autocatalytic effect. Of course, trapping of acidic byproducts in polymeric scaffolds can have detrimental consequences on cell health. Porous degradable polymers, such as PGA, polylactic acid, PLGA, and polycaprolactone (PCL) degrade via nonenzymatic random hydrolytic breaking of ester linkages. Sung et al. (61) studied the degradation of PLGA and PCL scaffolds *In vitro* and *In vivo*. They found a significant decrease in the molecular weight of these polymers within 1 month *In vitro* and, as expected, at a much faster rate *In vivo*.

Specifically, the influence of acidic byproducts from these polymeric scaffolds on cell health was investigated by measuring the pH of the media in which the polymers resided compared to the media in which tissue culture polystyrene (TCPS) was cultured. Changes in the media pH occurred only for PLGA (reducing it by 5) whereas no significant changes were measured during TCPS or PCL culture for up to 28 days (61).

Moreover, an *In vivo* study by Hedberg et al. (62) determined that soluble acidic products from degradable polymers lead to an increased recruitment of inflammatory cells compared to that induced by the scaffold itself. This was evidenced by the fact that a minimal inflammatory response was observed at the site of bone growth juxtaposed to the surface of polymeric scaffolds, whereas a major inflammatory response was observed in the scaffold where there was significant degradation. However, Sung et al. (61) suggested that an inflammatory response can be beneficial towards angiogenesis that is highly desirable to remove harmful degradation products from the interior of a polymer scaffold. This clearly demonstrates the need for controlling polymer degradation products in order to elicit a desirable response from host tissue (63). Collectively, such studies highlight the necessity for a better understanding of material degradation products on cell health.

In addition, according to Wu and Ding (64), the molecular weight of a porous PLGA scaffold decreases during degradation, which not only creates a more acidic local environment, but also leads to other changes. In their study, degradation was divided into three stages, marking distinct characteristics in mechanical properties (Fig. 3). In the first stage (I), the mechanical strength increased as the porous scaffold dimensions decreased while the weight remained constant; this can be interpreted simply as the change of porous biomaterial dimensions resulting in mechanical property increases. Increased elastic modulus of porous PLGA scaffolds with degradation time was also observed in another study by Zhang and Ma (65) who contributed this to decreased porosity of the foams with time. In the second stage (II), a dramatic decrease in mechanical properties were observed, which was correlated with an increased presence of low molecular weight degradation products (64). The third stage (III) was characterized by the breakdown of the scaffold's structural integrity and associated rapid weight loss due to pH decreases from acidic degradation products. Understanding of these three distinct phases of mechanical property degradation for every proposed porous biomaterial is imperative. In addition, more studies are needed that correlate cell function at each stage of mechanical property changes in porous biomaterials as they degrade.

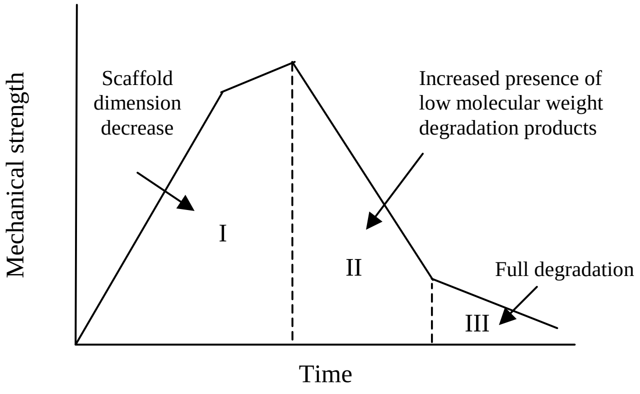


Figure 3. Three stages of mechanical strength degradation in porous biomaterials. (Adapted and redrawn from Ref. 64.)

The Role of Porosity, Pore Size, and Interconnectivity

Among other properties (e.g., the aforementioned mechanical properties), porosity can also influence how cells behave in a scaffold. Open pore structures are desirable in most tissue engineering applications, providing enhanced cell seeding, cell attachment, proliferation, extracellular matrix production, and tissue ingrowth. For example, for orthopedic applications, both *In vitro* and *In vivo* studies demonstrated exceptional osteoblast proliferation and differentiation leading to new bone growth in PLGA foams with 90% porosity (66,67). In addition, a study by Sherwood et al. (68) reported that osteochondral composite scaffolds with 90% porosity at the cartilage portion allowed full incorporation of the chondrocytes (cartilagesynthesizing cells) into the scaffold.

Permeability, or high interconnectivity of pores is a crucial property for a porous biomaterial due to its influence on cellular communication (16), adaptation to mechanical stimulation (45), and prevention of the accumulation of acidic degradation byproducts (69,70). It also allows for uniform cell seeding and distribution, as well as proper diffusion of nutrients and metabolic wastes. Studies have shown that when tissues become thicker than 100–200 μm, the oxygen supply to cells becomes limited in a static environment (71,72). Thus, interconnectivity of pores is an extremely important design consideration to increase tissue growth into porous biomaterials.

In addition, as mentioned in the section above, the increased tortuosity present in porous biomaterials will influence protein interactions and, thus, manipulate cellular functions. Specifically, because of altered initial protein interactions, certain cell types (e.g., chondrocytes) perform much better on porous compared to flat (or nonporous) biomaterials (42). Moreover, macroporosity (pore diameters > 50 μm) influences the type of cells adhering to a polymeric scaffold. For example, large pores (100–200 μm diameter) have been shown to enhance bone ingrowth compared to smaller pores (10–75 μm diameter) in which undesirable fibrous soft tissue formation has been observed (73). Yuan et al. (74) added that pore sizes < 10 μm promotes bone ingrowth due to optimal initial protein adsorption events possibly because of their greater surface areas. Furthermore, Bignon et al. (55) demonstrated greater cell spreading on biomaterials with micro (pore diameters < 10 μm) compared to macroporosity. Importantly as well, pore wall roughness is influenced by pore size that may be providing greater roughness to promote cell functions. Studies are needed to carefully control pore wall roughness to make accurate comparisons between scaffolds of various degrees of pore sizes. Since very small topographical changes have been shown to alter cell functions (75), surface roughness values in the nanometer regime could also be incorporated into porous biomaterials regardless of their pore sizes to significantly enhance protein–surface and protein–cell interactions (42,76). Fabrication methods that can provide for the manipulation of pore properties is further emphasized here.

POROUS BIOMATERIAL FABRICATION METHODS

Various methods for fabricating porous biomaterials have been explored to date. Examples for forming polymeric porous scaffolds include solvent casting with particulate leaching, gas-foaming processes, emulsion freeze drying, freeze-extraction, electrospinning, rapid-prototyping, and thermally induced phase separation. Although polymers receive the most attention in porous biomaterial applications, porous ceramics, and metals have been recently receiving much attention. This is mostly because ceramics and metals have a long history of implantation, so methods that can improve their cytocompatibility properties (e.g., by creating pores) are highly desirable. For ceramic and metallic porous biomaterials, electrophoretic deposition, salt leaching, microsphere (polymer) melt out, and annodization have been commonly employed. These methods will be briefly described in the following sections.

Cellular Solids

Current methods, such as solvent casting, gas foaming, vacuum drying, and thermally induced phase separation (TIPS) in conjunction with particulate leaching techniques create cellular solids (77). These methods can create porous constructs easily and in an inexpensive manner (78,79). In solvent casting, a pellet or powder form of a polymer is dissolved in a solvent. Then, water-soluble salt particles (e.g., sodium chloride, sodium citrate) or other particulate materials [e.g., gelatin, paraffin (79)] are added to the polymer solution. The solvent is removed through evaporation or lyophilization and then particles are leached out through the use of water or another solvent (depending on the particle chemistry) to create the desired porous structure. The advantages of these methods include simplicity and the ability to control pore size and porosity. However, the pore shape is limited to the shape of the porogen and the pore interconnectivity is poor; thus, the porogen may not be completely removed from the construct (80). Furthermore, uneven dispersion or settling of the particles within the constructs may occur. Lastly, these first generation approaches rarely provided the succinct spatial ability to control protein adsorption necessary for the next generation of more successful biomaterials.

For the gas-foaming process, a gas, usually carbon dioxide (CO_2), is utilized instead of using an organic solvent at high pressures to create a highly porous structure (80–82). Again, these techniques are easy to implement and are inexpensive. However, a polymer with highly amorphous fractions can be processed with this technique even though the interconnectivity of the pores is very low, only $\sim$ 10–30% (18).

Thermally induced phase separation produces a highly porous material using a solvent at elevated temperatures followed by lowering the temperatures to separate the solution into liquid–liquid or solid–liquid phases. Then, the unwanted solvent is removed through sublimation (65,83). Although high mechanical strength may be obtained with this technique, the pore size created with TIPS normally ranges from 10–100 μm, which does not satisfy the permeability requirements for the removal and entry of cellular wastes and nutrients, respectively.

The emulsion freeze-drying method was developed by Whang et al. (84). A porous structure is obtained through homogenization of a polymer, organic solvent, and water mixture; rapidly cooling the mixture to maintain the liquid state structure; and then removing the solvent, and water by freeze-drying (80). In Whang's study, 90% or greater porosity and up to 200 μm diameter pores were created. However, this method is user and technique sensitive, meaning that pore structures and associated interconnectivities greatly depend on the processing method. The freeze-extraction method is a modified version of the freeze-drying technique, in which the solvent in the frozen polymer solution is replaced with a nonsolvent at temperatures below the freezing point of the polymer solution. This procedure removes the solvent before the drying stage (85).

Electrospinning Technique

In electrospinning, an electric field directs polymer fibers to form and deposit onto a substrate of interest (86,87). Specifically, an electric potential is applied as the polymer solution is injected, which ultimately forms an electrically charged jet of polymer landing on the target substrate. The solvent evaporates and porous polymer fibers are formed. Fibrous polymer scaffolds with diameters of several hundred nanometers can be fabricated using this method, thus, simulating the physiological fibrous structure of such proteins like collagen that comprise tissues. Only films and cylindrical shapes of the porous material have been created through this technique, therefore, further investigations are needed. But in addition to creating biologically inspired nanometer fibers an advantage of this process is its ability to coat an existing implant material. Thus, this technique could be used to modify the surface properties of currently used implant materials to promote cell functions.

Rapid Prototyping

Rapid prototyping is a computer-guided manufacturing system that can produce complex designs rapidly. One of the prototyping techniques is called 3D printing (3-DP) and it has been used to fabricate biodegradable polymer scaffolds for tissue engineering purposes (88). This technique produces porous biomaterials by ink-jet printing a binder onto sequential powder layers. Importantly, growth factors, proteins, cells, and other biological factors can be incorporated into the porous biomaterial without risking inactivation because the process is performed at room temperature. However, a disadvantage of this process so far includes porous biomaterial size limitations (due to the size of the ink jet). This can also limit the creation of desirable fine details or nanostructures on the polymer.

Microsphere Burnt Out

The microsphere burnt out method is similar to the previously described salt leaching method except that polymer microspheres are utilized instead of a salt porogen. This method is useful for ceramic materials that require a sintering process at very high temperatures (approaching 1000°C) at which point the polymer melts. As a very simple and easy method, it also has the disadvantages of the need for large amounts of the microspheres to create high pore interconnectivity; this results in poor mechanical properties.

Electrophoretic Deposition

Another attractive method for creating porous ceramics is electrophorectic deposition (or EPD). Due to a relatively simple setup and accommodation of complex designs and shapes, EPD has received much attention for processing fine particles, especially for coating applications (89). For this process, Ma used a graphite cathode and a stainless steel anode in the EPD cell while a current was applied to induce deposition of the particles onto a designated material. In this study, hydroxyapatite 3D porous biomaterials were fabricated. Hydroxyapatite is the main inorganic component of bone and, thus, has experienced wide spread use in orthopedic applications. This simple powder consolidation method requires no additives and high pore interconnectivity can be achieved with sufficient mechanical strength. However, this process can be costly when designing a large sample.

Anodization

Although not many methods exist to create porous metals, anodization is one that is gaining in popularity. Anodization involves the application of a voltage to a metal submerged in an electrolyte solution. Anodization has been used to create various pore sizes (from 10 nm to 1 μm) and shapes on two popular orthopedic metal chemistries: titanium and aluminum. In both cases, compared to respective unanodized metals, increased osteoblast functions have been reported on anodized titanium and aluminum (90,91). In addition, a study by Chang et al. (90) demonstrated that under certain anodization conditions porous nanotubes were created in titanium that further increased osteoblast adhesion. Although more testing is required, these studies highlight the fact that anodization is a fast and inexpensive method for creating pores in metals necessary for promoting bone growth.

Chemical Vapor Deposition

Another technique used to create porous metals is chemical vapor deposition. Chemical vapor deposition has been mostly used to fabricate porous tantalum for orthopedic applications. Tantalum is a new metal to the orthopedic field that possesses exception cytocompatibility properties. Tantalum porous biomaterials have been synthesized using vitreous carbon as the skeleton structure material (92). Tantalum was then coated onto the template and the template was removed by either chemical or heat treatments. Chemical vapor deposition is a common technique used in the coating industry and can easily be utilized for the fabrication of porous materials as long as a template or a mold is provided.

FUTURE DIRECTIONS IN THE DESIGN OF MORE EFFECTIVE POROUS BIOMATERIALS

Although there are numerous avenues, investigators are pursuing to improve the efficacy of porous biomaterials, one approach that involves the incorporation of nanotechnology seems to be working. Nanotechnology embraces a system whose core of materials is in the range of nanometers (1 nm). The application of nanomaterials for medical diagnosis, treatment of failing organ systems, or prevention and cure of human diseases can generally be referred to as nanomedicine. The commonly accepted concept refers to nanomaterials as that material with the basic structural unit in the range of 1–100 nm (nanostructured), crystalline solids with grain sizes 1–100 nm (nanocrystals), extremely fine powders with an average particle size in the range 1–100 nm (nanopowders), and fibers with a diameter in the range 1–100 nm (nanofibers). There have been many attempts to improve health through the use of nanotechnology, but perhaps the closest to clinical applications involves nanostructured biomaterials.

The greatest advantage of nanobiomedical implants in a biological context centers on scientific activities that seek to mimic the nanomorphology that proteins create in natural tissues. As seen in Figs. 4 and 5, bone and vascular tissue possesses numerous nanometer surface features due to the presence of entities like collagen and other proteins (93). Dimensions of some additional proteins found in the extracellular matrix of numerous tissues are found in Table 2 (94). As can be seen, the fundamental dimensions of these proteins (and all proteins) are in the nanometer regime. Clearly, when assembled into an extracellular matrix that comprises a tissue, these proteins provide a diverse surface with numerous nanostructured features for cellular interactions. Since some of these proteins are also soluble and present in bodily fluids, they will initially adsorb to implanted materials to provide for a highly nanostructured surface roughness for cellular interactions. It is for these reasons that cells of our body are accustomed to interacting with nanostructured surfaces. This is in stark contrast to most conventional porous biomaterials that are smooth at the nanoscale.

Aside from mimicking the surface roughness of natural tissues, there are other more scientific reasons to consider porous nanostructured biomaterials for tissue regeneration. Specifically, surface properties (e.g., area, charge, and topography) depend on the surface feature sizes of a material (95,96). In this respect, nanophase materials that, by their very nature, possess higher areas with increased portions of defects [e.g., edge/corner sites and grain or particle boundaries (95,96)] have special advantageous properties that are being exploited by porous biomaterial scientists for applications involving proteins and cells. As mentioned, proteins have complex structures and charges. Thus, surfaces with biologicallyinspired nanometer roughness provide control over protein interactions that were not

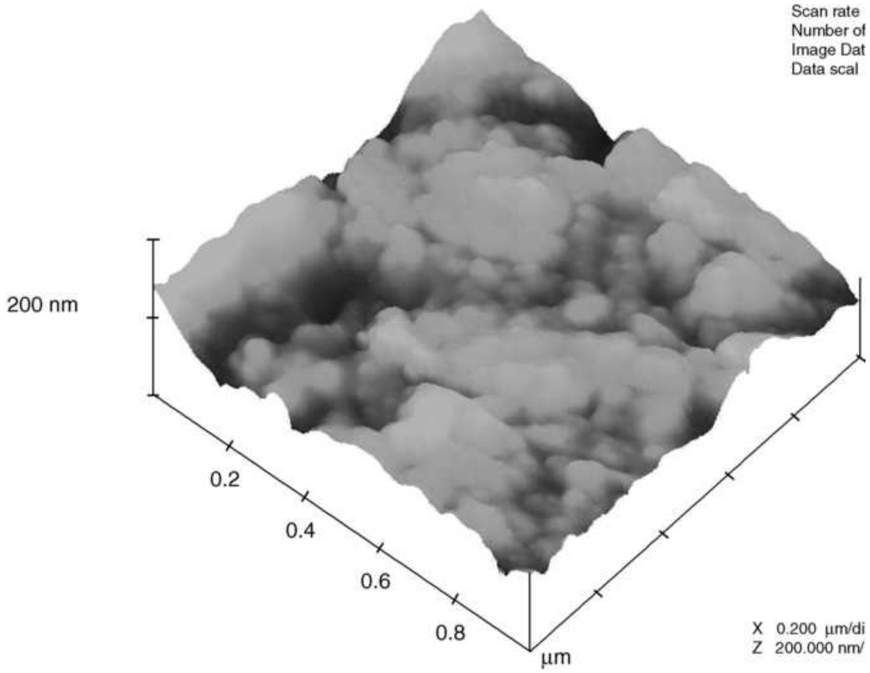


Figure 4. An AFM image of the surface of bovine cortical bone. Numerous nanometer features of bone duplicated in porous biomaterials are showing progress in orthopedic applications.

possible with conventional porous materials. Advances of porous nanostructured biomaterials pertinent to orthopedic, cartilage, vascular, central and peripheral nervous systems, and bladder applications will be briefly discussed in the sections below.

Orthopedic Applications

Nanophase ceramics (including alumina, titania, and hydroxyapatite), metals (e.g., titanium, titanium aluminum alloys, and cobalt chromium alloys), polymers (specifically, PLGA, polyether urethane, and polycaprolactone), and composites thereof have been explored for orthopedic applications (22,97,98). For these studies, nanophase surface features in ceramics and metals were created by using constituent nanometer particles, whereas nanostructured polymers were created using chemical etching techniques. In all of these studies, regardless of the manner in which the materials were synthesized, results indicated that nanophase materials enhanced osteoblast functions (e.g., attachment, proliferation, production of extracellular matrix proteins, and deposition of bone) compared to their respective conventional formulations (Fig. 6).

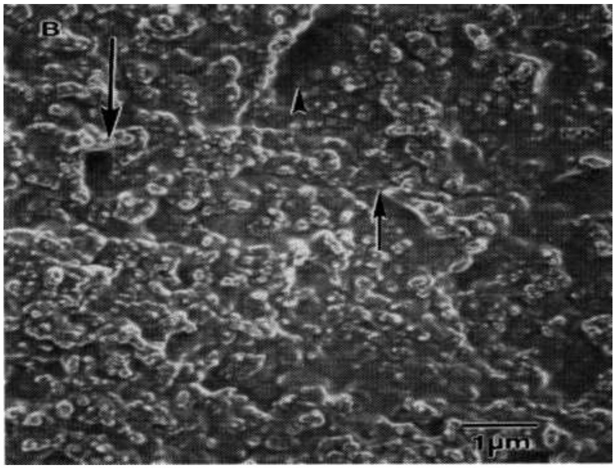


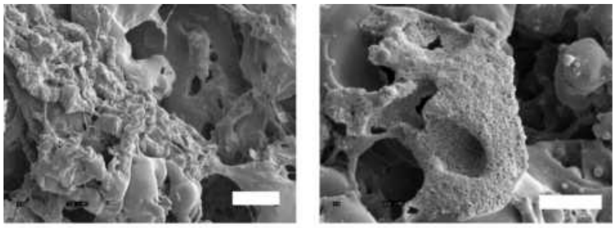
Figure 5. Cast replica of vascular tissue demonstrating nanometer roughness. (Adapted from Ref. 93.) Vascular tissue has numerous irregular nanometer features that when duplicated in porous biomaterials show progress in vascular applications.

In addition, other porous biomaterials with nanostructured surface features [e.g., carbon nanotubes in polymer composites (Fig. 7) and porous helical rosette carbon nanotubes (Fig. 8)], increased osteoblast functions over conventionally used PLGA scaffolds (99,100). Interestingly, as opposed to conventional porous biomaterials, helical rosette carbon nanotubes self-assemble into a porous biomaterial, which when heated to temperatures only slightly above body temperature solidify (100); thus, these materials could be formulated immediately before implantation to match the dimensions of any bony defect. These novel porous helical rosette nanotubes also allow for optimal pore interconnectivity for the transfer of nutrients and waste to and from cells (100).

Table 2. Nanometer Dimensions of Extracellular Matrix Proteins[a]

Protein	Characteristic Dimensions
Fibronectin	Dimmer of two identical subunits; 60–70 nm long; 2–3 nm wide
Vitronectin	Linear molecule 15 nm long
Laminin	Cruciform configuration with one 50 nm long arm and two 35 nm short arms; total length 50 nm; total width 70 nm
Collagen	Triple helical linear protein consisting of 2 $\alpha(1)$-chains and one $\alpha(2)$; 300 nm long; 0.5 nm wide; 67 nm periodicity

[a] See Ref. 94.



PLGA: Conventional TiO_2 70:30 wt. %

PLGA: Nanophase TiO_2 70:30 wt. %

Figure 6. The scanning electron microscopy (SEM) images of PLGA composites containing either conventional or nanophase titania. Increased bone regeneration has been measured in polymer composites containing nanometer compared to conventional ceramics. Bar = 10 μm.

Cartilage Applications

Such pore interconnectivity is also crucial for cartilage forming cells, chondrocytes, since chondrocytes reside far apart from each other and their main communication is through their extracellular matrix. Recently, a porous biomaterial matrix fabricated via solvent casting and particulate leaching to create nanometer surface roughness was tested for cartilage applications (42). The polymer used was PLGA and it was modified to possess nanometer surface features through soaking for 10 min in 10 *N* NaOH (Fig. 9). Compared to conventional PLGA, results showed increased chondrocyte adhesion, proliferation, and synthesis of a cartilage extracellular matrix (as noted by collagen and glycosaminoglycan synthesis) (42).

Vascular and Bladder Applications

Not only do osteoblasts and chondrocytes interact better with nanophase materials, but so do other cells such as vascular (including endothelial and vascular smooth muscle cells) and bladder cells. For example, Miller et al. (43) and Thapa et al. (44) created nanometer surface features on PLGA films by developing novel molds of NaOH treated PLGA (Fig. 10). When compared to PLGA films without nanometer surface features, vascular smooth muscle cell, endothelial cell, and bladder smooth muscle cell functions were enhanced on the nanostructured PLGA. For bladder applications, Pattison et al. (76) created NaOH induced nanofeatures onto 3D PLGA scaffolds and also observed greater bladder smooth muscle cell adhesion, proliferation, and collagen synthesis. Their studies have further demonstrated increased fibronectin and vitronectin adsorption on nanostructured PLGA compared to conventional PLGA, thus, providing a key mechanism for why vascular and bladder cell adhesion is enhanced on nanostructured PLGA surfaces (43). In addition, PCL and polyurethane have been modified to possess nanostructured surface features by NaOH and HNO_3 treatments, respectively; increased vascular and bladder cell functions have been measured on these treated compared to nontreated polymers (43,44). Such studies highlight the versatility of modifying numerous polymers to possess nanostructured features for enhanced vascular and bladder applications.

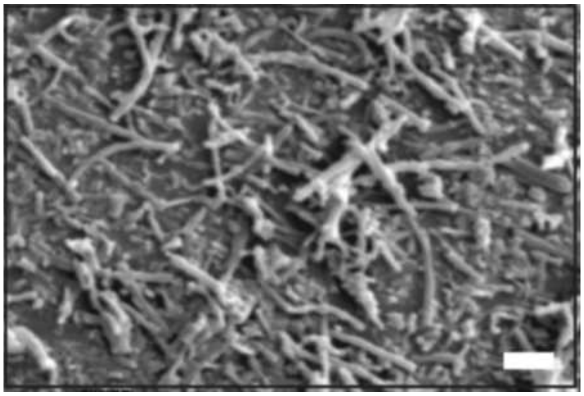


Figure 7. The SEM image of a polyether urethane composite containing nanophase carbon fibers. Increased bone regeneration has been measured in polymer composites containing nanometer compared to conventional carbon fibers. Bar = 1 μm.

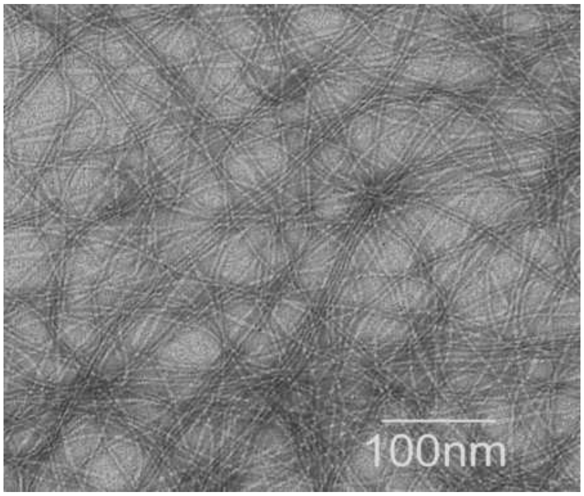
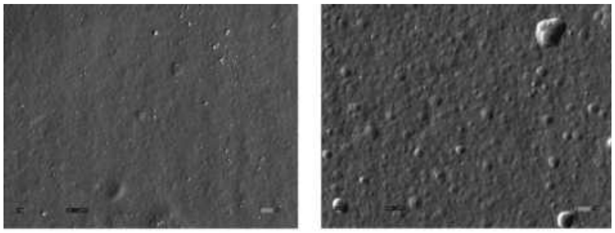


Figure 8. The transmission electron microscopy (TEM) micrograph of porous helical rosette carbon nanotubes. Individual outer-tube diameters are 4.6 ± 0.09 nm. Increased bone regeneration has been measured in helical rosette nanotubes compared to currently used titanium implants.



Conventional PLGA

Nano structured PLGA

Figure 9. The SEM images of PLGA possessing conventional and nanoscale surface roughness. Nanoscale surface roughness was created by fabricating molds of PLGA etched in 10 *N* NaOH for 10 min. Increased functions of chondrocytes, vascular cells, and bladder cells have been measured on polymers with nanoscale compared to conventional surface features. Bar = 1 μm.

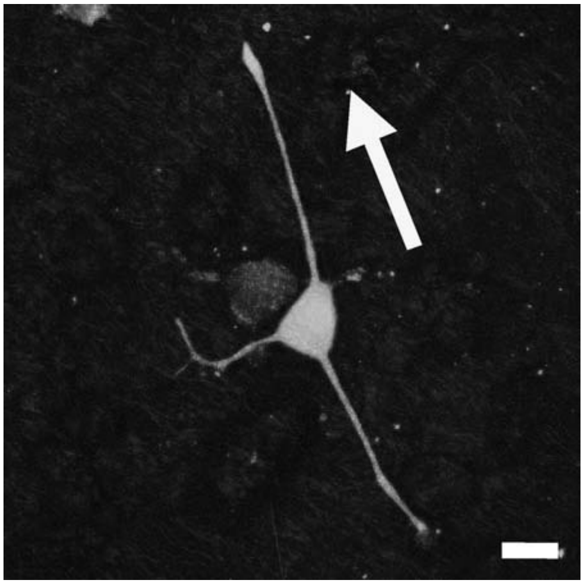


Figure 10. Fluorescent microscopy image of neuron axon alignment corresponding with aligned carbon nanofibers in polyether urethane composites. The arrow indicates carbon nanofiber alignment. Bar = 20 μm.

Central and Peripheral Nervous System Applications

Finally, McKenzie et al. (101) also provided evidence that biomaterials created to have numerous nanometer features can decrease glial scar tissue formation while at the same time increase interactions with neurons. These materials were created by combining carbon nanofibers to polyether urethane. In addition, these investigators have aligned carbon nanofibers in such porous structures to control the direction of axon extension from neurons. In this manner, such porous nanostructured biomaterials could be used to regenerate electrical activity in damaged areas of the brain.

CONCLUSIONS

All these indications point to the conclusion that a successful implantable porous biomaterial should possess properties and structures that simulate the formation of an extracellular matrix similar to that of the target organ it intends to replace. Equally as important for porous biomaterials are appropriate mechanical strength, mechanical structural integrity, degradation rate, permeability, porosity, pore structure, pore interconnectivity, surface energy, surface roughness, and surface chemistry. These all play a role in the function of an optimal porous biomaterial to regenerate tissue. Importantly, to date, to address some of these material properties, several processing techniques have been developed. Although much more needs to be learned concerning the most important aspects of tissue regeneration on porous biomaterials, proper attachment of the appropriate cell is crucial. This is mediated by initial protein interactions that must be the focus of future endeavors to design more effective porous biomaterials. Recent evidence has been provided that porous biomaterials with nanostructured surface roughness might just control initial protein interactions pertinent for enhancing cell functions necessary to improve the efficacy of orthopedic, cartilage, vascular, central and peripheral nervous system, and bladder applications.

BIBLIOGRAPHY

1. Davis ME. Ordered porous materials for emerging applications [Review]. Nature (London) 2002;417(6891):813–821.
2. Niklason LE, Langer R. Advances in tissue engineering of blood vessels and other tissues. Transplant Immunol 1997;5:303–306.
3. Praemer A, Furner S, Rice SD. Musculoskeletal Conditions in the United States, Park Ridge, IL: American Academy of Orthopaedic Surgery; 1992.
4. D' Angelo K, Austin T. The moving target: Understanding why arthritis patients do not consider total joint replacement. American Academy of Orthopaedic Surgeons: News Release July 6, 2004.
5. Arthritis Today. 52 Ways to bite back–5 astounding figures. The Arthritis Foundation. Available at http://www.arthritis.org/resources/arthritistoday/2003_archives/2003_09_10_51_ways_5figures.asp [2005, March 31].
6. American Parkinson's Disease Association. Available at http://www.apdaparkinson.org/user/AboutParkinson.asp. [2005, March 31]. Nakamura H, Ozawa H. Immunohistochemical localization of peharan sulfate proteoglycan in rat tibiae. J Bone Mineral Res 1994;9:1289–1299.
7. National Center for Chronic Disease Prevention. Available at www.cdc.gov/diabetes/pubs/costs/figure.htm figure1, June 1998.
8. National Institute of Health. Spinal cord injury: emerging concepts. NIH Proc Sept 30–Oct. 1, 1996.
9. Greenlee RT, Murray T, Bolden S, Wingo PA. Cancer statistics, 2000. CA Cancer J Clin 2000;50:7–33.
10. Melekos MD, Moutzouris GD. Intravesical therapy of superficial bladder cancer. Curr Pharm Des 2000;6:345–359.
11. Highly MS, Oosterom AT, Maes RA, De Bruijn EA. Intravesical drug delivery. Pharmacokinetic and clinical considerations. Clin Pharmacokinet 1999;37:59–73.
12. Dalbagni G, Herr HW. Current use and questions concerning intravesical bladder cancer group for superficial bladder cancer. Urol Clin North Am 2000;27:137–146.
13. Bianco FJ Jr, et al. Matrix metalloproteinase-9 expression in bladder washes from bladder cancer patients predicts pathological stage and grade. Clin Cancer Res 1998;4:3011–3016.
14. Lebret T, et al. Recurrence, progression and success in stage Ta grade 3 bladder tumors treated with low dose bacillus Calmette-Guerin instillations. J Urology 2000;163:63–67.
15. Hutmacher DW. Scaffold design and fabrication technologies for engineering tissues–state of the art and future perspectives [Review]. J Biomater Sci Polym Ed 2001;12(1):107–124.
16. Sander EA, et al. Solvent effects on the microstructure and properties of 75/25 poly(D,L-lactide-co-glycolide) tissue scaffolds. J Biomed Mater Res Part A 2004;70A(3):506–513.
17. Holy CE, Fialkov JA, Davies JE, Shoichet MS. Use of a biomimetic strategy to engineer bone. J Biomed Mater Res 2003;65A:447–453.
18. Peters MC, Mooney DJ. Synthetic extracellular matrices for cell transplantation. Mater Sci Forum 1997;250:43–52.
19. Gomes ME, et al. Effect of flow perfusion on the osteogenic differentiation of bone marrow stromal cells cultured on starch-based three dimensional scaffolds. J Biomed Mater Res 2003;67A:87–95.
20. Kasemo B, Gold J. Implant surfaces and interface processes [Review]. Adv Dental Res 1999;13:8–20.

21. Vogler EA. Structure and reactivity of water at biomaterial surfaces. Adv Colloid Interface Sci 1998;74:69–117.
22. Webster TJ, Hellenmeyer EL, Price RL. Increased osteoblast functions on theta+delta nanofiber alumina. Biomaterials 2005;26(9):953–960.
23. Horbett TA. Proteins: structure, properties and adsorption to surfaces. In: Ratner BD, Hoffman AS, Schoen AS, Lemmons JE, editors. Biomaterials Science: An Introduction to Materials in Medicine. New York: Academic Press; 1996. p 133.
24. Horbett TA. Chapter 13 Principles underlying the role of adsorbed plasma proteins in blood interactions with foreign materials. Cardiovas Pathol 1993;2(137S):137–148.
25. Hlady V, Buijs J. Protein adsorption on solid surfaces. Curr Opin Biotechnol 1996;7:72–77.
26. Norde W. Driving forces for protein adsorption at solid surfaces. Macromol Symp 1996;103:5–18.
27. Webster TJ, et al. Enhanced functions of osteoblasts on nanophase ceramics. J Biomed Mater Res 2000;51:475.
28. Horbett TA. Techniques for protein adsorption studies. In: Williams DF, editor. Techniques of Biocompatibility Testing, Boca Raton, FL: CRC Press; 1986. p 183.
29. Price RL, Haberstroh KM, Webster TJ. Improved osteoblast viability in the presence of smaller nanometer dimensions carbon fibres. Nanotechnology 2005;15(8):892–900.
30. Kramer RH, Enenstein J, Waleh NS. Integrin structure and ligand specificity in cell matrix interactions. In: Rohrbach DJ, Timpl R, editors. Molecular and Cellular Aspects of Basement Membranes, New York: Academic Press; 1993. p 239–258.
31. Hynes RO. Integrins: versatility, modulation, and signaling in cell adhesion. Cell 1992;69:11–25.
32. Schwartz MA. Transmembrane signaling by integrins. Trends Cell Biol 1992;2:304–308.
33. Puleo DA, Bizios R. Mechanisms of fibronectin-mediated attachment of osteoblasts to substrates *In vitro*. Bone Mineral 1992;18:215–226.
34. Dalton BA, et al. Polymer surface chemistry and bone cell migration. J Biomater Sci Polym Ed 1998;9(8):781–799.
35. Schakenraad JM. Cell: their surfaces and interactions with materials. In: Ratner BD, Hoffman AS, Schoen AS, Lemmons JE, editors. Biomaterials Science: An Introduction to Materials in Medicine, New York: Academic Press; 1996. p 141.
36. Webster TJ. Nanophase ceramics: the future orthopedic and dental implant material. In: Ying JY, editor. Advances in Chemical Engineering, Vol. 27, New York: Academic Press; 2001. p 125.
37. Sinha RK, Tuan RS. Regulation of human osteoblast integrin expression by orthopedic implant materials. Bone 1996; 18(5):451–457.
38. Davies JE. The importance and measurement of surface charge species in cell behavior at the biomaterial interface. In: Ratner BD, editor. Surface Characterization of Biomaterials: Progress in Biomedical Engineering, Vol. 6, New York: Elsevier; 1988. p 219.
39. Brunette PM. The effect of surface topography of cell migration and adhesion. In: Ratner BD, editor. Surface Characterization of Biomaterials: Progress in Biomedical Engineering, Vol. 6, New York: Elsevier; 1988.p 203.
40. Luck M, et al. Analysis of plasma protein adsorption on polymeric nanoparticles with different surface characteristics. J Biomed Mater Res 1998;39:478–485.
41. Degasne I, et al. Effects of roughness, fibronectin and vitronectin on attachment, spreading, and proliferation of human osteoblast-like cells (Saos-2) on titanium surfaces. Calcif Tissue Int 1999;64(6):499–507.
42. Park GE, Pattison MA, Park K, Webster TJ. Accelerated chondrocyte functions on NaOH-treated PLGA scaffolds. Biomaterials 2005;26(16):3075–3082.
43. Miller DC, Haberstroh KM, Webster TJ. Mechanism(s) of increased vascular cell adhesion on nanostructured poly(lactic-co-glycolic acid) films, J Biomed Mat Res 2005;73(4):476–484.
44. Thapa A, Miller DC, Webster TJ, Haberstroh KM. Nanostructured polymers enhance bladder smooth muscle cell function. Biomaterials 2003;24(17):2915–2926.
45. Agrawal CM, Ray RB. Biodegradable polymer scaffolds for musculoskeletal tissue engineering. J Biomed Mater Res 2001;55:141–150.
46. Carver SE, Heath CA. Influence of intermittent pressure, fluid flow, and mixing on the regenerative properties of articular chondrocytes. Biotechnol Bioeng 1999;65:274–281.
47. Yaszemski MJ, et al. Evolution of bone transplantation: molecular, cellular, and tissue strategies to engineer scaffold human bone. Biomaterials 1995;17:175–185.
48. Hutmacher DW. Scaffolds in tissue engineering bone and cartilage. Biomaterials 2000;21:2925–2943.
49. Goulet RW, et al. The relationship between the structural and orthogonal compressive properties of trabecular bone. J Biomech 1994;27:375–389.
50. Kwon IK, Kidoaki S, Matsuda T. Electrospun nano- to microfiber fabrics made of biodegradable copolyesters: structural characteristics, mechanical properties and cell adhesion potential. Biomaterials 2005;26(18):3929–3939.
51. Maroudas NG. Chemical and mechanical requirements for fibroblast adhesion, Nature (London) 1973;244:363–364.
52. Pelham RJ, Wang YL. Cell locomotion and focal adhesions are regulated by substrate flexibility. Proc Natl Acad Sci USA 1997;94:13661–13665.
53. Katz BZ, et al. Physical state of the extracellular matrix regulates the structure and molecular composition of cellmatrix adhesions. Mol Biol Cell 2000;11:1047–1060.
54. Ohya S, Kidoaki S, Matsuda T. Poly(*N*-isopropylacrylamide) (PNIPAM)-grafted gelatin hydrogel surfaces: interrelationship between microscopic structure and mechanical property of surface regions and cell adhesiveness. Biomaterials 2005;26:3105–3111.
55. Bignon A, et al. Effect of micro- and macroporosity of bone substitutes on their mechanical properties and cellular response. J Mater Sci Mat Med 2003;14:1089–1897.
56. Wilkinson CDW, et al. The use of materials patterned on a nano- and micro-metric scale in cellular engineering. Mater Sci Eng 2001;19:263.
57. Clark P, et al. Topographical control of cell behaviour. II. Multiple grooved substrata. Development 1999;108:635.
58. Tranquillo RT. Self-organisation of tissue equivalents: the nature and role of contact guidance. Biochem Soc Symp 1999;65: 27.
59. Li SM, Garreau H, Vert M. Structure-property relationships in the case of the degradation of massive poly(hydroxyl acid) in aqueous media,part 2:degradation of lactide–glycolide copolymers. J Mater Sci Mater Med 1990;1:131–139.
60. Grizzi I, Garreau H, Li S, Vert M. Hydrolytic degradation of devices based on poly(d,l-lactic acid): size dependence. Biomaterials 1995;16:305–311.
61. Sung HJ, Meredith C, Johnson C, Galis ZS. The effect of scaffold degradation rate on three-dimensional cell growth and angiogenesis. Biomaterials 2004;25:5735–5742.
62. Hedberg EL, et al. *In vivo* degradation of porous poly(propylene fumarate)/poly(DL-lactic-co-glycolic acid) composite scaffolds. Biomaterials 2005;26:4616–4623.
63. Perugini P, et al. PLGA microspheres for oral osteopenia treatment: preliminary "*In vitro*"/"*In vivo*" evaluation. Int J Pharm 2003;256:153–160.
64. Wu L, Ding J. *In vitro* degradation of three-dimensional porous poly(D,L-lactide-co-glycolide) scaffolds for tissue engineering, Biomaterials 2004;25:5821–5830.

65. Zhang R, Ma PX. Processing of polymer scaffolds: Phase separation. In: Atala A, Lanza R, editors. Methods of Tissue Engineering, San Diego, CA: Academic Press; 2001. p 715–724.
66. Ishaug SL, et al. Bone formation by three-dimensional stromal osteoblast culture in biodegradable polymer scaffolds. J Biomed Mater Res 1997;36:17–28.
67. Ishaug-Riley SL, et al. Ectopic bone formation by marrow stromal osteoblast transplantation using poly(DL-lactic-co-glycolic acid) foams implanted into the rat mesentery. J Biomed Mater Res 1997;36:1–8.
68. Sherwood JK, et al. A three-dimensional osteochondral composite scaffold for articular cartilage repair. Biomaterials 2002;23:4739–4751.
69. Athanasiou KA, Schmitz JP, Agrawal CM. The effects of porosity on in vitro degradation of polylactic acid-polyglycolic acid implants used in repair of articular cartilage. Tissue Eng 1998;4:53–63.
70. Agrawal CM, McKinney JS, Lanctot D, Athanasiou A. Effects of fluid flow on the in vitro degradation kinetics of biodegradable scaffolds for tissue engineering. Biomaterials 2000;21:2443–2452.
71. Lightfoot EN. Transport phenomena and living systems. New York: John Wiley & Sons, Inc.; 1974.
72. Colton CK. Implantable biohybrid artificial organs. Cell Transplant 1995;4:415–436.
73. Hulbert SF, et al. Potential of ceramic materials as permanently implantable skeletal prostheses. J Biomed Mater Res 1970;4(3):433–456.
74. Yuan H, et al. A preliminary study on osteoinduction of two kinds of calcium phosphate ceramics. Biomaterials 1999;20(19):1799–1806.
75. Turner S, et al. Cell attachment on silicon nanostructures. J Vas Sci Technol B 1997;15:2848–2854.
76. Pattison MA, Wurster S, Webster TJ, Haberstroh KM. Three-dimensional, nano-structured PLGA scaffolds for bladder tissue replacement applications. Biomaterials 2005;26(15): 2491–2500.
77. Gibson LJ, Ashby MF. Cellular Solids: Structure and Properties. 2nd ed. Cambridge University Press; 1997.
78. Ma PX, Langer R. Fabrication of Biodegradable Polymer foams for cell transplantation and tissue engineering. In: Yarmush M, Morgen J, editors. Tissue Engineering Methods and Protocols, Totowa, NJ: Humana Press; 1998. p 47–56.
79. Ma Z, Gao C, Gong Y, Shen J. Paraffin spheres as porogen to fabricate poly(L-lactic acid) scaffolds with improved cytocompatibility for cartilage tissue engineering. J Biomed Mater Res Part B 2003;67(1):610–617. Mikos AG, et al. Preparation and characterization of Poly(L-lactic acid) foams. *Polymer* 1994;35:1068–1077.
80. Liu X, Ma PX. Polymeric scaffolds for bone tissue engineering [Review]. Ann Biomed Eng 2004;32(3):477–486.
81. Harris LD, Kim BS, Mooney DJ. Open pore biodegradable matrices formed with gas foaming. J Biomed Mater Res 1998;42:396–402.
82. Mooney DJ, et al. Novel approach to fabricate porous sponges of poly(D,L-lactic-co-glycolic acid) without the use of organic solvents. Biomaterials 1996;17:1417–1422.
83. Nam YS, Park TG. Porous biodegradable polymeric scaffolds prepared by thermally induced phase separation. J Biomed Mater Res 1999;47:8–17.
84. Whang K, Thomas CH, Healy KE, Nuber G. A novel method to fabricate bioabsorbable scaffolds. Polymer 1995;36(4):837–842.
85. Ho MH, et al. Preparation of porous scaffolds by using freeze-extraction and freeze-gelation methods. Biomaterials 2004;25:129–138.
86. Matthews JA, Wnek GE, Simpson DG, Bowlin GL. Electrospinning of collagen nanofibers. Biomacromolecules 2002;3: 232–238.
87. Reneker DH, Chun I. Nanometre diameter fibres of polymer, produced by electrospinning. Nanotechnology 1996;7:216–223.
88. Giordano RA, et al. Mechanical properties of dense polylactic acid structures fabricated by three dimensional printing. J Biomater Sci-Polym Ed 1996;8:63–75.
89. Ma J, Wang C, Peng KW. Electrophoretic deposition of porous hydroxyapatite scaffold. Biomaterials 2003;24(20):3505–3510.
90. Chang, et al. 2005.
91. Popat KC, et al. Influence of nanoporous alumina membranes on long-term osteoblast response. Biomaterials 2005;26(22): 4516–4522.
92. Shimko DA, et al. Effect of porosity on the fluid flow characteristics and mechanical properties of tantalum scaffolds. J Biomed Mater Res Part B Appl Biomater 2005;73(2):315–324.
93. Goodman SL, Sims PA, Albrecht RM. Related Articles, Links Three-dimensional extracellular matrix textured biomaterials. Biomaterials. 1996;17(21):2087-2095.
94. Ayad S, et al. The extracellular matrix factsbook, San Diego, CA: Academic Press; 1994.
95. Baraton MI, Chen X, Gonsalves KE. FTIR study of a nanostructured aluminum nitride powder surface: Determination of the acidic/basic sites by CO, CO_2 and acetic acid adsorptions. Nanostruct Mater 1997;8(4):435–445.
96. Klabunde KJ, et al. Nanocrystals as stoichiometric reagents with unique surface chemistry. J Phys Chem 1996; 100:12142–12153.
97. Gutwein LG, Webster TJ. Increased viable osteoblast density in the presence of nanophase compared to conventional alumina and titania particles. Biomaterials 2004;25(18):4175–4183.
98. Webster TJ, Ejiofor JU. Increased osteoblast adhesion on nanophase metals: Ti, Ti6Al4V, and CoCrMo. Biomaterials 2004;25(19):4731–4739.
99. Price RL, et al. Osteoblast function on nanophase alumina materials: Influence of chemistry, phase, and topography, J Biomed Mater Res 2004;67(4):1284–1293.
100. Chun AI, Moralez JG, Fenniri H, Webster TJ. Helical rosette nanotubes: a more effective orthopaedic implant material. Nanotechnology 2004;15(4):S234–S239.
101. McKenzie JL, Waid MC, Shi R, Webster TJ. Decreased functions of astrocytes on carbon nanofiber materials. Biomaterials 2004;25(7–8):1309–1317.

See also BIOMATERIALS: TISSUE ENGINEERING AND SCAFFOLDS; ORTHOPEDICS, PROSTHESIS FIXATION FOR; VASCULAR GRAFT PROSTHESIS.

POSITRON EMISSION TOMOGRAPHY

GEORGE KONTAXAKIS
Universidad Politécnica de
Madrid
Madrid, Spain

INTRODUCTION: FROM MEDICAL TO MOLECULAR IMAGING

Medical imaging conventionally refers to the non invasive or minimally invasive techniques employed to view internal organs of the body, typically for diagnosing disease. In a broader sense, it refers to a field that enables acquisition, processing, analysis, transmission, storage, display, and archiving of images of internal body parts for interpretation and patient management (diagnosis, disease staging and evaluation, treatment planning and follow-up). Medical imaging was practically born with the discovery of the X rays by W. C. Roentgen in 1895 and has since based its success on observation and the accumulated experience of the examining physician.

Molecular imaging is a natural out grown of the medical imaging field. Recent advances in molecular biology have resulted in an improved understanding of many disease and natural processes. Consequently, molecular imaging links the empirical diagnostics and experimentally tried treatment management protocols with the fundamental understanding of the underlying processes that generate the observed results. As discoveries of the molecular basis of disease unfold, one top research priority is the development of imaging techniques to assess the molecular basis of cell dysfunction and of novel molecular therapy. Molecular imaging techniques are ideally based on technologies that have an intrinsically high resolution (spatial and temporal) and allow the detection of low concentrations of target biomolecules involved, such as nuclear medicine imaging (Positron Emission Tomography, PET; Single-Photon Emission Tomography, SPET), magnetic resonance imaging (MRI) and spectroscopy (MRS), optical tomography, autoradiography, or acoustical imaging.

The examination of biochemical processes with an imaging technology is of vital importance for modern medicine. As, in most cases, the location and extent of a disease is unknown, the first objective is an efficient means of searching throughout the body to determine its location. Imaging is an extremely efficient process for accomplishing this aim, because data are presented in pictorial form to the most efficient human sensory system for search, identification, and interpretation: the visual system. Recognition depends on the type of information in the image, both in terms of interpreting what it means and how sensitive it is to identifying the presence of disease.

PET stands in the forefront of molecular imaging and allows the quantitative evaluation of the distribution of several pharmaceuticals in a target area *in vivo*. PET is a unique diagnostic imaging technique for measuring the metabolic activity of cells in the human body. It produces images of the body's basic biochemistry and biological activity in a noninvasive way, combining techniques applied in nuclear medicine with the precise localization achieved by computerized image reconstruction. PET is therefore a powerful diagnostic test that is having a major impact on the diagnosis and treatment of disease, as well as on patient management.

PET images can demonstrate pathological changes and detect and stage tumors long before they would be revealed with other conventional imaging modalities. Traditional diagnostic techniques, such as X rays, computerized tomography (CT) scans, or MRI, produce anatomical images of what the internal organs look like. The premise with these techniques is that a visible structural change exists in anatomy caused by disease. However, biochemical processes are also altered with disease and may occur before a change in gross anatomy occurs. Furthermore, PET can provide medical doctors with important early information about very subtle changes of function in the brain and heart, due to disease-related modifications in tissue perfusion, cell metabolic rates heart disease, or neurological disorders (Alzheimer's, Parkinson's, epilepsy, dementia, etc.), allowing physicians to diagnose and treat these diseases earlier and, consequently, more efficiently and accurately, according to the axiom "the earlier the diagnosis, the better chance for treatment." PET can also help physicians monitor a patient's response to treatment, as well as identify distant metastases that can affect treatment, helping curtail ineffective treatments and reduce unnecessary invasive procedures. The field of PET has been emerging today into clinical diagnostic medicine and is approved by many insurance carriers for coverage.

HISTORY OF PET

The positron emission and detection of the radiation produced was a known technique that dates back to the early days of the twentieth century. However, it is only in the last few decades, with the booming development of fast electronic circuits and powerful computer systems, that this knowledge could be used in practice as a valuable diagnostic tool: The electronic circuits used in PET should be able to detect the coincidental arrival of two high energy photons (a timing resolution of the order of few nanoseconds), and the image reconstruction requires modern computer systems in order to produce an accurate image of the activity distribution within a clinically reasonable time.

In the beginning of the 1950s, researchers at the Massachusetts General Hospital (MGH) in Boston and the Duke University in Durham proposed the idea that, in spite of the short half-lives of the, by that time recently discovered, positron-emitting radionuclides, they offered an attractive method for the regional study of metabolism due to their commonality. A single-detector pair brain probe was then developed at MGH and used in experiments. However, it was not until the early 1960s that these positron-emitting radionuclides began to gain popularity, when a number of centers such as the MGH in Boston, the Sloan Kettering Institute in New York, Ohio State University, and the University of California at Berkeley began to use cyclotrons. At the same time, the first image reconstruction techniques were proposed by researchers at MGH, and, in the early 1970s, the concept of computerized tomography (CT) was presented by Hounsfield, who later was awarded with the Nobel Prize.

In the early 1970s, the first PET scanners were developed at the MGH, the Brookhaven National Laboratory, the Washington University, and the Montreal Neurological Institute in Canada, used then as research tools. At the same time, a private company (EG&G OTREC, Oak Ridge, TN, USA) got involved in the developments of the first ring PET scanners, joined in the market a couple of years later by TCC (The Cyclotron Corporation, Berkeley, CA, USA), and in 1976 the first commercial PET scanner was delivered at the University of California, Los Angeles (UCLA). A year later, Scanditronix from Sweden brought Europe into PET. The first PET scanners used single slices when performing tomographs, with transaxial resolution greater than 2 cm full-width half-maximum (FWHM) and used NaI crystal material. Such systems were installed at several research institutions, apart from the ones mentioned above, like the University of California at Berkeley, the Lawrence Berkeley Laboratory, and the University of Pennsylvania.

By the end of 1970s, PET had shown its potential for application to clinical medicine. The following generation of PET scanners reduced detector size and added additional rings to allow for simultaneous acquisition of multiple slices. The slice resolutions improved from greater than 2 cm FWHM to less than 1 cm FWHM. As time progressed, more detectors and photomultiplier tubes (PMTs) were added to these machines to increase their sensitivity and resolution. In the mid-1980s, the first BGO pixelated detector blocks were presented. At the same time, the first dedicated medical PET cyclotron units with automated radiopharmaceutical delivery systems were commercially available.

At the end of 1980s, the major medical imaging companies (mainly Siemens with CTI PET, Inc., and General Electric with Scanditronix) began investing in PET. The first whole-body PET scanners have been presented and research in new detector materials led to significant discoveries (LSO, etc.) in the beginning of the 1990s. Since then, PET has shown a steady increase in acceptance for clinical application, both medically and administratively, and PET centers are being installed worldwide at an increasing pace. PET is now a well-established medical imaging technique that assists in the diagnosis and management of many diseases.

More details on the history of PET instrumentation and the related developments can be found in References 1 and 2.

PHYSICAL PRINCIPLES OF PET

PET images molecules of substances with a specific biological activity. In order to monitor their distribution, kinetic characteristics, and behavior of (pharmaceuticals) within the body, these substances are tagged with radioactive compounds (with short half-life and at extremely low concentrations) (3). These radiopharmaceuticals are chosen to have a desired biological activity, depending on the metabolic activity of the organ under study, and are introduced to the subject by injection or inhalation.

The most commonly used radionuclides are listed in Table 1 and are compounds that constitute, or are consumed by, the living body, like carbon, nitrogen, and oxygen. They are isotopes of biologically significant chemical elements that exist in all living tissues of the body and in almost all nutrients. Therefore, the above radionuclides are easily incorporated in the metabolic process and serve as tracers of the metabolic behavior of the body part, which can be studied *in vivo*.

Table 1. The Most Commonly Used Radionuclides in PET

Radionuclide	Half-life
Carbon-11	20.3 min
Nitrogen-13	9.97 min
Oxygen-15	2.03 min
Fluorine-18	1.83 h
Gallium-68	1.83 h
Rubidium-82	1.26 min

Table 2. Major PET Radiopharmaceuticals and their Specific Medical Applications

Agent	Images
F-18 fluorodeoxyglucose	Regional glucose metabolism
F-18 sodium fluoride	Bone tumors
C-11 methionine	Amino acid uptake/protein synthesis
C-11 choline	Cell membrane proliferation
C-11 deoxyglucose	Regional brain metabolism
O-15 oxygen	Metabolic rate of oxygen use/OEF
C-11 carbon monoxide	Cerebral blood volume
O-15 carbon monoxide	Cerebral blood volume
O-15 water	Cerebral blood flow
O-15 carbon dioxide (Inhaled)	Cerebral blood flow
C-11 butanol	Cerebral blood flow
C-11 N-methylspiperone	Dopamine D2 and Serotonin S2 receptors
F-18 N-methylspiperone	D2 and S2 receptors
C-11 raclopride	D2 receptors
F-18 spiperone	D2 receptors
Br-76 bromospiperone	D2 receptors
C-11 carfentanil	Opiate mu receptors
C-11 flumazenil	Benzodiazepine (GABA) receptors

Table 2 shows a list of the major radiopharmaceuticals used as PET agents with their specific medical applications. The most common radiopharmaceutical used in PET studies today is fluorodeoxyglucose (FDG) (4), a chemical compound similar to glucose, with the difference that one of the -OH groups has been replaced by F-18. Carbon-11 can also be used as a radiotracer to glucose. The short half-lives of these particles allow the subject and the people handling them to receive only a low radiation dose.

The identification and detection of the presence of the molecules of the radiotracer in a specified location within the source (i.e., the body under study) is performed by a chain of events, based on physical principles and data processing techniques, which are schematically depicted in Fig. 1 and briefly described below.

A positron is emitted during the radioactive decay process, annihilates with an electron, and, as a result, a pair of γ rays is emitted (two high energy photons of 511 keV each). The two γ rays fly off in almost opposite directions (according to the momentum conservation laws), penetrate the surrounding tissues, and can be recorded outside the subject's body by scintillation detectors placed on a circular or polygonal detector arrangement, which forms a PET tomograph. When the γ ray hits a scintillation detector material, it then deposits its energy in that crystal by undergoing photoelectric effect, which is an atomic absorption process where an atom totally absorbs the energy of an incident photon (5). This energy is then used to eject an orbital electron (photoelectron) from the atom and is,

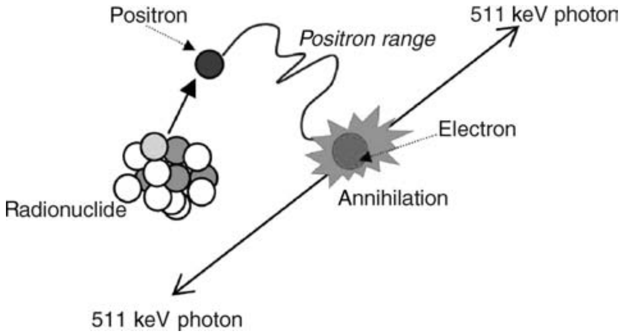


Figure 1. This schematic depicts the chain of events that described the physical properties of high energy gamma pair emission from positron-emitting radioisotopes. All radioisotopes used with PET decay by positron emission. Positrons are positively charged electrons. Positron emission stabilizes the nucleus of unstable radioisotopes by removing a positive charge through the conversion of a proton into a neutron. An emitted positron travels a short distance (*positron range*, which depends on the energy of the positron) and collides with an ordinary electron of a nearby atom in an annihilation reaction. When the two particles annihilate, their mass turns into two 511 keV gamma rays that are emitted at 180° to each other. When detected, the 180° emission of two gamma rays following the disintegration of positronium is called a coincidence line. Coincidence lines provide a unique detection scheme for forming tomographic images with PET.

therefore, transformed in visible light. This light can be detected by specialized devices (photomultiplier tubes, PMT) that capture and transform it into an electronic signal, shaped at a later stage by the electronic circuits of the tomograph to an electronic pulse, which provides information about the timing of the arrival of the incident γ ray and its energy. Figure 2 summarizes the principles of gamma ray event detection in PET described here.

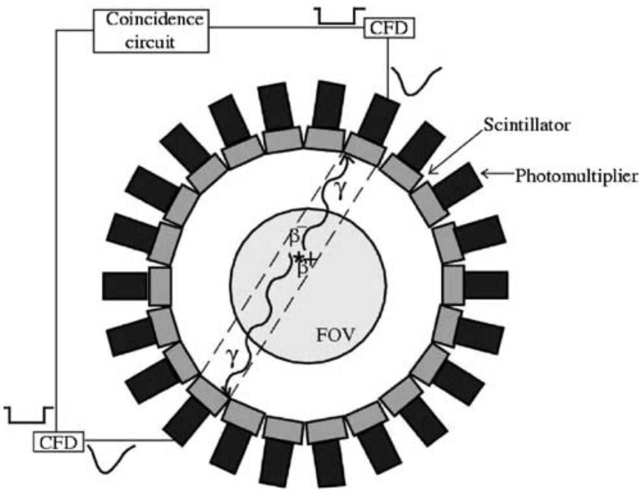


Figure 2. Scintillation detectors coupled to photomultiplier tubes are placed around the detector ring of the scanner. An annihilation event (*) inside the field of view (FOV) produces two γ rays that get detected by a pair of detectors. The event is identified to occur inside a specific detector tube (dashed stripe) by the electronic devices (constant fraction discriminators, CFD, and the coincidence detection circuit) that connect every pair of detectors.

By measuring a coincidence photon, the detector array in a PET system identifies that an annihilation event occurred inside the volume defined between the surfaces of the pair of detectors that registered the coincidence event. At the end of a PET scan, for each pair of detectors, a number of coincidence events that have been identified exist. This information represents the radioactivity in the subject viewed at different angles, when sorted in closely spaced parallel lines. In order to reconstruct the activity density inside the source from its projections (events registered at each detector pair), a mathematical reconstruction algorithm is applied by computer. The collected data are corrected for scatter, attenuation, and accidental coincidences; normalized for the differences in detector efficiencies, and reconstruct the spatial distribution of the radioactivity density inside the organ or the system under study in the form of a 2D or 3D image. The result is a digital image of the source, where the value of each picture element (pixel) or, in modern 3D tomograph systems, volume element (voxel) is proportional to the activity density inside the source at the area (or volume) that corresponds to this pixel/voxel. This image can be directly displayed on a screen. Further analysis of the data and processing of the produced images can be carried out with the use of a computing system.

A high energy photon produced by an annihilation event can deviate from its original trajectory if it gets involved in Compton scattering inside the subject's body, a collision between a photon and a loosely bound outer-shell orbital electron of an atom. In this case, because the incident photon energy greatly exceeds the binding energy of the electron to the atom, the interaction can be considered as a collision between the photon and a "free" electron. The photon does not disappear in Compton scattering, but it is deflected through a scattering angle θ and some of its energy is transferred to the electron (recoil electron) (5). In the case this ray gets detected in coincidence with the second gamma produced at the same event, then this event will be counted to have occurred in a detector tube that will not contain the original annihilation site: This is an erroneous event (scattered event).

It is also possible that this ray will never reach a detector crystal and, therefore, get lost. This type of Compton scattering, along with photoelectric absorption of the produced gamma rays inside the source, where they have been generated, are the major sources of attenuation of the emitted radioactivity.

The physics of positron emission allow for attenuation correction of the collected data, which can produce a quantitatively (but also qualitatively) accurate image that may resolve small lesions, especially when these lie deep within the body. In order to correct for attenuation, two additional measurements are typically performed: the blank scan and the transmission scan. The blank scan is recorded using an external source without the patient, representing the unattenuated case. For the transmission scan, the patient and the bed are placed into the scanner and the attenuated data are measured using the external source. The attenuation correction factors (ACF) can be calculated as the ratio of the measured counts without and with the attenuating object. The disadvantages of attenuation correction are that it requires more time for image acquisition and the potential exists to add noise to the image if the attenuation measurements become misaligned by patient motion or if inadequate statistics in the transmission scan are collected. As a result of noise, transmission measurements are usually smoothed prior to the division. Otherwise, the noise in the ACF propagates to the corrected emission sinogram. The drawback of smoothing is that the resulting blurring of ACFs propagates to the emission sinogram as well. Techniques for the reduction of noise propagation include, as an example, classification techniques for the main tissue categories observed in the transmission images (segmentation) or the use of iterative methods for the reconstruction of the transmission images (6).

Compton scattering can also occur inside the detector crystal before the ray undergoes (the desirable) photoelectric effect. In that case, it is possible that the ray will escape the detector material and deposit its energy in an adjacent scintillator, causing the detected event to be mispositioned. Another source of erroneously counted events is the coincidental arrival at the detector ring of two single gamma rays coming from two different annihilation events (random or accidental coincidence). When three or more γ rays arrive at the detector ring within the time coincidence window set by the electronic circuitry of the scanner for the coincidence detection, then these gammas must be rejected, because it is not possible to recognize, in that case, the pairs of photons that came from the same annihilation event (7).

The high energy gamma rays have increased penetrating abilities and can be detected coming from deep-lying organs better than α particles or electrons (β particles), which can penetrate only a few millimeters of tissue and, therefore, cannot get outside the body to the radiation detector (5). Imaging system detectors must, therefore, have good detection efficiency for γ rays. It is also desirable that they have energy discrimination capability, so that γ rays that have lost energy by Compton scattering within the body can be rejected and a good timing resolution to accurately measure the time difference of the arrival of two photons. Sodium iodide (NaI), BaF_2 (barium fluoride), and BGO (bismuth germanate oxide) provide both of these features at a reasonable cost (5). Research for new scintillator materials, like LSO (lutetium oxyorthosilicate) (8), GSO (germanate oxide) (9), $PbCO_3$ (lead carbonate) (10), $PbSO_4$ (lead sulfate) (11), CeF_3 (cerium fluoride) (12), YalO (13), and LuAlO (14), is very active in an effort to produce faster detector crystals with good stopping power and light output.

Table 3 summarizes some of the main physical properties of the scintillators used for PET: NaI(Tl), BGO, BaF_2, CsF, GSO, and LSO. In order to interpret this table, assume the following:

- An elevated density guarantees a high stopping power for the high energy 511 keV annihilation photons and consequently assures elevated detection efficiency. High stopping power also allows the use of crystals of small dimensions, which means an improved spatial resolution of the tomograph.
- High scintillation efficiency, due to a good intrinsic energy resolution of the crystal, leads to a good energy resolution of the detection system, which leads to a better discrimination of scatter.
- A fast scintillation (described by a short scintillation constant decay time) translates to a low dead time of the system and, therefore, to good count rate performance. Moreover, this property directly influences the temporal resolution (uncertainty of the moment of detection), on which depends the choice of the length of the time coincidence resolution window and, therefore, the rate of accidental coincidences.

The comparison of the characteristics of scintillation crystals shows that the ideal scintillator for PET must have the temporal characteristics (decay time) of BaF_2, the density (stopping power) of BGO, and the scintillation efficiency (light output) of NaI(Tl) (15). It also reveals that the newest crystals GSO and LSO are very promising for PET applications.

Originally, NaI was the detector of choice for nuclear medicine imaging cameras and is still in use by some manufacturers of gamma cameras, SPET, and even PET systems. NaI is a scintillation crystal discovered in 1949 with very high scintillation efficiency but a stopping power too low for high energy photons; therefore, NaI has very low sensitivity. In the 1980s, BGO emerged as the detector of choice for PET scanners, a material with considerably lower light output than NaI but, on the other side, twice as dense and, therefore, able to detect high energy photons more effectively. LSO was discovered in the early 1990s and exhibits a very fast scintillation time (40 ns), which provides significantly reduced detector dead time and consequently higher count-rate capabilities, which is essential in clinical PET imaging in order to use the injected activity most efficiently and to make the emission scan time as short as possible, meaning the patient spends less time immobile on the tomograph's bed without compromising the image quality.

In the optimization of the design of a PET tomograph, an important aspect is the way crystals are assembled and the

Table 3. Scintillation Crystal Characteristics

	NaI	BGO	BaF_2	CsF	LSO	GSO
Density (g/cm^3)	3.67	7.13	4.87	4.64	7.40	6.71
Relative scintillation efficiency	100	20	16	6	75	30
Decay constant (ns)	250	300	0.6	2.5	40	60
Hygroscopic	Yes	No	No	Yes	No	No

way they are coupled to the photomultiplier tubes. Various strategies have been developed, including:

- one-to-one connection crystal-PMT (5);
- detector blocks, where a crystal array (mainly BGO or LSO) is coupled to a smaller number of PMTs (15,16);
- NaI(Tl) crystals of large dimensions coupled to a grid of PMT (Anger logic, common to gamma cameras) (17);
- the most recent design of a system of GSO crystals coupled to light guides to a PMT grid (18).

Scintillation detectors have been the dominant element in high energy gamma ray detection for PET. However, other technologies have also been applied, explored, and developed for this purpose. One of the oldest alterantive technologies is the High Density Avalanche Chamber (HIDAC) PET system (19), which consists of a Multiwire Proportional Chamber (MWPC) with the provision of laminated cathodes containing interleaved lead and insulating sheets and mechanically drilled with a dense matrix of small holes. Ionization resulting from photons interacting with the lead is trapped by, amplified in, and extracted from, the holes by a strong electric field into the MWPC. On arrival at an anode wire, further avalanching occurs. Coordinate readout may be obtained from orthogonal strips on the cathodes. The result is precise, 2D localization of the incident gamma rays. Every hole on the cathodes acts as an independent counter. By stacking these MWPCs, millions of these counters are integrated to form a large-area radiation camera with a high spatial resolution.

The resolution of a PET scanner primarily depends on the size of the detectors and on the range of positrons in matter (distance traveled by the positron in the tissue before interacting with a free electron, see also Fig. 1). For most of the positron emitters, the maximum range is 2–20 mm. However, the effect on spatial resolution is much smaller, because positrons are emitted with a spectrum of energies and only a small fraction travel the maximum range, and, in addition, in case of 2D acquisitions, the range of the third dimension is compressed. Another limitation in the resolution is that the paired annihilation photons are not emitted precisely 180° from each other, because the e^+–e^- system is not at complete rest. Other components of the system resolution are the sampling scheme used, the interactions between more than one crystal due to intercrystal scatter, the penetration of annihilation photons from off-axis sources to the detector crystals, the reconstruction technique used, the filters applied, and the organ and patient motion during the scan.

Three types of spatial resolution exist in a typical ring PET system, defined by a full-width at half-maximum (FWHM): the radial, tangential, and axial resolutions. The radial, or in-slice, resolution deteriorates as we move from the center of the FOV and is best at the center. The same happens for the tangential resolution, which is measured along a line vertical to a radial line, at different radial distances. In systems with more than one detector ring, the axial resolution, or slice thickness, is measured along the axis of the tomograph.

A major source of error during the coincidence detection is the fact that not all the annihilation events are registered correctly as mentioned earlier. Additional accidental coincidences can result from poor shielding or backscatter and from ordinary γ rays from the radionuclide administered. The random and scattered coincidences are registered together with the true coincidences, obtained when a pair of gammas is correctly identified and classified to the appropriate detector tube, and are sources of background noise and image distortion.

In order to keep the number of scattered coincidences low, a discriminator should be used. A discriminator primarily generates timing pulses upon the arrival of a photon, but also can verify the total energy of the illuminating ray is above a preset energy threshold. Scattered rays have already deposited part of their energy and, therefore, can be identified.

Furthermore, the choice of the appropriate time coincidence (or coincidence resolving time) window is essential: It has to be narrow enough to keep the number of random coincidences as low as possible but also wide enough to include all valid coincidence pulses. In the existing PET units, the timing accuracy is of the order of tenths of nanoseconds.

A PET scanner can be designed to image one single organ, such as the brain or the heart, or can be able to image any organ in the body, including whole-body scans. Whole-body studies with F-18-FDG consist of repeated PET acquisitions at contiguous bed positions in order to provide 3D images (axial, sagittal, coronal, and oblique cut planes) covering one considerable portion of the patient's body (Fig. 3), which facilitates the search for metastases in oncological diagnostics (20).

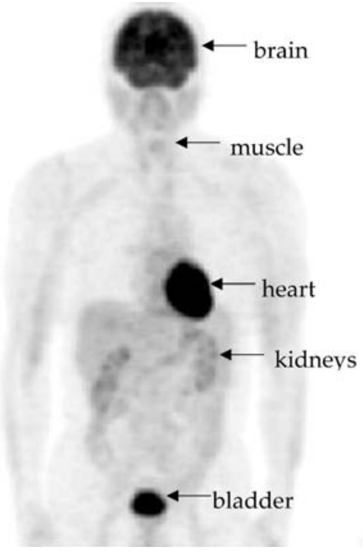


Figure 3. A whole-body F-18-FDG PET image of a normal subject (no pathological situation diagnosed). Areas with high metabolic activity (brain, myocardium) or with high concentration of the radioactive tracer (bladder) are visible. [Courtesy of A. Maldonado and M.A. Pozo from the Centro PET Complutense, Madrid, Spain.]

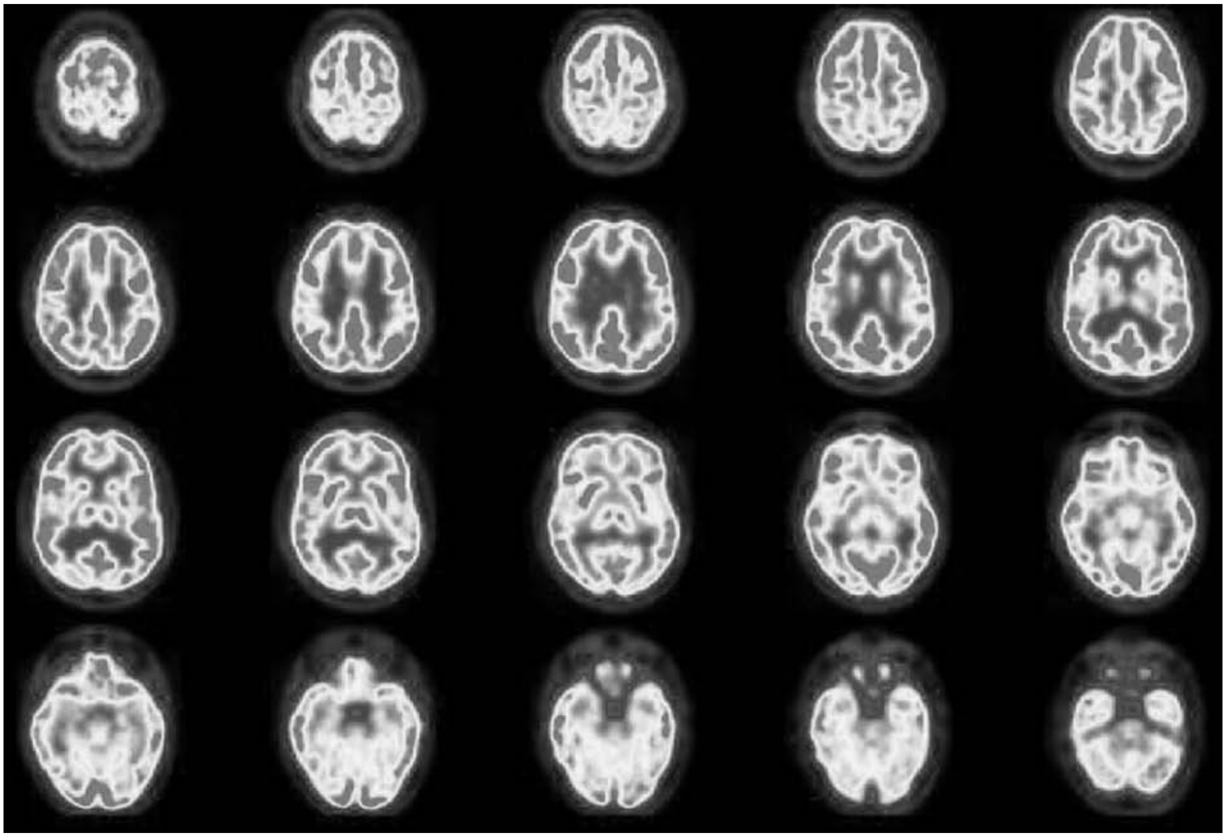


Figure 4. Sequential images from an F-18-FDG PET brain study of a normal individual. Red-yellow areas correspond to the high metabolic activity in the gray matter (cortex). [Courtesy of A. Maldonado and M.A. Pozo from the Centro PET Complutense, Madrid, Spain.]

Most PET systems today are whole-body systems (i.e., they have a typical transaxial FOV of 60 cm). This FOV is adequate to handle most patients. The axial FOV of most PET systems today is limited to approximately 10–15 cm (21). This relatively narrow axial FOV imposes some limitation on the imaging procedures that can be performed clinically. It also requires more accurate positioning of the patient in comparison with conventional nuclear medicine procedures. For a clinical system, it would be desirable to extend the axial FOV to 15–20 cm, which would, for instance, allow full brain (Fig. 4) and heart imaging in a single frame and more efficient whole-body imaging. As the detectors contribute a significant portion of the total cost of the scanner, however, this would bring into question what would be an acceptable cost for the PET scanner.

MANUFACTURING OF RADIOPHARMACEUTICALS

A cyclotron is a particle accelerator that produces positron-emitting elements or short-lived radioisotopes. These radioisotopes can then be incorporated into other chemical compounds that are synthesized into a final product that can be injected into a person. These radioisotopes are used to "label" compounds so it can later be identified where in the body the radiopharmaceutical is being distributed. The compounds that are being labeled are organic molecules normally used in the body, such as sugar, neurotransmitters, and so on (22).

First, the cyclotron bombards nonradioactive elements in the target with accelerated particles, which converts these elements into positron-emitting radioactive isotopes of fluorine, nitrogen, oxygen, or carbon. The major radioactive isotope produced at almost all sites is fluorine-18 (F-18), which has a half-life of 110 min. F-18 thus produced from the cyclotron is delivered to a chemical synthesis unit called the chemical processing unit, which is where F-18 is incorporated into a precursor to produce the final product FDG, the labeled sugar molecule. This entire process is fully automated and performed in the cyclotron lab. When a dose is needed, it is transported to the PET scan room by various means, depending on the distance between the production site and the PET tomograph and ranging from a dedicated pneumatic tube system to long-distance transport via air or road.

APPLICATIONS OF PET

Molecular imaging opens the way for medical doctors to successfully pursue the origin of disease. As long as disease is of unknown origin, more tests and exams are needed, something that means increased health-care costs, in addition to the patient's discomfort and pain. PET can accurately identify the source of many of the most common cancers, heart diseases, and neurological diseases, eliminating the need for redundant tests, exploratory surgeries, and drug overload of the patient. PET produces powerful images of the body's biological functions and reveals the mysteries of health and disease (23).

PET can be used to obtain information about the tissue perfusion using inert tracers (e.g., O-15 labeled water), the metabolism with metabolically active tracers (e.g., F-18-FDG), or the kinetic of a cytostatic drug (e.g., F-18-Fluorouracil).

In cardiology (22), this imaging technique represents the most accurate test to reveal coronary artery disease or rule out its presence. Traditionally, when a patient shows signs or symptoms of heart disease, his or her physician will prescribe a thallium stress test as the initial diagnostic study. The conventional thallium stress test, however, is often not as accurate as a PET scan. PET images can show inadequate blood flow to the heart during stress that can pass undetected by other noninvasive cardiac tests. A PET study could enable patients to avoid cardiac catheterization when a conventional perfusion or echocardiographic stress test is equivocal. A PET scan shows myocardial viability in addition to perfusion abnormality. More specifically, PET exams for metabolism and perfusion of the heart tissues can determine the need for heart transplant, in case both are absent in a large area of the heart, or confirm with certainty that simple bypass surgery would be enough, when metabolism is maintained even if blood flow is significantly reduced. As metabolism indicates that tissue is still alive, complicated heart transplantation can be avoided and coronary bypass would have great chances to improve cardiac function. Documented studies have shown that thallium stress testing overestimates irreversible myocardial damage in at least 30% of cases, which can result in the patient being placed on the transplant list rather than receiving bypass surgery or angioplasty. No other diagnostic test can more precisely assess myocardial viability than PET. The most recent developments in cardiac PET have been summarized in Reference 3.

PET can reveal abnormal patterns in the brain and is, therefore, a valuable tool for assessing patients with various forms of dementia (3,22). PET images of the brain can detect Parkinson's disease: A labeled aminoacid (F-DOPA) is used as tracer at a PET examination in order to determine if the brain has a deficiency in dopamine synthesis. If it does not, Parkinson's disease can be ruled out and possible tremors in the patient's muscles will be treated in a different manner. Although the only definitive test for Alzheimer's disease (AD) is autopsy, PET can supply important diagnostic information. When comparing a normal brain versus an AD-affected brain on a PET scan, a distinctive and very consistent image pattern appears in the area of the AD-affected brain, where certain brain regions have low metabolism at the early stages of the disease, allowing early detection several years before diagnosis can be confirmed by a physician. PET can also help to differentiate Alzheimer's from other confounding types of dementia or depression (29). Conventionally, the confirmation of AD was a long process of elimination that averaged between two and three years of diagnostic and cognitive testing. PET can help to shorten this process by identifying distinctive patterns earlier in the course of the disease. Furthermore, PET allows the accurate identification of epileptogenic brain tissue (because of its reduced glucose metabolic rates) and can successfully lead the surgical removal of the epileptic foci.

In oncology (3,22), in which the clear majority the total PET examinations refer, this technique inspects all organs and systems of the body to search for cancer in a single examination. PET is very accurate in distinguishing malignant tumors from benign growths. It can help detect recurrent brain tumors and tumors of the lung, breast, lymph nodes, skin, colon, and other organs. The information obtained from PET studies can be used to determine what combination of treatment is most likely to be successful in treating a patient's tumor, as it can efficiently determine the resistance of a specific cancer to the drugs applied and, consequently, can dynamically optimize the treatment management and follow-up of the patient on an individual basis. With this technique, it is possible to evaluate if a tumor has been successfully destroyed after therapy, as anatomical follow-up imaging is often not in the position to assess if a residue is still active or has definitely been eliminated after chemotherapy, radiation, or surgery.

A summary of the current status and future aspects of PET for cancer detection, as it has been recently presented by the Health Technology Advisory Committee is as follows (23):

Brain Cancer: F-18-FDG PET in brain tumor imaging may be useful, but its clinical application has yet to be established. F-18-FDG PET does not appear to be able to define tumor histology. Additional studies are warranted regarding the value of F-18-FDG PET in detecting Central Nervous System (CNS) and nonCNS brain metastasis, differentiating malignant from nonmalignant lesions, detecting disease recurrence in subjects who have undergone intensive radiotherapy, and in pediatric brain tumors. As a result of the paucity of data on radiotracers other then F-18-FDG, further studies will be required to validate the use of PET brain scanning with these radiotracers.

Head and Neck Cancer: Studies suggest that F-18-FDG PET is superior to MRI but comparable with CT in identifying the presence, absence, or recurrence of cancer.

Pituitary Cancer, Thyroid Cancer, Urinary Cancer, Kidney Cancer: The paucity of data on the use of PET in pituitary tumors, thyroid tumors, urinary cancer, and kidney cancer prevents conclusions regarding its value at this time.

Lung Cancer: Numerous studies evaluating PET for lung cancer applications demonstrate that PET, using F-18-FDG as a radiotracer, is effective and may be more effective than other noninvasive techniques, particularly CT, in differentiating benign and malignant pulmonary lesions. Thus, F-18-FDG PET appears to be an effective means of diagnosing lung cancer, whether a primary disease or a secondary metastatic disease, and detecting disease recurrence following lung cancer therapy.

Breast Cancer: Preliminary data suggest that F-18-FDG PET can differentiate benign from malignant breast lesions, when used in breast cancer staging, and can determine the presence of axillary node involvement. Although data are scarce regarding the use of PET in monitoring the effects of breast cancer therapy, available data suggest that both F-18-FDG PET and C-11-MET PET may be useful for breast cancer and may show response earlier than conventional methods. Regardless, due to the small study samples and limited amount of available data, further studies will be required to confirm the efficacy of PET for breast cancer imaging.

Esophageal Cancer: F-18-FDG PET may be valuable in the staging of esophageal cancer. Evidence is limited by the small number of subjects in each study and the lack of additional trials.

Pancreatic Cancer: Studies indicate that PET may have a role in the imaging of pancreatic tumors, but further study is needed to verify this indication.

Renal Cancer: F-18-FDG PET shows promise for evaluating renal masses, but confirmation is required.

Ovarian Cancer: Preliminary data suggest a potential role for F-18-FDG PET in ovarian cancer; however, further studies are required to confirm these findings.

Prostate Cancer: Although F-18-FDG PET has been used in certain prostate cancer cases, it is possible that the use of radiotracers other than F-18-FDG may be of more value. However, insufficient data exists at this time to draw conclusions regarding the use of PET in prostate cancer.

Testicular Cancer: With limited data, no conclusions can be made at this time.

Malignant Melanoma: Additional studies are needed to determine the role of PET in the imaging of malignant melanoma.

Colorectal Cancer: F-18-FDG PET may be a valuable tool for colorectal cancer in diagnosis, preoperative staging, and monitoring for recurrent disease or treatment response. However, further study is required to confirm these findings.

Neuroendocrine Gastrointestinal Cancer: PET proved superior to CT in detecting, delineating, and visualizing lesions. The study claimed that PET had a superior role, but further study is required to confirm this finding.

Malignant Lymphoma: Studies comparing F-18-FDG PET with alternative techniques found PET to be more accurate than CT, 99mTc-MIBI SPET, and 111In-somatostatin scintigraphy in detecting untreated and treated lymphoma. Supportive evidence is limited to a few trials that are hampered by small study samples. No conclusions can be drawn at this time regarding the efficacy of PET for malignant lymphoma.

A major use of PET is its ability for kinetic imaging analyses. This term refers to the measurement of tracer uptake over time. An image of tracer activity distribution is a good starting point for obtaining more useful information such as regional blood flow or regional glucose metabolism. The process of taking PET images of radioactivity distribution and then using tracer kinetic modeling to extract useful information is termed image analysis. The tracer kinetic method with radiolabeled compounds is a primary and fundamental principle underlying PET and autoradiography. It has also been essential to the investigation of basic chemical and functional processes in biochemistry, biology, physiology, anatomy, molecular biology, and pharmacology. Tracer kinetic methods also form the basis in *in vivo* imaging studies in nuclear medicine (24).

Besides its direct clinical applications, PET imaging is emerging as a powerful tool for use by the pharmaceutical industry in drug discovery and development (25). The role of small animal PET imaging (26) studies in rodents for the discovery of PET tracers for human use is significant, as it has the potential for permitting higher throughput screening of novel tracers in transgenic mice as well as the confounding effects resulting from potential species differences on receptor affinity, blood-brain barrier (BBB) transport, metabolism, and clearance. This setting is expected to allow new and unique experimental laboratory studies to be performed.

Other recent developments include dedicated mammography devices (known as positron emission mammographs, PEM) for breast functional imaging (27). Furthermore, the first PET/CT tomographs have made their way to the market (28). These are devices that house a positron tomograph and a CT scanner in a single device, allowing the acquisition and visualization of registered images detailing both anatomy and biological processes at the molecular level of internal organs and tissue, without the need of multiple examinations and further image processing to achieve similar results.

IMAGE INTERPRETATION

One of the final steps in the processing chain of the PET study is to produce a final layout of the images for the diagnosing physician. The conventional way of presenting the image data is to produce a transparency film (X-ray film) of the images on the computer display. In addition to the image data, the film should also be labeled with demographic data about the study, such as patient name and scan type. As this information is usually stored in the image files together with the image data, the labeling and layout of the images on the display can be automated in software. With the rapid development of local area networks, films may soon no longer be necessary. Instead, the images can be read from a display system located in the reading room, which has access to the PET image data through a computer network. Referring physicians do, in most cases, require a hard copy of the study, which can be accomplished using X ray films. With recent improvements in printer technology, high quality color output may also be a low cost alternative to the traditional film.

PROCEDURE FOR A PET SCAN

Most patients will be in the PET center for 2 or 3 h, depending on the type of study being conducted. The patient is informed as to when to stop eating before the test. Drinking lots of water is recommended before the scan. The patients also need to inform the PET center if they are diabetic or claustrophobic. In general, before the scan is performed, a catheter is placed in the arm so that the radioactive tracer can be injected. A glucose test will also be performed. Depending on the type of study conducted, scanning may take place before and after the injection is given. After the tracer is given, the patient waits for approximately 40–60 min before the final scan is done.

PET SCAN AND ASSOCIATED RISKS

The radiation exposure of PET is similar to that of having a CT scan or any other standard nuclear medicine procedure involving heart or lung scans. No pain or discomfort results from the scan. The half-life of F-18 is so short that by the time the patient leaves the PET center, almost no activity remains in the body. Patients typically do not experience any reactions as a result of the PET scan, because the tracer material is processed by the body naturally. Therefore, no side effects are expected. Of course, as with any other nuclear medicine procedure, when breast-feeding or pregnant, a PET scan must be performed under special conditions.

CURRENT STATUS AND FUTURE ASPECTS IN PET INSTRUMENTATION

Technological developments and research in the field of PET instrumentation are currently marking a fast evolution (30). New PET systems have been designed and developed with whole-body scanning capabilities. These systems are clearly designed for oncological studies (currently almost entirely performed in the clinical practice with the use of F-18-FDG), which represent maybe more than 80% of the total PET examinations performed worldwide. Therefore, a clear shift has occurred from the earliest systems, which were then mostly oriented to neurological applications.

The main requirement, which drive current R&D activities both in academia and in industry, is to increase the diagnostic accuracy (lesion detectability) of the technique and, at the same time, decrease the cost of a PET system installation, operation, and maintenance, which would allow the widespread use of PET in the clinical practice. In order to achieve this goal, an optimal balance should be found between high performance specifications and cost efficiency for the newly designed tomographs.

In particular, very high resolution 3D PET imaging (with applications in brain imaging, positron emission mammography, as well as small animal imaging) has demanded further advances in scintillation detector development, image reconstruction, and data correction methodology.

Table 4 lists the major performance characteristics of some last-generation tomographs for human whole-body studies, based on different design architectures and operating in 3D acquisition mode. For 2D acquisitions, lead or tungsten septa are placed between the detectors to absorb scattered radiation (out of slice activity). The septa reduce the amount of scatter to 10–15% of the total counts acquired and improve image contrast. For 3D acquisitions, the septa are removed and each individual detector is sensitive to radiation from a much larger area (30). This mode allows a significant increase of the detection efficiency of the order of a factor 5–6 over the 2D mode and therefore, provides an increase of the SNR in the produced images, an aspect of extreme importance in whole-body studies. 3D PET imaging can, in addition, significantly reduce the amount of tracer activity needed for the exam and shorten the acquisition time, thus reducing the time during which the patient must remain immobile on the tomograph's bed.

A limitation of the 3D mode, however, is an increase of the scatter component (almost one out of every two of the detected events has been scattered in the source or even inside the scintillation detectors) as well as of the number of the detected accidental coincidences (randoms) (30). A good energy resolution is therefore imperative in 3D PET systems, in order to reduce the scatter component (by correctly identifying detected γ rays with deposited energy lower than 511 keV). Furthermore, these systems must be able to manage high count rates in order to match the radioactivity present in the FOV. High temporal resolution in PET (high count rate) also permits dynamic imaging (repeated studies at short time intervals). With high count rates, pulses receiving a detector block can "pile-up" and the detector may become paralyzed, which decreases the sensitivity and detection efficiency of the tomograph. In addition, when scanning in a high counting rate environment, the random counting rate increases much more rapidly than does the true counting rate as a function of radioactivity in and near the FOV. In general, in 3D mode, an increased number of random events is detected, which degrades the image quality. Appropriate scatter and randoms corrections must therefore be applied to 3D-acquired data (31). Considering the nature of the scatter correction process and the heterogeneity of the activity distribution in the thoracic and abdominal areas (which are of particular interest for whole-body F-18-FDG PET studies), the use of scatter correction techniques is not yet consolidated and their effectiveness regarding the quality and quantitative accuracy of whole-body PET studies demands more research work.

The performance of a 3D-enabled PET tomograph is, therefore, the result of a compromise between the various physical parameters considered (spatial resolution,

Table 4. Performance Characteristics of PET Scanners in 3D Mode (15)

	Philips C-PET	GE Advance	ECAT HR+	ECAT Accel	Philips Allegro
Crystal	NaI	BGO	BGO	LSO	GSO
Crystal dimensions (mm)	500 × 300 × 25	4.0 × 8.2 × 30	4.0 × 4.4 × 30	6.8 × 6.8 × 20	4.0 × 6.0 × 20
Spatial resolution FWHM, mm (10 cm)	6.4	5.4	5.4	6.7	5.9
Efficiency (kpcs/Ci/cc)	450	1060	900	900	> 800
% Scatter fraction	25	35	36	36	25
50% Dead time (kcps/Ci)	0.2	0.9	0.6	-	0.6

detection efficiency, energy resolution, and linearity of count rate). In the modern design of such systems, the primary objective evolves from the optimization of the spatial resolution and the efficiency (typical of tomographs for cerebral applications) to the optimization of the balance between energy resolution and count rate performance. The size of the scintillation detector crystals, which together with the photomultiplier tubes constitute the main elements in the design of a PET system, determines the intrinsic spatial resolution of the tomograph. The volume of each crystal has a minimum, defined by the current technological limitations, but, at the same time, should be large enough to include a sufficient mass of material so that a significant number of the incident high energy γ rays are absorbed and converted to visible (detectable) light. A very small detector crystal could result transparent to γ rays, which would decrease the system's sensitivity.

Spatial resolution, an area of interest in the design of PET systems, refers to the development and implementation of techniques for the correction of the effect of "depth of interaction" (DOI) parallax error, which limits the uniformity of the spatial resolution in the FOV for PET tomographs with rings of detector block arrays (18). In such systems, the length of the detector crystals is about ten times as long as their width in order to improve detection efficiency. Therefore, PET measurements exhibit shift-variant characteristics, such as broadened sensitivity functions of each detector pair from center to edge of FOV and for oblique lines of response. Spatial uniformity can be restored if the DOI of the incident photons is known. A number of techniques for deriving DOI information from PET detectors have been proposed, including the use of a phoswich technique (32) (detector arrangements, composed from scintillation crystal layers; e.g., LSO/GSO phoswich block detector, where the distinct temporal characteristics of the crystals used allow to identify the DOI), extracting the DOI information by controlling the light-sharing between two crystals, coupling of two ends of the detection crystals to separate photodetectors, and extracting DOI information from a 3D matrix detector (33). Other approaches include the application of a light-absorbing band around each crystal, the introduction of a light-absorbing material between sections of the detector or use of a Multipixel Hybrid Photomultiplier (M-HPD) (34). When fully available in commercial tomographs, the implementation of correction techniques for DOI will allow the improvement of the spatial resolution and an ultimate optimization in the design of scintillation detection systems.

In order to draw a full advantage from the increase of the detection efficiency offered by 3D PET, developments in the field of the image formation are equally necessary. The acquired PET data are not an image of the activity distribution in the source but rather projections of it. The unknown image must be estimated from the available data computationally. Great interest is turned nowadays to completely 3D iterative image-reconstruction techniques (21). The more interesting feature of iterative techniques consists of the possibility to incorporate to the reconstruction process the statistical model of the process of acquisition and detection. In spite of their high computational cost, iterative techniques offer greater flexibility in the data processing, particularly of data with elevated statistical noise. The implementation of these reconstruction algorithms on clusters of workstations, grid platforms, or other high performance computing systems is an area of state-of-the-art research (35).

In spite of the fact that the clinical impact of the attenuation correction for whole-body F-18-FDG PET studies is still under discussion and study, iterative image-reconstruction techniques combined with correction for measured attenuation seem to offer various advantages:

- anatomical localization and spatial definition of lesions are improved,
- the geometric distortions observed can be compensated and corrected (requirement for being able to proceed to the co-recording with anatomical images – CT, MRI, and so on),
- the tracer update can be quantified.

An issue of greater technological interest for its major impact on oncological diagnosis is the development of integrated multimodality systems PET/CT (36). A PET/CT system consists of a PET tomograph and a CT tomograph, both of the last generation, assembled in a single gantry, controlled from a single workstation, with one unique patient bed. A PET/CT system allows the acquisition of PET and CT images in a unique examination with significant advantages:

- reduction of the examination time,
- integrated diagnosis by means of combined use of information from PET and CT,
- accurate interpretation of the PET functional images based on anatomical CT images (functional-anatomic correlation),
- improvement of the PET functional image quality using the anatomical information from CT (reconstruction with iterative techniques of the PET data with the use of the anatomical CT information as a priori information, for attenuation correction, and for accurate scatter correction, and for the correction of the partial volume effect),
- elimination of the radionuclide source for transmission scanning and elimination of the need for periodic replacement of decayed transmission sources.

The development of commercial PET/CT systems is quite recent, and the number of such systems installed and operational is still limited. Beyond the evaluation of the clinical effectiveness of such systems, various technical aspects still demand additional studies based on the clinical experience. The techniques of patient positioning must be optimized (arm position, etc.). The correction for attenuation based on CT studies must be validated (calibration of the attenuation-correction coefficients based on CT to the 511 keV energy window). The alignment of CT and PET studies must be verified, in particular regarding the acquisition protocols (conditions of apnea in CT studies and free respiration in PET studies). Furthermore, the performance of these complex and expensive systems should be compared with the performance of currently available software-based solutions for the co-registration and fusion of multimodality images (PET with CT, but also PET with MRI, ultrasound, etc.), which are shown to produce very accurate results, at least for brain studies.

Apart from whole-body human examinations, a challenging area of state-of-the-art research at the limits of current PET technology is the development of dedicated tomographs for small animal studies (25). In such systems, spatial resolution plays a crucial role as they are applied in the investigation of new pharmaceuticals and the development of new PET probes, as well as in the field of modern molecular biology, a scientific area that is currently focusing its interest toward imaging of laboratory mice and rats. As both the resolution and the sensitivity of small animal PET scanners are still limited by detector technology, image reconstruction algorithms, and scanner geometry, significant improvements may be expected in the performance of small animal PET scanners, whether prototype or commercial systems. In addition, multimodality imaging systems that will provide biological and anatomical information in an integrated setting, according to the model of PET/CT (or even PET/MR, etc.) systems already commercially available for human studies, should soon become available. The role of small animal PET in modern biology and pharmaceutical discovery and evaluation is in the process of being established, and it is likely that in vivo information of great value will be obtained. In addition, it is probable that the demanding requirements that small animal studies place on PET will result in technical advances and new technologies, which will dramatically improve the spatial resolution and image quality of clinical PET scanners for humans.

People today expect quality medical care at a reasonable cost, with accurate diagnosis and treatment, without having to undergo multiple exams and painful surgical exploration, and with fast and reliable results. Molecular imaging techniques, such as PET, display the biological basis of function in the organ systems of the human body unobtainable through any other means (37). PET is changing the way doctors manage care of their patients for some of today's most devastating medical conditions.

BIBLIOGRAPHY

1. Nutt R. The history of positron emission tomography. Mol Imag Biol 2002;4(1):11–26.
2. Brownell GL. A history of positron imaging. Online. 1999. Available at http://www.mit.edu/~glb/.
3. Phelps ME. PET Molecular Imaging and Its Biological Applications. New York: Springer; 2004.
4. Gambhir SS, Czernin J, Schwimmer J, Silverman DHS, Coleman E, Phelps ME. A tabulated summary of the FDG PET literature. J Nucl Med 2001;42:1S–93S. Online. Available at http://www.petscaninfo.com/zportal/portals/phys/clinical/jnmpetlit.
5. Sorenson JA, Phelps ME. Physics in Nuclear Medicine, 2nd ed. Orlando, FL: Grune and Stratton Inc.; 1987.
6. Zaidi H, Hasegawa B. Determination of the attenuation map in emission tomography. J Nucl Med 2003;44(2):291–315.
7. Turkington TG. Introduction to PET instrumentation. J Nucl Med Tech 2001;29(1):4–11.
8. Melcher CL, Schweitzer JS. Cerium-doped lutetium oxyorthosilicate: A fast, efficient new scintillator. IEEE Trans Nucl Sci 1992;39:502–505.
9. Ishibashi H, Kurashige K, Kurata Y, Susa K, Kobayashi M, Tanaka M, Hara K, Ishii M. Scintillation performance of large Ce-doped Gd_2SiO_5 (GSO) single crystal. IEEE Trans Nucl Sci 1998;45(3):518–521.
10. Moses WW, Derenzo SE. Lead carbonate, a new fast, heavy scintillator. IEEE Trans Nucl Sci 1990;37(1):96–100.
11. Moses WW, Derenzo SE, Shlichta PJ. Scintillation properties of lead sulfate. IEEE Trans Nucl Sci 1992;39(5):1190–1194.
12. Moses WW, Derenzo SE. Cerium fluoride, a new fast, heavy scintillator. IEEE Trans Nucl Sci 1989;36(1):173–176.
13. Ziegler SI, Rogers JG, Selivanov V, Sinitzin I. Characteristics of the new $YAlO_3$:Ce compared with BGO and GSO. IEEE Trans Nucl Sci 1993;40(2):194–197.
14. Moses WW, Derenzo SE, Fyodorov A, Korzhik M, Gektin A, Minkov B, Aslanov V. $LuAlO_3$:Ce-a high density, high speed scintillator for gamma detection. IEEE Trans Nucl Sci 1995; 42(4):275–279.
15. Gilardi MC. Tomografi PET: Attualitá e prospettive (in italian). XI Nat. Course on Professional Continuing Education in Nuclear Medicine (Pisa, 29-31/10/2001) Online. Available at http://www.aimn.it/ecm/pisa_01/Gilardi.pdf.
16. Casey ME, Nutt R. A multislice two dimensional BGO detector system for PET. IEEE Trans Nucl Sci 1986;33:460–463.
17. Karp JS, Muehllehner G, Mankoff DA, Ordonez CE, Ollinger JM, Daube-Witherspoon ME, Haigh AT, Beerbohm DJ. Continuous-slice PENN-PET: A positron tomograph with volume imaging capability. J Nucl Med 1990;31:617–627.
18. Surti S, Karp JS, Freifelder R, Liu F. Optimizing the performance of a PET detector using discrete GSO crystals on a continuous lightguide. IEEE Trans Nucl Sci 2000;47:1030–1036.
19. Jeavons A, Parkman C, Donath A, Frey P, Herlin G, Hood K, Magnanini R, Townsend D. The high-density avalanche chamber for Positron Emission Tomography. IEEE Trans Nucl Sci 1983;30:640–645.
20. Phelps ME, Cherry SR . The changing design of positron imaging systems. Clin Positron Imag 1998;1(1):31–45.
21. Tarantola G, Zito F, Gerundini P. PET instrumentation and reconstruction algorithms in whole-body applications. J Nucl Med 2003;44(5):756–768.
22. Let's Play PET. 1995, May 1. Online. Available at http://laxmi.nuc.ucla.edu:8000/lpp/lpphome.html.
23. Health Technology Advisory Committee. 1999, March. Positron emission tomography (PET) for oncologic applications. Online. Available at http://www.health.state.mn.us/htac/pet.htm.
24. Phelps ME. Positron emission tomography provides molecular imaging of biological processes. Proc Nat Acad Sci 2000;97(16): 9226–9233.
25. Fowler JS, Volkow ND, Wang G, Ding Y-S, Dewey SL. PET and drug research and development. J Nucl Med 1999;40: 1154–1163.
26. Chatziioannou AF. Molecular imaging in small animals with dedicated PET tomographs. Eur J Nucl Med 2002; 29(1):98–114.
27. Kontaxakis G, Dimitrakopoulou-Strauss A. New approaches for position emission tomography (PET) in breast carcinoma. In: Limouris GS, Shukla SK, Biersack H-J, eds. Radionuclides for Mammary Gland–Current Status and Future Aspects. Athens, Greece: Mediterra Publishers; 1997. p 21–36.
28. Beyer T, Townsend DW, Brun T, Kinahan PE, Charron M, Roddy R, Jerin J, Young J, Byars L, Nutt R. A combined

PET/CT scanner for clinical oncology. J Nucl Med 2000; 41(8):1369–1379.
29. Reba RC. PET and SPECT: Opportunities and challenges for psychiatry. J Clin Psychiatry 1993;54:26–32.
30. Fahey FH. Data acquisition in PET imaging. J Nucl Med 2002; 30(2):39–49.
31. Castiglioni I, Cremonesi O, Gilardi MC, Bettinardi V, Rizzo G, Savi A, Bellotti E, Fazio F. Scatter correction techniques in 3D PET: A Monte Carlo evaluation. IEEE Trans Nucl Sci 1999; 46(6):2053–2058.
32. Schmand M, Eriksson L, Casey ME, Andreaco MS, Melcher C, Wienhard K, Flugge G, Nutt R. Performance results of a new DOI detector block for a high resolution PET-LSO research tomograph HRRT. IEEE Trans Nucl Sci 1998;45(6):3000–3006.
33. Shao Y, Silverman RW, Farrell R, Cirignano L, Grazioso R, Shah KS, Vissel G, Clajus M, Tumer TO, Cherry SR. Design studies of a high resolution PET detector using APD arrays. IEEE Trans Nucl Sci 2000;47(3):1051–1057.
34. Meng LJ, Ramsden D. Performance results of a prototype depth-encoding PET detector. IEEE Trans Nucl Sci 2000; 47(3):1011–1017.
35. Kontaxakis G, Strauss LG, Thireou T, Ledesma-Carbayo MJ, Santos A, Pavlopoulos S, Dimitrakopoulou-Strauss A. Iterative image reconstruction for clinical PET using ordered subsets, median root prior and a We-based interface. Mol Imag Biol 2002;4(3):219–231.
36. Townsend DW. From 3-D positron emission tomography to 3-D positron emission tomography/computed tomography: What did we learn? Mol Imaging Biol 2004;6(5):275–290.
37. Phelps ME. PET: The merging of biology and imaging into molecular imaging. J Nucl Med 2000;41:661–681.

See also COMPUTED TOMOGRAPHY; IMAGING DEVICES; RADIOPHARMACEUTICAL DOSIMETRY.

PROSTATE SEED IMPLANTS

MARCO ZAIDER
Department of Medical Physics,
Memorial Sloan Kettering
Cancer Center
New York

DAVID A. SILVERN
Medical Physics Unit, Rabin
Medical Center
Petah Tikva, Israel

INTRODUCTION

Prostate seed implantation is a radiation therapy procedure by means of which small radioactive sources (colloquially referred to as “seeds”) are permanently implanted in the tumor-bearing tissue. In the absence of more specific information concerning the location within the prostate of tumor deposits, the goal of the implant is to deliver a minimum dose to the entire prostate gland while minimizing the dose to any adjacent healthy tissues, in particular, the urethra and rectum. As cause-specific death in prostate cancer is predominantly the result of distant metastasis (and not local failure), the *raison d'être* of prostate implantation must be the notion of some causal link, as opposed to mere association, between local control and distant disease (1–3). Whether this is indeed the case, remains at this time controversial (4).

The absorbed dose in the target as well as its spatial and temporal configuration is the only treatment tool available to the radiation oncologist. Consequently, patient eligibility for permanent prostate implants is based on the physician's ability to physically deliver the dose to the target, in other words, placing the seeds at relevant locations within or near the gland. Guidelines for patient selection consist of stage (T1–T2 disease) and prostate volume (less than about 50 cm^3). (*Staging* refers to the size and location of the tumor, whether tumor cells have spread to the lymph nodes, whether cancer cells have metastized to other parts of the body and to the abnormality of the tumor cells – this latter is referred to as *grade* and quantified by the Gleason score. Thus, T1 refers to a clinically inapparent tumor (not visible or palpable), and T2 refers to a low-grade tumor confined within the prostate. Pretreatment with androgen-ablation therapy is sometimes used to reduce the prostate volume.) Counter-indications for brachytherapy are short life expectancy (<5 years), the presence of metastatic disease, prior transurethral resection of the prostate (TURP), prostatitis, acute voiding symptoms, and inflammatory bowel disease (5).

The treatment may be delivered as monotherapy (implant alone) or in combination with external beam radiation therapy (EBRT) depending on whether the disease is confined (in which case monotherapy is administered) or extends beyond the prostate. Indications for combined treatment are extracapsular extension (ECE) and/or seminal vesicle invasion (SVI). It has been suggested that the patient's prognosis category as defined by pretreatment stage, Gleason score, and prostate specific antigen (PSA) may be taken as an indication of the likelihood that the disease did not spread outside the prostate; essentially, the lower the risk, the larger the probability of a confined tumor. Two classification schemes are currently in use. According to one proposal, low-risk patients are those with T1–T2b stage, Gleason 2–6, and PSA of less than 10 ng/mL. Intermediate-risk patients have one unfavorable factor: PSA larger than 10 ng/mL, Gleason score 7 or larger, or T2c or greater. High-risk patients have at least two unfavorable risk factors. The other sorting idea considers the risk low when PSA $\leq$ 10 ng/mL, Gleason <7, and T1a or T2a; intermediate when $10.1 < \text{PSA} \leq 20$, Gleason $= 7$ or T2b, and high otherwise (6,7).

The implant is a three-step process: First, using an imaging study of the prostate, the treatment planner calculates the number and location of seeds within the prostate volume that will result in a dose distribution in agreement with the prescribed constraints. The implantation is performed as a one-off outpatient surgical procedure. A post-implant evaluation, based on which the dosimetric quality of the implant is assessed, follows the treatment.

As with the other treatment choices (radical prostatectomy, EBRT), the survival benefits of transperineal permanent interstitial implantation among men with early, localized prostate cancer are uncertain (8,9). This state of affairs is compounded by the fact that the treatment of prostate cancer by either modality is accompanied by the risk of (occasionally permanent) rectal and urinary toxicity, as well as sexual disfunction (10–12). (Brachytherapy may be less likely to result in impotence in urinary incontinence than other forms of treatment.) Thus, for many patients with early, localized prostate cancer (the typical brachytherapy candidate), the decision to undergo a treatment of questionable benefit yet tangibly impacting on their quality of life (QOL) is understandably difficult; as a result, diminishing the risk of complications, and at the same time maintaining good dosimetric coverage of the tumor, remains the overriding concern in prostate brachytherapy.

In this article, we shall provide a step-by-step tutorial on permanent prostate implants, (by necessity) as practiced at the Memorial Sloan Kettering Cancer Center (MSKCC).

PREPLANNING OR INTRAOPERATIVE PLANNING?

Two modalities are currently in use for planning prostate implants. Preplanning refers to the situation where the treatment plan is completed several weeks before the actual implantation. The American Brachytherapy Society (ABS) discourages this approach, essentially because of well-recognized problems that may develop at the time the plan is implemented (5). For instance:

1. It is difficult to duplicate the patient position in the operating room (OR) to match the patient position during the preplanning simulation.
2. Patient geometry can change over time; for instance, urine in the bladder or feces in the rectum may swell the prostate.
3. A pretreatment plan may prove impossible to implement due to the needle site being blocked by the pubic symphysis.

The alternative to preplanning, which we and others strongly advocate, is to perform the plan in the OR using ultrasound (US) images of the treatment volumes acquired with the patient on the operating table in the implantation position (13–19). The ability to obtain a dose-optimized plan within a reasonable time (say, 10 minutes or less) is the key to intraoperative approach. Clearly, a manual (i.e., trial and error) approach to finding optimal seed positions in the OR will not do. The *sine qua non* condition of OR-based planning is then the availability of a computer-optimized planning technique, as described below.

ISOTOPE SELECTION AND DOSE PRESCRIPTION

The two isotopes commonly used for permanent prostate implants are ^{125}I (mean photon energy $E_\gamma = 27$ keV, half life $T_{1/2} = 60$ days) and ^{103}Pd ($E_\gamma = 21$ keV, $T_{1/2} = 17$ days). An important property these isotopes share is their low effective energies. At these energy levels, practically all decay radiation emitted by the implanted sources is absorbed in the patient's body. This enables the patient to be discharged shortly after the implant procedure without fear of posing a radiation hazard to the general public or to the patient's family.

Dose prescription in prostate brachytherapy makes use of the concept of *minimum peripheral dose* (mPD), which is defined as the largest isodose surface that completely surrounds the clinical target (20). The total dose prescription for patients treated at MSKCC is 144 Gy (mPD) for ^{125}I and 140 Gy for ^{103}Pd (21). The initial dose rate $\dot{D}(0)$ corresponding to these values can be calculated from

$$D_{\text{total}} = \frac{T_{1/2}}{\ln(2)} \dot{D}(0) \tag{1}$$

Thus, for ^{125}I, one has $\dot{D}(0) = 7$ cGy/h, whereas for ^{103}Pd, the corresponding value is 24 cGy/h. Based on radiobiological considerations it was hypothesized that the higher initial dose rate of ^{103}Pd may be more appropriate for rapidly proliferating tumor cells. Gleason score is taken as marker for such cells, hence, the notion that ^{103}Pd should preferentially be used for high-grade tumors. Retrospective studies have failed to demonstrate any clinical (22,23) or dosimetric difference between the two isotopes. (One may also note that in the United States, the price of a typical ^{125}I seed is about half the price of a ^{103}Pd seed; as well, a typical implant would use, say, 75 ^{125}I seeds as against some 100–110 ^{103}Pd seeds.)

A second difference between the two isotopes is the potential for somewhat larger relative biological effectiveness (RBE) of the ^{103}Pd compared with ^{125}I (24,25).

THE PHYSICAL CHARACTERISTICS OF ^{125}I AND ^{103}Pd

^{125}I is produced in a reactor by irradiating ^{124}Xe with neutrons to form ^{125}Xe. ^{125}Xe has a 16.9 h half-life and decays to ^{125}I via electron capture. ^{125}I in turn decays to an excited state of ^{125}Te via electron capture. The excited ^{125}Te nucleus immediately releases its excess energy in the form of a 35.5 keV gamma photon in 6.66% of the transformations, with the energy of the remaining 93.34% of the transformations being released as internal conversion electrons (26). The rearrangement of the orbital electrons results in the emission of characteristic X rays and Auger electrons. The Auger electrons are blocked by the encapsulation of the source and do not directly contribute to patient dose. As a result of the lower energy characteristic X-ray emissions, the average photon energy of ^{125}I is 28 keV.

^{103}Pd is produced in a reactor by irradiating stable ^{102}Pd with neutrons. ^{103}Pd decays with a 17 day half-life to excited states of ^{103}Rh via electron capture. The excited ^{103}Rh nuclei lose nearly all of their excess energy via internal conversion (26). The rearrangement of the orbital electrons results in the generation of characteristic X rays and Auger electrons. As is the case with ^{125}I sources, the Auger electrons are blocked by the encapsulation and do not directly contribute to patient dose.

The radioactive sources must be fabricated to high-quality control standards. The sources must remain biocompatible *in vivo* for decades. To prevent the internal contents of the sources from diffusing into the body tissues, the integrity of the source encapsulation must also remain sound for a period of decades. The physical size of the sources must be sufficiently small to allow their interstitial implantation without causing undue tissue trauma or interfering with physiologic function. In addition to biocompatibility issues, the sources must be physically strong enough to maintain their shape and integrity during sterilization, routine handling, and implantation. As well as to the need for physical ruggedness, the encapsulation must be made of a material that will not absorb an undue fraction of the emitted photons. For the purpose of source localization on radiographs and computed tomography (CT) images, it is desirable for the sources to be radio-opaque while causing minimal imaging artifacts.

For meeting the aforementioned requirements, titanium is the material of choice for source encapsulation. Titanium is as strong as steel but 45% lighter. It is only 60% heavier than aluminum but twice as strong. This metal is also highly resistant to sea water. As human tissues have a high degree of salinity, titanium can maintain its integrity *in vivo* for the remainder of the patient's life. Titanium has an atomic number of 22, low enough not to cause serious artifacts on CT images. Due to the high strength of titanium, the encapsulation can be made thin, minimizing absorption of the emitted photons. The main drawback of titanium is that it is expensive, significantly adding to the cost of the radioactive source.

The actual internal structure of the radioactive sources is vendor specific. The thickness of the titanium encapsulation ranges from 0.04 to 0.06 mm (26). In the classic Amersham 6711 source, the radioactive material is adsorbed to the surface of a silver rod. A schematic representation of this source is shown in Fig. 1. The silver serves as a radio-opaque marker. In other source models, the radioactive material is adsorbed onto resin beads, impregnated in ceramic material, or coated by other means onto various substrates. Most ^{125}I and ^{103}Pd radioactive sources include radio-opaque marker material. Gold, silver, tungsten, and lead are the commonly employed marker materials used in ^{125}I and ^{103}Pd sources. The differences in the internal structures of the radioactive sources result in vendor-specific dose distributions. These differences result in differing photon energy spectra, source anisotropy, and self-absorption properties. The dosimetric properties of ^{125}I and ^{103}Pd sources are discussed in the next section.

DOSIMETRY

Before any radioactive source can be employed in a clinical implant procedure, the dose distribution around the source in question must be known. One obvious requirement is that the accuracy of the dose calculations be as high as possible. Another useful requirement is wide acceptance of the dosimetric formalism applied. A universal dose formalism simplifies comparison of treatment outcomes of implants performed by different institutions and helps establish universal treatment protocols.

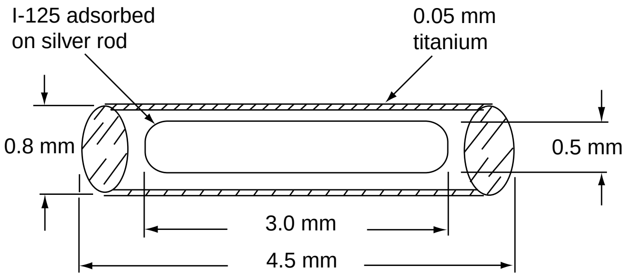


Figure 1. Schematic representation of an Amersham 6711 ^{125}I source.

To assure accurate dosimetric calculations, the following physical properties must be ascertained:

1. Photon interactions with structures inside the source and encapsulation.
2. Geometric distribution of radioactive material within the source.
3. Photon interactions with the tissues encompassing the source.
4. Reduction of radiation intensity as a function of distance from the source.

Photon interactions within the source have a noticeable effect on the shape, intensity, and energy spectrum of the dose distribution surrounding the source. These internal photon interactions affect the anisotropy of the emitted radiation. (Source anisotropy is a measure of angular dependence of the dose distributions surrounding the source.) For a point source with no asymmetric self-absorption, dose would only be a function of distance from the source without any angular dependence. Photon interactions inside the source also influence the energy spectrum of the emitted photons. Depending on the physical construction and materials used to fabricate the source, the emission spectra will vary among different source models. Lower energy photons will be preferentially attenuated compared with higher energy photons. The degree of filtration is dependent on the construction of the source. Another alteration to the intrinsic spectrum of the radioactive material is the generation of characteristic X rays by photoelectric interaction with the materials inside the source. One noteworthy example of spectral variations among different ^{125}I source models is the use of silver as the radio-opaque marker used by certain manufacturers. Silver has a K-edge that occurs at 25 keV (27,28). As the intrinsic photon emissions of ^{125}I are in this energy range, the photoelectric cross section for interaction with the silver marker is high. As a result, I^{125} sources using silver as the radio-opaque marker will have a local peak around 25 keV in their emission spectra. I^{125} sources using another element for radio-opacity will not exhibit a 25 keV peak in their emission spectra. Differences in source construction also influence the degree of Compton scattering, further altering the emission spectra. The geometric distribution of the radioactive material inside the source will affect the angular dependence of the dose distributions around the source as well as the dose reduction as a function of distance from the source. This effect will be investigated when source geometry factors are discussed later in this section.

Another important factor that must be calculated by a dose formalism is tissue attenuation as a function of distance from the source. Tissue attenuation is a function of how the emission photons interact with the tissues. When using I^{125} or ^{103}Pd sources, the predominant interactions are Compton scattering and photoelectric effect. In soft tissue, the probabilities of photoelectric and Compton interactions are equal at a photon energy of about 25 keV (28). The dose decreases with distance from the source due to these photon interactions.

The simplest dosimetric formalism is to assume point-source geometry, neglecting angular dependencies on the dose distributions. In this formalism, the dose is strictly a function distance from the source. Two components are assumed to contribute to the resulting dose at a given point. One component is the dose falloff due the geometric shape of the radioactive source. For a point source, this falloff component is the inverse square law, namely $1/r^2$. This dose falloff occurs irrespective of any photon interactions in the medium surrounding the source. It is solely dependent on the photon fluence intensity being geometrically reduced by the inverse square law. The second component of dose falloff results from photon interactions in the surrounding medium. One analytical approach to modeling this phenomenon is to assume that the dose reduction follows a simple exponential function, namely $F(r) = e^{-\mu r}$ where μ is an average linear attenuation coefficient for the energy spectrum of the emitted photons. It is also assumed that dose is directly proportional to the source activity A. The equation for calculating the dose using this formalism is

$$D(r) = \Gamma A\ f_{\text{med}} T_{\text{av}} \frac{e^{-\mu r}}{r^2} \tag{2}$$

where D represents the dose at distance r and A is the source activity. The factors Γ and f_{med} are the exposure rate constant and the tissue f-factor, respectively. The f-factor converts the exposure in air to absorbed dose in a small piece of tissue just large enough to assure electronic equilibrium. T_{av} is the average life an isotope atom exists before undergoing a nuclear transformation and μ is the effective linear attenuation coefficient. This analytical equation suffers from two drawbacks, the first being that this exponential equation is only rigorously correct for narrow-beam geometries. In deriving this equation, it is implicitly assumed that all photons that interact with the medium are removed from the beam and that no further interactions of scattered photons occur in the path of the beam. The geometry of a radioactive source in a medium is clearly not narrow-beam. Compton photons do in fact interact with the medium in the beam and hence contribute to dose. A second drawback of using an exponential is the implicit assumption that the photons are mono-energetic, clearly not the case for either ^{125}I or ^{103}Pd emissions. Although this formalism is not rigorously correct for the reasons stated, it has been used for many years in brachytherapy treatment planning. By empirically determining values for Γ and μ, the discrepancies of calculated and measured doses could be made acceptably small. In the years that this equation was used, modern instrumentation was not available for precise dose measurement, computers were slow, and the standards of conformance were not as stringent as today.

The accuracy of the formalism can be improved by replacing the exponential equation with a data table. The values in the table consist of experimentally measured doses in water at known distances, multiplied by the square of the distance. Photon interactions in water are similar to those in soft tissue. The dose at any arbitrary distance from the source is calculated via linear interpolation of the tabulated data and by dividing by the square of the distance. Using tabulated data solves the two problems associated with the exponential function. As tabulated data were derived from measurements in the true broad-beam geometry of the source, the formalism inherently accounts for the dose occuring from Compton-scattered photons as well as the primary photons. As a table could be created for any source model, the data will inherently account for the energy spectrum of the emitted photons. An added benefit provided by the table is that it accounts for the selective tissue filtration of lower energy photons with increasing depths. As the emission spectra vary among different source models, the tissue filtration will also likewise vary.

A more sophisticated formalism can be developed by accounting for the geometric distribution of the radioactive material inside the source. The most natural extension of the formalism is the assumption that the radioactive material is distributed as a line source. Although the distribution of the radioactive material in many commercially available sources is not a true uniform line, the line source model is still a more accurate and realistic representation than the point source model. Applying a line source model significantly complicates the dosimetric computations. When applying the point source model, only the distance from the source to the calculation point need be known to uniquely determine the dose. By contrast, using a line source model requires that the angular orientation in addition to the distance of each dose calculation point in relation to the source be known. The angular orientation of the calculation point with respect to the source needs to be calculated using analytic geometry.

The angular dependence of the dose distribution around a source is the result of two separate processes. The first results from the distribution of the radioactive material inside the source, and the second results from the angular dependence of attenuation of the photons within the source. To develop a formalism based on a line source, it is easiest to start by calculating the dose to a differential piece of tissue in empty space. It is assumed that there is no attenuation in the tissue and that electronic equilibrium exists. As tissue effects will be introduced at a later stage, this assumption does not compromise the rigor of the development of the formalism. At this stage, the self-absorption and tissue attenuation will not be considered.

The angular dependence of dose resulting from the geometric distribution of the radioactive material can be modeled by a line source. A schematic representation of a line source is shown in Fig. 2.

For any point P one can define a two-dimensional coordinate system as shown in Fig. 2. With such a coordinate system defined, the dose to any point P from the line source can be calculated. It is seen from Fig. 2 that the following formulation holds:

$$X = -r\cos\theta; \quad Y = r\sin\theta; \quad r = Y/\sin\theta; \quad \tan\theta = -Y/X; \quad X = -Y/\tan\theta \tag{3}$$

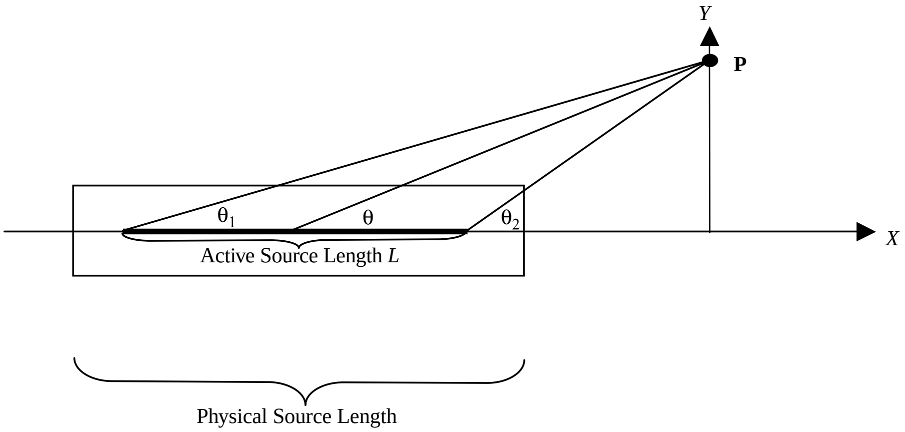


Figure 2. Dose to a point *P* from a line source. Physical length includes encapsulation. Active length only includes length of radioactive material.

The differential length dX is calculated to be

$$dX = Y(\tan\theta)^{-2}(\sec\theta)^{-2}d\theta = Y(\sin\theta)^{-2}d\theta \qquad (4)$$

The total activity A of the line source is assumed to be uniformly deposited along the active length L of the source. Under ideal conditions, uthe dose to a differential piece of tissue at point P in free space is given by

$$D(r) = A\Gamma\, f_{\text{med}}\frac{T_{\text{av}}}{r^2} \qquad (5)$$

To calculate the dose from the line source shown in Fig. 2, the source can be considered a continuum of differential point sources resulting in

$$dD = \frac{A}{L}\Gamma \quad f_{\text{med}}T_{\text{av}}\frac{dX}{r^2} = \frac{A}{L}\Gamma \quad f_{\text{med}}T_{\text{av}}\frac{d\vartheta}{Y} \qquad (6)$$

The total dose to point P is thus

$$D = \frac{A}{L}\Gamma \quad f_{\text{med}}T_{\text{av}}\int_{\vartheta_1}^{\vartheta_2}\frac{d\vartheta}{Y} = \frac{A}{L}\Gamma \quad f_{\text{med}}T_{\text{av}}\frac{\vartheta_2 - \vartheta_1}{r\sin\vartheta} \qquad (7)$$

By using equation 7, it is now possible to account for the linear distribution of radioactive material inside the source. In clinical brachytherapy, however, one is generally interested in knowing the dose when the source is in the patient and not in free space. This means that the photon interactions with the surrounding tissues must be factored into the analysis. Additionally, the self-absorption of the source must be taken into account. Unlike the linear distribution of the radioactive material, the self-absorption and tissue interactions are not amenable to a rigorous treatment. These data must be derived either from direct measurement or from Monte Carlo simulation, or a combination of the two. The data tables would need to be two-dimensional to account for the distance from the calculation point to the source as well as the angular orientation of the calculation point to the center of the source. As the active length of the source is known, the values of θ_1 and θ_2 are uniquely determined. Once these data tables are available, the component of the dose due to geometric distribution of the radioactive material can be removed from the data by dividing the tabulated doses by the ideal line source dependency. Reviewing equation 7, one can separate the terms involving the linear distribution of the radioactive material. Doing this yields the following:

$$G = \frac{\vartheta_2 - \vartheta_1}{Lr\sin\vartheta} \qquad (8)$$

where G is referred to as the *geometry factor*. Dividing the tabulated doses by G removes the line source contribution. Removing the geometry factor reduces the dependency of the data on distance from the source, enabling the use of smaller data tables for attaining the same degree of accuracy. The data tables still account for the tissue interactions and self-absorption. For permanent implant brachytherapy, the data tables can be in the form of total dose per unit source activity (or air kerma strength) because the implant time is infinite and the isotope half-life is known. The new formalism can be summarized as follows:

1. For each point P in space, define a two-dimensional coordinate system and compute the distances to the centers of each implanted source and compute the respective angles θ, θ_1, and θ_2.
2. Use equation 8 to calculate the geometry factors at point P for each implanted source.
3. Interpolate the data tables based on the distances and angles for each implanted source calculated in step 1.
4. For each implanted source, multiply the respective geometry factor, source activity (or air kerma strength), and interpolated table value together. Each product represents the dose to point P for each implanted source.
5. Sum the individual doses to obtain total dose at point P.

These calculations are tedious and time consuming if performed by hand. The only feasible way forward is to code the formalism in software and have the computer perform the computations. In this way, a finely spaced three-dimensional data grid of calculation points can be rapidly calculated. The dose at any arbitrary point in space can be calculated by interpolating the three-dimensional grid of calculated doses. This will be discussed in a later section devoted to treatment planning.

The dose formalisms discussed thus far are but a small sample of the calculation methods used in brachytherapy treatment planning. They were discussed to serve as illustrative examples and to provide a brief introduction to the physics involved in the development of dosimetry formalisms. Over the years, many different dosimetry formalisms have been proposed and implemented. By the early 1990s, treatment planning computers have become ubiquitous in radiotherapy departments. With many such systems in use, questions began to arise regarding the accuracy, consistency, and general agreement of the calculated doses generated by these systems. It was obvious that there were differences in the values calculated by these treatment planning systems as each system implemented its own dosimetric algorithm. During this same period, researchers also proposed new formalisms based on physical quantities not widely used in brachytherapy treatment planning. As a result of these issues, the American Association of Physicists in Medicine (AAPM) decided that the time had come to develop a standardized brachytherapy formalism. The adoption of such a formalism would standardize the dosimetric calculations, reducing the differences in values calculated by different treatment planning systems. The doses calculated among different institutions will be in closer agreement, enabling more realistic comparisons of different brachytherapy protocols.

The AAPM Radiation Therapy Committee Task Group 43 established a new recommended formalism. This formalism was published in 1995(26). The Task Group 43 (TG43) formalism is a radical departure from most established methodologies used up to that time. In following the new formalism, changes in dosimetric values in some instances were of sufficient magnitude to mandate changes in prescription doses for maintaining clinical consistency with older formalisms.

The most fundamental change recommended by TG43 is to use air kerma strength S_k in lieu of activity. The activity of a radioactive source is the number of disintegrations per unit time. As a fundamental unit, activity in and of itself does not provide any information regarding the nature of the energy deposition of the decay emissions. Traditional dosimetry formalisms need to use exposure rate constants and *f*-factors to convert from activity to exposure in air to dose deposited in tissue. The exposure rate constant is also isotope-specific.

Air kerma strength is defined as follows. A mass of air, dm, is placed a distance r_{ref} from the source along the perpendicular bisector. The source and air mass are in a vacuum. *Air kerma strength* is the kerma rate in mass dm multiplied by the square of the distance r_{ref}. Unlike activity, which is only applicable to radioactive isotopes, air kerma strength can be applied to any source of uncharged particles. For example, the radiation output from a linear accelerator or X-ray tube can also be quantified in terms of air kerma strength.

Another new quantity used by TG43 is the *dose rate constant* Λ defined as the dose rate to water at a distance of 1 cm from the perpendicular bisector of the source. Using the dose rate constant represents another departure from basing absorbed dose on exposure to air. It is a trend in the medical physics community to move toward basing dosimetric calculations and measurements on dose to water. The dose rate constant replaces the exposure rate constant used in older brachytherapy formalisms.

For modeling the contribution of the geometric distribution of the radioactive material inside the source, TG43 endorses the use of either a point source representation (inverse square law) or a line source representation (equation 8). These geometric factors have been in use before the advent of the TG43 report.

The interactions of the emitted photons in tissues are modeled by a new function defined by TG43, namely the radial dose function $g(r)$. The function $g(r)$ is given by the following equation:

$$g(r) = \frac{dD(r, \vartheta_0)/dt\, G(r_0, \vartheta_0)}{dD(r_0, \vartheta_0)/dt\, G(r, \vartheta_0)} \qquad (9)$$

where D is the dose at a point along the perpendicular bisector at a distance r from the source and G is the geometry factor that is equal to either the inverse square law or equation 8. In equation 9, θ_0 is equal to $\pi/2$. The radial distance r_0 is the reference distance to which all TG43 parameters are normalized. In practice, r_0 is taken to be 1 cm. It must be stressed that $g(r)$ is only defined along the perpendicular bisector of the source. It can be seen that when r is equal to r_0, $g(r)$ equals 1. The factors in equation 9 involving G remove the dependency of the geometric distribution of radioactive material inside the source on $g(r)$. The radial dose function models the interactions of the photons with tissues along the transverse axis of the source. In practical implementations of TG43, $g(r)$ is based on tabulated values, which in turn are based on measured data or Monte Carlo simulations. In some implementations, analytic functions are fit to the data, whereas in others, the value of the radial dose function is calculated via interpolation.

Another phenomenon that needs to be modeled is the anisotropic nature of the radiation resulting from photon interactions inside the source. If there were no interactions of the radiation within the source, all of the angular dependence of the radiation pattern would result from the geometry factor, assuming the source is in a homogeneous medium. In reality, however, self-absorption is significant, especially at the low energies of ^{125}I and ^{103}Pd.

In the original TG43 report, three methods of modeling source anisotropy were proposed. The most general of these models is the source anisotropy function $F(r, \theta)$. This function is defined as follows:

$$F(r, \vartheta) = \frac{dD(r, \vartheta)/dt\, G(r, \vartheta_0)}{dD(r, \vartheta_0)/dt\, G(r, \vartheta)} \qquad (10)$$

This function in effect is the ratio of the dose rate at an arbitrary distance from the source r and angle θ multiplied by the ratio of the geometry factor at a distance r but on the transverse axis and the geometry factor at the location r and θ. This function quantifies the angular variation of the dose distribution removing the contribution of the geometry function. In most practical implementations, the anisotropy function is calculated via bilinear interpolation of a two-dimensional data table.

A second method for modeling source anisotropy is to represent source anisotropy as a function of only the distance from the source. In this case, the dose rate is averaged over all values of solid angle from 0 through 4π steradians. The one-dimensional function is referred to as the anisotropy factor $\phi_{\mathrm{an}}(r)$. By taking advantage of the cylindrical symmetry of radioactive sources, the solid angle integral for defining this factor reduces to the following:

$$\phi_{\mathrm{an}}(r) = \frac{1}{2dD(r,\vartheta)/dt} \int_{0}^{\pi} \frac{dD(r,\vartheta)}{dt} \sin\vartheta \, d\vartheta \tag{11}$$

One fundamental difference between defining $\phi_{\mathrm{an}}(r)$ and $F(r,\theta)$ should be pointed out. The geometry factor is not removed from $\phi_{\mathrm{an}}(r)$ as is the case for $F(r, \theta)$. From a numerical standpoint, the average value of the geometry factor taken over 4π steradians is nearly equal to the nominal value of $G(r,\theta_0)$. This is especially true for distances greater than the active length of the source. Moreover, $\phi_{\mathrm{an}}(r)$ is an average value that will result in an inevitable error in the actual dose calculation. Factoring out the geometry factor would not significantly improve the accuracy of the dosimetry. If the geometry factor contribution was in fact removed, two anisotropy factors would need to be calculated, one for point source geometry and the other for line source. The main motivating factor for defining $\phi_{\mathrm{an}}(r)$ is to accommodate existing treatment planning systems that do not support two-dimensional dose calculations. Given a choice, using $F(r,\theta)$ is preferred as it is rigorous.

A third anisotropy correction method prescribed by the original TG43 report is the use of an anisotropy constant ϕ_{an} that is an average value independent of distance. The original TG43 report was updated in 2004. In the updated report, use of anisotropy constants was made obsolescent and is no longer considered consistent with current standards of practice. A more detailed discussion of the TG43 update is discussed below.

Application of the TG43 formalism can be summarized by the following equations where all of the terms have been discussed:

$$dD(r,\vartheta) = S_K \Lambda \frac{G(r,\vartheta_0)}{G(r_0,\vartheta_0)} g(r) F(r,\vartheta) \tag{12}$$

Equation 12 is the rigorous implementation of the original TG43 formalism. A simpler implementation of equation 12 for use in treatment planning systems not supporting two-dimensional dose calculations is the following:

$$dD(r,\vartheta) = S_K \Lambda \frac{G(r,\vartheta_0)}{G(r_0,\vartheta_0)} g(r) \phi_{\mathrm{an}}(r) \tag{13}$$

The third equation using the anisotropy constant is as follows:

$$dD(r,\vartheta) = S_K \Lambda \frac{G(r,\vartheta_0)}{G(r_0,\vartheta_0)} g(r) \phi_{\mathrm{an}} \tag{14}$$

Use of equation 18 is no longer recommended in the updated TG43 protocol. All treatment planning systems need to be capable of performing one-dimensional calculations. It is thus always possible to use the anisotropy factor $\phi_{\mathrm{an}}(r)$. In some treatment planning systems, the product of the radial dose function and the anisotropy factor may need to be used if only one distance-dependent term is used for calculating the dosimetry.

To end this section, a brief discussion of the updated TG43 formalism is presented. This update was published in 2004, 9 years after the original TG43 report (29,30). In the original formalism, the same radial dose functions were used irrespective of geometry factor. In the new formalism, two sets of radial dose functions are recommended. One radial dose function is to be used for point source geometry, whereas the other is to be used for line source geometry. Although the updated and original reports define the radial dose function using the same equation, the published values could be used for both geometry factors. To improve the consistency of the dose calculations, two sets of radial dose functions are now recommended. The choice of geometry factor dictates which radial dose function is to be used. Further refinements were also made in the tabulated values presented. In the intervening 9 years between the original report and the update, several new radioactive sources were introduced. The updated report includes published data for use with the newer source models. In the new report, the use of anisotropy constants is no longer recommended.

ULTRASOUND-GUIDED IMPLANTATION TECHNIQUE

The implantation proceeds as follows. The patient is placed in the extended lithotomy position. An ultrasound probe is positioned in the rectum, and needles are inserted along the periphery of the prostate using a perineal template as a guide. Trans-axial images of the prostate are acquired at 0.5 cm spacing (from base to apex), transferred to the treatment planning system using a PC-based video capture system, and calibrated. For each US image, prostate and urethra contours as well as the anterior position of the rectal wall are entered. Needle positions are identified on the ultrasound images and correlated with the US template locations. The contours, dose reference points, needle coordinates, and data describing the isotope/activity available along with predetermined dose constraints and their respective weights serve as input for the dose optimization algorithm.

TREATMENT PLANNING

A commissioned treatment planning system is the minimum equipment required for performing brachytherapy treatment planning. State-of-the-art treatment planning systems are based on modern computer hardware and software. Practically all modern brachytherapy treatment systems implement the TG43 dosimetry formalism for low-energy radioactive sources. These systems provide interfaces to imaging scanners such as CT, US, and magnetic resonance imaging. Having a scanner interface presents the opportunity of superimposing the dosimetric information on the anatomical images. It is a simple matter to superimpose dosimetric and anatomic information on a series of images. In most situations, the anatomical images are parallel to one another. The treatment planning computer calculates the doses to a three-dimensional grid of

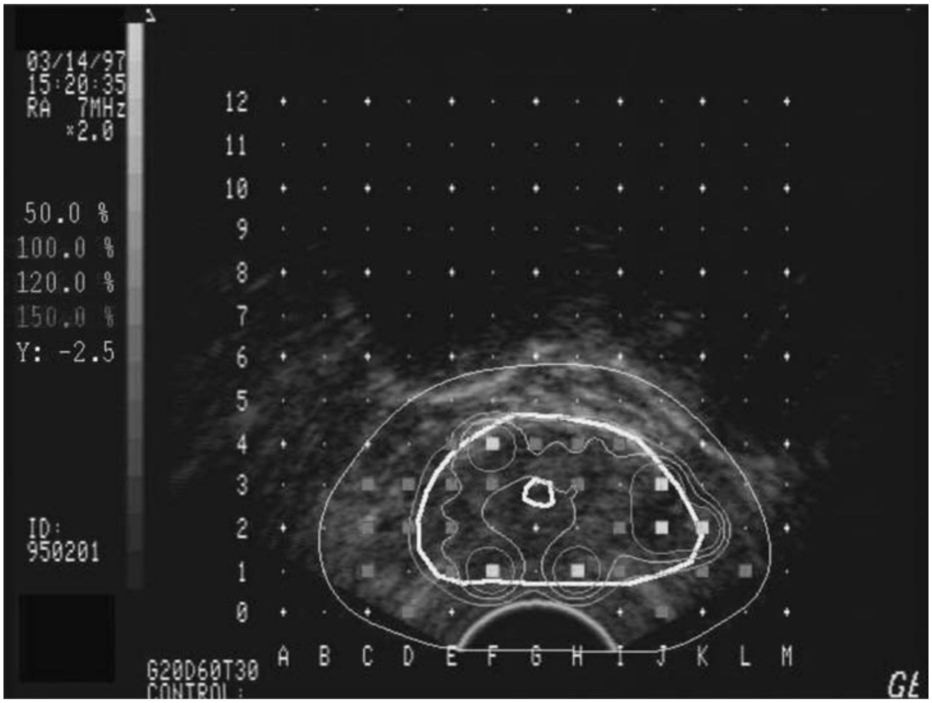


Figure 3. Isodoses superimposed on an ultrasound image of the prostate. The thick white contour line delineates the prostate. The inner isodose lines depict higher dose levels than outer contours. Note the conformity of the green prescription isodose with the shape of the prostate. The inner white contour is the urethra. Note avoidance of the 120% isodose contour with the urethra, a radiosensitive critical structure.

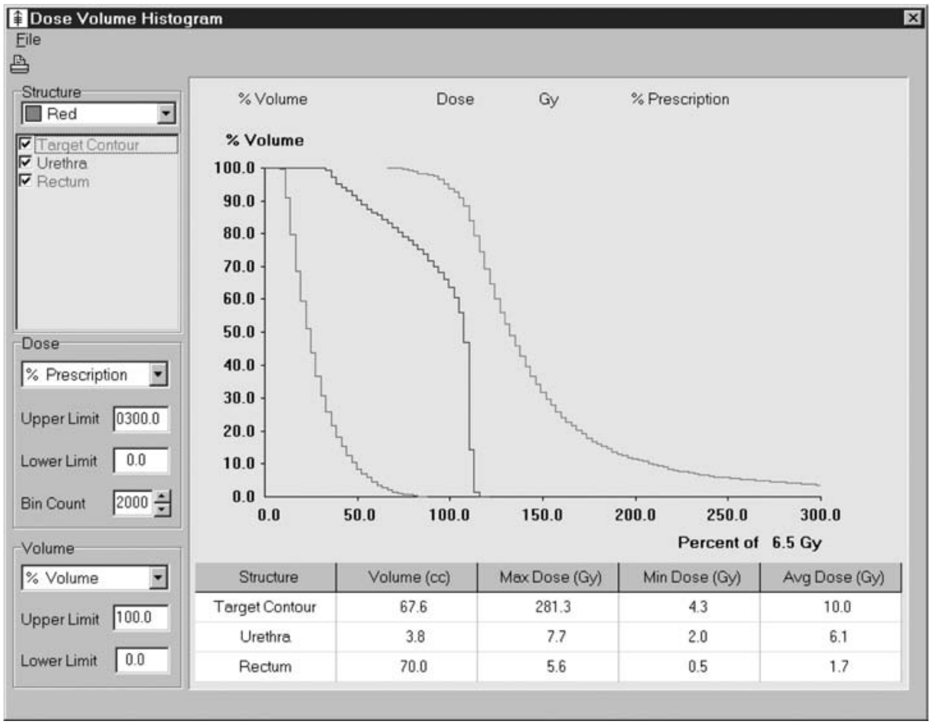
points spatially registered with the anatomic images. After the dosimetric calculations are completed, the dose distributions are represented as a series of colored contour lines commonly known as isodoses. Isodoses are analogous to the contour lines of a mountain on a topographical map. In the case of a mountain, the contour lines represent locations of equal elevation. In a similar manner, isodoses represent the locus of points of equal dose. An example of isodose contour lines superimposed on an axial ultrasound of the prostate is shown in Fig. 3.

In Fig. 3, the outermost yellow contour line represents 50% of the prescribed dose. This is the lowest dose level shown in the figure. Similarly, the innermost lavender contour represents 150% of the prescribed dose. Usually, but not always, the inner isodoses represent higher doses than the outer contours. On a topographical map of a mountain, the inner contour lines represent higher elevations than the outer contours.

Viewing isodose contours superimposed on anatomical images provides a means for visually evaluating the conformity of the dose distributions with the target anatomy. Ideally, the prescription isodose line should exactly conform to the shape of the target volume. This goal is not generally achievable although it can be approached. In Fig. 3, the green contour is the prescription isodose line. There is close conformity of the green isodose line with the prostate. In addition to being able to visually evaluate the conformity of the prescription isodose with the anatomy, it is also possible to ascertain doses received to critical structures. In the case of the prostate, the critical structures where excessive doses should be avoided are the rectum and urethra. Referring to Fig. 3, it is observed that the urethra receives less than 120% of the prescription dose. To properly evaluate a treatment plan, the isodoses on all images must be reviewed. Some slices may manifest excellent conformity and satisfactory critical structure doses, whereas reviewing other images may reveal unsatisfactory dose distributions.

In traditional prostate brachytherapy treatment planning, the physicist manually selects source and needle locations in an attempt to optimize the treatment plan. This entails maximizing the conformity of the prescription isodose with the prostate and minimizing the doses received by the critical structures. This is a tedious, time-consuming process because the conformity and critical structure conditions must be simultaneously met on all slices. Often, improvement on one slice degrades the dosimetry on another. There is no universal standard regarding what constitutes an acceptable treatment plan. There are differing opinions both among individual physicians and treatment centers. These differences generally pertain to acceptable dose values covering the prostate and critical structures. Despite these differences, however, the aim of treatment planning is to maximize the coverage of the prescription dose to the prostate and minimize the doses to the critical structures.

A popular graphical method for evaluating treatment plans is known as a cumulative dose volume histogram, commonly referred to as a DVH. A cumulative DVH is a graph of the volume of a target or critical structure as a function of minimum dose received by the volume. A typical cumulative DVH plot is shown in Fig. 4. Referring to this figure, it is observed that at low doses, the volume covered is essentially 100%. This shows that the target as well as the adjacent critical structures receive a significant dose. In Fig. 4, the blue graph represents the DVH for the urethra. It is observed in this graph that around 95% of the urethra receives doses exceeding 50% of the prescription. As the urethra traverses the prostate, it is physically impossible for the urethra not to receive a significant dose



Structure: Red

☑ Target Contour
☑ Urethra
☑ Rectum

Dose: % Prescription
Upper Limit 0300.0
Lower Limit 0.0
Bin Count 2000

Volume: % Volume
Upper Limit 100.0
Lower Limit 0.0

Structure	Volume (cc)	Max Dose (Gy)	Min Dose (Gy)	Avg Dose (Gy)
Target Contour	67.6	281.3	4.3	10.0
Urethra	3.8	7.7	2.0	6.1
Rectum	70.0	5.6	0.5	1.7

Figure 4. Typical cumulative DVH for a prostate implant.

without sacrificing the dose to the prostate. In well-designed treatment plans, the urethral doses are minimized. The specific criteria depend on the preferences of the physician or institution.

As was briefly indicated, manually optimizing prostate treatment plans is a time-consuming process. To address this issue, the authors developed an automated treatment planning and optimization system. The optimizer used in this system is based on a genetic algorithm (GA) (13,19). GAs are a genre of optimization algorithms that attempt to find the optimal solution of a problem based on evolutionary natural selection (31–34). In the case of the intraoperative optimizer, a population of bit streams is used to represent a population of treatment plans. Each bit stream is a one-dimensional array of numbers whose values are either one or zero. The length of the arrays is equal to the total number of potential source locations in the implant. The three-dimensional coordinates of each potential source location are known and fixed. If a source is present at a potential site, the bit for the corresponding site location is one. Conversely, if a source is not present at a potential site, the corresponding bit in the stream is set to zero. In this way, the treatment planning problem is encoded in a form amenable to genetic optimization.

To determine "how good" is the plan represented by a bit stream, an objective function is defined. The job of the objective function is to calculate a score, representing a figure of merit. The larger the score, the better the treatment plan, as determined by the criteria defined in the objective function. For instance (19), let PD be the prescription dose; the *prostate score* is the number of (uniformity) points that satisfy $\mathrm{PD} \leq D \leq 1.6\ \mathrm{PD}$, the *urethral score* is the number of points in the urethra for which $D \leq 1.2\ \mathrm{PD}$, and the *rectal score* sums up all points in the rectum for which $D \leq 0.78\ \mathrm{PD}$. From this one obtains a *raw score*:

$$\text{Raw score} = 5 \times (\text{prostate score}) + 35 \times (\text{rectal score}) + 50 \times (\text{urethral score})$$

and a *final score* (used in the optimization algorithm), which is a linear function of the raw score [$= A \times (\text{raw score}) + B$].

The objective function is defined to increase the score for greater coverage of the prescription dose to the prostate. On the other hand, as the doses to the rectum and urethra increase, the score is decreased. As can be observed, there are conflicting requirements and the score reflects all of these conditions.

At the beginning of execution, the bit streams are randomly assigned arbitrary bit patterns. During execution of the GA, these bit streams are gradually "evolved" into higher scoring streams, indicating improvements in the treatment plans they represent. In an attempt to improve the scores of a population of treatment plans, the GA mimics biological evolutionary processes. Crossover, mutation, and survival-of-the-fittest are simulated by the GA in an attempt to maximize the scores received by the population. Two bit streams in the population are selected to act as "parents." Crossover is implemented by interchanging corresponding randomly selected bits between the parents. Mutation is affected by inverting a very small (~1%) number of bits in the two bit streams. The modified bit streams are evaluated and replace the two lowest-scoring chromosomes in the population.

In biological evolution, the fitter species has a higher likelihood of survival compared with the lesser fit species. This means that although the lesser fit species has a nonzero probability of survival, species that are better suited to their environments have higher survival probabilities.

Natural selection is simulated in the GA by favoring higher scoring bit streams to act as parents. In 50% of the parent selections, the two highest scoring bit streams are selected to act as parents. In the other 50% of the selections, two bit streams are randomly selected, irrespective of their score. The parents undergo crossover and mutation and replace the two lowest-scoring bit streams in the population. After this process is repeated several thousands of times, a population of high scoring bit streams is generated. This in turn translates into a population of optimized treatment plans. The highest scoring treatment plan is selected to be used for the implant.

The GA can generally generate a treatment plan in under 5 min. Actual execution times depend on the size of the prostate and the number of needles used for implanting the radioactive sources.

Other optimization algorithms used in prostate brachytherapy are simulated annealing (35,36) and branch-and-bound (13–15).

SECONDARY DOSE VERIFICATION

Once the plan is approved by the clinician a second physicist must independently verify and formally approve the plan. An estimate of the total number of seeds can be obtained with the following equations, which give—for a given average dimension of the volume to be implanted, d_{avg}—the total source strength S_K required (37):

For a ^{125}I *permanent* implant and a prescription dose of 144 Gy:

$$\frac{S_k}{U} = \begin{cases} 5.709 \ \left(\dfrac{d_{\text{avg}}}{\text{cm}}\right) & d_{\text{avg}} \leq 3\,\text{cm} \\ 1.524 \ \left(\dfrac{d_{\text{avg}}}{\text{cm}}\right)^{2.2} & d_{\text{avg}} > 3\,\text{cm} \end{cases} \tag{15}$$

For a ^{103}Pd *permanent* implant and a prescription dose of 140 Gy:

$$\frac{S_k}{U} = \begin{cases} 29.41 \ \left(\dfrac{d_{\text{avg}}}{\text{cm}}\right) & d_{\text{avg}} \leq 3\,\text{cm} \\ 5.395 \ \left(\dfrac{d_{\text{avg}}}{\text{cm}}\right)^{2.56} & d_{\text{avg}} > 3\,\text{cm} \end{cases} \tag{16}$$

One evaluates the total number of radioactive sources needed by dividing the total required source strength S_K by the single-seed strength used in that implant. In general, one seeks agreement of 10% or better between the planned and the expected number of seeds.

The plan verification must also determine that the correct prescription dose was used and that input data to the planning software was correctly entered.

DOSE ESCALATION TO PROSTATE SUBVOLUMES

If information is available on tumor burden in specific subvolumes of the prostate, dose escalation to these voxels is recommended. For instance, it has been hypothesized that regions of the prostate where the choline/citrate ratio, as determined by magnetic resonance spectroscopy (MRS), is elevated may contain clinically significant cancer (38–42). In general, "clinically significant" refers to cancer cells that are fast proliferating and/or radioresistant (features associated with local failure) or of high grade and thus with a potentially larger probability of distant dissemination. There is limited evidence of a correlation between choline levels and histological grade (Gleason score). As well, biochemical arguments have been invoked to support the expectation that a larger value of this ratio is expected to reflect an increased rate of cell proliferation although no direct proof exists yet.

As applied at MSKCC, the implementation of an MRS-based dose escalation amounts to increasing the dose to the MRS-positive voxels to 200% of the prescription dose (with no upper limit) while keeping the urethral and rectal dose within the usual range of constraints (14).

POST-IMPLANT ANALYSIS

At MSKCC, post-implant evaluation is performed the day of the implant. The number of seeds implanted is confirmed with a pair of planar radiographic films. A CT study (3 mm slices) is then obtained, and anatomical structures are marked out on each slice. The coordinates of the center of each seed are determined by using appropriate computer software (Interplant Post-implant Analysis System, Version 1.0: Burdette Medical Systems, Inc., Champaign, IL). With the seeds thus identified, (DVHs) are calculated for each structure of interest and compared with the original plan.

CLOSING WORDS: KNOWN PROBLEMS AND POSSIBLE FIXES

Permanent brachytherapy prostate implants are now a well-accepted treatment modality for early stage prostate cancer. The two major limitations of this procedure are higher incidence of urethral complications (when compared with external-beam radiotherapy) and — for some patients — lower than prescribed delivered doses. In this section, we list several unresolved issues that may be responsible for this state of affairs and suggest possible solutions.

Post-implant evaluations of permanent prostate implants often indicate significant differences between the intended plan and its actual implementation. Although an experienced physician can minimize the magnitude of these differences, many factors controlling execution of the plan (e.g., bleeding, tissue swelling) are subject to random fluctuations. This often leads to a higher than intended dose to urethra and rectum and/or lower or higher doses to the prostate, especially at the periphery of the gland. In our view, this discrepancy represents the most important obstacle and challenge that currently needs to be overcome to achieve consistent application of a low urethral and rectal dose range and thereby reduce morbidity after prostate brachytherapy. In a series of recent articles the concept of intraoperative dynamic dosimetric optimization has been proposed (43–46). The idea is to re-optimize the plan several times during the implantation based on the actual positions of the seeds already implanted. The key problem is obtaining in real time (and within a reasonable time interval — 5 min or less) the coordinates of the implanted seeds in the system of reference used for planning. Visualization of the actual seed positions on the intraoperative ultrasound image is difficult, if not impossible, to achieve because of significant artifacts noted on the ultrasound image from needles and/or hemorrhage within the gland.

One can reconstruct seed coordinates from fluoroscopic images taken at three different angles (43,44) or, equivalently, from a CT study obtained in the OR for instance, using a C-arm with CT capabilities. CT-based systems that perform seed segmentation do exist (e.g., Variseed from Varian Medical Systems, Inc, Charlottesville, VA; Interplant Post-implant Analysis System, Burdette Medical Systems, Inc., Champaign, IL), but at this time they do not seem to have the capability of performing this task on the fly and at the same time maintain the required seed-detection reliability.

A second problem concerns the effect of changes in prostate volume (edema shrinkage) as well as seed migration after implantation and the effect this has on a treatment plan that is based on the geometry of the target at the time of implantation. A method of planning that incorporates temporal changes in the target-seed configuration during dose delivery and makes use of the concept of effective volume has been developed by Lee and Zaider (47).

The preceding enumeration of problems has been brief and (admittedly) selective, but we hope to motivate the reader to take a careful look at these important issues. The desideratum of dosimetric conformality in permanent prostate implants remains a topic of active interest in the brachytherapy community, and no doubt the last word on this subject has not yet been spoken.

BIBLIOGRAPHY

1. Coen JJ, Zietman AL, Thakral H, Shipley WU. Radical radiation for localized prostate cancer: Local persistence of disease results in a late wave of metastases. J Clin Oncol 2002;20:3199–3205.
2. Zagars GK, vonEschenbach AC, Ayala AG, Schultheiss TE, Sherman NE. The influence of local-control on metastatic dissemination of prostate-cancer treated by external beam megavoltage radiation-therapy. Cancer 1991;68:2370–2377.
3. Valicenti R, Lu JD, Pilepich M, Asbell S, Grignon D. Survival advantage from higher-dose radiation therapy for clinically localized prostate cancer treated on the radiation therapy oncology group trials. J Clin Oncol 2000;18:2740–2746.
4. Logothetis C. Challenge of locally persistent prostate cancer: An unresolved clinical dilemma. J Clin Oncol 2000;20:3187.
5. Nag S, Shasha D, Janjan N, Petersen I, Zaider M. The American Brachytherapy Society recommendations for brachytherapy of soft tissue sarcomas. Int Radiat Oncol Biol Phys 2000;49:1033–1043.
6. D'Amico AV, Whittington R, Malkowicz SB, Fondurulia J, Chen MH, Kaplan I, Beard CJ, Tomaszewski JE, Renshaw AA, Wein A, Coleman CN. Pretreatment nomogram for prostate-specific antigen recurrence after radical prostatectomy or external-beam radiation therapy for clinically localized prostate cancer. J Clin Oncol 1999;17:168–172.
7. Moul JW, Connelly RR, Lubeck DP, Bauer JJ, Sun L, Flanders SC, Grossfeld GD, Carroll PR. Predicting risk of prostate specific antigen recurrence after radical prostatectomy with the center for prostate disease research and cancer of the prostate strategic urologic research endeavor databases. J Urol 2001;166:1322–1327.
8. Albertsen PC, Hanley JA, Gleason DF, Barry MJ. Competing risk analysis of men aged 55 to 74 years at diagnosis managed conservatively for clinically localized prostate cancer. JAMA 1998;280:975–980.
9. Chodak GW. Comparing treatments for localized prostate cancer-persisting uncertainty. JAMA 1998;280:1008–1010.
10. Jani AB, Hellman S. Early prostate cancer: Clinical decision-making. Lancet 2003;361:1045–1053.
11. Sandhu AS, Zelefsky MJ, Lee HJ, Lombardi D, Fuks Z, Leibel SA. Long-term urinary toxicity after 3-dimensional conformal radiotherapy for prostate cancer in patients with prior history of transurethral resection. Int J Radiat Oncol Biol Phys 2000;48:643–647.
12. Zelefsky MJ, Hollister T, Raben A, Matthews S, Wallner KE. Five-year biochemical outcome and toxicity with transperineal CT-planned permanent I-125 prostate implantation for patients with localized prostate cancer. Int J Radiat Oncol Biol Phys 2000;47:1261–1266.
13. Lee EK, Gallagher RJ, Silvern D, Wuu CS, Zaider M. Treatment planning for brachytherapy: An integer programming model, two computational approaches and experiments with permanent prostate implant planning. Phys Med Biol 1999;44:145–165.
14. Zaider M, Zelefsky MJ, Lee EK, Zakian KL, Amols HI, Dyke J, Cohen G, Hu Y, Endi AK, Chui C, Koutcher JA. Treatment planning for prostate implants using magnetic-resonance spectroscopy imaging. Int J Radiat Oncol Biol Phys 2000;47:1085–1096.
15. Gallagher RJ, Lee EK. Mixed integer programming optimization models for brachytherapy treatment planning. Proc/ AMIA Annu Fall Symp 1997; 278–282.
16. Lee EK, Gallagher RJ, Silvern D, Wuu CS, Zaider M. Treatment planning for brachytherapy: An integer programming model, two computational approaches and experiments with permanent prostate implant planning. Phys Med Biol 1999;44:145–165.
17. Lee EK, Zaider M. Mixed integer programming approaches to treatment planning for brachytherapy. Ann Operat Res Optimizat Med Ann Operat Res 2002;119:147–163.
18. Lee EK, Zaider M. Intraoperative dynamic dose optimization in permanent prostate implants. Int J Radiat Oncol Biol Phys 2003;56:854–861.
19. Silvern DA. Automated OR prostate brachytherapy treatment planning using genetic optimization. 1998.
20. Nag S, Shasha D, Janjan N, Petersen I, Zaider M. The American Brachytherapy Society recommendations for brachytherapy of soft tissue sarcomas. Int J Radiat Oncol Biol Phys 2001;49:1033–1043.
21. Zelefsky MJ, Cohen G, Zakian KL, Dyke J, Koutcher JA, Hricak H, Schwartz L, Zaider M. Intraoperative conformal optimization for transperineal prostate implantation using magnetic resonance spectroscopic imaging. Cancer J 2000;6:249–255.
22. Potters L. Permanent prostate brachytherapy: Lessons learned, lessons to learn. Oncol-New York 2000;14:981–991.
23. Cha CM, Potters L, Ashley R, Freeman K, Wang XH, Waldbaum R, Leibel S. Isotope selection for patients undergoing prostate brachytherapy. Int J Radiat Oncol Biol Phys 1999;45:391–395.
24. Ling CC, Li WX, Anderson LL. The relative biological effectiveness of I-125 and Pd-103. Int J Radiat Oncol Biol Phys 1995;32:373–378.

25. Wuu CS, Zaider M. A calculation of the relative biological effectiveness of 125I and 103Pd brachytherapy sources using the concept of proximity function. Med Phys 1998;25:2186–2189.
26. Nath R, Anderson LL, Luxton G, Weaver KA, Williamson JF, Meigooni AS. Dosimetry of interstitial brachytherapy sources - recommendations of the aapm radiation-therapy committee task group no 43. Med Phys 1995;22:209–234.
27. Handbook of Chemistry and Physics. Boca Raton, FL: CRC Press; 1981.
28. Huda W, Slone R. Review of Radiologic Physics, 1995.
29. Rivard MJ, Butler WM, DeWerd LA, Huq MS, Ibbott GS, Li ZF, Mitch MG, Nath R, Williamson JF. Update of AAPM task group No. 43 report: A revised AAPM protocol for brachytherapy dose calculations. Med Phys 2004;31:3532–3533.
30. Williamson JF, Butler W, DeWerd LA, Huq MS, Ibbott GS, Li Z, Mitch MG, Nath R, Rivard MJ, Todor D. Recommendations of the American Association of Physicists in Medicine regarding the impact of implementing the 2004 task group 43 report on dose specification for Pd-103 and I-125 interstitial brachytherapy. Med Phys 2005;32:1424–1439.
31. Lance Chambers, Practical Handbook of Genetic Algorithms Boca Raton, FL: CRC Press; 1995.
32. Grefenstette JJ. American Association for Artificial Intelligence, Beranek a. N. i. Bolt, Naval Research Laboratory (U.S.), Genetic Algorithms and Their Applications Proceedings of the Second International Conference on Genetic Algorithms, July 28-31, 1987 at the Massachusetts Institute of Technology. Cambridge, MA: Hillsdale, NJ; 1987.
33. Man KF, Tang KS, Kwong S. Genetic Algorithms Concepts and Designs. London: 1999.
34. Zalzala AMS, Fleming PJ. Genetic Algorithms in Engineering Systems. London: 1997.
35. Sloboda RS. Optimization of brachytherapy dose distribution by simulated annealing. Med Phys 1992;19:964.
36. Pouliot J, Tremblay D, Roy J, Filice S. Optimization of permanent I-125 prostate implants using fast simulated annealing. Int J Radiat Oncol Biol Phys 1996;36:711–720.
37. Cohen GN, Amols HI, Zelefsky MJ, Zaider M. The Anderson nomograms for permanent interstitial prostate biplants: A briefing for practitioners. Int J Radiat Oncol Biol Phys 2002;53:504–511.
38. Wefer AE, Hricak H, Vigneron DB, Coakley FV, Lu Y, Wefer J, Mueller-Lisse U, Carroll PR, Kurhanewicz J. Sextant localization of prostate cancer: Comparison of sextant biopsy, magnetic resonance imaging and magnetic resonance spectroscopic imaging with step section histology. J Urol 2000;164: 400–404.
39. Kurhanewicz J, Vigneron DB, Males RG, Swanson MG, Yu KK, Hricak H. The prostate: MR imaging and spectroscopy — Present and future. Radiol Clin North Am 2000;38:115.
40. Scheidler J, Hricak H, Vigneron DB, Yu KK, Sokolov DL, Huang LR, Zaloudek CJ, Nelson SJ, Carroll PR, Kurhanewicz J. Prostate cancer: Localization with three-dimensional proton MR spectroscopic imaging — Clinicopathologic study. Radiology 1999;213:473–480.
41. Kurhanewicz J, Vigneron DB, Hricak H, Narayan P, Carroll P, Nelson SJ. Three-dimensional H-1 MR spectroscopic imaging of the in situ human prostate with high (0.24-0.1-cm(3)) spatial resolution. Radiology 1996;198:795–805.
42. Kurhanewicz J, Vigneron DB, Nelson SJ, Hricak H, MacDonald JM, Konety B, Narayan P. Citrate as an in-vivo marker to discriminate prostate-cancer from benign prostatic hyperplasia and normal prostate peripheral zone — detection via localized proton spectroscopy. Urology 1995;45:459–466.
43. Todor DA, Cohen GN, Amols HI, Zaider M. Operator-free, film-based 3D seed reconstruction in brachytherapy. Phys Med Biol 2002;47:2031–2048.
44. Todor DA, Zaider M, Cohen GN, Worman MF, Zelefsky MJ. Intraoperative dynamic dosimetry for prostate implants. Phys Med Biol 2003;48:1153–1171.
45. Tubic D, Zaccarin A, Pouliot J, Beaulieu L. Automated seed detection and three-dimensional reconstruction. I. Seed localization from fluoroscopic images or radiographs. Med Phys 2001;28:2265–2271.
46. Tubic D, Zaccarin A, Beaulieu L, Pouliot J. Automated seed detection and three-dimensional reconstruction. II. Reconstruction of permanent prostate implants using simulated annealing. Med Phys 2001;28:2272–2279.
47. Lee EK, Zaider M. On the determination of an effective planning volume for permanent prostate implants. Int J Radiat Oncol Biol Phys 2001;49:1197–1206.

See also BRACHYTHERAPY, HIGH DOSE RATE; NUCLEAR MEDICINE INSTRUMENTATION.

PTCA. See CORONARY ANGIOPLASTY AND GUIDEWIRE DIAGNOSTICS.

PULMONARY MECHANICS. See RESPIRATORY MECHANICS AND GAS EXCHANGE.

PULMONARY PHYSIOLOGY

JOHN DEMENKOFF
Mayo Clinic, Dept. of Anesthesia
Scottsdale, Arizona

INTRODUCTION

Present day pulmonary function testing is available in all hospitals and in a less sophisticated form in many physicians' offices. Such was not the case until the 1940s, when the fruits of physiological research dating back 150 years blossomed on the heels of World War II. This so-called golden age of pulmonary physiology spurred many of the currently available lung function tests which are used for diagnosis and treatment of existing lung disease; screening for early pulmonary disease; evaluation of respiratory symptoms such as cough and shortness of breath; performance of disability evaluations; preoperative assessment of thoracic and other surgical patients; determination of level of cardiopulmonary fitness; monitoring of adverse pulmonary effect of certain drug therapies.

Over the years, many have contributed to an understanding of the lung and how it works in health as well as in disease. These discoveries have provided building blocks of knowledge which form the basis of current modern pulmonary function testing.

PRE-1940S

Leonardo DaVinci: This genius drew detailed anatomical illustrations clearly depicting a bellows function of the respiratory muscles.

John Malysed: In 1674, he constructed a model of the chest and lungs with a bladder enclosed inside a simple bellows with the neck outside. With a glass plate on one side, one could watch the bladder inflate and deflate when the bellows operated.

John Hutchinson: In 1848, he developed a spirometer and measured the vital capacity in thousands of normal subjects (1). He also differentiated normal and abnormal results quantitatively, thus ushering in a diagnostic use for pulmonary testing.

Humphrey Davey: Discoverer of hydrogen gas in the early 1800s. This led the way for measuring various lung volumes and compartments other than Hutchinson's vital capacity. Davey built his own spirometer, filled it part way with hydrogen, and breathed it back and forth "for seven quick breaths", finally exhaling fully into the spirometer. Then by measuring the amount and concentration of hydrogen in the spirometer and assuming an equal concentration in his lungs, he calculated the amount of air in his lungs at the end of full exhalation, known today as the residual volume. Modern day lung volume determinations use the inert gas helium with a slightly different protocol, but the fundamental principles remain the same.

Marie Krogh: Prior to 1915, many eminent physiologists believed that oxygen was actively secreted by the lungs into the blood stream. Marie Krogh challenged this popular notion with her diffusion experiments using carbon monoxide. She devised a single breath test in which a subject first fully exhaled to residual volume, then inspired deeply from a spirometer containing 1% carbon monoxide. After an initial exhalation and a six second breath hold, the subject completed a full exhalation. Krogh measured the alveolar gas before and after the six-second breath hold and calculated the uptake of carbon monoxide by the bloodstream.

The amount of CO transfer was noted to be entirely by the process of diffusion and proportional to the pressure differential across the alveolar capillary membrane ($P_1 - P_2$). Because CO binds so tightly to the hemoglobin molecule, P_2 is small. The driving pressure P_1 can be easily calculated. Krogh took advantage of these factors in devising her test, which confirmed the importance of diffusion, not secretion, in the lung.

For many reasons open to speculation, the importance of Krogh's work was not fully appreciated nor developed clinically until the 1940s. The reader is referred to a delightful discourse on such medical curiosities in Ref. 2.

POST-1940S AND THE GOLD AGE OF PULMONARY PHYSIOLOGY

What occurred in the 1940s was a combination of intellect and pluck driven by military contracts and government funds. The resulting research and understanding of lung physiology paved the way for development of current-day pulmonary function laboratories. Some of the more brilliant, resourceful and ingenious researchers of this time are listed below.

Julius Comroe

Chairman, Department Physiology and Pharmacology, University of Pennsylvania, 1946–1957. Director of Cardiovascular Research Institute, University of California, San Francisco, 1957–1983. At both of these institutions, Dr. Comroe developed and fostered world-renowned faculty who studied multiple facets of pulmonary physiology. While at the University of Pennsylvania, Comroe demonstrated his ingenuity by adapting a used surplus bomber nose cone as a body plesthymograph. He wanted to apply Boyle's Law to the measurement of lung volumes, air flow, and airway resistance. His work ushered in modern-day plethysmography. His text, *Physiology of Respiration* (2) remains a classic.

Herman Rahn, Wallace O. Fenn, Arthur Otis

These remarkable men formed the core of a research effort at the University of Rochester. An account of this creative ground work is found in Ref. 3, and is rich in historical facts. In the 1940s, pneumotachographs had to be fabricated by individual research groups. In the Rochester group's first model, a cluster of soda straws encased in a brass tube served as the flow resistance element. In later versions, they used as resistive elements glass wool enclosed in a lady's hair net. Their contributions are evident today, as many of their postdoctoral fellows and research associates have gone on and taught the next generation of pulmonary specialists.

Andre Frederick Cournand, Dickinson Woodrow Richards

Both rshared shared the 1956 Nobel Prize in medicine and physiology, and formed the famous Bellevue Hospital Cardiopulmonary Lab at Columbia University. Their observations regarding prolonged nitrogen washout in the lungs of emphysematous patients fostered the clinical use of diagnostic pulmonary function tests. They also established normal values and formulated testing protocols. They pioneered catheterization of the right heart, making way for analysis of mixed venous blood and more accurate cardiac output and pulmonary blood flow via the direct Fick technique.

$$\text{Pulmonary blood flow} = \frac{O_2\ \text{consumption}}{\text{Arterial-mixed venous}\ O_2\ \text{difference}}$$

Current interventional cardiology, and the understanding of complex interrelatedness of pulmonary diseases on the heart, stem from these studies done in the 1950s at Columbia.

Since the mid-1960s, pulmonary function testing has evolved more slowly. Tests that are reproducible, well tolerated by patients, and offer helpful clinical information have been further refined by advances in instrumentation and computerization.

With the advent of rapidly responding gas analyzers, highly accurate and calibrated pneumotachographs, and sophisticated computer software, the study of lung function during exercise has become possible. The complex interactions of metabolic-cardiopulmonary systems is discussed below, in the section on exercise physiology. While a boon to performance-minded athletes, these tests also shed light on limitation of exercise tolerance due to diseases of the heart and lung.

From the time of DaVinci to the present, great strides have been made in the understanding of lung function and its measurement. Now simple acts, such as blowing out a candle or coughing, are known to be dependent on elastic recoil of the lungs and complex airways dynamics. Both properties of the lung are measured with pulmonary function testing.

Each test discussed in the following text carries with it a rich historical and intellectual story line.

PHYSIOLOGICAL PRINCIPLES UNDERLYING MODERN PULMONARY FUNCTION TESTS

The interpretation and analysis of pulmonary function tests is often conveyed in physiological terms rather than as specific medical diagnoses. As such important underlying physiological concepts are presented that will provide a deeper understanding of pulmonary function test results. Many of these concepts have been developed and refined over time and represent a legacy of scientific achievement.

SINGLE BREATH NITROGEN WASHOUT

Aptly named, this test measures the nitrogen concentration of a normal exhalation after a deep inhalation of 100% oxygen. It was developed by Fowler in 1948 to measure the anatomic dead space $V_{\mathrm{danatomical}}$.

During normal tidal breathing, a part of each breath remains in the conducting airways of the upper airway and tracheobronchial tree. It never reaches the alveoli; therefore it does not participate in gas exchange and is referred to as anatomical dead space. The fraction of total ventilation (O_{E}) that reaches gas exchanging space of alveoli is called alveolar ventilation or O_{A}.

$$O_{\mathrm{A}} = O_{\mathrm{E}} - f \times V_{\mathrm{danatomical}}$$
$$\text{where } f = \text{respiratory frequency}$$

In Fowlers method (Fig. 1), a simultaneous recording of nitrogen and exhaled volume is made after a deep inhalation of pure oxygen.

At the start of expiration, the gas comes from the anatomical dead space, which contains no nitrogen. Along the course of the S-shaped N_2 washout, a front between the alveolar air and dead space air can be determined (see Graphical depiction below).

The anatomical dead space is related to body weight and is ~ 150 mL for a normal man. The extrathoracic fraction, mouth and pharynx, contributes 66 mL with a range of 35–105 mL, depending on jaw and neck position. Anatomic dead space represents an inefficiency of the design of the respiratory system. With each breath, a significant volume of air must be moved, requiring work, for which no benefit is derived.

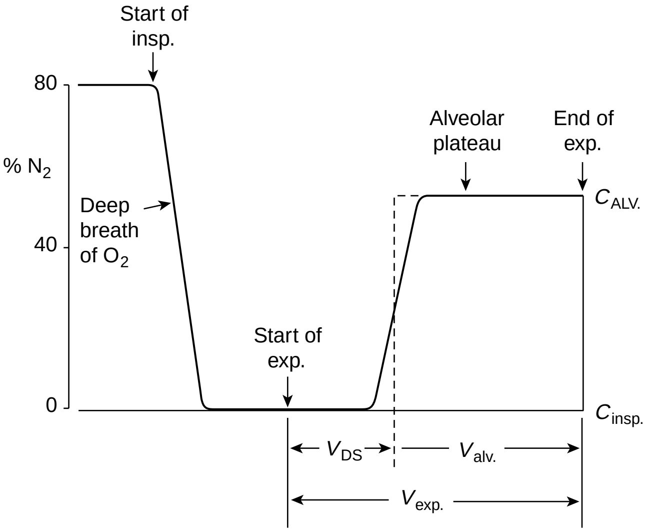


Figure 1. Fowler's method. For determination of anatomic dead space. See text for description. The flat portion of the curve is called the alveolar plateau and represents pure alveolar gas.

PHYSIOLOGICAL DEAD SPACE BOHR EQUATION

Inefficiencies also occur at the level of the alveoli where some air reaches the alveoli, but gas exchange never occurs. For example, the upper lobes in a normal resting upright lung are well-ventilated, but not perfused.

This is wasted ventilation, and when added to the anatomic dead space is designated as the physiological dead space. This can be measured by application of the Bohr equation (4). If one complete expiration is collected in a bag, the amount of carbon dioxide is $[F_{\mathrm{E}}\mathrm{CO}_2 \times V_{\mathrm{T}}]$. This volume of CO_2 comes partly from the nonexchanging dead space, which has a volume from the inspired air $[V_{\mathrm{D}} \times F_{\mathrm{I}}\mathrm{CO}_2]$, plus the volume from alveolar gas $F_{\mathrm{A}}\mathrm{CO}_2 \times [V_{\mathrm{T}} - V_{\mathrm{D}}]$.

$$[F_{\mathrm{E}}\mathrm{CO}_2 \times V_{\mathrm{T}}] = F_{\mathrm{I}}\mathrm{CO}_2 \times V_{\mathrm{D}} + F_{\mathrm{A}}\mathrm{CO}_2[V_{\mathrm{T}} - V_{\mathrm{D}}]$$

$$F_{\mathrm{A}}\mathrm{CO}_2 = \frac{V_{\mathrm{T}} \times F_{\mathrm{E}}\mathrm{CO}_2 - V_{\mathrm{D}}F_{\mathrm{I}}\mathrm{CO}_2}{V_{\mathrm{T}} - V_{\mathrm{D}}}$$

$$V_{\mathrm{D}} = \frac{[F_{\mathrm{A}}\mathrm{CO}_2 - F_{\mathrm{E}}\mathrm{CO}_2]V_{\mathrm{T}}}{F_{\mathrm{A}}\mathrm{CO}_2 - F_{\mathrm{I}}\mathrm{CO}_2}$$

If inspired, CO_2 is zero then $F_{\mathrm{I}}\mathrm{CO}_2 = 0$.

Hence,

$$V_{\mathrm{D}} = \frac{F_{\mathrm{A}}\mathrm{CO}_2 - F_{\mathrm{E}}\mathrm{CO}_2}{F_{\mathrm{A}}\mathrm{CO}_2} \quad V_{\mathrm{T}}$$

$$\frac{V_{\mathrm{D}}}{V_{\mathrm{T}}} = \frac{F_{\mathrm{A}}\mathrm{CO}_2 - F_{\mathrm{E}}\mathrm{CO}_2}{F_{\mathrm{A}}\mathrm{CO}_2}$$

$F_{\mathrm{A}}\mathrm{CO}_2$ = Alveolar CO_2 fractional concentration measured by obtaining an alveolar sample.

F_ECO_2 = Mixed expired CO_2 fractional concentration measured from a collection of expired air; Douglas bag or mixing chamber.

The parameter V_D/V_T is called the dead space to tidal volume ratio where the dead space is physiologic and includes the anatomic dead space. It is an efficiency rating, typically 0.3 in normals and up to 0.5 or so in patients with emphysema. In the latter case, 50% of the breath is wasted and does not participate in gas exchange.

FORCED EXPIRATION AND DYNAMIC COMPRESSION

The most familiar maneuver that utilizes a maximal expiratory effort is a cough. The resulting dynamic airway compression facilitates clearance of bronchial secretions. Even at maximal exercise, such flow rates are not attained, thus demonstrating an impressive reserve in flow characteristics of the lung.

The complex mechanics of forced exhalations were elucidated by the work of Hyatt, Schilder, and Fry. They described a maximal expiratory flow volume curve, where instantaneous expiratory flow is plotted against volume instead of time (as is done with FEV_1). Flow reaches a maximum at 80% of vital capacity and reaches zero at residual volume. The curve is shown to be effort-dependent >75% of VC and effort-independent <75%.

Once dynamic compression occurs, the lung behaves as a Starling Resistor (Fig. 2). The flow then depends on the elastic recoil of the lung and airway resistance upstream from the compressed lung segment. Under these conditions, an increase in effort produces no increase in flow.

DIFFUSION AND DIFFUSING

The purpose of the lung is to deliver oxygen to the blood stream and remove the byproduct of metabolism, carbon dioxide. This process begins with mass transport of oxygen down conducting tubes of diminishing caliber called bronchi, bronchioles, terminal bronchioles, and finally air sacs or alveoli. Simple diffusion then occurs at the interface between the walls of the alveoli and pulmonary capillaries. The 300 million or so alveoli in the human lung create a surface area for diffusion of 85 m^2. The rat at which gas, either oxygen or carbon dioxide that transverses this membrane follows Fick's law of diffusion and is proportional to the surface area of the sheet and inversely proportional to its thickness.

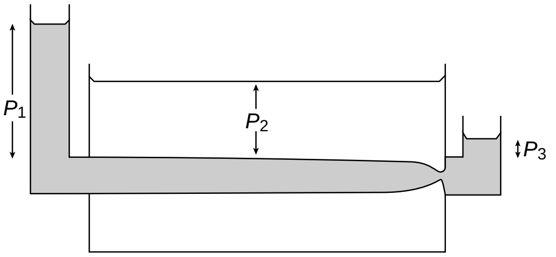


Figure 2. A Starling Resistor that is a mechanical analog for dynamic compression of the airways. The collapsible tubing in the chamber represents small airways. The pressure P_2 is pleural pressure during a forced vital capacity maneuver which collapses the airways at the equal pressure point, that is, $P_2 > P_3$. The pressure P_1 represents the elastic recoil of the lungs. Flow is proportional to P_1–P_2.

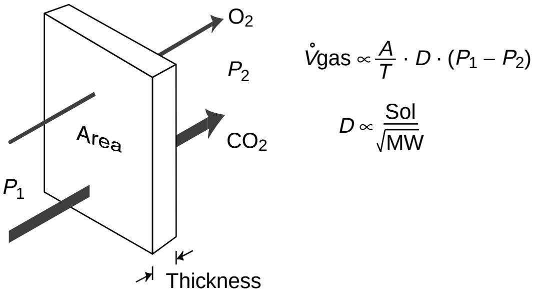

O_2
P_2
CO_2
Area
P_1
Thickness
$\overset{\circ}{V}gas \propto \frac{A}{T} \cdot D \cdot (P_1 - P_2)$
$D \propto \frac{Sol}{\sqrt{MW}}$

Figure 3. The process of diffusion and Fick's law of diffusion. Within the lung P_1 would represent the alveolar space and P_2 the capillary space.

The diffusion coefficient is proportional to the solubility of the gas and inversely proportional to the square root of the molecular weight (Fig. 3).

$$\dot{V}_{gas} = \frac{A}{T} \times D(P_1 - P_2)$$
$$D = \frac{Sol}{\sqrt{MW}}$$

where MW = molecular weight

When applying this formula to the special case of oxygen diffusion the partial pressure of oxygen in the alveolus (P_AO_2) or driving pressure (P_1) is 100 mmHg (13.2 kPa) while that in the pulmonary capillary (P_2) is 40 mmHg (5.3 kPa). The amount of oxygen transferred depends in part on this pressure differential ($P_1 - P_2$); 100 mmHg (13.2 kPa) minus 40 mmHg (5.3 kPa), which results in a diffusion gradient of 60 mmHg (7.9 kPa). In addition the thickness of the alveolar-capillary membrane (normally 0.3 μ) appears to be of equal importance. When this membrane is thickened by disease states, such as pneumonia, pulmonary fibrosis, asbestosis, or silicosis, oxygen transfer is considerably impaired.

The complexity of oxygen diffusion becomes apparent when one considers that red blood cells carrying the hemoglobin molecule typically spend only 0.75 s in the pulmonary capillary. Given a normal driving pressure (P_1–P_2) of 60 mmHg (7.9 kPa) and a healthy alveolar-capillary membrane (0.3 μm thick) equilibrium ($P_1 = P_2$) will occur in 0.25 s. In other words, as blood exits the gas exchange space it will have gone from a PO_2 of 40 mmHg (5.3 kPa)–100 mmHg (13.2 kPa) rapidly.

Various lung diseases can compromise the elegant process of diffusion outlined above. In some the alveolar oxygen level is reduced thereby diminishing the driving pressure. In others diffusion is impaired by a thickened alveolar capillary membrane. And finally, with exertion, as blood flow increases, the time available to load oxygen onto the hemoglobin molecule is reduced. Any one or combination of these factors can lead to a reduction in the diffusion of oxygen.

Having said this, the actual measurement of the diffusing capacity of the lung for oxygen as a clinically useful pulmonary function test has proved difficult to develop. This is due to the fact that the capillary oxygen pressure is ever increasing as P_2 approaches P_1 creating a back pressure thus slowing diffusion as red cells travel along the pulmonary capillary bed. The parameter P_2 can be calculated through a very complex and cumbersome integration method. In the final analysis, oxygen transfer then is actually significantly dependent on total flow of pulmonary capillary blood rather than diffusion alone.

Unlike oxygen, carbon monoxide transfer is diffusion-limited because it binds so tightly to hemoglobin that the partial pressure of CO in pulmonary capillary blood is low. There is little back pressure, so the amount of carbon monoxide transferred is related only to the driving pressure P_1 ($P_2 \approx 0$), which is the alveolar pressure P_A of carbon monoxide.

$$\underline{O_{\text{gas}}} = \left(\frac{A}{T} \times D\right)(P_1 - P_2)$$

$$\frac{O_{\text{gas}}}{P_1} = \frac{A}{T} \times D$$

The above equation is simplified as follows, where D_L is called the diffused capacity of the lung and includes the area, thickness, and diffusing properties of the sheet and the gas concerned.

$$D_{\text{LCO}} = \frac{A}{T} \times D = \frac{O_{\text{CO}}}{P_1\text{CO}} = \frac{O_{\text{CO}}}{P_A\text{CO}}$$

$$P_1 = \text{Driving pressure} = P_A\text{CO} = \text{Alveolar CO}$$

$$P_2 = 0$$

$$D_{\text{LCO}} = \frac{O_{\text{CO}}}{P_1 - P_2} \qquad D_{\text{LCO}} = \frac{O_{\text{CO}}}{P_A\text{CO}}$$

GAS LAWS

By convention, pulmonary function test results are expressed either at body temperature and ambient pressure, saturated (BTPS), ambient temperature and pressure, saturated (ATPS), or standard temperature and pressure, dry (STPD). A working knowledge of gas laws is essential for accurate conversion from one state to another.

Gas inside the lung is at BTPS. The pressure is the barometric pressure (P_B), and saturated refers to the saturated water vapor pressure (P_W), which is a function of temperature. At normal body temperature (37 °C), P_W is 47 mmHg (6.2 kPa).

Gas measured in the equipment is at ambient temperature, dry (ATD) if the expired water vapor is absorbed prior to the measurement or if inspired gas is from a cylinder. Alternately, it is called ATPS if expired gas is collected, but the water vapor was not absorbed. At normal room temperature (25 °C), P_W is 22 mmHg (2.9 kPa).

Inspired gas from the atmosphere is ordinarily between ATPD and ATPS. Since buildings typically are at 50% relative humidity, P_W is 50% of 22 mmHg (1.4 kPa) or 11 mmHg (1.4 kPa).

Boyle first published his ideal gas law in 1662. It states that, for a given mass of gas at constant temperature, the volume varies inversely with the pressure:

$$PV = RT$$

Charles's law states that, for a given mass of gas at constant pressure, the volume varies directly with the absolute temperature. Thus V/T is a constant, where T designates a temperature on the absolute or kelvin scale.

Combining both these gas laws gives an approximation of real gases under various conditions.

$$\frac{P_1V_1}{T_1} = \frac{P_2V_2}{T_2}$$

Usually called the ideal gas law.

Suppose that a patient expires into a Douglas bag, which is then transferred to a laboratory at 20 °C and squeezed through a dry gas meter. From a knowledge of the number of expirations collected and the respiratory frequency, the volume of gas at 20 °C corresponding to the minute volume ventilation O_E can be calculated. If the minute volume was 6 L at ATPS then, by use of the combined gas law equation, volumes can be adjusted to BTPS and STPD.

Assume that the patient's body temperature is 37 °C, the saturated water vapor pressure is 18 mmHg (2.4 kPa) at 20 °C and 47 mmHg (6.2 kPa) at 37 ° C and that the barometric pressure is 760 mmHg (101 kPa). Because water vapor does not follow the ideal gas law, its partial pressure is subtracted.

$$\frac{(760 - 18)}{273 + 20} \times 6 = \frac{(760 - 47) \times V_2}{273 + 37}$$

So that

$$V_2 = \frac{742 \times 6 \times 310}{713 \times 293} = 6.61\,\text{L BTPS}$$

The same volume of 6 L measured under atmospheric conditions would represent 5.45 L under STPD, that is, 760 mmHg (101 kPa) and 0 °C.

$$\frac{(760 - 18)}{273 + 20} \times 6 = \frac{760}{273} \times V_3$$

$$V_3 = \frac{742 \times 6 \times 273}{760 \times 293} = 5.45\,\text{L ATPD}$$

BTPS is used for lung volumes and ventilation O_E, ATPS for maximal inspiratory and expiratory flow, and STPD for oxygen consumption and carbon dioxide output.

CALCULATION OF OXYGEN UPTAKE

Oxygen uptake is the difference between oxygen breathed in and the amount in the exhaled air.

$$O_{O_2} = (O_I \cdot F_I\text{O}_2) - (O_E \cdot F_E\text{O}_2)$$

Where O_{O_2} is the oxygen uptake in liter per minute; O_I is the inspired minute volume (L·min^{-1}), $F_E\text{O}_2$ is the mixed expired oxygen fraction, and $F_I\text{O}_2$ is the inspired oxygen fraction. Because the volume of inspired air is slightly greater than expired air (more O_2 consumed than carbon dioxide, CO_2, is produced), a correction factor using measured nitrogen is used.

$$O_I = O_E\left(\frac{F_EN_2}{F_IN_2}\right)$$

This has been attributed to the British researcher and is referred to as the Haldane Transformation. It is used to calculate the inspired volume when only O_E is measured, the latter being much easier to measure than the former.

Substituting this correction factor into the original equation,

$$O_{O_2} = \left(O_E \times \left(\frac{F_EN_2}{F_1N_2}\right) \times F_1O_2\right) - (O_E \times F_EO_2)$$

since $F_EN_2 = (1 - F_EO_2 - F_ECO_2)$, this becomes

$$O_{O_2} = \left(O_E \frac{(1 - F_EO_2 - F_ECO_2) \times 0.2093}{0.7904}\right) - O_E \times F_EO_2$$

reducing to

$$O_{O_2} = O_E((1 - F_EO_2 - F_ECO_2) \times 0.265) - (F_EO_2)$$

By convention, O_{O_2} is expressed under standard conditions (STPD).

During a standard cardiopulmonary exercise stress, all the variables on the right side of the equation are measured as follows:

O_E	Douglas bag for collection or Pneumotachograph interfaced with a computer for exercise testing
F_EO_2	Measured from Douglas bag at rest or Mixing chamber for exercise testing
F_ECO_2	Measured from Douglas bag at rest or Mixing chamber for exercise testing

Calculation of carbon dioxide output O_{CO_2}

$$O_{CO_2} = O_E \times F_ECO_2$$

Because there is little CO_2 in inspired air this, calculation becomes much simpler. Again by convention, O_{CO_2} is also expressed under STPD.

Respiratory Exchange Ratio (R).

$$R = \frac{O_{CO_2}}{O_{O_2}}$$

This value is typically 0.8 during the steady state of respiration and represents the ratio of CO_2 produced to oxygen consumed by the metabolic pathways of the cell. The value of R is fixed depending on the primary source of fuel being metabolized. Pure carbohydrate gives a ratio of 0.7 and fat burns at a ratio of 1.0. A typical ratio is 0.8 and represents a mixture of the two food groups being metabolically consumed.

In the nonsteady state, the amount of CO_2 exhaled rapidly changes based on the level of hyper or hypoventilation, so R may vary from 0.6 to 1.4. In addition, CO_2 produced by bicarbonate buffering of lactic acid adds to the O_{CO_2} produced by metabolism during peak exercise. This will be discussed further in the section on cardiopulmonary exercise testing.

The measurement of O_{O_2}, O_{CO_2}, and the ratio O_{CO_2}/O_{O_2} provide important information on assessing overall lung function, at rest and especially during exercise testing.

INSTRUMENTATION

Volume Measuring Devises

In order to calculate minute ventilation (O_E), and other derived variables such as O_{O2} and O_{CO_2} the expired volume over time is collected in a Douglas bag or meteorological (Mylar) balloon. So collected, the expired gas is then connected and emitted into a large spirometer, such as the 120 L Tissot spirometer, and the volume is measured by use of a calibration factor. The Tissot spirometer is a typical water-filled spirometer, but due to its size and the considerable inertia of the bell, it is not used for measuring tidal breathing. Smaller water-filled spirometers (9–13.5) liters have a lower airway resistance and an appropriate response time (up to 20 Hz) needed to measure forced exhalation. All water-sealed spirometers, regardless of size, are configured similarly and operate on the same principles (Fig. 4). A bell is sleeved between the inner

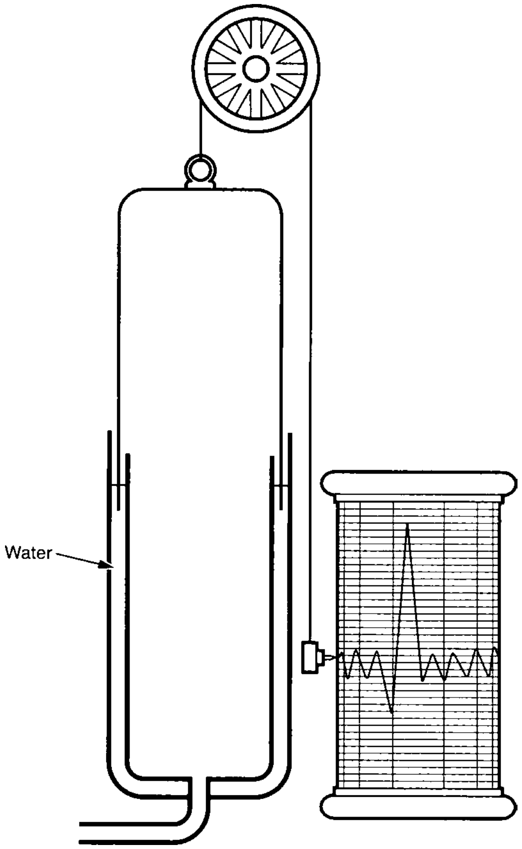


Figure 4. Water filled spirometer connected to a rotating Kymograph.

and outer housing. Water fills the space between the inner cylinder and outer housing, providing an airtight seal for air entering the bell. Rigid tubing connects the inner cylinder to exhaled air from the patient or collection bag. A CO_2 absorbant is placed in circuit when rebreathing maneuvers are carried out, such as resting metabolic rate or FRC determinations. When forced maneuvers, such as MVV, FVC, FEV_1 or PEFR are accomplished, the absorbant is removed, thereby reducing resistance in the expiratory system.

As air enters or leaves the spirometer the chain-suspended bell rises and falls. These movements are recorded by means of pens moving in parallel. A kymograph drum turns at a preselected pace, adding a time dimension to the volume changes. This allows measurement of the based variables such as MVV, FEV_1, FEF_{25-75}, PEF.

Another type of spirometer is the so-called dry rolling seal, also called the Ohio spirometer. A horizontal cylinder is attached to a flexible rolling seal. As air enters, the rolling seal allows the cylinder to move horizontally. Linear transducers attached to the cylinder are interfaced with a computer, allowing measurement of volume over time and flow.

Dry gasmeters typically measure inspired air to avoid accumulation of moisture on a bellows mechanism. The movement of the bellows is transmitted to a circular dial that is labeled with appropriate volumes.

A spirometer that uses a wedge-shaped bellows is called a wedge spirometer (5). The bellows expands and collapses as gas moves in and out. One side is stationary, while the other side moves a pen that records the changes. Pressure activation moves the chart horizontally, giving a time domain to the recording.

The peak flow meter (6) is a spirometer that works on a completely different principle from other spirometers. It is known as a variable orifice meter (Fig. 5), popularized as rotameter gas flow meter on anesthesia machines (7). As air enters the flow meter, a bobbin or light-weight marker is entrained in the vertical column of air. The flow meter has a variable inner orifice dimension that increases with height. The bobbin records the peak flow M, which corresponds to a particular inner orifice (r). The original peak flow meter was developed by F. M. Wright of England in 1959 and is often referred to as the Wright peak flow meter (6). It is based on Poiseuille's equation.

$$M = \frac{\pi P r^4}{8 n \ell} \qquad \begin{array}{l} M = \text{flow} \\ r = \text{radius} \end{array}$$

FLOW AND VOLUME TRANSDUCERS

Instantaneous flow signals generated from flow transducers discussed below can be integrated with respect to time, thereby obtaining volume measurements. Harmonic analysis of respiratory flow phenomena has shown significant signals out to 20 C.P.S., requiring all devices to respond with fidelity at this frequency (8).

The Fleish Pneumotach (9) quantifies airflow by measuring the pressure drop across an in-line obstruction, such as mesh or porous membrane (Fig. 6). The pressure drop follows Poiseuille's law and is for laminar or nonturbulent flow. To prevent nonlaminar flow, various size pneumotachographs are used for different settings, such as studying children or exercising adults.

A Pitot tube that utilizes the Venturi effect is another type of flow meter (Fig. 7). The pressures of two tubes, one facing and one perpendicular to the air stream, is measured with a differential pressure transducer. Air flow velocity is proportional to the density of the gas and to the square of the pressure difference. They do not depend on laminar flow, typically are low weight and as opposed to pneumotachograms are low resistance breathing circuits.

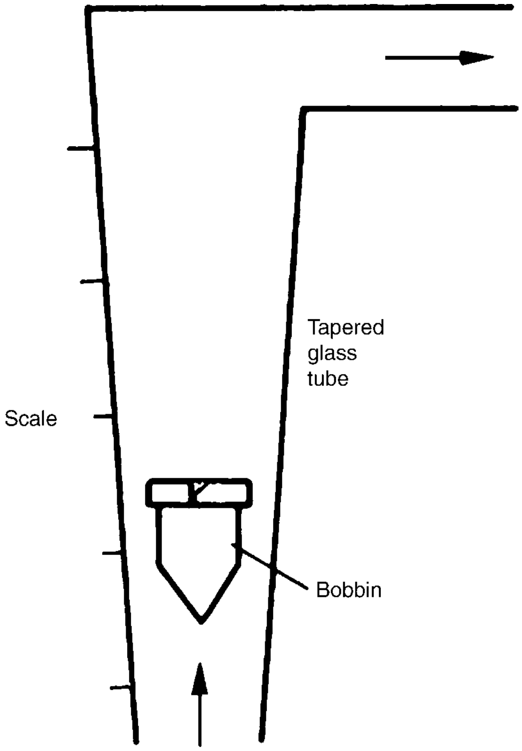


Figure 5. A variable orifice flow meter.

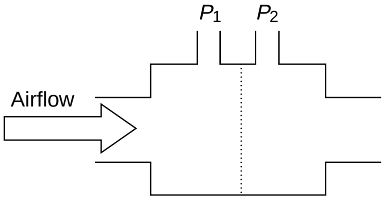


Figure 6. Fleisch pneumotachograph.

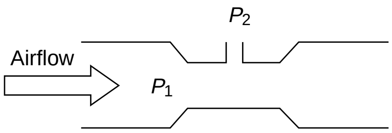


Figure 7. Pitot tube.

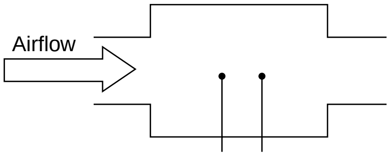


Figure 8. Hot wire anemometer.

Hot-wire anemometers measure mass flow by detecting the increase in current needed to heat a hot wire placed in the air stream as air flows over and cools the wire (Fig. 8).

A turbine transducer uses a low mass helical impeller mounted on bearings. As the impeller blade spins with airflow, an interposed light beam is broken and digital signals proportional to the flow are sent to the pressure sensor (Fig. 9).

The accuracy of each of these flow meters is potentially affected by the temperature, viscosity and density of the gas measured as well as the flow character (laminar or turbulent). When proper calibration is maintained, these devices produce a ± 3% accuracy, as recommended by the American Thoracic Society Guidelines (10).

GAS ANALYSIS

Rapidly responding (<100 ms) gas analyzers have made breath-by-breath analysis possible. Such measurements of expired oxygen and carbon dioxide give dense data useful in interpretation of cardiopulmonary exercise stress tests. When speed of analysis is not essential, chemical analysis by the Scholander or Haldane methods provide accurate results and are considered the gold standard. Gases measured by this method are used to validate other calibrating gases.

OXYGEN ANALYZERS

Discrete oxygen analyzers commonly used are paramagnetic, fuel cell or zirconium oxide. Each are calibrated over the expected range of measurement (e.g., 12–21%) by validated control gases. Of the three mentioned, the later two respond very quickly and so are used in breath-by-breath analysis. Paramagnetism is a distinctive property of oxygen: The molecules aligning in a magnetic field and thus enhancing it. A typical use of this slower responding analyzer is measuring oxygen concentration in large collecting bags or mixing chambers. Of note, oxygen does not have suitable absorption bands in the ultraviolet (UV), infrared (IR), or visible wavelengths. The following sections discuss gases that do have these properties.

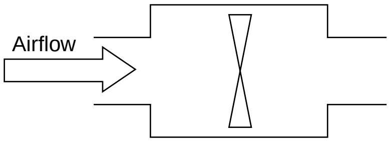


Figure 9. Turbine.

NONDISPERSIBLE INFRARED GAS ANALYZERS

This instrumentation (7,11) is used for multiple polyatomic gases including CO_2 and CO, commonly measured in pulmonary function testing. An IR beam is directed alternately through a reference and measurement cell. By means of a chopper wheel, a detector senses the alternating change in absorption of selected IR wavelengths. This signal is amplified with a high input impedance ac amplifier, rectified and displayed on a meter or digital recorder.

NITROGEN METER

The fact that nitrogen molecules can be excited in a low pressure electric discharge to emit visible light in the purple region forms the basis of the nitrogen meter (7,11). A 1500 V electric potential difference is maintained and optical filters select appropriate wavelengths in the violet range. The resulting light intensity is measured by a photocell with an amplifier.

MASS SPECTROMETER

Used primarily in research labs, mass spectrometry is capable of analyzing any gas with speed, specificity, sensitivity, and accuracy unmatched by any other method. Molecules of the sample are ionized at low pressure by a beam of electrons, and the ions are deflected in a circular path by a magnetic field. The stream of particles splits into beams of different molecular weight, any one of which can be detected by a suitably placed collector. Due to expense, mass spectometry is not a typical part of clinical (hospital or office-based) pulmonary function testing.

PULMONARY FUNCTION TESTS

Pulmonary function tests do not provide a complete diagnostic picture. At best, they support a clinical impression that is formed by a thorough history, physical exam, and X-ray studies. Given the myriad of lung function tests available, an informed decision on the most important ones to order maximized their usefulness.

SPIROMETRY

A forced exhalation maneuver after a deep inhalation is recorded by a moving kymograph on a small water-filled spirometer. A pneumotachograph with computer interfacing could be used with equally acceptable results. A tracing of volume over time is obtained and the following measurements are derived (Fig. 10). This is called a forced vital capacity maneuver.

FVC	Forced vital capacity. This value is effort dependent and depends on the full cooperation of the patient.
FEV_1	From the tracing, the amount of air exhaled in the first second is measured. Patient effort is required to obtain a reliable value.

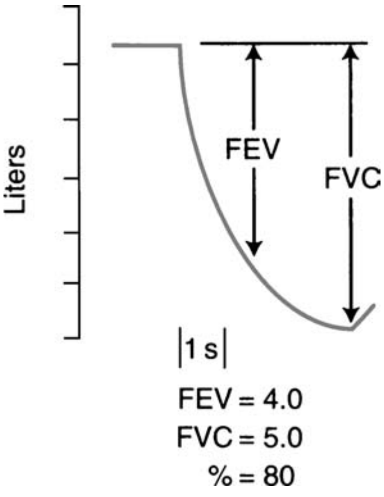


Figure 10. A normal spirographic tracing volume versus time.

FEV_1/FVC	This ratio is helpful in ascertaining airflow obstruction and is typically 70% in normal people.
FEF_{25-75}	This is a flow rate and represents the slope $\Delta V/\Delta T$ during the mid-section of the spirometer tracing. It is where dynamic compression of airways occurs (Fig. 11).
PEFR	Peak expiratory flow rate. The steepest slope on the curve, typically at 80% of the vital capacity maneuver.
MMEF	Mid-maximal expiratory flow rate is a slope taken at 50% of the vital capacity maneuver. It reflects smaller airways airflow.

All along the spirometric curve, an infinite number of slopes can be determined, from which a flow–volume curve could be constructed. However, a flow signal from a pneumotachograph plotted against time is the preferred method of generating this data.

SVC	Slow Vital Capacity. Instead of a forced maneuver, the vital capacity may be performed with less effort. In normals, the FVC and SVC are equal. Because of air trapping, patients with emphysema more fully exhale with the SVC.

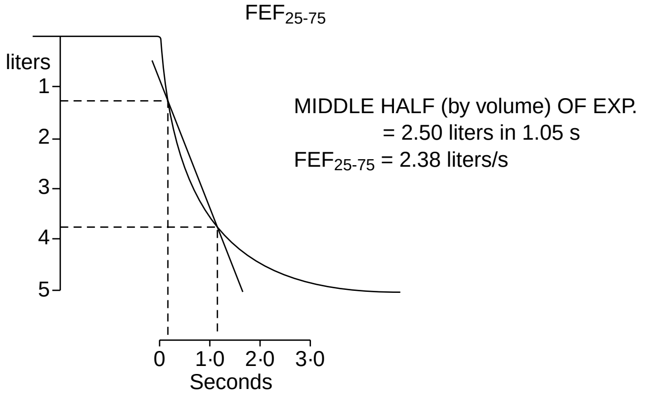


Figure 11. The FEF_{25-75} Slope $\Delta V/\Delta T$ derived from spirogram and expressed in liters per second.

FLOW–VOLUME CURVES

Spirometers and flow-volume curves contain the same information, displayed differently. Below is a typical flow–volume curve generated by a pneumotachograph (Fig. 12). A forced expiratory maneuver is performed. Pneumotachographs can measure peak inspiratory flow volume curves, which are important in diagnosing obstructive supraglottic lesions. Expiratory flow-volume curves can unmask central airway tumors not appreciated on simple spirometry at 50% FVC nor seen on standard chest X ray (12).

This curve reflects phenomena in small airways where dynamic compression occurs. Normally, flow here is dependent on gas density. This fact forms the basis for flow–volume measurements after inhalation of 80% helium–20% oxygen mixtures. When overlayed, the flow–volume curves of smokers show little difference after breathing low density helium–oxygen compared to room air (79% nitrogen–21% oxygen). The density independence reflects increased resistance in diseased small airways and is a very sensitive early indication of smoking-induced lung disease.

LUNG VOLUME

There are three methods for measuring the lung volume FRC or functional residual capacity: body plethysmography, nitrogen washout(13), or helium dilution. Once the FRC is measured, residual volume can be determined by asking a patient to exhale completely from FRC to residual volume (RV). The air left in the lung is the residual volume. The total lung capacity is then determined by measuring a deep inhalation from RV (inspiratory capacity) and adding

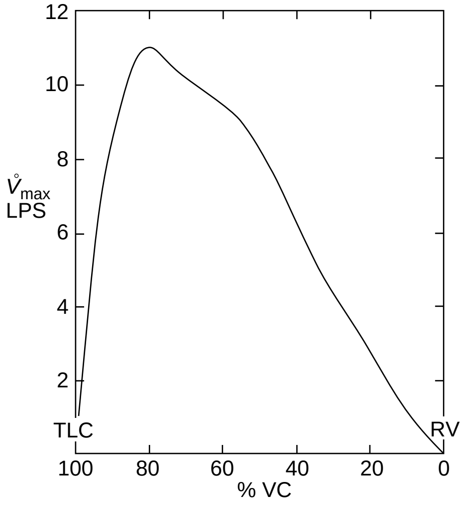


Figure 12. Flow volume curve. Forced exhalation from TLC to RV = FVC recorded by a pneumotachograph.

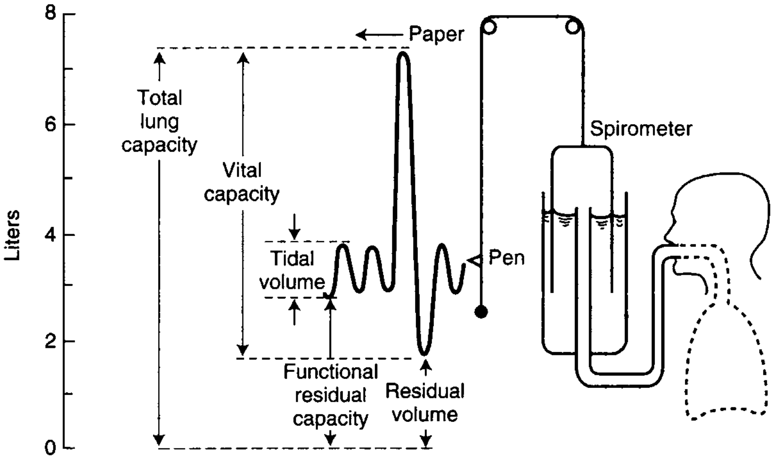


Figure 13. Lung volume. Measurements by a water filled spirometer.

that to the residual volume already measured (Fig. 13).

$$\mathrm{RV} = \mathrm{FRC} - \mathrm{ERC}$$

(ERC = expiratory reserve capacity)

$$\mathrm{TLC} = \mathrm{RV} + \mathrm{IC}$$

(IC = inspiratory capacity)

FRC BY PLETHYSMOGRAPHY

A plethysmograph is an airtight box of known volume, similar to a telephone booth, in which a patient sits. A mouthpiece connects the patient to air outside the apparatus and pressure sensors are located within the box and within the breathing capacity. At the end of a normal tidal breath, a shutter on the mouthpiece closes and the subject is asked to make respiratory efforts. As the subject tries to inhale, the volume of the lung expands slightly while the pressure drops due to the chest (lung) expansion (Fig. 14). Applying Boyle's law, if the pressures in the box before and after the inspiratory effort are P_1 and P_2, respectively, V_1 is the preinspiratory box volume and ΔV is the change in volume of the box (or lung) ΔV can be obtained from the equation $P_1V_1 = P_2(V_1 + \Delta V)$.

Applying Boyle's law to the gas in the lung, $P_3(V_2) = P_4(V_2 + \Delta V)$, where P_3 and P_4 are the mouth pressures before and after the inspiratory effort and V_2 is the FRC. Thus FRC can be obtained.

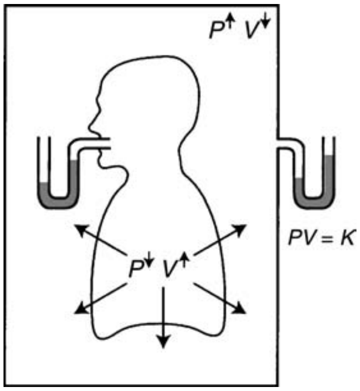


Figure 14. The FRC by whole body plethysmography utilizing Boyle's law.

The body plethysmograph measures the total volume of compressible gas, including any that is trapped behind closed or poorly communicating airways.

The other two methods measure only volumes based on gas communicating with and open to airways. This is not an issue in normal subjects, but in diseased lungs considerable amounts of gas are trapped and do not communicate freely. Therefore the FRC values differ depending on methodology.

NITROGEN WASHOUT

If a subject quietly breathes 100% oxygen for several minutes, all the nitrogen emptied from the lung can be determined by multiplying the exhaled volume by the exhaled nitrogen concentration. Since the initial lung concentration of N_2 is 80%, the measured volume of nitrogen exhaled multiplied by 1/0.8 equals the volume of the lung prior to 100% oxygen breathing. The value of FRC can be underestimated if significant parts of the lung communicate poorly or not at all with the inspired oxygen (13).

HELIUM DILUTION

As a subject breathes from a spirometer with a known concentration of helium, after several normal breaths the helium concentration in the lung and spirometer equilibrate (Fig. 15), Since helium is insoluble in blood, none of it is absorbed, so the final equilibrium concentration is a reflection of dilution only. The amount of helium before equilibration is $(C_1 \times V_1)$ and equals that after equilibration $C_2\ (V_1 + V_2)$ solving for V_2, $V_2 = (C_1V_1/C_2) - (V_1)$. During the equilibration period oxygen is added to the spirometer and carbon dioxide is absorbed.

Although many other measurements of lung function are perhaps more useful, knowledge of the lung volumes is essential in other complex measurements such as diffusing capacity.

DIFFUSING CAPACITY

Given the challenges with measuring the diffusing capacity of the lung for oxygen, carbon monoxide as originally

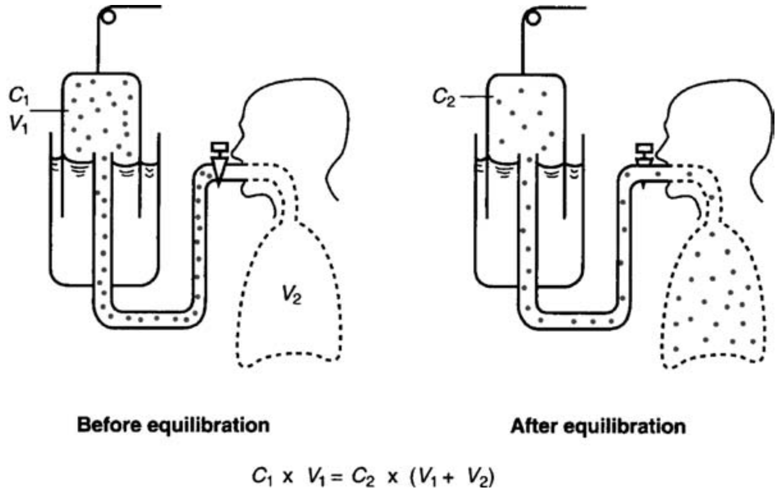


Figure 15. Helium equilibration. Technique for FRC determination.

used by Marie Krogh in 1914 is used for current day measurements. There are at least seven variations of this method that have since been developed.

The most common of these is the single-breath modified Krogh technique attributed to Kety and Fowler. Both helium and carbon monoxide are inhaled. After a period of breath hold (15 s), the alveolar portion of the exhaled gas is collected and the concentration of carbon monoxide and helium is measured.

The initial alveolar carbon monoxide concentration is calculated thus:

$$F_{\mathrm{I}}\mathrm{CO} \times \frac{\text{He\%in expired alveolar sample}}{\text{Inspired helium percentage}} = F_{\mathrm{A}}\mathrm{CO}$$

$$D_{\mathrm{L}}\mathrm{CO} = \frac{\text{Alveolar volume STPD} \times 60}{\text{Seconds of breath hold} \times \text{PB} - 47} \times \ln\left(\frac{F_{\mathrm{I}}\mathrm{CO\ alv}}{F_{\mathrm{E}}\mathrm{CO\ alv}}\right)$$

In the above equation, the alveolar volume is measured by a helium dilution technique similar to that in lung volume determinations.

CLOSING VOLUME

This sensitive test detects early changes in lung function and reflects pathology in the small airways. Smokers have an abnormally high closing volume prior to any other pulmonary function test changes. In this test, the subject inhales a breath of 100% O_2 to TLC. During the subsequent exhalation, the nitrogen is measured through the alveolar plateau to an abrupt rise in exhaled nitrogen, so-called phase 4 (Fig. 16). This signals closure of airways in the base of the lungs and preferential emptying of upper airways. Less of the 100% oxygen inhalation is distributed to the upper lung, making it richer in nitrogen. It is this fact that creates phase 4. In some lung diseases, the closing volume is above the FRC. This means that airways close even during normal breathing and is an indication of advanced disease.

MAXIMAL VOLUNTARY VENTILATION

This test measures the volume of air moved during 15 s of repetitive forced deep maximal exhalations. A water filled spirometer is used for measurement with a time kymographic tracing. Pneumotachographs with real time computer graphics may be used. The main requirement for accurate test results is a low resistance breathing circuit and avoidance of resonance in the system. Both problems have been overcome by modern spirometers, valves and tubing. Although this is formally a lung test, nonpulmonary factors such as motivation, muscular strength and endurance are very important and must be taken into consideration when interpreting the test. The results are expressed in liters per minute BTPS.

PEAK FLOW

Peak flow meters have the distinct advantage of being handheld, self-contained and thus very portable. This is an effort-dependent test, yet it is an excellent reflection of airways function. Its main utility is quickness and simplicity, and it is often used for the management of asthma. Much like a home glucose meter in a diabetic, the peak flow meter can give objective assessment of airways function throughout the day to help guide treatment and presage severe attacks.

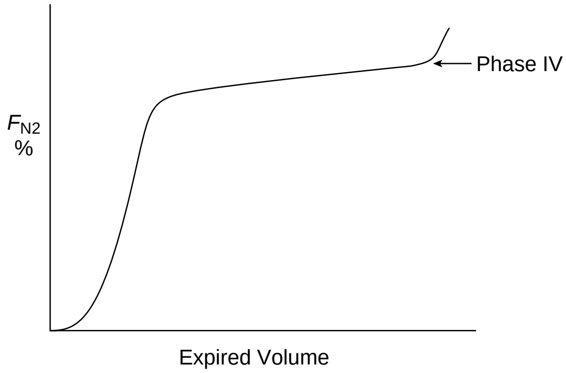


Figure 16. Abrupt rise at the end of exhalation called phase (IV) correlates with the closing volume.

STANDARDIZATION OF PULMONARY FUNCTION TESTS

The American Thoracic Society published standards on spirometry in 1979 at the Snowbird Workshop. An official update was issued in 1994, which details equipment and procedural guidelines to enhance accuracy and reproducibility.

Without meticulous attention to these recommendations, the utility of all pulmonary function tests is compromised.

CARDIOPULMONARY STRESS TESTING

Cardiopulmonary stress testing (14) uncovers disorders of the respiratory system as it functions in the integrated cardiopulmonary response to the metabolic demands of increasing incremental work loads. During the course of a cardiopulmonary stress test, the following variables are measured in real time:

$W^{\cdot}$	Work
f_c	Heart rate
O_{O_2}	Oxygen consumption
O_{CO_2}	Carbon dioxide production
O_E	Minute Ventilation
R	O_{CO_2}/O_{O_2}
O_E/O_{O_2}	Efficiency of ventilation
O_E/O_{CO_2}	Efficiency of ventilation
$P_{ET}O_2$	End tidal P_{O_2}
$P_{ET}CO_2$	End tidal P_{CO_2}
P_ECO_2	Mixed expired CO_2

The instrumentation required to make these breath-by-breath and time-averaged measurements are rapid responding CO_2 and O_2 meters; pneumotachograph for flows, ventilation; mixing chamber for collection of exhaled gas; ECG; ergometer cycle, arm, treadmill; computer interface with full graphics in real time.

Limits to exercise appear often in clinical medicine with the presenting problem of dyspnea or shortness of breath on exertion. The genesis of this symptom may reside within either the respiratory or cardiac systems or both simultaneously. Analysis of data obtained from cardiopulmonary stress testing can aid in clinical diagnosis.

Testing begins with an incremental ergometric work load that creates a systemic metabolic response measured as oxygen uptake or consumption. There is a strong reproducible linear relationship between the work load and oxygen consumption. It is derived from the coupling of work and mitochondrial metabolic pathways for cellular energy generation.

With increasing work loads and oxygen consumption, the total ventilation of the lung increases. This requires higher airflow which in normals, even at peak work loads, never reaches flows measured in maximal flow–volume curves. Such is not the case in obstructive lung disorders, where reduced flow is a hallmark of the disease. Exercise is limited due to the inability to generate flows capable of sustaining the metabolic demands.

As the exercise test progresses real-time analysis of dead space can be measured. Typically the $P_ECO_2 = P_ACO_2 = P_aCO_2$ remains constant with the mixed expired CO_2 P_E CO_2 increasing. This is reflected in a decrease in the O_E/OCO_2 marking increased efficiency as more CO_2 is exhaled per breath. By making use of the Bohr equation:

$$V_D/V_T = \frac{F_ACO_2 - F_ECO_2}{F_ACO_2}$$

it appears that V_D/V_T or wasted ventilation decreases with increased exercise. This occurs because lung apical units not perfused but ventilated at rest now are fully perfused and participate in gas exchange. If V_D/V_T does not decrease with exercise, then it is likely based on a structural disease such as emphysema. The work of breathing at any given work load is higher in these patients in part because V_D/V_T (wasted ventilation) remains abnormally high.

As the level of incremental work load increases, the delivery of oxygen fails to meet the metabolic demands of tissues and anaerobic metabolism becomes prominent. Lactic acid is dumped into the blood stream and is quickly buffered by bicarbonate, which generates more carbon dioxide. A dramatic upsurge in OCO_2 marks this point and is called the anaerobic or metabolic threshold. The parameter $R(OCO_2/OO_2)$ values that previously were 0.8 are now 1.2–1.4, indicating a combined metabolic and buffer source of carbon dioxide. Ventilation (O_E) is driven by the chemical stimulation of lactic acidosis in addition to the demands of oxygen delivery.

Data from such a stress test yields much clinically useful information and allows one to differentiate a pulmonary from a cardiac cause for exercise limitations. Cardiopulmonary deconditioning has a distinct pattern as does obesity. Low peak OO_2 and low A.T. as a percent of OO_2 max are good indications of these two conditions.

Future of Pulmonary Function Testing

No doubt, advances in instrumentation and computers will continue to refine pulmonary function testing. Miniaturization of testing equipment allows complex measurements not only in the laboratory but also in the wild. The burgeoning science of sleep medicine is an example of this.

Epidemiological studies will explore the relationship of pulmonary function to health and uncover what makes the vital capacity so vital to life. A fundamental role for FEV_1 in total mortality independent of cigarette smoking has been proposed. Whether reduced lung function leaves an individual open to oxidative stress is unknown.

If pulmonary function proves to be a long-term predictor for overall survival rates in both genders, it could be used as a tool in general health assessment.

The search for tests that implicate early potentially reversible lung disease will continue. The benefit to asymptomatic patients and society as a whole is obvious.

The long-term effects of air pollution and impact of air quality on lung health will always be of prime concern, not only to the general public, but to government officials who set air quality standards.

Perhaps pulmonary function testing will ultimately guide and protect us all.

TERMINOLOGY—DEFINITIONS—EQUATIONS

Spirometer	A measuring device for determining lung volume, its subcompartments, and expiratory flow rates.
FRC	Functional residual capacity: The volume in the lung after a normal exhalation. At this volume the recoil pressure of the lungs inward is exactly balanced with the outward recoil pressure of the chest wall.
FEV_1	Forced expiratory volume in the first second: The amount of air expired in the first second of a forced expiratory maneuver.
TLC	Total lung capacity.
RV	Residual volume: The volume of air left in the lungs after a full exhalation.
FVC	Forced vital capacity: The amount of air exhaled during a complete exhalation.
$\mathrm{FEF}_{25\text{–}75}$	Forced expiratory flow: The mean expiratory flow measured between 75 and 25% of the vital capacity during forced exhalation.
PEFR	Peak expiratory flow rate during forced exhalation.
MVV	Maximal voluntary ventilation expressed in liters per min.
$\mathrm{DL}_{\mathrm{CO}}$	Diffusing capacity for carbon monoxide.
Flow–volume curve	A maximal exhalation measuring flow versus volume.
$P_{\mathrm{A}}\mathrm{O}_2$	Alveolar oxygen partial pressure.
$P_{\mathrm{A}}\mathrm{CO}_2$	Alveolar carbon dioxide partial pressure.
$P_{\mathrm{ET}}\mathrm{O}_2$	End tidal oxygen partial pressure.
$P_{\mathrm{ET}}\mathrm{CO}_2$	End tidal carbon dioxide partial pressure.
$P_{\mathrm{E}}\mathrm{O}_2$	Mixed expired oxygen partial pressure.
$P_{\mathrm{E}}\mathrm{CO}_2$	Mixed expired carbon dioxide partial pressure.
$O\mathrm{o}_2$	Volume of oxygen take up per minute.
$O\mathrm{co}$	Volume of carbon dioxide output per minute.
O_{E}	Minute ventilation: Total volume of air expressed per minute from the lungs.
V_{T}	Tidal volume: The volume of a single breath.
V_{D}	The volume of physiological dead space.
$V_{\mathrm{D}}/V_{\mathrm{T}}$	The ratio between dead-space volume and tidal volume. This ration indicates the efficiency of ventilation.
V_{A}	Volume of alveolar gas in the tidal volume.
General gas law	$PV = RT$
Boyle's law	$P_1V_1 = P_2V_2$ (temperature constant)
Charles's law	$\frac{V_1}{V_2} = \frac{T_1}{T_2}$ (pressure constant)
Poiseuille's law	$V^{\cdot} = \frac{P\pi r^4}{8n}$ P = Pressure difference across length ℓ and radius r n = Coefficient of viscosity
Bohr equation	$V_{\mathrm{D}}/V_{\mathrm{T}} = \frac{P_{\mathrm{A}}\mathrm{CO}_2 - P_{\mathrm{E}}\mathrm{CO}_2}{P_{\mathrm{A}}\mathrm{CO}_2}$

BIBLIOGRAPHY

1. Hutchinson J. On the capacity of the lungs, and on the respiratory functions, with a view of establishing a precise and easy method of detecting disease by the spirometer. Med Chir Trans (London) 1846;29:137.
2. Comroe JH Jr. Retrospectoscope. Insights into medical discovery. Menlo Park, (CA): Von Gehr Press; 1977.
3. Otis AB, Rahn H. Developments of Concepts in Rochester, New York, in the 1940's. In: West JB, editor. Pulmonary Gas Exchange Volume 1. New York: Academic; 1980. pp. 33– 65.
4. Bates DV, Macklem DT, Christie RV. Respiratory Function in Disease. 2nd ed. Philadelphia: Saunders; 1971.
5. Horton GE, Phillips S. The expiratory ventilagram: application of total and time vital capacities and maximal expiratory flow rate, as obtained by a bellows apparatus, for bedside and office use. Am Rev Respir Dis 1959 Nov; 80:724–731.
6. Wright BM, McKerrow CB. Maximum forced expiratory flow rate as a measure of ventilatory capacity: with a description of a new portable instrument for measuring it. Br Med J 1951 Nov. 21; 5159:1041–1046.
7. Hill DW. Physics Applied to Anesthesia. New York: Appleton-Century-Crofts; 1972.
8. McCall CB, Hyatt RE, Noble FW, Fry DL. Harmonic content of certain respiratory flow phenomena of normal individuals. J App Physiol 1957 Mar; 10(2):215–218.
9. Bouhuys A. The clinical use of pneumotachography. Acta Med Scand 1957 Nov. 15; 159(2):91–103.
10. Standardization of Spirometry. Official statement of the American Thoracic Society. Am J Respir Crit Care Med. 1995;152:1107–1136.
11. Gaensler EA. Evaluation of pulmonary function: methods. Annu Rev Med 1961;12:385–408.

12. Miller DR, Hyatt RE. Obstructing lesions of the larynx and trachea: clinical and physiologic characteristics. Mayo Clinic Proc. 1966 Mar; 44:145–161.
13. Emmanuel G, Briscoe WA, Cournand A. A method for the determination of the volume of air in the lungs: measurements in chronic pulmonary emphysema. J Clin Invest Feb 40:329–337.
14. Cooper CB, Storer TB. Exercise testing and interpretation. Cambridge (MA): Cambridge University Press; 2001.

Reading List

West JB. Respiratory Physiology the Essentials. 7th ed. Philadelphia: Lippincott Williams & Wilkins; 2004.

See also HEART-LUNG MACHINES; LUNG SOUNDS; RESPIRATORY MECHANICS AND GAS EXCHANGE; VENTILATORY MONITORING.

PUMPS, INFUSION.

See DRUG INFUSION SYSTEMS.

Q

QUALITY CONTROL, X-RAY. See X-RAY QUALITY CONTROL PROGRAM.

QUALITY-OF-LIFE MEASURES, CLINICAL SIGNIFICANCE OF

CELIA C. KAMATH
JEFFREY A. SLOAN
Mayo Clinic
Rochester, Minnesota
JOSEPH C. CAPPELLERI
Pfizer Inc
Groton, Connecticut

INTRODUCTION

The field of patient-reported outcomes, particularly health-related quality of life (QOL), has burgeoned in the last few years (1,2). The importance assigned to the study of these outcomes has been attributed to the aging of the population and consequently higher prevalence of chronic diseases, along with the reality that medical treatment often fails to cure the disease, but may affect QOL (3). Health-related quality of life has gained attention in research and clinical trial settings (3,4).

The increasingly important role assigned by patients and clinicians to QOLs role in medical decision making has resulted in greater attention paid to the interpretation of QOL scores, particularly as it relates to clinical significance (5–7). Clinical significance relates to the clinical meaningfulness of intersubject or intrasubject changes in QOL scores. Clinical significance has been difficult to determine, in part due to the development of a myriad of QOL instruments over the past decade (8,9). Some of these have had little or no psychometric validation (1,2,6,10,11) or clinical validation (9,12). Moreover, relative to traditional clinical endpoints, like survival and systolic blood pressure, QOL as a clinical endpoint is relatively unfamiliar especially in regard to interpretation and relevance of changes in QOL scores (13).

Why is clinical significance of QOL scores important? It aids in the design of studies by helping to determine sample size calculations. Evidence of clinical significance may be used by regulatory agencies for drug approval, by clinicians to decide between treatment alternatives, by patients to make informed decisions about treatment, by the healthcare industry for formulary and reimbursement decisions, and by healthcare policy makers to make policy decisions regarding resource allotment. Early evidence of the clinical implications of QOL is evident in the links between survival and QOL components (e.g., patients' fatigue levels, social support, and group counseling) (14–17). Even a simple, single-item measure of patient global QOL can be related to patient survival (18). Changes in QOL scores can also be linked to positive economic (19,20) and social (21) outcomes.

HISTORICAL BACKGROUND

Statistical significance as measured by a P-value is influenced by sample size and data variability. While statistical significance can be considered a prerequisite for clinical significance, only clinical significance assigns meaning to the magnitude of effect observed in any study. Historically, Cohen (22) was responsible for proposing one of the earliest criteria for identifying important change, which can be construed as clinically significant. He suggested that a small effect size (defined later in this article) was 0.2 standard deviation units, a medium effect size was 0.5, and a large effect size was 0.8. Although his intention was to provide guidance for sample size calculations in the social and behavioral science, Cohen's benchmarks have extended to healthcare research to decide whether or not a change in QOL scores is important. Current research suggests that a moderate effect size of one-half a standard deviation unit (effect size $= 0.5$) is typically important (23). A more recent and popular definition of clinical significance uses an anchor-based approach based on an external standard that is interpretable and appreciably correlated to the target QOL measure, in order to elucidate the meaning of change on the target QOL measure.

Embedded under the rubric of clinical significance is the minimum important difference, a lower bound on clinical significance. One definition of a minimum important difference (MID) is "the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side-effects and excessive cost, a change in the patient's management"(24). Some researchers prefer to use the term "minimally detectable difference"(25,26); other approaches have sprouted (e.g., the empirical rule effect size, ERES, method) (27,28).

No single solution to the challenging topic of assessing clinical significance exists. Nevertheless, a series of proposals has engendered understanding and appreciation of the topic. Special issues in *Statistics in Medicine* (1999, Vol. 18) and the *Journal of Consulting and Clinical Psychology* (1999, Vol. 67) have been dedicated to the topic of clinical significance of QOL and other clinical measures. Proceedings from a meeting of an international group of ~30 QOL experts were published recently in a special issue of the *Mayo Clinic Proceedings* (2002, Vol. 77) (29–35), which provides practical guidance regarding the clinical significance of QOL measures.

CHAPTER OUTLINE

This article draws largely from recent scientific literature, including the *Mayo Clinic Proceedings* (29–35) and other sources (4), to provide an overview on the clinical significance of QOL measures.

The following section on Design and Methodology covers the different perspectives and existing methods to determine clinical significance. The next section on Examples illustrates trials in which researchers attempted to define the concept on specific QOL measures. Then the section on Recent Developments highlights new methods to determine clinical significance. Finally, the section on Concluding Remarks discusses some future directions for research.

DESIGN AND METHODOLOGY

Perspectives for Determining and Interpreting Clinical Significance

Clinical significance involves assigning meaning to study results. The process of establishing such meaning can be conceptualized in two steps: (*1*) understanding what changes in score mean to the concerned stakeholder (e.g., patient, clinician, clinical researcher, policy maker) and (*2*) making results of clinical studies interpretable and comprehensible to such stakeholders or decision makers (30,36). The term clinical in relation to significance has different meaning and implications for different stakeholders such as patients, clinicians, and society.

From the patient's perspective, clinical significance can be defined as the change in QOL scores that patients perceive as beneficial (or detrimental) and important that prompts them to seek health care or request changes in their treatment (33), or that induces patients to determine that the intervention has been successful (24). From the clinician's perspective, it can be defined as the diagnosis of the clinician as to the amount of change in QOL scores that would mandate some form of clinical intervention (37). From the societal or population perspective, clinical significance is based on the values of the group surveyed, where importance is defined by the outcomes that are deemed worthy of society's resources. Any or all of these perspectives for defining clinical significance may be applicable, but they may not always be in agreement (4).

An equally important issue is the different perspectives for interpreting clinical meaningfulness of changes in reported QOL (35). For example, a clinician may use QOL data to explain the treatment alternatives to a patient, while a health policy maker may describe to elected officials the financial impact on a patient population whose QOL has changed. Similarly, a regulatory agency and pharmaceutical company may ascertain the appropriate level of evidence for a successful research study (35). Thus QOL results must be framed, analyzed, and presented in a way that is meaningful to the pertinent audience and its respective needs. Only then will the concept be meaningful and gain greater acceptance and use over time.

METHODS TO EXPLAIN THE CLINICAL SIGNIFICANCE OF HEALTH STATUS MEASURES

Two common approaches used to establish the interpretability of QOL measures are termed anchor and distribution based. The characteristics of each approach are described below. Several examples will be given later in the section on Examples. Interested readers are encouraged to read Lydick and Epstein (1993) [Lydick, 1993 #40] Crosby et al. (4), and Guyatt et al. (30) for an expanded discussion of the concepts presented here.

Anchor-based approaches are used to determine clinically meaningful change via cross-sectional or longitudinal methods involve comparing measures of QOL to measures with clinical relevance (4). Cross-sectional methods include several forms: (*1*) comparing groups that are different in terms of some disease-related criterion (38,39); (*2*) linking QOL to some external benchmarking criteria (40–42); (*3*) eliciting preference-based ratings on a pairwise basis, where one person's ratings state serves as an anchor to evaluate the other person's ratings (43); and (4) using normative information from dysfunctional and functional populations (6). Longitudinal methods involve the comparison of changes in QOL scores across time with the use of (*1*) global ratings of change as "heuristics" to interpret changes in QOL scores (5,24,38,44); (*2*) significant future medical events for establishing difference thresholds (45); and (*3*) comparisons of changes in HRQOL to other disease-related measures of outcome across time (46). Anchor-based methods are cataloged in Table 1.

Anchor-based methods require two properties (30): (*1*) anchors must be interpretable, else they will hold no meaning to clinicians or patients; and (*2*) anchors must share appreciable correlation with the targeted QOL measure. The biggest advantage of anchor-based approaches is the link with a meaningful external anchor (4), akin to establishing the construct validity of the measure (49). Potential problems, however, exist with this approach. These include recall biases (50), low or unknown reliability and validity of the anchor measure (51), low correlation between anchor and actual QOL change score (52–55), and complex relationships between anchors and QOL scores (56) and the challenge of defining a meaningful change in the anchor itself.

Hays and Wooley (57) recommend caution in the indiscriminate dependence and use of a single minimum important difference (MID) measure. They also list several problems in estimating MIDs: the estimated magnitude could vary depending on the distributional index (57,58), the external anchor (59), the direction of change (improvement vs. decline) (60), and the baseline value (61). In general, longitudinal methods are preferable because of their direct link with change (4).

Distribution-based approaches for determining the importance of change are based on the statistical characteristics of the obtained sample, namely, average scores and some measure variability in results. They are categorized as (*1*) those that are based on statistical significance using p-values (i.e., given no real change, the probability of observing this change or a more extreme change), which include the paired *t*-statistic (62) and growth curve analysis (63); (*2*) those that are based on sample variation (i.e., those that evaluate mean change in relation to average variation around a mean value), which include effect size (22,64), standardized response mean (SRM) (44), and responsiveness statistic (65); and (*3*) those that are based on the measurement precision of the instrument

Table 1. Anchor-Based Methods of Determining Change[a]

Type	Method	Examples	HRQOL evaluated in relation to:	Advantages	Disadvantages
Cross-sectional	Comparison to disease-related criteria	39, 47	Disease severity or diagnosis	Can be standardized Easy to obtain	May not reflect change Groups may differ in other key variables
	Comparison to nondisease-related criteria	40, 41	Impact of life events	Easy to obtain Provides external basis for interpretation	May no reflect change Groups may differ on other key variables Relationship to HRQOL not clear
	Preference rating	43	Pairwise comparisons of health status	All health states are compared	May not reflect change Hypothetical, artificial Time Consuming
	Comparison to known populations	6	Functional or dysfunctional populations	Uses normative information	Normative information not always available Amount of change needed not specified
Longitudinal	Global ratings of change	5, 24, 38, 44	Patients' or clinicians' ratings of improvement	Easy to obtain Best measure from individual perspective Can take into account a variety of information	Does not consider measurement precision Unknown reliability Influenced by specific rating scale and anchors
	Prognosis of future events	45	Those experiencing and not experiencing some future event	Prospective Provides evidence of predictive validity	Does not consider measurement precision Difficult to obtain
	Changes in disease related outcome	48	Changes in clinical outcome	Tied to objective outcome measure Known psychometric properties	Does not consider measurement precision Assumes strong HRQOL-outcome correlation

[a] Reprinted with permission from Ref. 4.

(i.e., evaluate change in relation to variation in the instrument as opposed to variation of the sample), which includes the standard error of the mean (SEM) (7) and the reliable change index (RC) (6). Distributed-based methods are catalogued in Table 2.

An advantage of the distribution-based methods is that they provide a way of establishing change beyond random variation and statistical significance. The effect size version of the distribution-based methods is useful to interpret differences at the group level and has benchmarks of 0.20 standard deviations units as a small effect, 0.50 as a moderate effect, and 0.80 as a large effect (22,64,66). The measure that seems most promising for the purpose of establishing clinical significance at the individual patient level are the SEM and the RC. These measures are based on the measurement precision of the instrument and incorporate the reliability of the instrument (e.g., Cronbach's alpha or test–retest reliability), and the standard deviation of scores. In principle, SEM and RC are sample invariant. Researchers have favored Cronbach's alpha over test–retest reliability to calculate reliability for the SEM (7,30,67), because this is more conveniently available over test–retest data.

Distribution methods are particularly helpful when used together with meaningful anchors, which enhances validity, and hence meaning to the QOL measure. There is some encouragement to know that anchor-based measures appear to coincide with distribution-based methods. Researchers have found a correspondence between SEM and anchor-based determinant of a minimum important difference across difference diseases (7,23,67,68). The 1 SEM benchmark corresponds with an effect size (ES) of ~0.5. Nonetheless, note that the SEM is moderated by the reliability of the measure, where measures with higher reliability are "rewarded" by lowering the effect size (ES) needed to achieve a minimally important difference. Thus the 1 SEM benchmark corresponds with an $ES = 0.5$, when reliability of the $scale = 0.75$; the correspondence shifts to 1 SEM is equivalent to an $ES = 0.33$, when reliability increases to 0.9, which is frequently attainable in focused assessments. A rationale for a SEM as a measure of MID is provided by Norman et al. (23) who assert that Miller's theory (69) of the limits of human discernment is linked to the threshold of 0.5 standard deviation units.

EXAMPLES

This section provides examples of studies used to determine clinical significance and presents general advice for defining and interpreting clinical significance in clinical studies. Table 3 includes several examples on the use of both anchor-based methods and distribution-based methods to establish clinical significance across a wide range of QOL measures. These examples span several disease groups, instruments, and methods for determining clinical significance. Readers are encouraged to review the cited papers for further details on these studies.

Table 2. Distribution-Based Methods of Determining Change [a]

Method	Reference	HRQOL evaluated in relation to:	Calculation	Advantages	Disadvantages
Paired *t*-statistic	62	Standard error of the mean change	$\dfrac{x_1 - x_0}{\sqrt{\dfrac{\sum(d_i - \bar{d})^2}{n(n-1)}}}$	None	Increases with sample size
Growth curve analysis	63	Standard error of the slope	$\dfrac{B}{\sqrt{V}}$	Not limited to pre-test and post-test scores Uses all of the available data	Increases with sample size Requires large sample sizes Assumes data missing at random
Effect size	22, 64	Pre-test standard deviation	$\dfrac{x_1 - x_0}{\dfrac{\sqrt{\sum(x_0 - \bar{x}_0)^2}}{n-1}}$	Standardized units Benchmarks for interpretation Independent of sample size	Decreases with increased baseline variability of sample Does not consider variability of change May vary widely among samples
Standardized response mean	44	Standard deviation of change	$\dfrac{x_1 - x_0}{\sqrt{\dfrac{\sum(d_i - \bar{d})^2}{n-1}}}$	Standardized units Independent of sample size Based on variability of change	Varies as a function of effectiveness of treatment
Responsiveness statistic	65	Standard deviation of change in a stable group	$\dfrac{x_1 - x_0}{\sqrt{\dfrac{\sum\left(d_{i\ stable} - \bar{d}_{stable}\right)^2}{n-1}}}$	Standardized units More conservative than effect size Independent of sample size Takes into account spurious change due to measurement error	Data on stable subjects frequently not available
Standard error of measurement	7	Standard error measurement	$\dfrac{x_1 - x_0}{\sqrt{\dfrac{\sum(x_0 - \bar{x}_0)^2(\sqrt{1-r})}{(n-1)}}}$	Relatively stable across populations Takes into account the precision of the measure Cutoffs based on confidence intervals	Assumes measurement error to be constant across the range of possible scores
Reliable change index	6	Standard error of the measurement difference	$\dfrac{x_1 - x_0}{\sqrt{2(SEM)^2}}$	Relatively stable across populations Takes into account precision of measure Cutoffs based on confidence intervals	Assumes measurement error to be constant across the range of possible scores

[a] Reprinted with permission from Ref. 4.

We begin with a classic paper by Jaeschke et al. (24), one of the first papers on clinically meaningful differences determined through the anchor-based approach. The magnitude of difference considered minimally significant was an average of 0.5 per item on a 7-point scale, which was confirmed by Juniper et al. (5) on the asthma quality of life questionnaire (AQLQ). A third study, by Kazis et al. (64), examined the difference between statistical significance and clinical significance.

Using several pain studies and the Pain Intensity numerical rating scale (PI-NRS) scale, Farrar et al. (70) found a reduction of 2 points or 30% on the 11-point pain scale to be clinically significant. Using the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) scale for osteoarthritis, Angst et al. (71) compared estimates derived from anchor- and distribution-based approaches; they determined sample sizes for specified changes signifying worsening and, separately, improvement. Using the chronic heart failure questionnaire (CHQ), Wyrwich et al. (7) also compared anchor and distribution-based approaches in determining that 1 SEM equals the MCID of 0.5 per item determined by the anchor-based approach. Finally, using the functional assessment of cancer therapy–Lung (FACT-L) questionnaire, Cella et al. (72) showed the convergence of three different complementary approaches on clinical significance.

Taken from Sprangers et al. (34), Table 4 provides a useful checklist of questions to help in interpretation of

Table 3. Examples of Studies for Determining Clinically Significant Change

<table>
<tr><th>Authors and Instrument–Anchor Used</th><th>Study Description</th><th>Study Results</th><th>Comments</th></tr>
<tr>
<td>Jaeschke, Singer and Guyatt (1989) (24)
Chronic Respiratory Questionnaire & Chronic Health Failure Questionnaire.
QOL dimensions:
individual domains of dyspnea, fatigue and emotion.
7 point scale
focus: change scores
Anchor:
Longitudinal within-patient rating of change since last visit
A Very great deal worse (−7) to
A Very great deal worse (+7)</td>
<td>75 patients
QOL questionnaires completed at baseline and at 2,6,12 &24 weeks.
Links QOL change scores with classification on global patient-rated anchor.
Classification of change on AQLQ
Somewhat worse: −3 to −1; +3 to +1.
Moderately or good deal worse: −4 or −5; +4 or +5
A great deal worse: −6 or −7; +6 or +7</td>
<td><table>
<tr><th></th><th colspan="4">Global Rating of change</th></tr>
<tr><th></th><th>None</th><th>Small</th><th>Moderate</th><th>Large</th></tr>
<tr><td>Dyspnea</td><td>0.10</td><td>0.43</td><td>0.96</td><td>1.47</td></tr>
<tr><td>Fatigue</td><td>0.12</td><td>0.64</td><td>0.87</td><td>0.94</td></tr>
<tr><td>Emotion</td><td>0.02</td><td>0.49</td><td>0.81</td><td>0.86</td></tr>
</table></td>
<td>Anchor-based approach
One of the first studies to determine the magnitude of change that is clinically meaningful</td>
</tr>
<tr>
<td>Juniper et al (1994)(5)
Asthma Quality of Life Questionnaire (AQLQ)
QOL Dimensions:
overall score and individual domains of activities, symptoms, emotional
7 point scale
focus: change scores
Anchor:
Longitudinal within-patient rating of change since last visit
“Worse”: −1 hardly any worse to −7 a very great deal worse
“Better”: +1 hardly any better to a +7 a very great deal better</td>
<td>39 subjects with asthma treatment
AQLQ completed before and after 8 week asthma treatment
Links QOL change scores with classification on anchor.
Classification of change on AQLQ
Small: −2 or −3: +2 or +3
Moderate: −4 or −5; +4 or +5
Large: −6 or −7; +6 or +7</td>
<td><table>
<tr><th></th><th colspan="4">Global Rating of change</th></tr>
<tr><th></th><th>0–1</th><th>2–3</th><th>4–5</th><th>6–7</th></tr>
<tr><td>Overall QOL</td><td>0.11</td><td>0.52</td><td>1.03</td><td>2.29</td></tr>
<tr><td>Activity</td><td>0.12</td><td>0.47</td><td>0.87</td><td>1.83</td></tr>
<tr><td>Symptoms</td><td>0.20</td><td>0.49</td><td>1.13</td><td>2.21</td></tr>
<tr><td>Emotional</td><td>0.20</td><td>0.58</td><td>1.51</td><td>2.70</td></tr>
</table></td>
<td>Anchor-based approach
Based in part on these results, Glaxo-Wellcome obtained a HqoL promotional claim for salmeterol for noctural asthma</td>
</tr>
<tr>
<td>Kazis, Anderson & Meenan (1989) (64)
Arthritis Impact Measurement Scale (AIMS).
64 items, of which 45 are health status questions.
9 scales
0–10 scale, with higher scores indicating worsehealth states.</td>
<td>299 patients with rheumatoid arthritis
AIMS completed at beginning and end of 5 year-period.</td>
<td><table>
<tr><th></th><th colspan="2">Change Score</th><th colspan="3">Effect Size</th></tr>
<tr><th>Scale</th><th>$\bar{x}_5-\bar{x}_1$</th><th>$\mathrm{SD}_{\mathrm{Difference}}$</th><th>$t$</th><th>$p$-value</th><th>SRM</th></tr>
<tr><td>Mobility</td><td>0.07</td><td>2.80</td><td>0.43</td><td>0.6</td><td>+0.02</td></tr>
<tr><td>Physical Activity</td><td>0.41</td><td>2.57</td><td>2.76</td><td>0.006</td><td>+0.17</td></tr>
<tr><td>Dexterity</td><td>0.42</td><td>3.63</td><td>2.00</td><td>0.046</td><td>+0.12</td></tr>
<tr><td>Activities of Daily Living</td><td>0.02</td><td>1.76</td><td>0.20</td><td>0.8</td><td>+0.01</td></tr>
<tr><td>Household Activities</td><td>0.05</td><td>1.61</td><td>0.54</td><td>0.6</td><td>−0.03</td></tr>
<tr><td>Anxiety</td><td>0.19</td><td>2.01</td><td>1.63</td><td>0.1</td><td>+0.09</td></tr>
<tr><td>Depression</td><td>0.25</td><td>1.71</td><td>2.53</td><td>0.012</td><td>+0.14</td></tr>
<tr><td>Pain</td><td>0.96</td><td>2.49</td><td>6.67</td><td>0.001</td><td>+0.42</td></tr>
<tr><td>Social Activity</td><td>0.36</td><td>2.11</td><td>2.95</td><td>0.003</td><td>+0.17</td></tr>
</table></td>
<td>Combines anchor-based and distributional approaches
Compares statistical significance with clinical significance</td>
</tr>
</table>

(continued)

Table 3. (*Continued*)

<table>
<tr><th>Authors, and Instrument/Anchor Used</th><th>Study Description</th><th>Study Results</th><th>Comments</th></tr>
<tr>
<td>Farrar JT et al. (2001) (70)
Pain Intensity Numerical Rating Scale (PI-NRS)
11-point pain scale: 0 = no pain to 10 = worst
baseline score = mean of 7 diary prior to drug
endpoint score = mean of last 7 diary entries
focus: change scores
Anchor:
Longitudinal within-patient Global Impression Change (PGIC)
Very much improved (1) to Very much worse (7).</td>
<td>10 chronic pain studies with 2724 subjects consisting of several placebo-controlled trials of pregabalin and covering several conditions (e.g., fibromylagia and osteoarthritis).
Links clinical improvement on PI-NRS with anchor.
Mean change among ‘much improved’ on PGIC
Receiver operating characteristic (ROC) curve
Favorable: much or very much improved
Not favorable: otherwise</td>
<td>Clinically important difference = Reduction of about 2 points on PI-NRS; reduction of about 30% on PI-NRS.
ROC curve analysis:
sensitivity = 77% and specificity = 78%
area under curve = 78%
Consistent relationship between change in PI-NRS and PGIC regardless of study, disease type, age, sex, study results or treatment group.
Higher base-line scores required larger raw scores for clinically important differences.</td>
<td>Anchor-based approach</td>
</tr>
<tr>
<td>Angst F. et al. (2001)
Western Ontario and McMaster Universities Osteoarthritis Index: (WOMAC)
QOL Dimensions:
global score and individual domains of pain, stiffness and physical function. emotional
10 point scale
focus: change scores for worsening and improving patients
Anchor:
Retrospective within-patient transition rating on health in general related to osteoarthritic joint 3 months ago
Much worse
Slightly worse
Equal
Slightly better
Much better</td>
<td>122 patients with osteoarthritis of lower extremities
Before and after rehabilitation (3 months)
Links MCID to sample size for future studies
Mean effect = mean difference from baseline to 3 months <i>within each transition group separately for global WOMAC and each domain</i>
MCID for improvement = <i>Mean Effect (“slightly better”)– Mean Effect (“equal”)</i>
MCID for worsening = <i>Mean Effect (“slightly worse”)– Mean Effect (“equal”)</i>
Effect size (ES) = MCID/SD (baseline)</td>
<td>MCID and Sample Sizes
<table>
<tr><th>WOMAC</th><th>Worsening</th><th></th><th></th><th>Improvement</th><th></th><th></th></tr>
<tr><th>(range 0 to 10)</th><th>MCID</th><th>ES</th><th><i>n</i>*</th><th>MCID</th><th>ES</th><th><i>n</i>*</th></tr>
<tr><td>Pain (5 items)</td><td>1.10</td><td>0.49</td><td>66</td><td>0.75</td><td>0.33</td><td>142</td></tr>
<tr><td>Stiffness (2 items)</td><td>0.51</td><td>0.19</td><td>431</td><td>0.72</td><td>0.27</td><td>216</td></tr>
<tr><td>Physical Function (17 items)</td><td>1.33</td><td>0.61</td><td>43</td><td>0.67</td><td>0.31</td><td>167</td></tr>
<tr><td>Global (24 items)</td><td>1.29</td><td>0.62</td><td>42</td><td>0.67</td><td>0.32</td><td>153</td></tr>
</table>
*Sample size per group, assumes 80% power, 0.05 significance level (two-tailed for two-sample t-test)</td>
<td>Combined anchor and distributional approaches
Lower values of MCID for improvement (except stiffness) than worsening; improvement may be subjectively easier to notice
Larger sample sizes needed for less responsive sub scale (i.e., stiffness)</td>
</tr>
</table>

Wyrwich et al.(1999) (67) Chronic Heart Failure Questionnaire: CHQ QOL domains *Dyspnea (patient-specific 3–5 items affected by chest pain) scored as one (extreme amount) to seven (none at all)* *fatigue (4 items), emotional function (7 items) scored on a 7 point scale: one (worst), seven (best).* *baseline to follow-up (6,12,18 mths)* *change scores combined* Anchor: *Retrospective within-patient global assessment of change over last 4 weeks.* "Worse": −1 hardly any worse to −7 a very great deal worse "Better": +1 hardly any better to a +7 a very great deal better	605 cardiac patients in an outpatient setting. Secondary analysis of data from a RCT. Anchor standard for MCID *comes from previous research on CHQ of about 0.5 average per item change for each domain* *Classification of change on anchor: improved, stable, declined*. Anchor-based method linked QOL change scores with classification on anchor. Classification of change on anchor Minimal clinically important: 1 to 3/ −3 to−1 Moderate clinically important: 4 to 5/−5 to −4. Large clinically important: 6 to 7/ −7 to −6.	1 SEM based on baseline SD and Cronbach's alpha 1 SEM (Dyspnea) = 2.41 CHQ points per 5 items *equates to 0.48 average per item* 1 SEM (Fatigue) = 2.10 CHQ points per 4 items *equates to 0.53 average per item* 1 SEM (Emot. Func.) = 2.90 CHQ points per 7 items *equates to 0.41 average per item* SEM concords highly with MCID standard of 0.50 per item.weighted kappa (1.0, 0.87, 0.91 for the 3 domains) used to assess degree of association between 1SEM and MCID.	Combines the anchor and distributional approaches. 1 SEM as MCID also found in independent study (7) using the Chronic Respiratory Disease Questionnaire
Cella et al. (2002) (22) Functional Assessment of Cancer Therapy–Lung Questionnaire: FACT-L. *7-item Lung Cancer Sub scale (LCS)* *its Trial Outcomes Index (TOI)* adds scores on physical well-being sub scale and functional well-being sub scale of FACT-L to LCS scores	Randomized trial with 599 patients *advanced non-small cell lung cancer* *3 chemotherapeutic regimens (no difference in FACT-L)* *measurements at baseline and week 12* Three Complementary Approaches to MCID 1) Group means based on baseline differences in LCS and TOI scores on following anchors: *prior 6-month weight loss* (<5% vs.≥5%) *performance status (normal vs. some symptoms)* *primary disease symptoms* (≤1 vs. <1) 2) Group means based on changes in LCS and TOI scores over time on following anchors: *response to treatment (complete/partial vs. stable vs. deterioration)* *time to disease progression (<median time, >median time)* 3) Distribution-based criteria *1/3 and 1/2 standard deviation change standard of baseline scores, at week 12 scores, change scores.* *one standard error of measurement (SEM) at baseline and at week 12.* *Patients classifies as "declined" or "improved" (if − or + change scores > 1SEM respectively) or "unchanged" (if change score<1 SEM)*	Approximate MCID for 3 approaches converged; Cohen's kappa used to compare between empirically derived categories (i.e., anchor-based) and distribution-based categories Clinically meaningful differences: 2 to 3 points for the LCS (on 0-to-28-point scale) _7.1 points (=2*100/28) on 0-to-100 point sale_ 5 to 7 points for the TOI (on 0-to-84-point scale) _6 points (=5*100/84) on 0-to-100 point scale_	Combines and compares anchor and distributional approaches. Use of multiple clinical anchors to validate clinically meaningful difference.

Table 4. Checklist for Assessing Clinical Significance over Time in *QOL*[a]

What are the characteristics of the population for whom changes in QOL are reported?

- What are their disease (e.g., tumor type), treatment (e.g., duration), socio-demographic and cultural (e.g., age, ethnicity), and behavioral (e.g., alcohol use) characteristics?
- To what extent are the QOL data applicable to your patients?
- Is actual QOL status of individual patients reported (e.g., by providing confidence intervals, standard deviations, subgroup data, individual data plots), thus documenting the amount of individual variation in response to treatment?

Is the QOL questionnaire relevant, reliable, valid, and responsive to change?

- Is the questionnaire appropriate given the research objective and the rationale for QOL assessment?
- Is the questionnaire appropriate given the domains included and in light of the disease and population characteristics?
- Is the questionnaire reliable and valid? Is this information reported in the article?
- Is the questionnaire responsive to change? Is this information reported in the article?
- Is the questionnaire appropriate given practical considerations (e.g., regarding respondent burden and the availability of different language versions)?
- Are patients' baseline QOL scores close to the extremes of the response scale? Do the treatment groups differ in baseline QOL?

Are the timing and frequency of assessments adequate?

- Is a baseline assessment included?
- Is QOL assessed at appropriate times for determining minimally important change given the natural course of the disease?
- Is QOL assessed long enough to determine a clinical effect, taking disease stage into account?
- Is QOL assessed at appropriate times to document treatment course, clinical events, and post-treatment effects?
- Are standard research design procedures followed (e.g., avoidance of respondent burden, collection of data prior to treatment or consultation)?
- Is the timing of the QOL assessments similar across treatment arms?

Is the study adequately powered?

- Is the sample size appropriate for the research questions (e.g., by providing a power calculation)?
- Is a rationale and/or source for the anticipated effect size specified?
- Does the power calculation take into account: the scale range of the anticipated effect, the score distribution (i.e., magnitude and form), the number of outcome measures, and research hypothesis (i.e., equivalence versus difference)?

How are multiple QOL outcomes addressed in analyses?

- Is the adopted approach of handling multiplicity explicitly described?
- Which approach is taken: limiting the QOL outcomes, use of summary measures, adjustment of p-values, and/or multivariate statistical analysis and modeling?
- Did the interpretation of the results take the problem of multiple outcomes into account?

How are multiple time-points handled?

- Are the data presented in a meaningful and suitable way enabling an overview of QOL changes over time?
- Do the tabular and graphical presentations take the problems inherent in the data into account (e.g., presence of floor and ceiling effects, patient attrition)?
- Are the data appropriately analyzed (e.g., are all time points included, are missing data taken into account, are pre-treatment co-variates included)?
- Does the article provide sufficient information on the statistical models selected?

Can alternative explanations account for the observed change or lack of observed change?

Are dissimilar baseline characteristics adequately accounted for?

- Is the baseline QOL score used as a co-variate?
- Are missing data handled adequately?
- Does the article indicate how missing items within a questionnaire are handled?
- Does the article report the number of missing questionnaires at each scheduled assessment?
- Does the article report the reasons for missing questionnaires?
- Is there an association between patients' health status and missing QOL data?
- If patients with incomplete data are excluded from the analysis (e.g., by using complete case methods), does the article document that these are non-ignorable missing data?
- In cases of nonignorable missing data, are several analytical approaches presented to address possible bias in conclusions based on this QOL data set?
- Is observed survival difference combined with QOL in evaluating change?
- If patients have died in the course of the study, is mortality accounted for in the evaluation of QOL?
- Are summary indices (e.g., QALYs, Q-TWiST) or imputation techniques used?

Did the patient's QOL perspective change over time?

- Are changes in patient's internal standards, values, and/or the conceptualization of QOL explicitly measured?
- Are insignificant or small changes in QOL reported despite substantial changes in patient's health status (i.e., deterioration or improvement)?
- How likely is it that patients have changed their internal standards, values, and/or their conceptualization of QOL as a result of adaptation to deteriorating or improving health?

How is statistical significance translated into meaningful change?

- Does the article provide some guidance regarding the clinical importance of the observed change in QOL?
- To what extent is the statement of clinical importance appropriate and empirically warranted?

---

[a] Reprinted with permission from Ref. 34.

longitudinal, patient-derived QOL results presented in clinical trials and the clinical literature. These questions are based on the premise that detecting meaningful change depends on the adequacy of the research design, measurement quality, and data analysis.

RECENT DEVELOPMENTS

The One-Half Standard Deviation Rule

It would be desirable to simply define, at least initially, what a clinical significant result is likely to be. Emerging research comparing anchor- and distribution-based estimates provides an evolving standard as to what to use as an initial estimate (23). The anchor-based estimates averaging 0.5 per item on a 7-point scale appear to converge with an estimate of one-half standard deviation (SD) units. This latter estimate is derived through distribution-based methods, such as the effect size approach (22,64), SEM (7,67,68), and the standardized response mean (73). Potential moderating factors that could impact these estimates upward or downward are the method used to determine minimum difference estimates, the reliability of the measure and whether patients were suffering from acute or chronic conditions (23,74)

Empirical Rule Effect Size

Sloan et al. (27,28) have taken this concept one step further in the form of the ERES by combining Cohen's effect size categorization (22) with the empirical rule from statistical theory (75). The ERES is based on Tchebyschev's theorem and states that the distribution of any QOL tool is contained within six SDs of the observed values. The ERES entails the estimation of QOL change scores in terms of SD estimates, expressed as units on the theoretical range of a QOL instrument. Thus small, moderate, and large effect sizes for comparing QOL treatment groups turn out to be 3, 8, and 13%, respectively, of the theoretical range of any QOL tool. This simple and intuitive rule to identify the magnitude of clinically significant changes is likely to be easy for clinical researchers to comprehend. The rule can facilitate the design of clinical trials in terms of sample size calculations and interim monitoring of clinical trials. The ERES framework for a priori establishment of effect sizes is sample independent and thus an improvement over sample-dependent methods (5,21,76).

However, the simplicity of the ERES method gives rise to some challenges and questions. The theoretical range of the instrument is rarely observed in its entirety, necessitating the modification of the theoretical range to more practical limits before calculating the ERES estimate for one SD as necessarily 16.7% (i.e., one-sixth of distribution of observed values) of the range. Similarly, truncated distributions, where the patient population is homogeneously ill or uniformly healthy, can be accommodated by incorporating this knowledge into the definition of the appropriate range. These guidelines for clinical treatments can be used in the absence of other information, but will need modification in their application to idiosyncratic or unique clinical settings. More research is needed to examine the generalizability of such benchmarks across baseline patient health, severity of illness, and disease groups.

Group Change versus Individual Change

Distinctions should be made in determining the significance of change at the group versus the individual level. Every individual in a group does not experience the same change in outcomes (group level outcomes are assigned a mean change value). There is higher variability in individual responses than those of the group. Depending on the distribution of individual differences, the same group mean can have different implications for an individual (77).

The traversing of group and individual level QOL data entails procedures for moving from one level to the other involving two distinctive scientific traditions: deductive and inductive (31). A deductive approach is employed when one addresses the extent to which group data can be used to estimate clinical significance at the individual level. An inductive approach is used when one evaluates the extent to which individual change data can be brought to the group level to define clinical significance. Related to this is the fact that the lower end of a MID estimate may be useful for powering studies to detect meaningful differences between groups (with ES as low as 0.2), while the higher end of the MID estimate, especially for more sensitive tools, can be used to power studies detecting change at the individual level. Readers are advised to read Cella et al. (31) for a more detailed account.

Quality of Life as a "Soft" Endpoint

The "softness" of QOL as an endpoint, relative to say survival and tumor response, is cited as a particular barrier to implementation and interpretation of results (13). However, methodological and conceptual strides made in defining and measuring QOL, and the growing familiarity with the interpretation and potential utility of QOL data, make those concerns increasingly outdated.

Psychometric advances have been made in QOL assessment tools across disease areas (8,78–81). Funding opportunities to study QOL endpoints have allowed for study designs that are large enough to have power to detect meaningful differences (13). Moreover, accumulated experience with analyzing QOL endpoints have resulted in the recognition that their statistical challenges are no different from those of "hard" endpoints.

CONCLUDING REMARKS

Several suggestion on clinical significance are offered. First, the application of multiple strategies for determining clinical significant is recommended. Doing so would enable better interpretability and validity of clinically significant change, add to existing evidence of the magnitude of change that constitutes clinical significance, and would provide indicators of distributional parameters that create convergence or divergence in estimation of clinical significance. For example, Kolotkin et al. (46) found convergence between anchor- and distribution-based methods at moderate level of impairment but wide disparities at mild and severe levels of impairment.

Second, more research in needed into the effect of psychometric properties (i.e., reliability, validity and responsiveness of QOL instruments) have in quantifying clinically meaningful change (4,62,82). Similarly, research into the psychometric properties of global rating and health transition scales used in anchor-based methods is also needed. Global ratings tend to be single item measures and may therefore fall short in terms of explaining complex QOL constructs. Anchoring assessment also tends to be positively correlated with post-treatment states but with near-zero correlation with pretreatment states, suggesting a recall bias (83) or response shift (84). More research is needed to address the cognitive process used by patients to retrospectively assess changes in health over time (30).

Third, baseline severity results in regression to the mean (RTM), an error-based artifact describing the statistical tendency of extreme scores to become less extreme at follow-up. Failure to take this into account may lead to false conclusions that patients with severe impairments at baseline have shown clinical significant change, when in fact this was just RTM. The RTM also has a greater impact upon data when the measure is less reliable (4,85). More research is also needed into the effect of baseline QOL impairment on magnitude of clinically meaningful change (4,47,48,66,86,87). Similar research is needed in terms of the generalizability of the standardized benchmarks for determining clinically meaningful change, especially for distribution-based methods (4,66). Specifically, how satisfactory are the evolving benchmarks (effect sizes of 0.2, 0.5, and 0.8 for small, moderate, and large change, respectively) across different dimensions of QOL (e.g., mental versus physical), different disease groups (e.g., arthritis versus cancer), respondents (e.g., patients versus clinicians), measures (e.g., generic versus disease specific) patient populations (e.g., older versus younger), or patient conditions (e.g., improving versus deteriorating)?

Finally, care must be taken in presenting results of studies in a way that is familiar to the user of the information. For example, translating clinical significance into a number needed to treat (NNT) and a proportion of patients achieving various degrees of clinical benefit relative to the control may provide a desirable way to present study results (30).

BIBLIOGRAPHY

Cited References

1. Aaronson N. Methodologic issues in assessing the quality of life of cancer patients. Cancer 1991;67(3 Suppl):844–850.
2. Cella D, Bonomi AE. Measuring quality of life. Oncology 1995;9(11 Suppl):47–60.
3. Berzon R. Understanding and using health-related quality of life instruments within clilnical research studies. Quality of Life assessment in clinical trials: methods and practice. Oxford UK; 2000. p 3–15.
4. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003;56:397–407.
5. Juniper EF, Guyatt GH, Willan A. Determining a minimal important change in a disease-specific quality of life questionnaire. J Clin Epidemiol 1994;47:81–87.
6. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol 1991;59:12–19.
7. Wyrwich KW, Tiemey WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health related quality of life. J Clin Epidemiol 1999;52:861–873.
8. Cella D. Quality of life outcomes: measurement and validation. Oncology 1996;10(11 Suppl):233–246.
9. Sloan JA, O'Fallon JR, Summan VJ. Incorporating quality of life measurements in oncology clinical trials. Proceeding of the Biometrics Section of the American Statistical Association, 1998. p 282–287.
10. Spilker B. Quality of life and pharmacoeconomics in clinical trials. New York: Lippincott Raven; 1996.
11. Osoba D. What has been learned from measuring health-related quality of life in clinical oncology. Eur J Cancer 1999;35(11):1565–1570.
12. Sloan JA, Symonds T. Health-related quality of life measurement in clinical trials: when does a statistically significant change become relavant?, in Unpublished manuscript. 2003.
13. Frost MHSJ. Quality of Life Measures: A soft outcome—or is it? Am J Managed Care 2002;8(18, Supp.):S574–579.
14. Degner L, Sloan JA. Symptom distress in newly diagnosed ambulatory cancer patients as a preditor of survival in lung cancer. J Pain Symptom Mange 1995;10(6):423–431.
15. Chochinov HM, Kristjanson L. Dying to pay: the cost of end-of-life care. J Palliat Care 1998;14(4):5–15.
16. Silliman RA, Dukes KA, Sullivan LM. Breast cancer care in older women: sources of information, social support, and emotional health outcomes. Cancer 1998;83(4):706–711.
17. Spiegel D, Bloom JR, Kraemer H. Psychological support for cancer patients. Lancet 1989;2(8677):1447.
18. Sloan JA, Loprinzi CL, Kuross SA. Randomized comparison of four tools measureing overall quality of life in patients with advanced cancer. J Clin Oncol 1998;16:3662–3673.
19. Patrick DL, Erickson P. Applications of health status assessment to health policy. Qual Life Pharmacoeco Clin Trials 1996; 717–727.
20. Gold MR, Patrick DL, Torrance GW. Identifying and valuing: Cost Effectiveness in Health and Medicine. 1996; 82–134.
21. Juniper EF. The value and quality of life in Asthma. Eur Resp J 1997;7:333–337.
22. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale (NJ): Lawrence Erlbaum Associates; 1998.
23. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care 2003;41(5):582–592.
24. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989;10(4):407–415.
25. Jones P. Interpreting thresholds for a clinically significant change in health status (quality of life) with treatment for asthma and COPD. Eur Resp J 2002;19:398–404.
26. Wright JG, The minimally important difference:who's to say what is important? J Clin Epidemiol 1996;49:1221–1222.
27. Sloan JA, et al. Detecting worms, ducks, and elephants: a simple approach for defining clinically relevant effects in quality of life measures. J Cancer Integrative Med 2003;1 (1):41–47.

28. Sloan JA. Practical guidelines for assessing the clinical significance of health-related QOL changes within clinical trials. Drug Inf J 2003;37:23–31.
29. Sloan JA, et al. Assessing clinical significance in measuring oncology patient quality of life: introduction to the symposium, content overview, and definition of terms. Mayo Clin Proc 2002;77:367–370.
30. Guyatt GH, et al. Methods to explain the clinical significance of health status measures. Mayo Clin Proc 2002;77:371–383.
31. Cella D et al. Group vs individual approaches to understanding the clinical significance of differences or changes in quality of life. Mayo Clin Proc 2002;77:384–392.
32. Sloan JA, et al. Assessing the clinical significance of single items relative to summated scores. Mayo Clin Proc 2002;77:479–487.
33. Frost MH, et al. Patient, clinician, and population perspectives on determining the clinical significance of quality-of-life scores. Mayo Clin Proc 2002;77:488–494.
34. Sprangers MAG, et al. Assessing meaningful change in quality of life over time: A users' guide for clinicians. Mayo Clin Proc 2002;77:561–571.
35. Symonds T. et al. The clinical significance of quality-of-life results: practical considerations for specific audiences. Mayo Clin Proc 2002;77:572–583.
36. Testa MA, Interpretation of quality-of-life outcomes issues that affect magnitude and meaning. Med Care 2000;38:II166–II174.
37. van Walraven CM, Moher JL, Bohm C, Laupacis A. Surveying physicians to determine the minimal important difference: implications for sample-size calculation. J Clin Epidemiol 1999;52:717–723.
38. Deyo RA, Inui TS. Toward clinical application of health status measures: sensitivity of scales to clinically important changes. Health Serv Res 1984;19:278–289.
39. Johnson PA, Goldman L, EJ O. Comparison of the medical outcomes study short-form 36-item health survey in black patients and white patients with acute chest pain. Med Care 1995;33:145–160.
40. Brook RH, Ware JE, Davies-Avery A. Conceptualization and measurement of health for adults in the health insurance study. 1979.
41. Testa M, Lenderking WR, Interpreting pharmacoeconcomic and quality-of-life clinical trial data for use in therpeutics. Pharmacoeconomics 1992;2:107.
42. Testa M, Simonson DC, Assessment of quality-of-life outcomes. New Engl J Med 1996;28:835–840.
43. Llewellyn-Thomas HA, Williams JI, Levy L. Using a trade-off techniques to assess patients' treatment preferences for benign prostatic hyperplasia. Med Decis Making 1996;16:262–272.
44. Stucki G, Liang MH, Fossel AH. Relative responsiveness of condition specific and health status measures in degenerative lumbar spinal stenosis. J Clin Epidemiol 1995;48:1369–1378.
45. Mossey JM, Shapiro E. Self-rated health: a predictor of mortaility among the elderly. Am J Public Health 1982;72: 800–808.
46. Kolotkin RL, Crosby RD, Kosloski KD. Development of a brief measure to assess quality of life in obesity. Obes Res 2001;9:102–111.
47. Deyo RA, Inui TS, LJ. Physical and psychosocial function in rheumatoid arthritis: clinical use of a self-adminstered health status instrument. Arch Intern Med 1992;142:879.
48. Kolotkin RL, Crosby RD, Williams GR. Integrating anchor-based and distribution-based methods to determine clinically meaningful change in obesity-specific quality of life. Qual Life Res 2002;11:670.
49. Lydick E, Epstein RS. Interpretation of quality of life changes. Qual Life Res 1993;2:221–226.
50. Schwartz N, Sudman S, Autobiographical memory and the validity of retrospective reports. New York: Springer-Verlag.
51. Wyrwich KW, Metz S, Babu AN. The reliability of retrospective change assessments. Qual Life Res 2002;11:636.
52. Mozes B, Maor Y, Shumueli A. Do we know what global ratins of health-related quality of life measure? Qual Life Res 1999;8:269–273.
53. Kirwan JR, Chaput de Sainttonge DM, Joyce CRB. Clinical judgment in rheumatoid arthritis. III. British rheumatologists' judgment of 'change in response to therapy.' Ann Rheum Dis 1984;43:686–694.
54. Cella D, Hahn EA, Dineen K. Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening. Qual Life Res 2002;11:207–221.
55. Guyatt GH, Jaeschke R. Reassessing quality of life instruments in the evaluation of new drugs. Pharmacoeconomics 1997;12:616–626.
56. Lydick F, Yawn BP. Clinical interpretation of health-related quality of life data. Quality of Life assessment in clinical trials: methods and practice. Oxford (UK); 1998. p 299–314.
57. Hays RD, Wooley JM. The concept of clinically meaningful difference in health-related quality-of-life research. How meaningful is it? Pharmacoeconomics 2000;18(5):419.
58. Wright J, Young NL. A comparison of different indices of responsiveness. J Clin Epidemiol 1997;50:239–246.
59. Barber B, Santanello NC, Epstein RS. Impact of the global on patient perceivable change in an asthma specific QOL questionnaire. Qual Life Res 1996;5:117–122.
60. Ware J, et al. SF-36 health survey: manual and interpretation guide. Boston: The Health Institute; 1993.
61. Baker DW, Hays RD, Brook RH. Understanding changes in health status: is the floor phenomenon merely the last step of the staircase? Medical Care 1997;35:1–15.
62. Husted JA, Cook RJ, Farewll VT. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol 2000;53:459–468.
63. Speer DC, Greenbaum PD. Five methods for computing significant individual client change and improvement rates: support for an individual growth curve approach. J Consult Clin Psychol 1995;63:1044–1048.
64. Kazis L, Anderson JJ, Meenan RS. Effect sizes for interpreting changes in health status. Med Care 1989;27(Suppl 3):S178–189.
65. Guyatt GH, Bombardier C, Tugwell PX. Measuring disease-specific quality of life in clinical trials. CMAJ 1986;134:889–895.
66. Samsa G, Edelman D, Rothman ML. Determining clinically important differences in health status measures: a general approach with illustration to the Health Utilities Index Mark II. Pharmacoeconomics, 1999;15:41–55.
67. Wyrwich KW, Nienaber NA, Tiemey WM. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. Med Care 1999;37:469–478.
68. Wyrwich KW, Tiemey WM, Wolinsky FD. Using the standard error of measurement to identify important changes on the Asthma Quality of Life Questionnaire. Qual Life Res 2002;11:1–7.
69. Miller GG, The magic number seven plus or minus two: some limits on our capacity for processing information. Psychol Rev 1956;63:81–97.
70. Farrar JT, et al. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. Pain 2001;94:149–158.
71. Angst F, Aeschlimann A, Stucki G. Smallest Detectable and Minimal Clinically Important Differences of Rehabilitation Intervention With Their Implication for Required Sample Sizes Using WOMAC and SF-36 Quality of Life Measurement

Instruments in Patients With Osteoarthritis of the Lower Extremities. Arthritis Care and Research 2001;45: 384–391.

72. Cella D, et al. What is a clinically meaningful change on the Functional Assessment of Cancer Therapy-Lung (FACT-L) Questionnaire? Results from Eastern Cooperative Oncology Group (ECOG) Study 5592. J Clin Epidemiol 2002;55:285–295.
73. McHorney C, Tarlov A. Individual-patient monitoring in clinical practice: are available health status measures adequate? Qual Life Res 1995;4:293–307.
74. Stewart AL, Greenfield S, Hays RD. Functional status and well-being of patients with chronic conditions: results from the medical outcomes study. JAMA 1989;262:907–913.
75. Pukelsheim F. The three sigma rule. Am Stat 1994;48:88–91.
76. Juniper EF, Guyatt GH, Feeny DH. Measuring quality of life in childhood asthma. Qual Life Res 1996;5:35–46.
77. Guyatt G, et al. Interpreting treatment effects in randomized trials. BMJ 1998;316:690–693.
78. Chassany O, et al. Patient-reported outcomes: the example of health-related quality of life - a European guidance document for the improved integration of health-related quality of life assessment in the drug regulatory process. Drug Inf J 2002;36:209–238.
79. Speilberger C. State-Trait Anxiety Inventory: STAI (Form Y). Palo Alto (CA): Consulting Psychologists Press Inc.; 1983.
80. Radloff L. The CES-D scale: a self-report depression scale for research in the general population. Appl Psychol Meas 1977;1:385–481.
81. McNair DM, Lorr M, Droppleman LF. Profile of mood states manual. EdiTS 1992.
82. Hays RD, Anderson R, Revicki D. Psychometric considerations in evaluating health-related quality of life measures. Qual Life Res 1993;2:441–449.
83. Norman GR, Stratford PW, Regehr G. Methodological Problems in the Retrospective Computation of Responsiveness to Change: The Lessons of Cronbach. J Clin Epidemiol 1997;50(8):869–879.
84. Schwartz CE, Sprangers MAG. Methodological approaches for assesing response shift in longitudinal health-related quality-of-life research. Soc Sci Med 1999;48:1531–1548.
85. Moser MT, Weis J, Bartsch HH. How does regression to the mean affect thresholds of reliable change statistics? Simulations and examples for estimation of true change in cancer-related quality of life. Qual Life Res 2002;11:669.
86. McHorney C. Generic health measurement: past accomplishments and a measurement paradigm for the 21st century. Ann Int Med 1997;127:743–750.
87. Stratford PW, Binkley J, Riddle DL. Sensitivity to change of the Roland-Morris Back Pain Questionnaire: part 1. Phys Ther 1998;78:1186–1196.

See also HEART, ARTIFICIAL; HOME HEALTH CARE DEVICES; STATISTICAL METHODS.

RADIATION DETECTORS. See RADIATION PROTECTION INSTRUMENTATION.

RADIATION DOSE PLANNING, COMPUTER-AIDED

JONATHAN G. LI
University of Florida
Gainesville, Florida
LEI XING
Stanford University
Stanford, California

INTRODUCTION

The term radiation dose planning in radiation therapy refers to the process of designing treatment strategies for optimally delivering a desired radiation dose to the intended volume while minimizing the dose to healthy tissues as much as possible, and estimating the radiation dose throughout the irradiated volume. In pursuit of better energy deposition characteristics, higher and higher energy photon and electron sources have been developed and used throughout the history of radiation therapy. Ionizing radiation dose can be delivered externally using electron accelerators or radioactive sources, or internally by implanting radioactive sources in the tumor volume. Heavy particles (e.g., protons, neutrons, and carbons) have also been used in radiation therapy, although the high cost of these machines have limited their widespread application. Our discussion will be restricted to radiation dose planning with external photon beams generated with medical megavoltage electron linear accelerators (linacs), which is by far the most widely used method of radiation cancer treatment.

Most medical linacs are isocentrically mounted, that is, they can rotate around a horizontal axis in 360°. Combined with the rotation of the treatment couch, radiation can be directed toward the patient from all possible directions. Two pairs of collimators or jaws moving in the orthogonal direction are usually built in the linac head to collimate the beam into a square or rectangular shape with continuously variable field sizes. Figure 1 shows a typical medical linac. A treatment usually involves several beams from different directions with different beam weights and beam-modifying devices in order to deliver a uniform dose to the target and to limit dose to surrounding normal tissues. Commonly used beam-modifying devices include custom-made blocks, which further collimate the radiation beam into any arbitrary shape, and wedge filters, which are wedge-shaped metal absorbers placed in the path of the beam to cause a tilt of the resulting isodose curves in the patient. To achieve optimal results, computer-aided radiation dose planning has played an increasing role in radiation therapy.

Many steps are involved in planning a radiation treatment. One of the first steps in the process is to establish a three-dimensional (3D) patient anatomy model based on the patient's image information. Toward this goal, one needs to delineate the areas to be treated (targets) and any dose-limiting normal structures. Developments in 3D imaging, digital imaging processing, and multimodality imaging have greatly aided this process. Treatment strategies are then developed where radiation beams are chosen for optimal target coverage without delivering excessive dose to critical structures. Radiation dose throughout the irradiated volume is calculated, and the plan evaluated. Several trial and error efforts are usually required before a clinically acceptable plan is generated. Most of the treatment planning is now performed using computers with dedicated software called a treatment planning system (TPS). With the availability of fast processors and large random access memory (RAM), voluminous patient data depicting accurate 3D geometry and anatomy from a computed tomography (CT) scanner can be manipulated in a TPS, giving radiation oncologists and treatment planners better visualization of the internal structures and greater ability to tailor the treatment to the particular circumstances. Dose display and plan evaluation tools have made it easier to compare different treatment plans. This article concentrates on new developments in radiation dose planning since the first edition of this encyclopedia (1). The widespread clinical implementation of 3D planning techniques (e.g., virtual simulation, image registration, model-based dose calculation algorithms, and treatment plan optimization) have made a major impact on the current practice of radiation therapy. All of the new developments are computationally intensive and require large RAM for image processing. Their increasing role in radiation dose planning has tied closely to the development and availability of fast computers and large RAM.

VIRTUAL SIMULATION

Virtual simulation is a process of delineating target and normal organs and designing treatment field arrangements and portals on a computer, based on a detailed 3D model of a patient built from a sequence of closely spaced transverse images from a CT scanner. Virtual simulation is now widely used and has replaced conventional simulation for most of the treatment planning except in some simple or emergency cases. Conventionally, the patient simulation is done using a simulator with the patient in the treatment position. A conventional simulator duplicates a linac geometry, but uses a diagnostic kilovoltage X-ray tube to enhance image contrast. Two-dimensional (2D) projection radiographs are taken from various gantry positions that have been chosen for treatment. Target volume and normal structures are drawn on the 2D simulation films and correct positioning of the fields and shielding blocks can be obtained in relation to

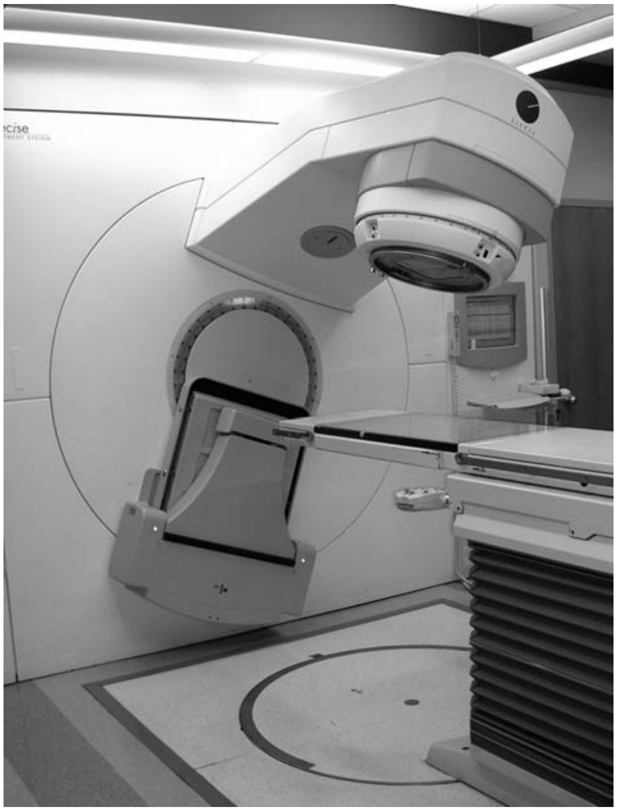
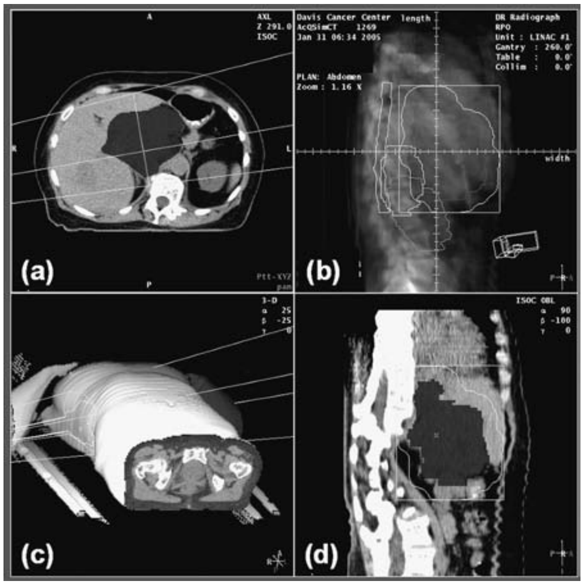


Figure 1. Photograph of an isocentric medical electron linear accelerator (Elekta Precise, Elekta Inc.).



AXL
Z 291.0
ISOC

Davis Cancer Center
AcQSimCT 1269
Jan 31 06:34 2005

PLAN: Abdomen
Zoom : 1.16 X

DR Radiograph
RPO
Unit : LINAC #1
Gantry : 260.0°
Table : 0.0°
Collim : 0.0°

3-D
α 25
β -25
γ 0

ISOC OBL
α 90
β -100
γ 0

Figure 2. A screen capture of a commercial virtual simulation software (AcQsim 4.9.1, Philips Medical Systems) depicting the various tools available for treatment planning. (a) Axial view through the isocenter of a right posterior oblique (RPO) beam. (b) Digitally reconstructed radiograph of the RPO beam with the beam shape (yellow) and various structures projected onto it. (c) A 3D rendering of the patient's external contour. (d) Sagittal view through the middle of the patient.

anatomical and external landmarks. These geometries are transferred to the linac for patient treatment. In contrast to conventional simulation, where the treatment fields are designed on 2D radiographs while the patient stays on the simulation table, virtual simulation relies on the 3D CT data acquired with the patient in the treatment position and using the same immobilization device as will be used for treatment. The volumetric CT data represent the "virtual" or digital patient. Target delineation and field design are then done off-line without the patient's presence using dedicated virtual simulation workstations or treatment planning systems. Figure 2 shows a screen capture of a commercial virtual simulation package with different views that aid in visualizing the patient's internal structures and in the selection of beams and beam portals for treatment.

A typical patient CT data set has > 100 axial slices, each of which contains 512×512 picture elements (pixels). With 16 bits per pixel, a CT data set can easily run over 50 MB. Manipulating, displaying, and storing such voluminous data sets require enormous computer resources and have only been made possible in the past two decades due to the dramatic advancements in computer hardware. Historically, the evolution of radiation therapy has been strongly dependent on the available computer and imaging technologies and this trend is expected to continue in the years to come as radiation therapy proceeds into an era of computer-controlled delivery and real time image guidance and feedback.

An important step in virtual simulation is the generation of digitally reconstructed radiographs (DRR) for treatment planning and verification (2). A DRR is a computer-generated beam's-eye-view image that simulates the X-ray attenuation property and projection geometry of a conventional simulator. It is obtained by tracing the divergent path of X rays from the radiation source through the 3D patient CT data set onto a plane beyond the data set and orthogonal to the central ray. Fast ray-tracing algorithms have been developed to calculate the radiological path through a CT data set (3,4). Compared with radiographs from a conventional simulator, DRRs offer several distinct advantages. Whereas tumors and normal anatomic structures are sometimes difficult to visualize on a conventional radiograph because of the overlapping effect, they can usually be discerned much better on an axial CT image. The contours of the tumor and normal structures drawn on the CT images can be projected onto the DRR. This greatly helps the treatment planner in selecting beam geometries that will irradiate the target while avoiding critical anatomic structures. The DRRs can be generated quickly for any angular projection through the body, whereas using a simulator and film requires many minutes of patient setup and film developing for each projection. The brightness and contrast of a DRR can be digitally manipulated to bring out certain anatomic features. This can be used for patient setup verification by comparing the DRR with a radiographic image of the treatment field (portal image) obtained on the treatment machine. Figure 3 illustrates a DRR with several structures and the treatment field shape projected onto it and the corresponding portal image of the same

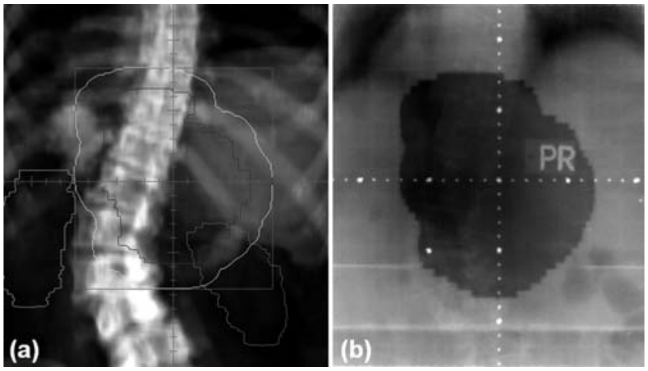


Figure 3. Comparison of a digitally reconstructed radiograph (a) and the corresponding portal film (b). The beam shape (yellow) was projected onto the DRR for treatment verification.

field. Such comparison verifies both the positioning of the patient and the treatment field shape.

IMAGE REGISTRATION

Radiotherapy treatment planning has been based mostly on CT images. Advantages of CT images include high spatial integrity, high spatial resolution, excellent bony structure depiction, and the ability to provide relative electron density information used for radiation dose calculation. However, CT images do not provide good soft tissue contrast. Moreover, CT images are anatomic in nature, that is, they provide snapshots of the patient's anatomy without any functional information of various organs or structures. Other imaging modalities, especially the magnetic resonance imaging (MRI) and positron emission tomography (PET), have been used increasingly in radiation therapy planning in conjunction with CT images. Magnetic resonance imaging provides better soft tissue contrast than CT images, and is the modality of choice when delineating treatment target of brain tumors. Positron emission tomography provides functional information about tumor metabolism and is a useful tool in tumor diagnosis, staging, target volume delineation and assessment of therapeutic response. However, current MRI and PET devices suffer from several drawbacks that make them unsuitable for radiotherapy planning as the sole modality. Imaging artifacts and geometric distortions exist in MR images. PET has a lower resolution than CT and contains no anatomy information of the normal structures. Information derived from MRI or PET needs to be fused or registered with the corresponding CT images for treatment planning.

Three-dimensional image registration aims at finding a geometric transformation that maps the volume elements (voxels) of one tomographic data set onto another tomographic data set of the same or different modality. Since different scans are done at different times and possibly with different patient immobilization devices, image registration is difficult to perform manually, and sophisticated computer algorithms have been developed for various registration applications (5). Some algorithms make use of common geometric features, such as points, lines, or surfaces identified on both image sets to determine the transformation. These features can be extracted either manually or automatically, and the accuracy in the identification of the common features directly affects the registration accuracy. Other image registration algorithms are based on information contained in the whole image sets (e.g., the image intensity values) and seek to maximize the amount of shared information in the two data sets. Image registration between images of different modalities based on shared information methods is especially challenging as different image modalities involve entirely different physical processes, and therefore voxels of the same tissues can appear very differently on different images. For example, high voxel intensity areas on a CT image (e.g., the bone) may correspond to dark areas on a MRI image. Figure 4 shows an axial CT image of a brain with the registered MR image. The registered secondary image (MR image in this case) is interpolated and resampled to give the axial view at the same axial position as the primary CT image. Contours of the tumor and normal structures drawn on one image are transferred and displayed automatically onto the other image.

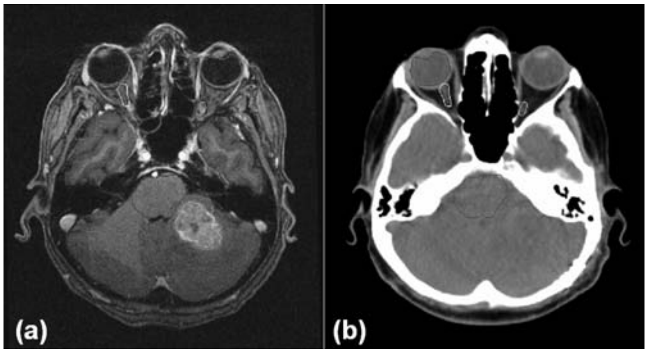


Figure 4. Registered MR (a) and CT (b) images of a brain tumor. The tumor (red contour) is readily seen on the MR image.

Various degrees of simplification have been assumed in the medical image registration. The simplest and widely used registration assumes that the transformation is rigid body, where only image translation and rotation is allowed with six degrees of freedom. Rigid body transformation preserves all distances in an image. The most important and successful application of rigid body registration is in the head, and particularly the brain. Image registration between MRI and CT has been the standard practice for the treatment planning of brain tumors in most cancer centers in the United States. In more general situations, where other parts of the body are involved and images are acquired under different conditions or using different modalities with the patient in different positions, more degrees of freedom are often needed. The simplest of the nonrigid body transformations, the affine transformation, introduces an additional six degrees of freedom with anisotropic scaling and skews. An affine transformation preserves collinearity (i.e., all points lying on a line initially still lie on a line after the transformation), parallelism (parallel lines stay parallel after the transformation), and ratios of distances in an image, but not necessarily angles or distances. Image registration with more degrees of freedom than the affine transformation is an active area of research and its application in radiation therapy has so far been limited (6–8). This most general type of image registration technique is refereed to as deformable image registration. Commonly used deformable models can be categorized into two categories: free form B-spline and biomechanical finite element methods. Clinically, the need for a robust deformable registration technique is ever increasing because of the recent development in image-guided radiation therapy and much research is being carried out in this area. In four-dimensional (4D) radiation therapy of breast cancer, for example, where time-resolved CT images were used to monitor anatomic changes due to breathing, reconstruction of dose to different organs relied on the registration between the voxels at different phases of the breathing cycle (9). However, a detailed discussion of the subject is clearly beyond the scope of this article and the readers are referred to the references cited above.

The drive for more accurate registration between PET and CT images has led to the development of PET–CT, a new imaging technology that combines high quality PET and CT scanners in a single device (10,11). With PET–CT, patients undergo CT and PET scans sequentially under the same immobilization with a table translation, therefore providing simultaneous anatomic and metabolic information under almost identical conditions. Image registration becomes a simple process of correcting for the known table translation. An added advantage of PET–CT over PET only is the faster scan time, where the CT scans, which only take a few minutes, are used for attenuation correction. With PET only, the attenuation correction is obtained from a transmission scan, which takes on the order of 30 min. Although incorporating functional imaging in radiotherapy treatment planning is relatively new, the interest is increasing steadily and many studies have shown that new functional information would change patient management decisions in many disease sites (12–14).

RADIATION DOSE CALCULATION

Dose calculation plays a pivotal role in radiation therapy treatment planning. To achieve the expected therapeutic results, radiation dose distribution throughout the irradiated volume should be known to a desired degree of accuracy. Generally speaking, there are two major types of dose calculation algorithms: correction and model based. Correction-based methods compute the dose distributions in patients by correcting the dose distributions of similar geometries in a homogeneous water phantom for the beam modifiers, patient contours, tissue heterogeneities, and volume scattering effect. There are several algorithms for heterogeneity corrections. Two simple one-dimensional (1D) methods are the ratio of TPR, in which only densities along primary photon path are considered, and the power-law (or Batho) method, which takes the depth of the heterogeneity with respect to the depth of the point of measurement into account. Sontag and Cunningham (15) implemented the first algorithm, often referred to as the equivalent tissue/air ratio method, to estimate scatter dose in three dimensions and took advantage of the detailed anatomical information derived from CT images. Wong and Purdy (16) examined eight methods of photon inhomogeneity correction for their photon transport approximations and improved correction-based algorithms by introducing more realistic transport models into the calculations. The volume scattering effects (scatter dose as a function of field size and shape) are often computed by using the equivalent square field method and/or Clarkson integration (17). Some pencil beam methods, like the finite-size pencil beam algorithm (18), are also classified as the correction-based methods. Model-based algorithms simulate the treatment situation from first principle and can directly calculate the dose distributions in a patient for a given beam energy, geometry, beam modifiers, patient contour, and tissue heterogeneities. The kernel-based convolution–superposition and Monte Carlo method are representatives of the kind. These commonly used algorithms are briefly summarized below.

Correction-Based Methods

Calculation of radiation dose is conventionally done by interpolating from measured data in a water phantom and correcting for any nonstandard situations. To serve this purpose, large amounts of beam data need to be collected that would allow for data interpolation with reasonable degree of accuracy. Measurement is usually done with a computer-controlled automatic scanning system. Dose as a function of depth along the central axis of the beam and off-axis distance along the transverse directions is measured for a large number of field sizes at a fixed source-to-water surface distance (SSD). These measurements need to be repeated for both open and wedged fields. Dose as a function of depth along the central axis, when normalized to a given depth (usually the depth of maximum dose for a reference field size), is called the percentage depth dose (PDD) and was an important quantity in the early days of radiation therapy when most of the treatment setups were at a fixed SSD. Figure 5a illustrates the definition of PDD and can be measured readily with a scanning system. Modern radiotherapy with isocentric-mounted gantries typically uses fixed SAD setups. As the gantry rotates around the patient with the isocenter near the center of the tumor, the SSD (and the tumor depth) changes. The quantity most useful for dose calculation in these cases is the tissue/phantom ratio (TPR), which is defined as the ratio of dose on the central axis at depth d in a phantom to dose at the same point and field size at a reference depth d_{ref} and is shown in Fig. 5b and c. When the depth of dose maximum is chosen as d_{ref}, TPR is sometimes called the tissue-maximum ratio (TMR). Whereas PDD for the same field size varies significantly with SSD, TPR is essentially independent of SSD. Therefore, a single TPR table can be used for all SSDs. Dose calculations based on TPR or other TPR derivatives have been discussed extensively (1), especially in the textbooks of Johns and Cunningham (19) and Khan (20).

The limitations of correction-based methods are numerous. Patient geometry and internal structures can deviate significantly from a flat and homogeneous water phantom.

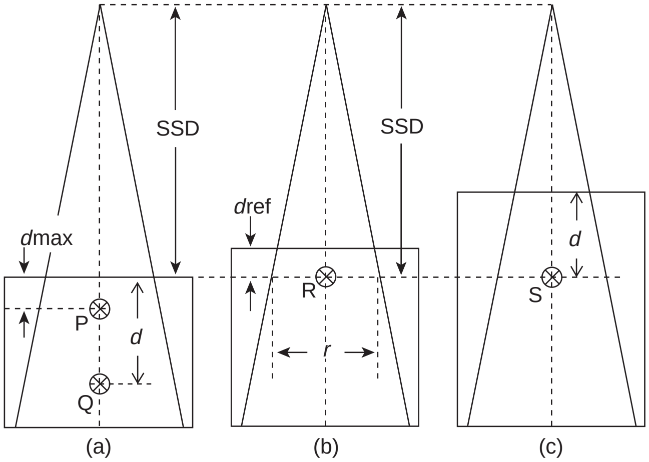


Figure 5. Schematic diagram illustrating the definition of the PDD and the phantom ratio (TPR). The PDD at a depth d is defined as the dose at the point, $D(Q)$, to the dose at point P, at the same SSD. It depends both on the SSD and the field size r. The TPR at depth d is defined as the ratio of dose at point S to dose at point R, at the same source-to-axis distance (SAD).

Although corrections can be made to account for the deviation, they are approximate and do not fully take into account the effect of secondary electron transport. This effect arises because secondary electrons ejected by megavoltage photons deposit a significant fraction of their energy far away from their point of origin. In regions where there is an imbalance between secondary electrons coming in and going out, such as near the interface between tissue–bone or tissue–air or in the beam penumbra, a condition known as charged particle disequilibrium exists. Correction-based methods are likely to produce erroneous results in these regions. Nowadays, correction-based dose calculation algorithms have mostly been replaced by model-based methods in commercial TPSs, where data measured in a water phantom under standard conditions are used to fit a few machine-specific model parameters, which in turn are used to calculate dose under the existing patient geometry. These include the convolution–superposition method and the Monte Carlo method. Nevertheless, correction-based methods are intuitive and have often been used as a quality assurance procedure to double-check the results of computer-generated treatment plans at one or a few calculation points for both conventional and intensity-modulated radiation therapy (21,22).

Convolution–Superposition Method

In the convolution–superposition techniques (23–27), the dose deposition is viewed as a superposition of appropriately weighted kernels of point irradiations and the superposition can be efficiently evaluated by means of convolution if the kernels are considered as spatially invariant. The kernels, representing the energy transport and dose deposition of secondary particles stemming from a point irradiation, can be calculated by Monte Carlo simulation (28). The Monte Carlo method computes dose distributions by simulating particle transport in a patient and will be described in the next section. Model-based algorithms are capable of accounting for electronic disequilibrium, and therefore are more accurate in dealing with tissue inhomogeneity and calculating dose in the electronic disequilibrium regions.

A thorough review of dose calculations in radiation therapy has been given by Ahnesjo and Aspradakis (27). There are a few different convolution–superposition methods, which can be divided into point kernel models and pencil kernel models. Mathematically, the dose at a spatial point, $D(r)$, comprises contributions from the shower of secondary particles resulting from primary interaction sites at r'. Assuming that the direction of all incident particles is parallel to the central axis throughout the beam, the total dose deposited by a monoenergetic beam irradiating a homogeneous medium can be expressed as a convolution operation:

$$D(r) = \int T(r\prime)A(r - r\prime)dr\prime$$

where $A(r - r')$ is the dose spread kernel that describes the mean fraction of energy deposited per unit volume at r per photon interaction at r', and $T(r')$ is the total energy released by primary X-ray interactions per unit mass, or TERMA. The above formula forms the basis for point kernel-based modes. Although the formulation of the point kernel model is simple and appealing, the demand on computer time is enormous due to the need for modeling various second-order beam characteristics. As emphasized by Ahnesjo and Aspradakis (27), there are three major issues that must be addressed for accurate dose calculation: broad primary beam spectral, beam divergence, and tissue density inhomogeneity. After all these factors are considered, the convolution–superposition dose calculation becomes a computationally intensive task. Therefore, their clinical implementation is tied closely to the availability of fast computers. The use of fast transform convolution techniques, such as the fast Fourier transform (FFT) method, to carry out the discrete convolution greatly facilitate the calculation process. Another widely accepted approach is the so-called collapsed cone convolution (26) based on an angular discretization of the kernel.

The pencil beam method is essentially a hybrid algorithm that fully accounts for beam modulations and field shapes, but relies on broad beam scaling/correction methods to handle heterogeneities and patient contour changes (18,27). The poly-energetic pencil beams are generally compiled from a linear combination of monoenergetic pencil beams within the constraints of a spectrum model to reproduce a set of depth-doses. The pencil beam kernels can also be determined by direct Monte Carlo calculation, or derived experimentally based on scatter factor differentiation.

Monte Carlo Method

Monte Carlo is a statistical simulation method that simulates the tracks of individual particles and all subsequently generated particles as they traverse the medium (28–30). The method takes into account all the relevant interactions and physical processes that take place in the medium. For each particle, the probability and types of interaction at a point, its path length, and its direction are sampled from probability distributions governing the individual physical processes using machine-generated pseudo-random numbers. These particles and all the daughter products are followed until they either are fully absorbed or escape from the region of interest. Dose distribution and other macroscopic quantities can be calculated by simulating a large number of particle histories. Provided that the physical models used in the simulation are accurate, Monte Carlo simulation can accurately predict radiation dose distribution as it simulates particle transport and energy deposition from first principles. In particular, Monte Carlo simulation can calculate dose in regions of charged particle disequilibrium more accurately than any other existing dose calculation algorithms. For a detailed discussion on the Monte Carlo method in radiotherapy dose calculation, see the chapter on "Radiation therapy treatment planning, Monte Carlo calculations".

An intrinsic limitation of the Monte Carlo method is that the results contain statistical noise. The statistical error of Monte Carlo calculation is proportional to $1/\sqrt{N}$, where N is the number of simulated particle histories. To obtain dose distributions with acceptably small statistical uncertainty, a large number of particle histories need to be simulated, which makes the Monte Carlo method computationally intensive. The prohibitively long computation time has been the main obstacle for its routine clinical application. However, the computer speed has been increasing exponentially (Moore's law) since the initial application of the Monte Carlo method in medical physics in the 1970s and this trend is expected to continue. With the rapid increase in computer speed and the development of innovative variance reduction techniques (31), Monte Carlo simulation is fast becoming the next generation dose calculation engine for radiation treatment planning of photon and electron beams. The first commercial TPS that employs Monte Carlo dose calculation engine has already been released (PEREGRINE, North American Scientific) with dose calculation time on the order of minutes on two 2.4 GHz Pentium Xeon processors with a grid size of $0.5 \times 0.5 \times 0.5$ cm^3. The clinical impact of using Monte Carlo dose calculations is a subject of considerable interest and needs to be evaluated carefully (32).

TREATMENT PLAN OPTIMIZATION

Many treatment-planning parameters affect the quality of a treatment plan. In conventional 3D conformal radiotherapy (3DCRT), the treatment parameters that are at the treatment planner's disposal include the beam modality and energy, number of beams, beam and treatment couch angles, wedge angles and orientations, radiation-defining blocks, and the weights of each beam. Since the number of beams used in 3DCRT is usually small (3~5), clinically satisfactory plans can be produced manually in a trial-and-error fashion. Disease- and site-specific treatment techniques developed over the decades help to reduce the number of adjustable parameters significantly. With the development and clinical implementation of intensity-modulated radiation therapy (IMRT), where the fluence of each radiation beam can be modulated arbitrarily in order to achieve a highly conformal radiation dose distribution, manual planning is not practical, and computer optimization algorithms have to be used to design complicated beam fluences (33). Figure 6 shows the fluence maps of a seven-beam IMRT plan from a commercial inverse planning system used for the treatment of a head and neck tumor. Each beam is divided into a grid of 1 × 1 cm beamlets, and each beamlet can take a fixed number of fluence levels. Such fluence modulation can be achieved with a computer-controlled multileaf collimator (MLC). The treatment planning process where the desired goals (the output) (e.g., the desired dose distribution and dose and/or dose–volume constraints, are specified first, and computer optimization is used to determine the needed beam fluences (the input) is sometimes termed inverse planning.

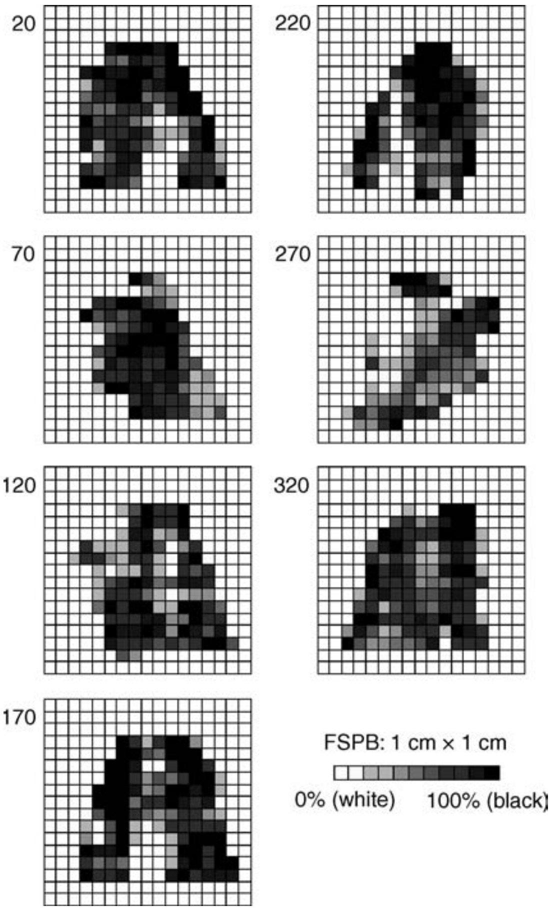


Figure 6. Fluence patterns of a seven-beam IMRT treatment plan of a head-and-neck case from a commercial inverse planning system (CORVUS 5.0, North American Scientific).

As with any mathematical optimization problems, inverse planning starts with the construction of an objective function. The objective function, with its value the sole measure of the quality of a plan, guides the optimization algorithm. Additional constraints can also be imposed which limit the solution space. The compromise between delivering a high dose to the target while limiting the dose to critical structures is implicitly built into the objective function. One of the simplest forms of the objective function that has been used often is the weighted least-squares function, which can be expressed as

$$F_{\text{obj}} = \frac{1}{N}\sum_{n=1}^{N} r_\sigma [d(n) - d_0(n)]^2$$

where $d_0(n)$ and $d(n)$ are the prescribed and calculated dose distributions, respectively, n is the voxel index, N is the total number of voxels, and r_σ is the relative importance factor of structure σ, which controls the tradeoffs between different structures. The calculated dose to voxel n can be obtained from the weighted sum of all the beamlets as

$$d(n) = \sum_{i=1}^{I}\sum_{j=1}^{J(i)} w_{ij} D_{ij}(n)$$

where I is the total number of beams, $J(i)$ is the total number of beamlets of the ith beam, $D_{ij}(n)$ is the relative dose contribution from the jth beamlet of the ith beam to voxel n, and w_{ij} is the weight of the beamlet (i, j). The parameter $D_{ij}(n)$ can be precalculated, so minimization of F_{obj} with respect to w_{ij} produces an optimal plan in the least-squares sense. The choice of the size of the dose calculation grid is an important consideration in IMRT plan optimization. While larger grid sizes reduce the model size and can speed up the computation considerably, too large a grid size will introduce aliasing artifacts. An information theory-based Fourier analysis of a 1 × 1 cm 6 MV photon beamlet from a medical linac predicted that an isotropic dose grid with < 2.5 mm spacing is sufficient to prevent dose errors larger than a percent (34). The tradeoffs between target coverage and critical structure sparing, which is controlled by the weighting factors r_σ, are usually not known *a priori*, and iterative adjustments are required to tailor the treatment plan to each particular circumstance. Algorithms aiming to automate the selection of the weighting factors have been proposed, which promises to significantly reduce the labor-intensive effort of the trial-and-error determination of the factors. In addition, the concept of intravoxel tradeoff has been introduced and an effective approach to model the tradeoff based on voxel specific weighting factors has been proposed (35,36). To obtain an adequate set of local weighting factors with a manageable amount of computing time, algorithms based on *a priori* dosimetric capability information and *a posteriori* adaptive algorithms were developed. With the introduction of intravoxel tradeoff, the IMRT dose distribution has been remarkably improved in comparison with the conventional plan obtained with structurally uniform weighting factors.

Clinical implementations of IMRT are dominated by dose- and/or dose–volume-based objective functions. Other forms of objective functions, particularly those employing biological indexes, such as the tumor control probability (TCP), normal tissue complication probability (NTCP), and equivalent uniform dose (EUD) (37,38), have also been applied to IMRT plan optimization. The use of biological indexes is especially appealing, as these are the ultimate measures of treatment outcomes. However, there is a lack of clinical and biological data to support these models, and their clinical use at present is not warranted. For example, current TCP models do not contain spatial information, and a cold spot would have the same adverse effect on the TCP irrespective of its location. In reality, the treatment outcome very much depends on the location of the cold spot, that is, whether it is in the periphery or in the middle of the target. Moreover, the TCP-, NTCP-, and EUD-based biological models as they are currently implemented are equivalent to voxel dose-based physical models in a multicriteria framework (39). It is expected that as radiation biology research leads to more robust models, biological based models will become more widely adopted in radiation therapy treatment plan optimization.

A practical approach to bridge the gap between biologically insensible physics-based formalism and a clinically impractical biology-based model is to establish a clinical outcome driven objective function by seamlessly incorporating a clinical endpoint to guide the treatment plan optimization process. Indeed, currently available dose-based objective functions do not truly reflect the nonlinear relationship between the dose and the response of tumors and tissues. On the other hand, biologically based inverse planning involves the use of a number of model parameters whose values are not accurately known and entails a prescription in terms of biological indexes. Recently, Yang and Xing proposed an effective method for formalizing the existing clinical knowledge and integrating clinical endpoint data into inverse planning (40). In their approach, the dose–volume status of a structure was characterized by using the effective volume in the voxel domain. A new objective function was constructed with incorporation of the volumetric information of the system so that the figure of merit of a given IMRT plan depends not only on the dose deviation from the desired dose distribution, but also the dose–volume status of the involved organs. The incorporation of clinical knowledge allows us to obtain better IMRT plans that would otherwise be unattainable.

The considerable interest in developing faster and more robust IMRT optimization algorithms has led to research collaboration between the radiation oncology community and the operations research (OR) community (41). The OR community has long investigated various optimization technologies and has the expertise in addressing large scale, complex optimization problems [see, e.g., the excellent textbooks of Winston (42) and Chong and Zak (43)]. Such collaboration is expected to enhance the IMRT model and algorithm development tremendously due to the large sizes of the problems in IMRT optimization combined with the clinical need to quickly solve each optimization for interactive plan review. For example, to avoid the nonconvex nature of conventional dose–volume constraints (44), Romeijn et al. introduced novel, convex dose–volume constraints that allowed them to formulate the IMRT fluence map optimization as a linear programming (LP) model (45). Using an industrial LP solver (CPLEX 8, ILOG

Inc.), a seven-field head-and-neck case with ~190,000 constraints, ~221,000 variables, and ~1,100,000 nonzero elements in the constraint matrix has been solved to global optimality in ~ 2 min of computation time on a 2.5 GHz Pentium 4 personal computer with 1 GB of RAM. While fluence map optimization has been studied extensively and clinically implemented, other challenging problems, such as finding the optimal number of beams and their angles, and fluence map optimization in the presence of organ motion, have not been solved satisfactorily. The problem of beam angle optimization has an enormous search space (46). The goodness of a chosen set of beam angles is not known until the fluence map optimization is performed. Current computer technology is not fast enough to solve such a nested optimization problem for routine clinical use. It is hoped that the collaboration with the OR community would help to develop novel and efficient algorithms in radiotherapy plan optimization.

DOSE DISPLAY AND PLAN EVALUATION

Treatment plans are evaluated based on the dose coverage and dose uniformity of the target, dose- to-sensitive normal structures, and magnitudes and locations of any hot and cold spots. This is best served by displaying the dose as isodose lines and superimposing them on the corresponding 2D image. Images can be displayed in the axial, sagittal, or coronal planes with the targets and normal structures delineated and dose distribution can be evaluated by going through the entire irradiated volume slice by slice. Figure 7 shows the isodose distribution of a head-and-neck IMRT plan in the three orthogonal planes overlaid on the corresponding CT images. Other graphic visualization tools, such as displaying the isodose in 3D as an iso-surface (47), have also been developed.

For quantitative plan evaluation, dose statistics can be calculated from the 3D dose distribution. Mean, maximum and minimum doses to the targets and mean and maximum doses to the critical structures can aid in plan selection. Dose uniformity throughout the target volume can be assessed from the standard deviations of the target dose distribution. A conformity index, defined as the quotient of the treated volume and the planning target volume (PTV) when the treated volume totally encompasses the PTV, has been introduced (48) to quantitatively evaluate the amount of normal tissue treated. Biological model-based TCP and NTCP are now available on some commercial TPS. However, the large uncertainty in the biological models combined with the lack of clinical experience limit their application in current clinical practice.

A very useful tool in 3D plan evaluation is the calculation of cumulative dose–volume histograms (DVHs) (49). The DVHs summarize 3D dose-distribution data into 2D histograms and are helpful in rapid screening of rival plans. An example of the DVHs of various structures of a head-and-neck IMRT plan is shown in Fig. 8. The DVH of a structure is calculated as

$$\%V(D) = \frac{\sum\limits_{d \geq D} v(d)}{V_0}$$

where $v(d)$ is the volume of a voxel in the structure receiving dose d and V_0 is the total volume of the structure. Therefore, each point $\%V(D)$ on a DVH curve represents the percent volume of the structure receiving doses $\geq D$. From the DVH, dose coverage and dose uniformity of the target(s) can easily be appreciated. The relevant dosimetric parameter to some critical structures that display serial nature such as the spinal cord and the brain stem is the maximum dose (48) and it can be read directly from the DVH. The functionality of many normal structures displays a dose–volume effect (50). For example, in trying to preserve salivary function for head-and-neck patients using IMRT, planning criteria of at least 50% volume of either parotid gland receiving doses < 30 Gy have been established (51,52). The development of $\geq$ Grade 2 pneumonitis in patients after radiation treatment for non-small cell lung cancer was found to be significantly correlated

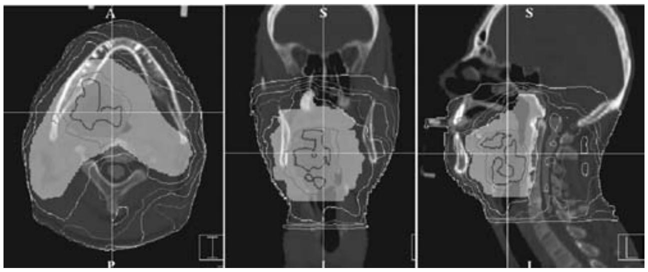


Figure 7. Overlay of isodose lines on an axial, coronal, and sagittal planes of a head-and-neck cancer. The IMRT plan was generated using a commercial inverse treatment planning system (CORVUS 5.0, North American Scientific). The isodose lines are, from inside out, at dose levels of 76, 72, 49.5, 40, 30, 20, and 10 Gy. The gross tumor volume (red) and the subclinical target volume (yellow) are shown as color washes to help evaluate the quality of the plan.

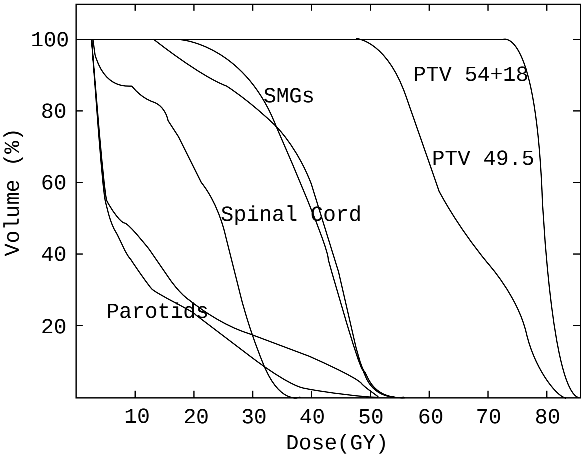


Figure 8. Cumulative DVH of different structures of a head-and-neck IMRT treatment plan. The prescribed doses to the targets are 72 Gy for the gross tumor volume (labeled “PTV 54 + 18”) and 49.5 Gy for the subclinical target volume (labeled “PTV 49.5”). Also shown are the DVHs of the left and right parotids, the left and right submandibular glands (SMGs), and the spinal cord.

with the percent volume of the total lung exceeding 20 Gy (53). These dose–volume relations can be obtained directly from the DVH. Dose–volume histograms from two or more competing plans can also be overlaid on top of each other to facilitate plan comparison. The increasing use of DVHs in routine clinical treatment planning over the past decade is a direct result of increased computer power and fast and more accurate dose calculation algorithms which make dose calculation in the entire irradiated volume possible.

The DVH should be used with caution. Since a DVH is a 2D histogram of a 3D dose distribution, it does not provide any spatial information. For example, Fig. 8 indicates that $\sim 2\%$ of the volume of the target labeled "PTV 49.5" did not receive the prescribed dose of 49.5 Gy. However, the locations of the underdosed volume is unknown. While underdosed areas near the periphery of the PTV may be acceptable, they are clearly unacceptable in the middle of the PTV where the gross tumor volume is located. Thus, DVHs must replace isodose distributions. Rather, it is a tool to enhance our ability to choose between different plans. This is especially true when evaluating IMRT plans, as the dose distributions of IMRT plans are spatially independent (54,55). A good DVH is therefore a necessary but not sufficient condition for a clinically acceptable treatment plan.

SUMMARY

In the first edition of this Encyclopedia > 17 years ago (1), Dr. Radhe Mohan envisioned that rapid developments in computer technology would make CT-based 3D treatment planning, including visualization of the anatomic structures and target region in 3D, fast and accurate radiation dose calculation in the entire irradiated volume, and sophisticated graphic and analytic tools for treatment plan display and evaluation, clinically routine. Radiation therapy has certainly gone through a series of revolutionary changes over the past 17 years. Three-dimensional treatment planning is now a standard practice and has mostly replaced 2D treatment planning. Monte Carlo dose calculation, which used to be considered as a luxury research tool, is now being incorporated into treatment planning systems. The fast development and widespread clinical implementation of IMRT over the past decade have provided the radiation oncology discipline a powerful tool to deliver highly conformal doses to the tumor volume while improving sensitive structure sparing. We are just entering an era of image-guided radiation therapy with exciting new developments. Recently, these have spurred efforts toward implementation of time-resolved or 4D imaging techniques, such as 4D CT (56–58) and 4D PET (59), into radiation oncology treatment planning and delivery. Currently, 4D imaging information is mainly being used as a tool for better definition of patient specific margins in the treatment of tumors in the thorax and upper abdomen. The ultimate goal is to establish a new scheme of 4D radiation therapy, where the 4D patient model is used to guide 4D planning and treatment. While there is still a long way to go toward this goal, much progress has been made, especially in the area of 4D inverse planning. Another important area that is under intense investigation is biologically conformal radiation therapy (BCRT) (60–62). Different from the current radiation therapy, which is aimed at producing a homogeneous target dose under the assumption of uniform biology within the target volume, BCRT takes the inhomogeneous biological information into account and produces customized non-uniform dose distributions on a patient specific basis. There are a number of challenges to accomplish the new radiation treatment paradigm, such as the determination of the distribution of biological properties of the tumor and critical structures, the prescription of the desired dose distribution for inverse planning, and the technique for inverse planning to generate most faithfully the prescribed nonuniform dose distribution. The most fundamental issue is, perhaps, how to extract the fundamental biological distribution for a given patient with biological imaging techniques and how to link the imaging information to the radiation dose distribution to maximize tumor cell killing. Hopefully, with the multidisciplinary efforts, the issues related to molecular imaging, image quantitation, planning, and clinical decision making would be resolved in the next decade. This will lead to truly individualized radiation therapy, and eventually individualized medicine when combined with the efforts in molecular medicine.

ACKNOWLEDGMENTS

JGL would like to thank Debbie Louis, CMD, for useful discussions and LX wishes to acknowledge support from the Department of Defense (PC040282) and the National Cancer Institute (5R01 CA98523-01).

BIBLIOGRAPHY

1. Mohan R. Radiation dose planning, computer-aided. In: Webster JG, editor. Encyclopedia of Medical Devices and Instrumentation. New York: John Wiley & Sons, Inc.; 1988. p 2397–2407.
2. Sherouse GW, Novins K, Chaney EL. Computation of digitally reconstructed radiographs for use in radiotherapy treatment design. Int J Radiat Oncol Biol Phys 1990;18:651–658.
3. Siddon RL. Prism representation: a 3D ray-tracing algorithm for radiotherapy applications. Phys Med Biol 1985;30:817–824.
4. Siddon RL. Fast calculation of the exact radiological path for a three-dimensional CT array. Med Phys 1985;12:252–255.
5. Hill DLG, et al. Medical image registration. Phys Med Biol 2001;46:R1–R45.
6. Wang H, et al. Implementation and validation of a three-dimensional deformable registration algorithm for targeted prostate cancer radiotherapy. Int J Radiat Oncol Biol Phys 2005;61:725–735.
7. Wu X, Dibiase SJ, Gullapalli R, Yu CX. Deformable image registration for the use of magnetic resonance spectroscopy in prostate treatment planning. Int J Radiat Oncol Biol Phys 2004;58:1577–1583.
8. Lu W, et al. Fast free-form deformable registration via calculus of variations. Phys Med Biol 2004;49:3067–3087.

9. Ding M, et al. Dose correlation for thoracic motion in radiation therapy of breast cancer. Med Phys 2003;30:2520–2529.
10. Beyer T, et al. A combined PET/CT scanner for clinical oncology. J Nucl Med 2000;41:1369–1379.
11. Schoder H, et al. PET/CT: a new imaging technology in nuclear medicine. Eur J Nucl Med Mol Imaging 2003;30:1419–1437.
12. Esthappan J, et al. Treatment planning guidelines regarding the use of CT/PET-guided IMRT for cervical carcinoma with positive paraaortic lymph nodes. Int J Radiat Oncol Biol Phys 2004;58:1289–1297.
13. van Der Wel A, et al. Increased therapeutic ratio by 18FDG-PET CT planning in patients with clinical CT stage N2-N3M0 non-small-cell lung cancer: a modeling study. Int J Radiat Oncol Biol Phys 2005;61:649–655.
14. MacManus MP, et al. F-18 fluorodeoxyglucose positron emission tomography staging in radical radiotherapy candidates with nonsmall cell lung carcinoma: powerful correlation with survival and high impact on treatment. Cancer 2001;92:886–895.
15. Sontag MR, Cunningham JR. The equivalent tissue-air ratio method for making absorbed dose calculations in a heterogeneous medium. Radiology 1978;129:787–794.
16. Wong JW, Purdy JA. On methods of inhomogeneity corrections for photon transport. Med Phys 1990;17:807–814.
17. Clarkson JR. A note on depth doses in fields of irregular shape. Br J Radiol 1941;14:265–268.
18. Bourland JD, Chaney EL. A finite-size pencil beam model for photon dose calculations in three dimensions. Med Phys 1992;19:1401–1412.
19. Johns HE, Cunningham JR. The physics of radiology, 4th ed. Springfield (IL): Charles C. Thomas; 1983.
20. Khan FM. The physics of radiation therapy, 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2003.
21. Kung JH, Chen GT, Kuchnir FK. A monitor unit verification calculation in intensity modulated radiotherapy as a dosimetry quality assurance. Med Phys 2000;27:2226–2230.
22. Xing L, et al. Monitor unit calculation for an intensity modulated photon field by a simple scatter-summation algorithm. Phys Med Biol 2000;45:N1–N7.
23. Mackie TR, Scrimger JW, Battista JJ. A convolution method of calculating dose for 15 MV X-rays. Med Phys 1985;12:188–196.
24. Boyer AL, Mok EC. A photon dose distribution employing convolution calculations. Med Phys 1985;12:169–177.
25. Mackie TR, et al. Generation of photon energy deposition kernels using the EGS Monte Carlo code. Phys Med Biol 1988;33:1–20.
26. Ahnesjo A. Collapsed cone convolution of radiant energy for photon dose calculation in heterogeneous media. Med Phys 1989;16:577–592.
27. Ahnesjo A, Aspradakis MM. Dose calculations for external photon beams in radiotherapy. Phys Med Biol 1999;44:R99–R155.
28. Nelson WR, Hirayama H, Rogers DWO. The EGS4 code system. Stanford Linear Accelerator Center Report 1985; SLAC-265.
29. Ma C-M, Jiang SB. Monte Carlo modeling of electron beams from medical accelerators. Phys Med Biol 1999;44:R157–R189.
30. Verhaegen F, Seuntjens J. Monte Carlo modeling of external radiotherapy photon beams. Phys Med Biol 2003;48:R107–R164.
31. Bielajew AF, Rogers DWO. Variance Reduction Techniques. In: Jenkins TM, Nelson WR, Rindi A, editors. Monte Carlo Transport of Electrons and Photons. New York: Plenum; 1988. p 407–419.
32. Boudreau C, et al. IMRT head and neck treatment planning with a commercially available Monte Carlo based planning system. Phys Med Biol 2005;50:879–890.
33. Shepard DM, Ferris MC, Olivera GH, Mackie TR. Optimizing the delivery of radiation therapy to cancer patients. SIAM Rev 1999;41:721–744.
34. Dempsey JF, et al. A Fourier analysis of the dose grid resolution required for accurate IMRT fluence map optimization. Med Phys 2005;32:380–388.
35. Yang Y, Xing L. Inverse treatment planning with adaptively evolving voxel-dependent penalty scheme. Med Phys 2004;31: 2839–2844.
36. Shou Z, et al. Quantitation of the a priori dosimetric capabilities of spatial points in inverse planning and its significant implication in defining IMRT solution space. Phys Med Biol 2005;50:1469–1482.
37. Wu QW, Mohan R, Niemierko A, Schmidt-Ullrich R. Optimization of intensity-modulated radiotherapy plans based on the equivalent uniform dose. Int J Radiat Oncol Biol Phys 2002;52:224–235.
38. Thieke C, Bortfeld T, Niemierko A, Nill S. From physical dose constraints to equivalent uniform dose constraints in inverse radiotherapy planning. Med Phys 2002;30:2332–2339.
39. Romeijn HE, Dempsey JF, Li JG. A unifying framework for multi-criteria fluence map optimization models. Phys Med Biol 2004;49:1991–2013.
40. Yang Y, Xing L. Clinical knowledge-based inverse treatment planning. Phys Med Biol 2004;49:5101–5117.
41. Langer M, et al. Operations research applied to radiotherapy, an NCI-NSF-sponsored workshop—February 7–9, 2002. Int J Radiat Oncol Biol Phys 2003;57:762–768.
42. Winston WL. Operations Research: Applications and Algorithms, 4th ed. Belmont: Thomson Learning; 2004.
43. Chong EKP, Zak SH. An Introduction to Optimization, 2nd ed. New York: John Wiley & Sons, Inc.; 2001.
44. Deasy JO. Multiple local minima in radiotherapy optimization problems with dose-volume constraints. Med Phys 1997;24: 1157–1161.
45. Romeijn HE, et al. A novel linear programming approach to fluence map optimization for intensity modulated radiation therapy treatment planning. Phys Med Biol 2003;48:3521–3542.
46. Stein J, et al. Number and orientations of beams in intensity-modulated radiation treatments. Med Phys 1997;24:149–160.
47. Schreibmann E, Theodorou K, Kappas C, Xing L. A software package for dose visualization in IMRT. XIVth Int Conf Comp Radiat Thera 2004; 700–703.
48. International Commission on Radiation Units and Measurements, Prescribing, recording and reporting photon beam therapy (Supplement to ICRU report 50), ICRU report 62, Bethesda (MD); 1999.
49. Drzymala RE, et al. Dose-volume histograms. Int J Radiat Oncol Biol Phys 1991;21:71–78.
50. Emami B, et al. Tolerance of normal tissue to therapeutic irradiation. Int J Radiat Oncol Biol Phys 1991;21:109–122.
51. Eisbruch A, Chao KSC, Garden A. Phase I/II study of conformal and intensity modulated irradiation for oropharyngeal cancer. Radiation Therapy Oncology Group protocol 0022, 2001.
52. Lee N, Garden A, Kramer A, Xia P. A phase II study of intensity modulated radiation therapy (IMRT) +/– chemotherapy for nasopharyngeal cancer. Radiation Therapy Oncology Group protocol 0225, 2003.
53. Graham MV, et al. Clinical dose-volume histogram analysis for pneumonitis after 3D treatment for non-small cell lung cancer (NSCLC). Int J Radiat Oncol Biol Phys 1999;45:323–329.
54. Xing L, Li JG. Computer verification of fluence map for intensity modulated radiation therapy. Med Phys 2000;27: 2084–2092.
55. Chao KSC, Blanco AI, Dempsey JF. A conceptual model integrating spatial information to assess target volume coverage for IMRT treatment planning. Int J Radiat Oncol Biol Phys 2003;56:1438–1449.

56. Pan T, Lee TY, Rietzel E, Chen GT. 4D-CT imaging of a volume influenced by respiratory motion on multi-slice CT. Med Phys 2004;31:333–340.
57. Keall PJ, et al. Acquiring 4D thoracic CT scans using a multislice helical method. Phys Med Biol 2004;49:2053–2067.
58. Low DA, et al. A method for the reconstruction of four-dimensional synchronized CT scans acquired during free breathing. Med Phys 2003;30:1254–1263.
59. Nehmeh SA, et al. Four-dimensional (4D) PET/CT imaging of the thorax. Med Phys 2004;31:3179–3186.
60. Alber M, Nusslin F. An objective function for radiation treatment optimization based on local biological measures. Phys Med Biol 1999;44:479–493.
61. Xing L, et al. Inverse Planning for Functional Image-Guided IMRT. Phys Med Biol 2002;47:3567–3578.
62. Ling CC, et al. Towards multidimensional radiotherapy (MD-CRT): biological imaging and biological conformality. Int J Radiat Oncol Biol Phys 2000;47:551–560.

See also PHANTOM MATERIALS IN RADIOLOGY; RADIATION DOSIMETRY FOR ONCOLOGY; RADIATION DOSIMETRY, THREE-DIMENSIONAL; RADIATION THERAPY TREATMENT PLANNING, MONTE CARLO CALCULATIONS IN; RADIOTHERAPY TREATMENT PLANNING, OPTIMIZATION OF.

RADIATION DOSIMETRY FOR ONCOLOGY

MALCOM MCEWEN
National Research Council
of Canada
Ontario, Canada

INTRODUCTION

Cancer is a disease that touches everyone, either directly or through a close friend or relative, and radiotherapy is one of the primary modalities for treating cancer. The intent may be a full cure or to relieve pain associated with the cancer and it is either used alone or in conjunction with other techniques, such as surgery or chemotherapy. The aim of radiotherapy is to use ionizing radiation (usually either high energy photons or electrons) to destroy the tumor while at the same time sparing healthy tissues. In the delivery of such treatments the quantity of interest is the absorbed dose (defined as the energy deposited per unit mass) as it can be used to estimate the biological effect of the radiation (i.e., cell killing). Too high a dose will kill all the cancerous cells, but will produce significant side effects due to damage to other organs. Too low a dose will leave some malignant cells alive, which can develop into a new tumor. One of the primary concerns in radiotherapy is therefore delivering the correct dose to destroy the tumor with the minimum of side effects and a fine line exists between under and over dosing. The allowable error in the delivered dose depends on many factors, such as the type and location of the tumor, but for some cancers can be as little as 3–4%.

The output of the machines that produce the radiation for radiotherapy (linear accelerators, X-ray tubes, radioactive sources) must be known to a very high accuracy and a great deal of dosimetry work is carried out in the radiotherapy clinic to monitor dose delivery. For example, before the linear accelerator (linac) can be used for patient treatment, daily checks are carried out to ensure consistency of output. *In vivo* dosimetry may be used to verify the patient dose and detailed measurements are made during weekly and monthly quality assurance sessions checking all aspects of treatment delivery. Audits are used to check that procedures are being followed correctly and to ensure national consistency.

Over recent years the number of treatment modalities has increased significantly as well as the complexities of treatment. The oncologist today can choose from an array of techniques including low energy X-ray tubes (usually used for superficial tumors); low doserate or high doserate brachytherapy, where different radioactive species are inserted or implanted in the body; external beam therapy using a linac (producing either photon or electron beams); protons and heavy ions; and neutrons.

The treatment can be simple, such as a single square field from a ^{60}Co unit, or complex, such as Image Guided Radiotherapy (IGRT) using a modern linac, where the patient is imaged immediately prior to treatment, the tumor volume verified and the dose delivered with a large number of shaped fields.

To cover all possible radiotherapy techniques is beyond the scope of this article, and therefore we will focus on external beam therapy using photon and electron beams, as this is most common and where dosimetry is most advanced. The aim is to give a basic grounding in radiation dosimetry for oncology together with a review of dosimetry techniques and an up-to-date bibliography where the reader can obtain further detail.

RADIATION MEASUREMENT AND QUANTITIES

In radiation oncology, we are interested in the relationship between biological damage to cells and the radiation producing the damage. Various attempts have been made to define biological dosimeters [e.g., deoxyribonucleic acid (DNA) strand breaks, chromosome aberrations], but none have resulted in a quantity that is reproducible and can be transferred from one situation to another: a primary requirement for a measurement quantity. Therefore physical quantities are used as a basis for estimating biological effects.

Fluence

The particle fluence, Φ, is defined (1) as dN/da: the number of particles dN, incident on a sphere of cross-sectional area da. The use of a sphere expresses the fact one considers the area perpendicular to the direction of each particle. The energy fluence Ψ (defined as the energy incident on a sphere of unit area) is generally of more interest for photons as it is more closely related to the dose deposited (see below).

Interaction Coefficients

The stopping power S of a material is defined as the energy lost by the charged particle (electron or positron) dE, along an increment of path dl. Ignoring energy losses due to nuclear reactions, stopping power has two principal components, namely, that due to collisions and that due to radiative losses. The collision component includes all energy losses in particle collisions that directly produce secondary electrons and atomic excitations. It also includes energy losses due to the production of Cerenkov radiation. The radiative component includes all energy losses of the primary electron that lead to bremsstrahlung production. The collisions of the primary electrons can produce high energy secondary electrons (δ-rays) that then become involved in independent interactions. The concept of restricted mass collisional stopping power L is therefore introduced to calculate the energy transferred to a localized region of interest. This region of interest is defined by a threshold (often denoted as Δ) for the energy transferred to the secondary (charged) delta particles. Highly energetic secondary particles with energy above this threshold escape the region of interest and do not contribute to the local absorbed dose and it is assumed that electrons with energy below Δ have negligible range. The restricted stopping power (L_Δ) is therefore always lower than the unrestricted stopping power and the choice of the energy threshold depends on the problem at hand. For problems involving ionization chambers, a frequently used threshold value is 10 keV since the range of a 10 keV electron in air is approximately 2 mm (the typical dimension of the air cavity of an ionization chamber). The parameter L_Δ is also known as the linear energy transfer (LET).

In practice, mass stopping powers (S/ρ, L/ρ) are generally used so that it is easier to compare the properties of materials with very different densities (e.g., air and water). A complementary quantity is the scattering power T, which describes the increase in the mean square scattering angle of the electron beam as it passes through a material.

For photon beams, there are a much larger number of possible interactions with the medium, the dominant ones in the energy range of interest being the photoelectric effect, Compton effect, pair production, coherent (Rayleigh) scattering, and nuclear photoeffect. The total interaction cross-section is simply the sum of all the individual cross-sections. The attenuation coefficient, μ, tends to be used rather than cross-sections as it describes the probability per unit thickness that a photon will undergo an interaction while traversing a material. As for stopping powers, the effect of density is removed and for dosimetric purposes two further coefficients are defined. The mass energy-transfer coefficient μ_{tr}/ρ relates the energy transferred from the photon to kinetic energy of charged particles and is used in the determination of kerma (see below). The mass energy absorption coefficient μ_{en}/ρ takes account of the fact that some of the energy transferred to charged particles is not deposited locally, but lost as bremsstrahlung.

Kerma

Kerma (kinetic energy released per unit mass), is introduced because neutral particles (photons and neutrons) deposit their energy in two steps: (*1*) interaction of the photon with an atom resulting in the transfer of energy to charged particles (predominantly electrons), and (*2*) deposition of that energy in the medium via Coulomb interactions (excitation and ionization). The dose contributed through direct interactions between photons or neutrons and the absorbing material will generally be negligible compared with this two-step process. Reference 1 gives the definition of kerma as:

$$K = \frac{dE_{\mathrm{tr}}}{dm} \tag{1}$$

where dE_{tr} is the kinetic energy transferred from photons to electrons in a volume element of mass dm. Total kerma can be split into two parts: collisional and radiative kerma. Collisional kerma, K_{col}, leads to the production of electrons that dissipate their energy as ionization near electron tracks in the medium. Radiative kerma, K_{rad}, leads to the production of bremsstrahlung as the charged particles are decelerated in the medium.

For a monoenergetic photon spectrum, energy E, with fluence Φ, equation 1 becomes

$$K = \Phi E \frac{\mu_{\mathrm{tr}}}{\rho} \tag{2}$$

where (μ_{tr}/ρ) is the mass energy transfer coefficient. For a polyenergetic photon beam, equation 2 becomes an integral over the full photon spectrum. As the photon energy increases, the maximum energy of the secondary electrons increases, the concept of a localized energy transfer begins to break down and kerma is therefore generally limited to photon energies below 3 MeV.

Absorbed Dose

The absorbed dose is defined as the mean energy imparted (absorbed) per unit mass. It is a nonstochastic quantity in that one is not measuring single events-the interaction between an incident photon or electron and a molecule—but the mean energy arising through the interaction of the radiation field with the material it passes through. As the mass of a sample decreases the energy per unit mass will become more random (stochastic). Whereas kerma is only defined for neutral particles, absorbed dose applies both to photon and electron beams.

Reference 2 applies this definition of absorbed dose in the situation where there is a small volume of the medium, which is thermally isolated from the remainder:

$$D_{\mathrm{i}} = \frac{dE}{dm} = \frac{dE_{\mathrm{h}}}{dm} + \frac{dE_{\mathrm{s}}}{dm} \tag{3}$$

where D_{i} is the mean absorbed dose in the absorber of material i, and mass dm; dE is the mean energy imparted to the absorber by the radiation beam (photons or electrons); dE_{h} is the energy appearing as heat; and dE_{s} is the energy absorbed by chemical reactions (which may be positive or negative). The left-hand relation is independent of the measurement technique while the right-hand relation represents one of the most common methods for determining dose: the measurement of heat. The unit of absorbed dose is the gray (Gy); 1 Gy = 1 J·kg^{-1}.

It can be inferred from the definitions above that collision kerma and absorbed dose should be related in someway, since they both deal with the deposition of energy in a localized area. If a state of charged particle equilibrium exists (and assuming no energy losses due to bremsstrahlung) then the absorbed dose will be equal to the kerma (conservation of energy). Charged particle equilibrium (CPE) exists at a point in the medium if the number and energy of charged particles entering a volume is equal to that leaving. True CPE only exists in the special case where there is no attenuation of the photon beam. In general there is always some photon attenuation, but there is said to be transient charged-particle equilibrium (TCPE), since the spectrum of charged particles changes very little as the photon beam penetrates the medium. Transient charged-particle equilibrium exists at the center of a broad photon beam at depths away from the surface (the depth at which TCPE is established depends on the incident photon energy and spectrum). For the general case, where TCPE exists and there are bremsstrahlung energy losses, the dose is given by

$$D = K_{\mathrm{col}} = K(1-g) \tag{4}$$

where g is the fraction of the energy that is lost to bremsstrahlung. For a ^{60}Co beam, g has a value of 0.003.

Absorbed dose is also related to the photon energy fluence at a point in a medium irradiated by a photon beam under conditions of transient charged particle equilibrium by

$$D = \Psi\left(\frac{\mu_{\mathrm{en}}}{\rho}\right)\beta \tag{5}$$

where β is the ratio of absorbed dose to collision kerma at a point. As written, equation 5 is valid for a monoenergetic photon beam; for a realistic (broad) photon spectrum, the mass–energy absorption coefficient must be averaged over the photon fluence.

There is a charged-particle analog of equation 5. Under the restrictive conditions that (*1*) radiative photons escape the volume of interest and (*2*) secondary electrons are absorbed on the spot (or there is charged-particle equilibrium of secondary electrons), the absorbed dose to medium is given by the electron fluence multiplied by the collisional stopping power.

Dose Equivalent

This quantity is useful where the effect produced by the same absorbed dose is dependent on the particle type "delivering" the dose. This is the case in biological damage: the principal pathway for cell killing is a double-strand break of the cell's DNA, which is much more likely for densely ionizing particles, such as protons, neutrons, and α-particles, than it is for electrons or photons. A radiation quality factor, w is therefore introduced to take account of this and the dose equivalent is defined as the absorbed dose multiplied by this quality factor. Values of w vary from 1 for photons and electrons to 20 for α-particles.

PRIMARY METHODS OF DETERMINING ABSORBED DOSE AND AIR KERMA

Due to the complexities of the measurements, the absolute determination of radiation quantities is almost exclusively the domain of national standards laboratories. Two primary International System of Units (SI) quantities are realised by national standards laboratories for radiotherapy dosimetry: air kerma and absorbed dose. Air kerma can only be measured using an air-filled ionization chamber but absorbed dose can be determined in a variety of ways.

The absolute measurement of absorbed dose has a number of problems (some fundamental, others practical) that limit the accuracy of the result and put constraints on the experimental techniques that can be used.

1. *Doses of interest are small*. The definition of absorbed is in terms of the energy absorbed in an amount of material. Radiotherapy dose levels are typically < 10 Gy (10 J·kg^{-1}), which represents a very small energy deposition. If one is trying to determine this energy absolutely by measuring the radiation-induced temperature rise (of the order of a few mK) there is a significant challenge in achieving uncertainties $< 0.1\%$.
2. *The quantity required is the dose in an undisturbed phantom*. In radiotherapy, the required end-point is the dose to the tumor. However, since radiation interactions are very material dependent a homogeneous phantom is the chosen medium for reference dosimetry. This immediately presents a problem in that any measuring instrument will perturb the phantom and affect the measurement one wishes to make.
3. *The quantity required is the dose at a point in this phantom*. For radiotherapy dosimetry, one is not interested in the average dose to the whole phantom (although mean dose or integral dose is required for radiation protection, when considering lifetime dose to organs, etc.). Radiotherapy treatments using photon and electron beams produce significant dose variations within a phantom; otherwise, healthy tissue could not be spared. It is therefore important to be able to measure these dose variations, which by implication requires a small detector. Such a detector will generally have a larger uncertainty than a larger detector. Care is required in designing a detector that samples the dose at a point and does not give some unwanted averaging.
4. *Scattered radiation contributes a significant proportion of the absorbed dose*. In a typical radiotherapy radiation field used for cancer treatment (e.g., a 6 MV photon beam), 15% of the dose at the point of interest is due to scattered, rather than primary, radiation. The experimental geometry is therefore very important and care must be taken in designing experiments, especially when comparing or calibrating dosimeters, so that scattered radiation is properly taken into account.

5. *Optimization of the measurement is difficult.* One of the biggest practical constraints is that in the measurement of absorbed dose one is not determining some fundamental constant or characteristic of a material. The dose is the effect of a particular radiation field at a point in a particular material and it is therefore not possible to optimise all aspects of a measurement. There are many "influence quantities" (material, energy spectrum, geometry) so that what may appear to be minor variations from the real measurement problem (dose to a tumor) can result in significant errors being introduced.

These issues also apply to the determination of air kerma.

Ionometry

An ionization chamber measures the ionization produced by the incident radiation beam in a mass of air. Historically, the first quantity to be measured was Exposure (symbol X) and is simply the charge, Q, liberated in a volume of air mass m_{air}. It is not a recommended SI unit but is related to Air Kerma by

$$K_{\mathrm{air}} = X \frac{W}{e} \frac{1}{1-g} \tag{6}$$

where W/e is the average energy required to liberate an ion pair and g has the same definition as in equation 4. The value of W/e has been measured by a large number of experimenters of many years and there is an agreed value of 33.97 eV/ion pair (3), which is constant over widely varying conditions (air pressure, electron energy, etc.).

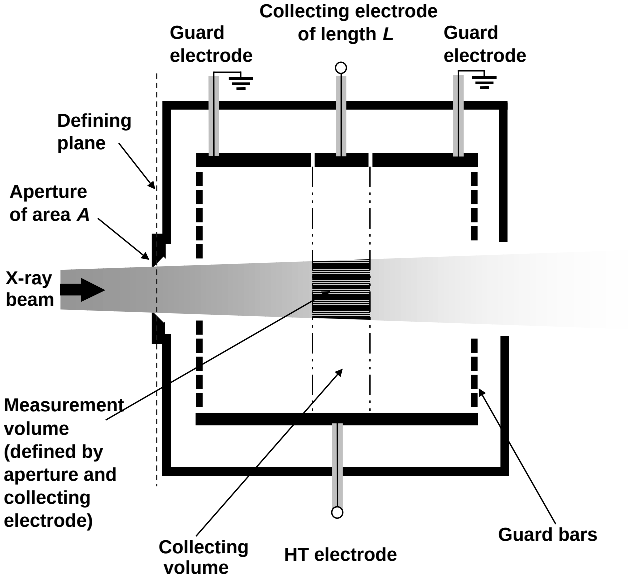
For low energy beams produced by X-ray tubes ($<$ 400 kVp) exposure is measured using a free-air chamber. A typical chamber is shown in Fig. 1. By careful design and precise manufacturing of the electrodes and entrance aperture it is possible to accurately define the volume (and thus mass) of air that is irradiated. The size of the free-air chamber scales with the incident energy and for a ^{60}Co beam a free-air chamber would be impractically large: several meters in each dimension.

At energies of ^{137}Cs (663 keV) and above, one must therefore use a cavity chamber, where the volume of the air cavity is small (typically a few cm^3). One must determine the mass of air in the cavity and various designs been employed: spherical, cylindrical, and "pancake" (Fig. 2).

Until 1990, all absorbed dose measurements in the radiotherapy clinic were based on air-kerma calibrations using protocols as seen in Refs. 4 and 5. In recent years, absorbed dose-based calibrations have become available from national standards laboratories and associated protocols produced (e.g., Refs. 6–8). However, air kerma standards are still required for the dosimetry of kilovolt X-ray beams, brachytherapy, and radiation protection.

The dose to the air volume of a cavity chamber irradiated by an X- or γ-ray beam is given by

$$D_{\mathrm{air}} = \frac{Q}{m_{\mathrm{air}}} \frac{W}{e} \tag{7}$$



Schematic diagram of the National Physics Laboratory (NPL) primary standard free-air ionisation chamber

Figure 1. Diagram of a free-air chamber. (Courtesy NPL.)

Bragg–Gray cavity theory is then used to relate the dose in the air cavity to the dose to the medium. The conditions for application of Bragg–Gray cavity theory are

1. The cavity must be small when compared with the range of charged particles incident on it so that its presence does not perturb the fluence of charged particles in the medium.
2. The absorbed dose in the cavity is deposited solely by charged particles crossing it (i.e., photon interactions in the cavity are assumed negligible and thus ignored).

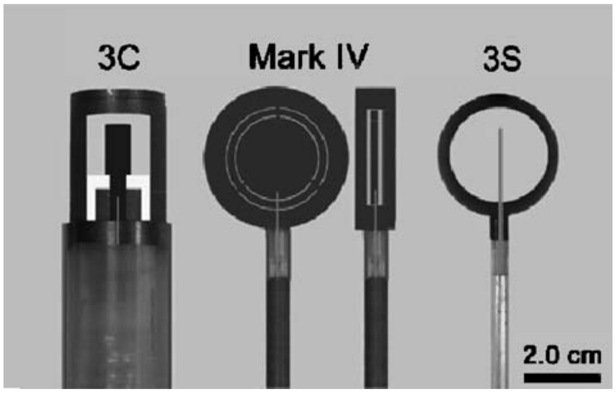


Figure 2. Schematics of three graphite-walled cavity ion chambers designed and operated at the National Research Council (NRC) - cylindrical (3C), parallel-plate (Mark IV), and spherical (3S). The 3S utilizes an aluminum electrode, while the other two chambers utilize graphite electrode.

Condition (2) implies that all electrons depositing the dose inside the cavity are produced outside the cavity and completely cross the cavity. Therefore, no secondary electrons are produced inside the cavity and no electrons stop within the cavity. The dose to the medium is obtained using a ratio of stopping powers:

$$D_{\text{med}} = \frac{Q}{m_{\text{air}}} \frac{W}{e} \left(\frac{S}{\rho}\right)_{\text{med,air}} \tag{8}$$

where $(S/\rho)_{\text{med,air}}$ is the mass stopping power ratio for the medium divided by that for air. The Bragg–Gray cavity theory does not take into account the creation of secondary (delta) electrons generated as a result of the slowing down of the primary electrons in the sensitive volume of the dosimeter. The Spencer–Attix cavity theory is a more general formulation that accounts for the creation of these electrons that have sufficient energy to produce further ionization on their own account.

Equation 8, in principle, gives a possible route to the absolute absorbed dose, if the stopping power ratio is known. The Bureau International des Poids et Mesures (BIPM) maintains such an ionometric standard for absorbed dose to graphite. This is a graphite walled ionization chamber, whose volume has been determined by mechanical means, and is described in detail by Boutillon and Peroche (9). Strictly speaking, however, this is not a primary device, as the value for the product of *W/e* and the stopping power ratio is taken from calorimeter measurements. Although there are independent measurements of *W/e*, there is a lack of measured stopping power data (see below) to provide a true measurement of absorbed dose absolutely.

As noted above, one of the difficulties in realising a primary standard cavity ion chamber is defining the effective volume of the chamber. This problem can be overcome to a certain extent by the use of an extrapolation, or gradient, chamber. In such a chamber, the absolute volume of the cavity is not known, but can be changed by a known amount, usually by changing the electrode spacing. Assuming that the chamber does not perturb the medium then the dose is given by

$$D_{\text{med}} = \frac{\Delta Q}{\Delta x} \frac{W}{e} \left(\frac{S}{\rho}\right)_{\text{med,air}} \frac{1}{A\,\rho_{\text{air}}} \tag{9}$$

where ΔQ is the change in the measured ionization charge for a change in the electrode spacing of Δx, A is the area of the electrode and ρ_{air} is the air density. The dose measurement becomes a relatively simple charge measurement and the problem of the determination of volume is reduced to that of determining the area of the collecting electrode. Klevenhagen (10) carried out some of the first work on gradient chambers for the determination of dose in megavoltage photon and electron beams and Zankowski and Podgorsak (11) describe a chamber where the entire device is manufactured from a plastic with similar radiation properties to water (see Fig. 3). It is a parallel plate chamber with a fixed radius and the plate separation is varied by way of a precision micrometer. The charge gradient $\Delta Q/\Delta x$ can be determined at the 0.2% level and the effective area of the collecting electrode (determined by a capacitance measurement) has an uncertainty of 0.1%. The limiting factor for this type of device is the uncertainty in *W/e* and the stopping power ratio at energies other than ^{60}Co.

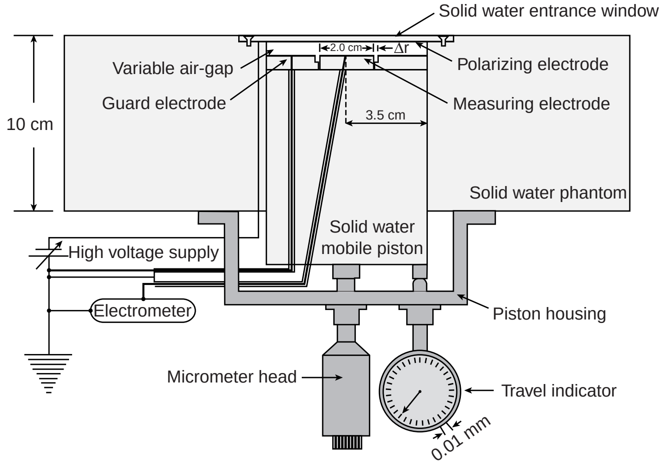


Figure 3. Extrapolation chamber design (11).

An extrapolation chamber is also the device of choice for β-ray brachytherapy sources (such sources are used for the treatment of ophthalmic cancers) and one can derive absorbed dose using the Bragg–Gray principle (12).

Calorimetry

A calorimeter directly measures the absorbed dose as it is defined above. Although care must be taken in the design of a calorimeter, in terms of geometry and material composition, no primary conversion factor (e.g., *W/e*, G-value) is required. However, the basic operating principle of calorimetry is that all the energy deposited by the radiation is expressed as heat. If this is not the case, then there is said to be a heat defect. The heat defect can be due to crystal dislocations, radiation-driven chemical reactions or some other mechanism and is strongly material dependent. It follows from equation 3 that if there is no heat defect then

$$D_{\text{i}} = c_{\text{p}} \Delta T \tag{10}$$

where c_{p} is the specific heat at constant pressure and ΔT is the temperature rise in the material (absorber). The size of the > element > that defines the measured dose will depend on the specific application as well as the design of the calorimeter and the material used. A material with a high thermal conductivity will require mechanically defined components (e.g., some small absorber thermally isolated from the rest of the material), while it is much easier to measure a point dose in a material with a low thermal conductivity. Inhomogeneities in the phantom will affect the scattering of the radiation beam, and therefore change the absorbed dose measured. A correction will be required to give the dose in a homogeneous medium.

The three challenges in calorimetry are therefore to (*1*) measure the radiation induced temperature; (*2*) measure a material of known specific heat capacity, and (*3*) make sure that what is measured is relevant to the particular application of the radiation beam. These problems have been addressed in a number of (often novel) ways over many years, but currently there are basically two types of calorimeter: graphite (e.g., see Refs. 13 and 14) and water (e.g., see Ref. 15). Graphite has some obvious advantages: it is solid and a graphite calorimeter can be made smaller and more robust than a water device. For example, McEwen and Duane (16) demonstrate a calorimeter designed to be taken routinely into radiotherapy clinics. There is no heat defect or convection to consider for graphite and the temperature rise pure unit absorbed dose is much larger than that of water due to the low specific heat capacity ($c_{\mathrm{p,graphite}} \sim 700$ $c_{\mathrm{p,water}} \sim 4200$ $\mathrm{J{\cdot}kg^{-1}{\cdot}{}^{\circ}C^{-1}}$). However, the high thermal conductivity is an issue and the quantity realized is absorbed dose to graphite so a conversion is required to obtain absorbed dose to water (see below). Since water is the standard reference material for radiation dosimetry the majority of standards laboratories are moving over to water calorimeters as the primary standard and the device operated at the National Research Council in Canada is shown in Fig. 4. The major problems in developing a water calorimeter are controlling convection and obtaining a stable (ideally zero) heat defect for the sample of water irradiated. The present state-of-the-art in calorimetry yields an uncertainty in absorbed dose to water of ~ 0.3% and for recent reviews of calorimetric development see Ross and Klassen (17), Williams and Rosser (19), and Seuntjens and DuSautoy (20).

Chemical Dosimetry

In chemical dosimetry, the absorbed dose is determined from some chemical change in an appropriate material and any well-characterized chemical reaction where the reaction product(s) can be measured with good precision may serve as the basis for the dosimeter. Chemical dosimeter systems were developed as early as 1927 and a wide range of systems have been studied. Although, in principle a chemical dosimeter is a secondary device, in that it does not directly measure the absorbed energy, it can be regarded as a primary device if the relation between absorbed energy and chemical change can be determined absolutely. This relationship is termed the radiation chemical yield and is expressed as the number of molecules or ions of product X liberated per unit absorbed energy, designated $G(\mathrm{X})$. Assuming that the G-value is known and is constant with dose then the absorbed dose is given by

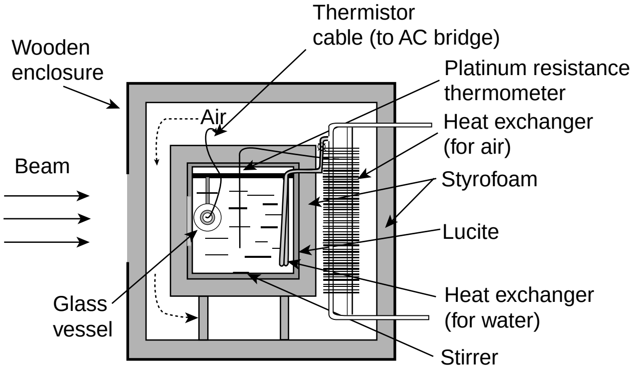

Thermistor cable (to AC bridge)
Wooden enclosure
Platinum resistance thermometer
Air
Heat exchanger (for air)
Beam
Styrofoam
Lucite
Heat exchanger (for water)
Glass vessel
Stirrer

Figure 4. Overview of the NRC sealed water calorimeter. The outer box, which provides thermal insulation, is ~ 80 cm on a side, while the water phantom is ~ 30 cm on a side. The inner-glass vessel provides a stable volume, where the purity of the water can be rigorously maintained (to control the heat defect). The radiation-induced temperature rise (typically a few mK) is measured using thermistor probes and the outer sys- tem controls the temperature at 4°C to minimize convection effects.

$$D = \frac{\Delta M}{G(\mathrm{X})\rho} \tag{11}$$

where ΔM is the volume concentration of molecules produced by the radiation absorbed and ρ is the density of the medium.

To act as a primary dosimeter, a chemical dosimeter should be dose, dose-rate, and LET independent. Aqueous systems are preferred as they are basically water equivalent, although this introduces a containment vessel whose effect must be taken into account.

The ferrous sulfate (21) dosimeter is the most widely used and longest established dosimetry system. It demonstrates the advantages and problems of chemical dosimeters. The reaction mechanism is the oxidation of ferrous to ferric ions, in aerated sulfuric acid. The oxidation proceeds via a number of reactions involving hydroxyl radicals, hydroperoxy (HO_2) radicals, and hydrogen peroxide. The ferric ion formation is directly proportional to energy absorbed as long as some oxygen remains in the solution, hence the requirement for aeration. All the reactions are fast (< 1 min), therefore there is no aftereffect under usual γ- or electron irradiations. However, great care must be taken in the preparation of the solutions, particularly with regard to water purity as organic impurities can have a significant effect. Spontaneous oxidation of the ferrous ions occurs that can be corrected for by the use of an unirradiated sample as a control.

The concentration of ferric ions may be determined by titration, but absorption spectroscopy is generally a more convenient technique, using ultraviolet (UV) wavelengths of 304 or 224 nm. The Fricke dosimeter is dose-rate independent for ^{60}Co radiation in the range 0.1–40 $\mathrm{Gy{\cdot}s^{-1}}$ (a range that covers both radiotherapy and industrial dosimetry applications) and for linac irradiations, $G(\mathrm{Fe}^{3+})$ production is linear up to a maximum dose-per-pulse of 2 Gy (significantly greater than radiotherapy linacs). The normal dose range is 5–350 Gy, although this can be extended by suitable modifications of the composition of the system, or of its analysis. With care, Fricke dosimetry is capable of 0.1% precision, but for absolute dosimetry one requires an accurate determination of the G-value. As with ionometry, this factor can be determined from calorimetry, but preferably one would like an independent measurement. Such a measurement is possible if one knows the total energy in the radiation beam. Roos and Hohlfeld (22) describe the system developed at the PTB (Physikalisch Technische Bundesanstalt) in Germany based on a microtron accelerator with a very well determined electron energy and beam current. With such a system an uncertainty in the *G*-value (the effective response of the Fricke is actually derived in this measurement) of $< 0.4\%$ is achievable and the overall uncertainty in measuring absorbed dose to water is $\sim 0.5\%$. Other chemical dosimeter systems include ceric sulfate, oxalic acid, potassium dichromate, and alanine, which cover higher dose ranges than Fricke. However, *G*-values for these systems are either unknown, or have a much larger uncertainty, and therefore cannot be regarded as primary dosemeters.

Conversion of Dose Between Materials

Since dose is material dependent, the primary device one uses to measure absorbed dose may not yield the quantity required. A conversion procedure is therefore needed. If the uncertainty on this conversion is sufficiently large then the usefulness of the primary device is called into question. The majority of effort in this area has concentrated on water and graphite since graphite is commonly used for primary standard calorimeters and water is the material of interest for radiotherapy dosimetry.

For electron beams, the conversion factor is a product of two factors: a ratio of stopping powers and a fluence correction (the latter takes account the differences in scattering power between the two materials). The most accurate values for stopping powers are those given in Ref. 23, but these are based on calculation alone and a quoted uncertainty of $\sim 1\%$ for each material is given. There have been a number of attempts to measure stopping powers (e.g., Ref. 24), but these did not have the accuracy to validate the calculations. One of the problems in measuring stopping powers is that it is only possible to measure relatively small energy losses and this significantly increases the precision required if the achieved overall uncertainty is to be $< 1\%$. Faddegon et al. (25) presented a new technique using a large sodium iodide detector to directly measure elemental stopping powers and McPherson (26) reports the results of such measurements. The standard uncertainty on these measurements (0.4–0.7%) is significantly lower than the previous attempts and is at a level where the calculated values in Ref. 23 can be tested. Fluence corrections are determined either through direct measurement in phantoms of different materials or using Monte Carlo simulations (see below).

For megavoltage photon beams, there is more than one method available for converting dose from one material to another. Burns and Dale (27) describe two methods: the first making use of the photon fluence scaling theorem (28) and the second based on cavity ionization theory. Nutbrown et al. (29) repeated the experimental work of Burns and Dale and applied a third method based on extensive Monte Carlo simulations. Fricke dosimeters can also be used to transfer the dose between materials as it can be assumed that the *G*-value is independent of the phantom material (27).

Monte Carlo: A Primary Technique for the Future?

There has been a rapid development of Monte Carlo techniques for radiation dosimetry in the last 10–15 years. A Monte Carlo calculation is based on radiation transport physics and tracks individual particles as they interact with the detector and phantom. By averaging over a large number of particles (typically > 10 million), statistical fluctuations can be reduced to an acceptable level. The big advantages of a Monte Carlo simulation are (*1*) there is no reliance on a physical artifact, such as a ion chamber or calorimeter, and (*2*) you are not constrained by many of the problems of physical measurement as outlined above and can derive the exact quantity you require. Calculations initially focused on determining correction factors, such as the ion chamber wall effect for air kerma standards and the effect of inhomogeneities in a medium. More recent Monte Carlo codes have included the accurate simulation of the radiation source (e.g., BEAM (30)) and the detector (e.g., EGSnrc (31)). The sophistication has reached the level where they may be considered as viable alternatives to measurements.

In considering the idea of Monte Carlo as a primary technique one can clearly not escape some absolute measurement for the primary realization of absorbed dose. For example, the absolute beam current produced by a linear accelerator or the total activity of a radioactive source would be required as an input to the simulation, but the dose itself would be calculated. If this measurement can be determined with high accuracy and the absolute uncertainties in the Monte Carlo can be reduced, then this offers a potential alternative to the present primary standards. The major limitation is the accuracy of input data for the physics models: interaction cross-sections, stopping powers, and so on are not known accurately enough. The high accuracy obtained in the determination of correction factors in dosimetry is because in those situations one does not rely in such a direct way on absolute interaction coefficients, but on differences (or ratios) in interaction coefficients, where one benefits from the cancellation of correlated uncertainties. To date, the majority of the effort has been in developing the Monte Carlo codes (improving efficiency and refining the physics modeled), but there are still significant gaps in the input data so it is not clear whether the potential for the absolute application of Monte Carlo techniques can be fulfilled.

REFERENCE OR SECONDARY DOSIMETERS

As with primary devices, there are number of different types of secondary dosimeter that are used in radiotherapy. Secondary dosimeters require calibration against a primary standard and are then used to realise absorbed dose on a more routine basis.

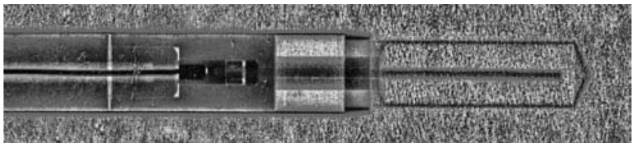


Figure 5. Radiograph of a NE2571 Farmer-type chamber. One can see the graphite outer wall of the cavity, aluminum central electrode, and internal construction of the stem.

Ion Chambers

Ionization chambers (particularly air-filled chambers) are the most widely used instruments in radiotherapy dosimetry. They offer a number of characteristics which are particularly suited to the measurement of therapy radiation beams - sensitivity, long-term stability, and ease-of-use. The most widely used chamber is the Farmer-type (33), an example of which is shown in Fig. 5. Chambers of this type show excellent stability (figures of ± 0.3% over 25 years are not uncommon) with the only disadvantage being a lack of waterproofing. A waterproof sleeve is therefore required for measurements in water. Such a sleeve should be thin (< 1 mm), close fitting (no air gaps), and made from some low Z material to minimize any additional perturbation effect [PMMA-poly(methyl-methacrylate) is commonly used].

Parallel-plate chambers, which are usually waterproof, are recommended for the dosimetry of electron beams. The NACP design (34) is one of the most widely used (Fig. 6).

Ionization chambers are usually vented to the atmosphere and therefore an air density correction, f_{TP}, is required to normalise for variations in air temperature and pressure (T_{air} and P_{air}, respectively):

$$f_{\mathrm{TP}} = \frac{273.15 + T_{\mathrm{air}}}{273.15 + T_{\mathrm{ref}}} \cdot \frac{P_{\mathrm{ref}}}{P_{\mathrm{air}}} \tag{12}$$

P_{ref} is taken to be 101.325 kPa, but there is no agreed value for T_{ref}. In Europe a value of 20 °C is used, while in North America the reference temperature is 22 °C. Care must therefore be taken when comparing results from different laboratories.

A correction for the humidity of the air in the chamber is not generally applied. The presence of water vapor affects the value of *W*/*e* (36), as well as stopping power and mass-energy absorption coefficients. However, for relative humidities between 10 and 90% (when comparing to a standard humidity of 50%) the effect is a maximum of 0.1%.

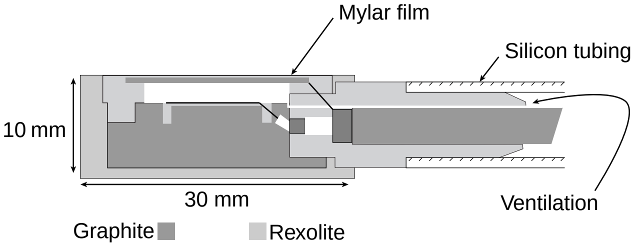


Figure 6. Schematic of the NACP parallel-plate chamber (taken from Ref. 35). Design features include waterproof construction, low Z materials to minimize the perturbation correction, and large guard ring to minimize in-scatter.

A correction is required to account for the incomplete collection of charge due to ion recombination in the chamber volume. The correction for ion recombination is the sum of two components: initial recombination and general or volume recombination. General recombination takes place when oppositely charged ions from different ionization tracks (i.e., created by different incident ionizing particles) recombine while they drift under the influence of the electric field toward their respective electrodes. Initial recombination takes place when oppositely charged ions from the same ionization track recombine; as the name suggests, this process takes place before the electric field is able to pull the track structure apart and therefore precedes general recombination. Initial recombination is independent of dose rate but general recombination depends on the ion density in the cavity. This ion density depends on the dose rate for continuous radiation and on the dose per pulse for pulsed beams. Initial recombination is typically small (~ 0.1–0.2% for the usual cylindrical and parallel-plate chambers employed in radiotherapy). General recombination is typically a small effect for continuous radiation (e.g., kilovoltage X-ray beams or ^{60}Co γ-ray beams) but for pulsed beams it can often be significant, especially so for modern linear accelerators that employ large dose-per-pulse values (recombination corrections of up to 5% have been reported).

The theoretical aspects of ion recombination for pulsed and continuous radiation have been well discussed in the literature (37–39). However, in recent years a number of authors (40–42) have presented recombination data that do not agree with the standard theory. A number of possible mechanisms have been proposed, including ion multiplication, air volume change, and direct collection of primary electrons, but at present there is no consensus as to which, if any, of these is the reason for these anomalous results. Dosimetry protocols recommend that a full $1/I$ against $1/V$ plot be measured where I is the measured Ionization current and V is the polarizing voltage to establish the range of linearity where the standard theory holds and the chamber then operated at voltages to remain within that range (Fig. 7).

Ion chamber measurements must also be corrected for the effects of polarity. The polarity effect is the difference in readings obtained in the same irradiation conditions, but taken with positive and negative polarizing voltages. Boag (43) identified a number of components of the polarity effect including secondary electron emission (due to the Compton effect) that produces a negative current independent of polarity; uneven distribution of the space charge due to a difference in the drift velocity of negative and positive ions; variation of the active volume due to space charge distortions; stopping of fast electrons in the collecting electrode not balanced by the ejection of recoil electrons; and collection of current outside the chamber volume due to leakage in solid insulators. In practice, it is difficult to identify the mechanisms acting in a particular

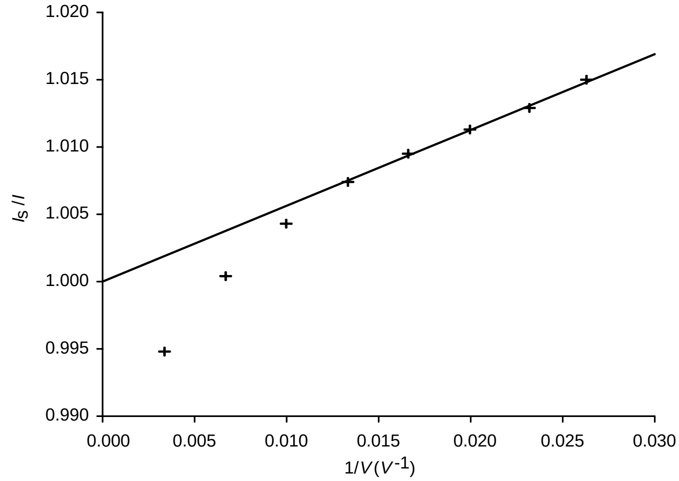


Figure 7. Plot from Burns and McEwen (41) showing the deviation from theory (straight line) of the recombination behavior for a NACP chamber. Without extensive measurements errors of up to 1% are possible.

situation, but measurements show that the polarity effect will vary with chamber type, beam energy and modality, measurement depth and can vary with other factors, such as field size.

The polarity correction is given by

$$f_{\text{pol}} = \frac{|M^+| + |M^-|}{2|M|} \qquad (13)$$

where the superscripts + or − indicate the reading (M) with collecting voltage positive and negative, respectively, and M in the denominator is the reading taken with the normal polarity used during measurements. Table 1 summarizes typical polarity corrections for chambers in different beams.

A variation on the air-filled ionization chamber is the liquid ion chamber. In this design, the air is replaced by a liquid, which offers two major advantages: a flat energy response and an increased carrier concentration (and therefore increased spatial resolution). Liquid ion chambers have been developed over many years, but their use as secondary dosimeters has been severely hampered by the volatility of the liquid (usually a short-chain hydrocarbon) resulting in loss of signal. However, recent results (47) show impressive stability and may indicate that liquid ion chambers have a role to play in reference dosimetry.

Table 1. Typical Polarity Corrections

Beam	Cylindrical Chambers	Parallel-Plate Chambers
Megavoltage photons	$< 0.2\%$ beyond $d_{\max}$, more variable in build-up region	Generally $< 0.3\%$, but can show variable behavior.
Megavoltage electrons	Up to 1% at lower end of recommended energy range (44)	$< 0.2\%$ for well-designed chambers (45). Can be significant for other chamber types (46)

Fricke

The Fricke dosimeter was described in detail in the section above on primary dosimeters. As a secondary dosimeter it is used in exactly the same way, except that the G-value is effectively measured for each batch of solution by comparison with a calorimeter (48). The big disadvantages of Fricke are (*1*) the care needed to produce 'good' solutions, and (*2*) the perturbation correction required for the vessel holding the Fricke solution (usually glass or quartz) is generally large. The NRC in Canada has used Fricke to transfer the dose from water calorimeter to ionization chambers (49).

TLD

Another class of systems is thermoluminescent dosimeters (TLDs). One of the obvious advantages of such dosimeters is that they can be made very small, and are therefore ideal for plotting dose distributions. The TLD material can be used as a powder or can be formed in various shapes (chips, rods, pellets, etc.). These materials have a wide dose range, from a few tens of μGy to ~ 1 kGy. The readout (measurement of the glow curve) is destructive, but the dosimeters can be reused. The equipment required is readily available and the production and readout of dosimeters is relatively simple, particularly compared to Fricke or alanine (Fig. 8).

Lithium fluoride is the most widely used system for radiotherapy applications as it has a mean atomic number close to that of tissue ($Z_{\text{eff}} = 8.2$, compared to 7.4 for tissue). It has a fairly flat response with energy (especially in the megavoltage region) and is therefore not particularly sensitive to variations in beam quality. Both CaF_2 and $CaSO_4$ are useful in that they have sensitivities 10–100 times greater than LiF but, because of their high Z values, they show a very rapid change in energy response at low energies. Lithium borate has a better tissue similarity ($Z_{\text{eff}} = 7.4$) but has a sensitivity of only one tenth of that of LiF. As for the other systems based on some chemical change, TLD materials require calibration against a primary dosimeter. It is not possible to determine any thermoluminescent equivalent of a G-value as the dose response depends on the annealing process and tends to be batch dependent. Typical reproducibility at the 1% level is possible routinely with an overall uncertainty of 2–3% (one standard deviation). However, Marre et al. (50) obtained a reproducibility of better than ± 0.5% and an overall standard uncertainty in measuring absorbed dose to water of ± 1.6%. These values are approaching those of ion chambers although the delay between irradiation and readout and the care required to achieve this level of precision limit the applications for this dosimeter.

TLD is an attractive dosimeter for the dosimetry of low doserate brachytherapy sources. The source strength is normally too low for small ionization chambers, and large volume chambers have poor spatial resolution. However, for ^{125}I, one of the commonly used isotopes in prostate

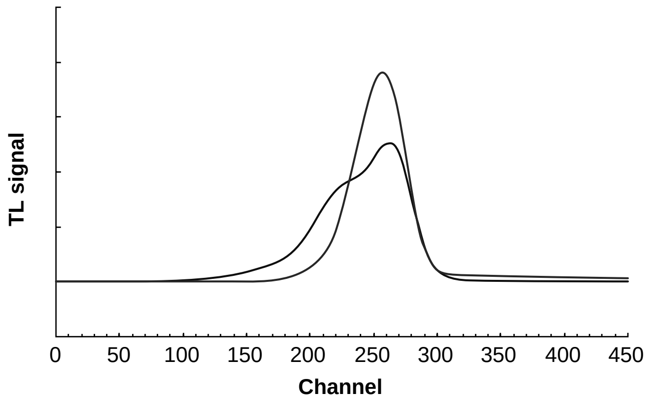


Figure 8. Glow curves from two different TLD materials. The temperature is slowly ramped to a maximum (in this case 240 °C) and the thermoluminescent intensity measured using a photomultiplier. The shape of the glow curve depends both on the material and the thermal pre-treatment (annealing).

treatment, the mean photon energy in only 27 keV and therefore the energy dependence of TLD needs to be known accurately (Fig. 9).

Since LiF is nontoxic, TLD can be used as an *In vivo* dosimeter, placed directly on the patient, to verify treatment delivery. It is a less invasive technique compared to diodes or MOSFET detectors (see below), as there are no trailing wires or associated equipment.

Alanine

Over recent years, alanine has become more widely accepted as a chemical dosimeter for radiotherapy dosimetry. It has a very wide dose range, showing a linear response from 10 Gy to 70 kGy. It is a solid dosimeter, with a density and atomic number close to that of water (close to zero perturbation) and the dosimeters are small: typically disks are 5 mm in diameter and 3 mm thick, but can be made as thin as 0.5 mm for measuring low electron energies. The energy dependence is very small. Zeng et al. (52) showed that any variation in the sensitivity of alanine is not more than ± 0.5% over the energy range from ^{60}Co to 25 MV X rays. The dosimeter is read out nondestructively using ESR (electron spin resonance) spectroscopy. This nondestructive read-out, together with the long-term stability of the radiation-induced signal means that alanine has a potential role as a dose record. The National Physical Laboratory in the United Kingdom offers a mailed dosimetry service for radiotherapy using alanine dosimeters.

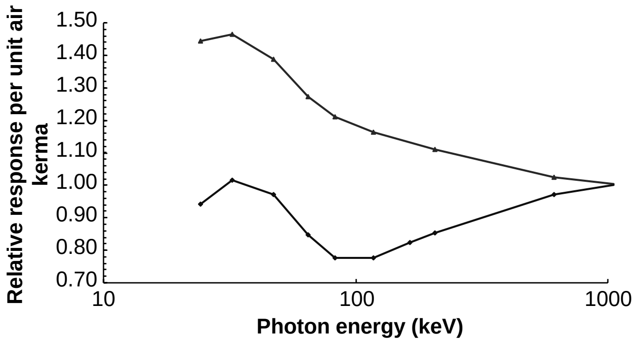


Figure 9. Energy dependence of two types of LiF TLD dosimeters (from Ref. 5). Triangles - TLD-100, diamonds - TLD-100H. LiF:Mg,Ti (TLD-100) has been a widely used dosimeter since the 1960s, LiF:Mg,Cu,P (TLD-100H) was developed in 1976, with 20–30 times greater sensitivity.

Summary

For reference dosimetry in radiotherapy clinics the system of choice is the ion chamber. Ion chambers are simple to use, offer high precision and accuracy and give an immediate reading. Integrating dosimeters (Fricke, alanine, and TLD) tend to be used as QA checks, either internally or within a wider framework of national or international comparisons (e.g., TLD is used for both the IAEA international mailed reference dosimetry service (53) and the RPC audit scheme in North America (54)). Generally, ion chambers are calibrated against primary standards and then used to calibrate other dosimetry systems within the clinic.

CALIBRATION OF SECONDARY DOSIMETERS

Basic Formalism

An ideal secondary dosimeter will have a zero energy dependence. Calibration against a primary standard (calorimeter) would then only need to be carried out at one beam quality (e.g., ^{60}Co). Energy independence also implies that the calibration coefficient is the same in photon and electron beams since the dose in a photon beam is dependent on the secondary electron spectrum generated in-phantom. In practice, the majority of secondary dosimeters commonly in use have some energy dependence. Ionization chambers, for example, show a variation of $>$ 3% over the energy range from ^{60}Co to 25 MV photons, with even larger variations at low X-ray energies.

The obvious method to calibrate an ion chamber in terms of absorbed dose is to compare a chamber with a primary device. However, although accelerators were being used from the 1950s for radiotherapy, there were no absorbed dose standards for megavoltage photon or electron beams until the 1970s. Absorbed dose measurements using ion chambers were therefore based on air-kerma calibrations derived at lower photon energies (either ^{60}Co or 2 MV X rays). Protocols were developed to enable users to obtain a measurement of the absorbed dose delivered by a linac in the clinic. Only in recent years have absorbed dose-based calibrations become available from national standards laboratories and associated protocols produced (e.g., Refs. 6–8). For the purpose of the following discussion, we will only deal with absorbed dose calibrations in megavoltage photon and electron beams, but the principles are basically the same for other situations (kV X rays, protons, etc.).

The basic formalism for the calibration of an ion chamber is simple. The chamber is compared against the ^{60}

primary device and a calibration coefficient ($N_{\mathrm{D,sec}}$) for that beam is derived

$$N_{\mathrm{D,sec}} = \frac{D_{\mathrm{std}}}{M_{\mathrm{sec}}} \tag{14}$$

The parameter D_{std} is the dose measured by the primary device and M_{sec} the chamber reading corrected for influence quantities. This calibration coefficient will be a function of the energy of the photon or electron beam and is given in terms of a beam quality specifier, Q_{ref}. The user then derives the calibration coefficient for the user beam quality, $N_{\mathrm{D,ref}}$ (Q_{user}). Some primary laboratories only supply calibration coefficients for ^{60}Co and thus correction factors are required, which are given in dosimetry protocols (e.g., Ref. 7). An alternative approach, as used in the United Kingdom's Code of Practice (6) is to obtain absorbed dose calibration coefficients in linac photon beams. In this case there is no need for the calculated conversion factors and an ion chamber is calibrated in a beam similar to what it will be used to measure.

A measurement is then made in the user's radiation beam to measure the absorbed dose, D_{user}:

$$D_{\mathrm{user}} = N_{\mathrm{D,sec}}(Q_{\mathrm{user}})M_{\mathrm{user}} \tag{15}$$

where M_{user} is the chamber reading.

Implied in equations 14 and 15 is the reference depth at which the measurement is carried out. The concept of the reference depth for a calibration is much more important for electrons than photons. In a phantom irradiated by a megavoltage photon beam the secondary electron spectrum (which determines the dose) varies only slowly with depth (for depths greater than the range of incident primary electrons). By contrast, in the situation of a primary electron beam, the electron spectrum seen by the detector constantly changes from the surface to the practical range. The choice of reference depth should be both clinically relevant and reliable in terms of transferring the dose from the primary laboratory to the user's beam. For photon beams there may be only one or two reference depths defined for all energies while for electron beam dosimetry all modern protocols define the reference depth as a function of energy.

Potential Problems with Beam Quality Specifiers

A typical radiotherapy linac accelerates electrons to energies in the range 4–22 MeV and can also produce bremsstrahlung X-ray beams over a similar energy range. In both cases, the detector calibration coefficient will be some function of this spectrum. Since it is not generally possible to measure the energy spectrum directly a beam quality specifier (Q) is used. This is obtained by measuring some property of the radiation beam (e.g., the penetration through a material). A "good" beam quality specifier is one such that a value of Q relates uniquely to the effect of a particular spectrum. A problem arises if Q is not a good beam quality specifier, that is, there is some ambiguity in the relation between Q and the effect of the incident spectrum.

Beam Quality Specifiers for Photon Beams. Typical photon depth–dose curves from a clinical linac are shown in Fig. 10 (it should be noted that MeV tends to be used as a label for electron beams and MV for photon beams). Over the years, a number of beam quality specifiers have been proposed for megavoltage photon beams, but all relate in some way to the penetration of the photons through some material.

The most widely used parameter has been $\mathrm{TPR}_{20,10}$ (tissue phantom ratio), which is defined as the ratio of ionization currents at measurement depths of 20 and 10 cm in water with a fixed-field size and source to chamber distance. The 10 and 20 cm points in Fig. 10 are on the downward portion of the curves, and therefore TPR is related to a measurement of the attenuation of the beam.

There has been much debate in recent years over the sufficiency of $\mathrm{TPR}_{20,10}$ as a beam quality specifier for

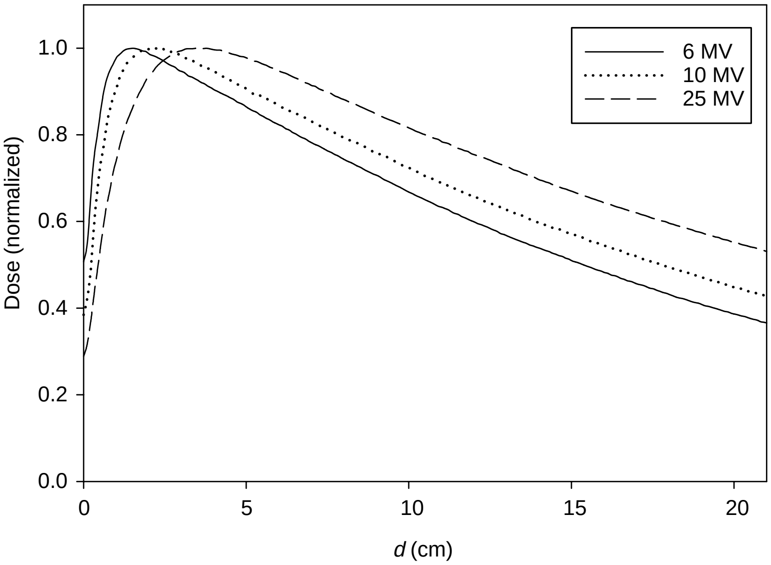


Figure 10. Depth dose curve for three photon beams.

the purpose of ion chamber calibration in terms of absorbed dose to water. Rosser et al. (55) found that an error of up to 0.6% could be introduced by the incorrect application of calibration coefficients using TPR. A number of other beam quality specifiers have been put forward as alternatives to TPR including: d_{80} (the depth at which the dose is 80% of the peak dose); the HVL of water and the percentage depth dose at a depth of 10 cm, $\%\mathrm{dd}(10)_X$ (where a 1 mm lead filter is used to correct for electron contamination). There is no consensus on this problem at the moment: the new IAEA Code of Practice (8) uses $\mathrm{TPR}_{20,10}$ while the AAPM absorbed dose protocol (7) uses $\%\mathrm{dd}(10)_X$. However, in practice there is no real controversy: Kalach and Rogers (56) showed that although $\%\mathrm{dd}(10)_X$ gave better agreement for a wide range of accelerators, for the heavily filtered beams produced by modern clinical linacs, $\mathrm{TPR}_{20,10}$ was an adequate beam quality specifier.

Beam Quality Specifiers for Electron Beams. It is potentially simpler to measure the electron spectrum from a Linac than a photon spectrum. The most accurate method is to use a calibrated magnetic spectrometer (57,58), but this technique tends to be rather time consuming and the necessary equipment is not always available. The mean energy of the electron beam can be determined via activation analysis (59,60) or the determination of the total charge and energy using a Faraday cup. However, all these systems tend to be rather complex so the actual electron spectrum (or even mean electron energy) is rarely measured.

As with photons, parameters derived from the penetration of electrons in a medium are used as a measure of electron energy. A typical electron depth–dose curve in water is shown in Fig. 11. The two most important parameters obtained from such a curve are R_{50}, defined as the depth at 50% of the peak dose; and R_p (the practical range), defined as the point where the extrapolation from the point of maximum gradient on the downward part of the curve meets the extrapolation of the bremsstrahlung background.

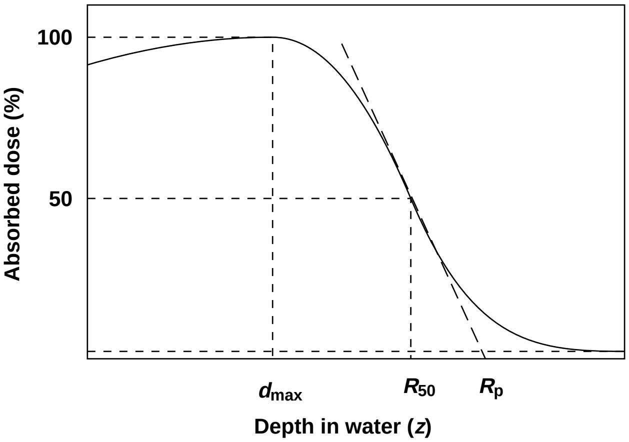


Figure 11. An electron beam depth–dose distribution in water showing the various range parameters. (From Ref. 61).

Considerable work has been done to relate these parameters to beam energy (see Ref. 2), and it is generally understood that R_{50} and R_p described different aspects of the incident electron spectrum. The parameter R_{50} relates to the mean electron energy, while R_p is directly related to the most probable energy. For a symmetrical, single-peaked spectrum the mean and most probable energies will be the same but, as shown by Klevenhagen (62), the spectrum incident on a phantom will be skewed towards lower energies due to scattering in air. Reference 2 shows depth–dose data for two spectra where the most probable energy is the same but with different mean energies (and different energy spreads). In this case, the two curves give the same value for the practical range, but different values for R_{50}. However, Burns et al. (63) collated a large amount of depth–dose data from a wide variety of linacs and showed that there was a direct relation between R_{50} and R_p, indicating that the majority of linacs in use today either generate symmetrical or very similar spectra.

RELATIVE DOSIMETRY AND QUALITY/VERIFICATION

For relative dosimetry or for quality (QA) measurements there are a wide range of dosimetry systems to choose from (including many discussed above as secondary dosimeters). The choice will depend on a number of factors including application (simple external beam therapy, intensity modulated radiotherapy (IMRT), brachytherapy); precision; spatial resolution and/or detector size; type of measurement (i.e., relative, QA, etc.); and immediacy (instant readout required?). However, one of the primary drivers will be practical issues such as cost, availability, complexity and setup time. With so many detectors to choose from it is difficult to give anything other than a very brief overview here.

Solid-State Detectors (1D)

Diodes. Semiconductor diodes offer increased sensitivity over air-filled ionization chambers due to the higher density of charge carriers. This means that the sensitive volume can be made ~ 100–1000 times smaller, giving excellent spatial resolution. The stopping power ratio silicon/water varies much less with energy than the air/water ratio, and therefore diodes are ideally suited for measuring dose distributions. The biggest problem with diodes is that the sensitivity is dose dependent and diodes need recalibrating approximately every few hundred gray. Dose diodes are available in two types: electron and photon. Photon diodes employ shielding around the sensitive volume to correct for the effects of scattered radiation in a photon beam. Uncorrected, an unshielded diode overestimates the dose at depth by as much as 15% for a 6 MV beam and Yin et al. (64) present a method to correct for the response of diodes in photon beams. Due to the potential for confusion as to the construction of a diode, dosimetry protocols usually recommend that diode measurements are validated using an ion chamber.

Diamond. Diamond detectors have been investigated for over 20 years (65,66). The spatial resolution of diamond detectors (1–6 mm^3) is comparable to that of commonly used silicon diode detectors with the added advantage of showing high resistance to radiation damage (0.05% $k \cdot Gy^{-1}$, > 100 times lower than typical diode values). As expected for a solid-state dosimeter, diamond detectors have a high sensitivity, but also a good long-term stability and low temperature dependence. Diamond has a reasonable tissue equivalence for both photon and electron beams. The majority of recent work with diamond detectors has focused on their use for small field IMRT and brachytherapy, where the high sensitivity and small size are highly advantageous. Mack et al. (67) compared a number of techniques for the dosimetry of small radiosurgery beams and found that a diamond detector gave very good results down to field sizes of 4 × 4 mm. However, a significant disadvantage of diamond detectors at present is the very high cost compared to other solid-state detectors. This is because at present natural diamonds are used and have to be carefully selected, since the dose linearity and polarization effects are very sensitive to impurities and defects in the crystal. However, the more recent availability of low cost, polycrystalline diamond specimens produced using chemical vapor deposition (CVD) offers the potential for improved diamond detectors with selectable size and impurities.

MOSFETs. The MOSFET dosimeter (68) is a more recent development. The dosimeter operates by measuring the threshold voltage of a MOSFET field effect transistor, which is permanently changed by the absorption of radiation (radiation damage). As an integrating detector it therefore has similar applications to TLD or alanine and the small detector area (only 0.2 × 0.2 mm) offers very high spatial resolution. The reproducibility is typically ± 2%.

Two-Dimensional Detectors

Radiographic Film. Film dosimetry has long been viewed as an attractive alternative to ionometric and thermoluminescent methods as an entire two-dimensional (2D) chart may be extracted from a single film exposure. In addition, film has the highest spatial resolution of any practical dosimeter and is easily set up and exposed. However, radiographic film dosimetry is not without problems - daily film calibrations are essential to obtain absolute dose results and care must be taken with film handling and processing not to introduce artifacts. A more fundamental problem is that the high atomic number of the silver halide film emulsion means that dose response relative to water varies significantly in the low energy photon range (< 200 keV), as can be seen in Fig. 12. This is also an issue for measurements in megavoltage photon beams where the scatter-to-primary ratio can change (e.g. off-axis). Having said this, radiographic film dosimetry is experiencing a renaissance in the radiation therapy community, driven by the need to verify the absorbed dose delivered with IMRT, where both detector resolution and 2D data acquisition are advantages.

As with any other secondary dosimeter, a calibration curve must be derived. The quantity measured using film is the optical density (measured using a scanner or densitometer) and ideally this should show a linear increase with dose (Fig. 13).

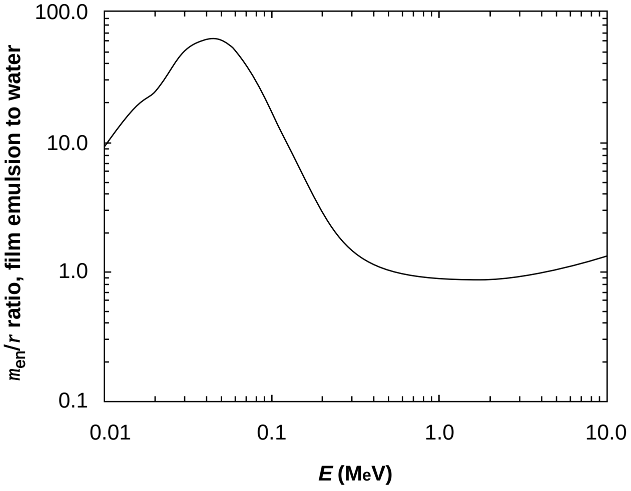


Figure 12. Energy dependence of radiographic film relative to that of water.

Although there are obvious problems with radiographic film, alternatives for 2D dose verification have their own limitations and therefore it is likely that film will continue to play a role in dosimetry. Radiochromic film is considered tissue equivalent and energy independent, but is expensive, limited in size, and prone to large dose response nonuniformities due to the manufacturing process. Commercially available 2D ion chamber and diode arrays provide a fast and accurate evaluation, but are still limited in resolution, and therefore better used to dose verification rather than commissioning IMRT. Electronic portal imagers have the advantage of being available in many modern therapy centres but by design, they measure fluence patterns, not dose distributions in phantoms, and therefore interpretations of delivery errors could be difficult.

Radiochromic Film. A radiation-induced color change is one of the simplest dosimeters one can think of and a number of systems were developed in the early 1900s. Although widely used in the radiation processing industry, where

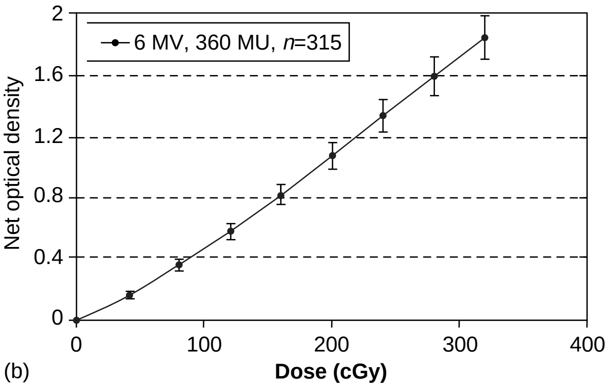
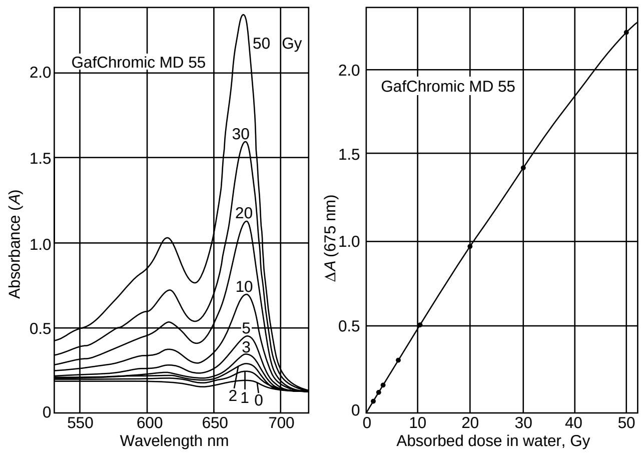


Figure 13. Dose linearity curve for Kodak EDR2 film (from Ref. 69). The response is very repeatable and linear for doses above 150 cGy.

the large doses can induce easy-to-detect color changes, radiochromic films have only recently begun to be used once again for radiotherapy dosimetry. The most promising to date is the GafChromic material. One of the main advantages of radiochromic films over radiographic is that they are essentially tissue equivalent so the energy response relative to water only changes very slowly with energy (Fig. 14).

As with any dosimeter there are problems in obtaining high accuracy dosimetric information. Klassen et al. (71) carried out a detailed investigation and showed that the precision was affected by the readout method, the readout temperature and wavelength as well as the polarization of the light source. However, with care, dosimetry with a relative uncertainty of $< 1\%$ is possible for doses of the order of 6 Gy.

EPIDs. Electronic portal imaging devices (EPIDs) have been gradually replacing conventional radiographic film for geometric verification in radiation therapy. The obvious advantage of using an EPID for dosimetry is that they are now standard equipment on most modern linacs. Early generations, employing liquid ion-chambers or camera-based fluoroscopy, generally produced poorer images compared to film, but it was shown that EPIDs could be used for IMRT quality assurance (e.g., leaf position verification for Multi- Leaf Collimators, or MLCs). The most recent class of EPID uses flat-panel photodiode arrays and with improved spatial resolution and higher detector efficiency are especially well suited for IMRT applications. However, to use any EPID for dosimetric IMRT requires calibration coefficients to relate pixel intensity to either fluence or dose. Calibration of the EPID is more involved than simple cross-calibration of pixel response with dose measurements made with an ion chamber in a homogeneous water phantom, but the ability to verify treatment "as it happens" is a significant advantage over other methods. Warkentin et al. (72) describe the use of a flatpanel detector for accurate pretreatment dosimetric verification of IMRT treatment fields.



(a) (b)

Figure 14. Performance of GafChromic MD-55 film. (From Ref. 70 reprinted with permission from Elsevier.) (a) Absorption spectra as a function of dose. (b) Dose response curve measured at the absorption band peak.

Three-Dimensional Detectors

The adoption of conformal radiotherapy techniques and, in particular, IMRT, where verification of the delivered 3D dose distribution is very important, has been a major driving force in the development of 3D detection systems. Presently, there are basically three options: TLD (either as individual pellets or in powder form), film stacks (radiographic or radiochromic), and gel dosimeters.

Gel Dosimetry. Although the use of radiation sensitive gels for dosimetry measurements was suggested as early as the 1950s, the use and development of this type of dosimeter has only grown significantly in the last decade. Gel dosimeters offer a number of advantages over other 3D techniques, such as TLD or film stacks, including resolution, number of data points, energy dependence, and water equivalence (Fig. 15).

There are currently two main types of gel dosimeter: (*1*) Fricke gel – ferrous sulfate solution is incorporated into aqueous gel matrices of gelatin, agarose or poly (vinyl alcohol) (PVA). As for the Fricke dosimeter, there is a conversion of Fe^{2+} ions to Fe^{3+} and this change in concentration is readout via magnetic resonance imaging (MRI) or optical tomography. One of the main drawbacks of Fricke gels is that there is a rapid diffusion of the ferric ions centers within the matrix, which tends to smooth out the dose distribution. (*2*) Polymer gels: This system is based on the polymerisation of certain materials. Initial work focused on the materials acrylamide (AA) and *N,N*-methylene-bis(acrylamide) (BIS) with readout again via MRI. One of the main problems with these systems is that they are sensitive to atmospheric oxygen contamination. A newer formulation named methacrylic and ascorbic acid in gelatin initiated by copper (MAGIC) is less sensitive to the presence of oxygen and looks promising as a gel dosimeter. Perhaps the biggest problem with gel systems is that they require

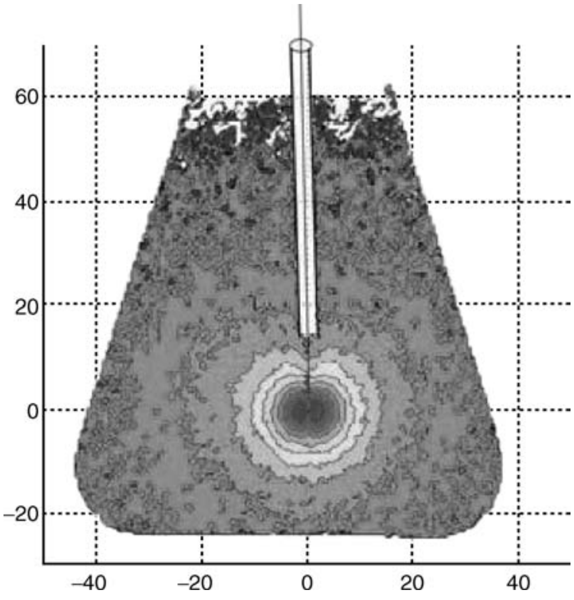


Figure 15. Dose distribution for a high dose rate ^{192}Ir brachytherapy source measured in a flask of polymer gel dosimeter. (From Ref. 73.)

a containment vessel, which can both perturb the dose measurement and introduce imaging artefacts.

There is a very active gel dosimetry community worldwide and development continues both on gel formulations (e.g., reduce diffusion or sensitivity to impurities) and readout (e.g., CT and ultrasound have been suggested as alternative readout methods to MRI). For a recent review of the subject see Baldock (74).

CONCLUSION

This article has outlined the basic theory of radiation dosimetry and the problems involved in measuring absorbed dose. A number of primary and secondary measurement techniques have been described together with the formalism for calibrating dosimeters. Since whole textbooks have been written on this subject, this can be no more than a brief introduction to the field. Readers are referred to the extensive bibliography for further detail.

BIBLIOGRAPHY

1. International Commission on Radiation Units and Measurements (ICRU) 1998 Fundamental Quantities and Units for Ionizing Radiation ICRU Report 60. Bethesda, MD: ICRU; 1998.
2. International Commission on Radiation Units and Measurements (ICRU) 1984 Radiation dosimetry: electron beams with energies between 1 and 50 MeV, ICRU Report 35. Bethesda, MD: ICRU; 1984.
3. Boutillon M, Perroche A M. Re-evaluation of the W value for electrons in dry air. Phys Med Biol 1987;32:213–219.
4. American Association of Physicists in Medicine (AAPM) AAPM TG-21: A protocol for the determination of absorbed dose from high-energy photon and electron beams. Med Phys 1983;10:741–771.
5. International Atomic Energy Agency (IAEA Absorbed dose determination in photon and electron beams: an international code of practice. International Atomic Energy Agency Technical Report 277. Vienna: IAEA; 1987.
6. Institute of Physical Sciences in Medicine (IPSM) Code of Practice for high-energy photon therapy dosimetry based on the NPL absorbed dose calibration service. Phys Med Biol 1990;35:1355–1360.
7. American Association of Physicists in Medicine (AAPM). AAPMs TG-51 protocol for clinical reference dosimetry of high-energy photon and electron beams Report of AAPM Radiation Therapy Committee Task Group No. 51. Med Phys 1999;26:1847–1870.
8. International Atomic Energy Agency (IAEA)Absorbed Dose Determination in External Beam Radiotherapy (IAEA Technical Reports Series No. 398). Vienna: IAEA; 2000.
9. Boutillon M, Peroche A-M. Ionometric determination of absorbed dose to water for cobalt-60 gamma rays. Phys Med Biol 1993;38: 439–454.
10. Klevenhagen SC. Determination of absorbed dose in high-energy electron and photon radiation by means of an uncalibrated ionization chamber. Phys Med Biol 1991;36:239–253.
11. Zankowski CE, Podgorsak EB. Calibration of photon and electron beams with an extrapolation chamber. Med Phys 1997; 24:497–503.
12. van der Marel J and van Dijk E 2003. Development of a Dutch primary standard for beat emitting brachytherapy sources Standards and Codes of Practice in Medical Radiation Dosimetry (Proc. Int. Symp. Vienna, 2002), IAEA, Vienna.
13. Domen SR, Lamperti PJ. J Res Natl Bur Stand (US) 1974; 78:595.
14. DuSautoy AR. The UK primary standard calorimeter for photon beam absorbed dose measurement. Phys Med Biol 1996;41:137.
15. Ross CK, Seuntjens JP, Klassen NV, Shortt KR. The NRC Sealed Water Calorimeter: Correction Factors and Performance. Proceeding of the Workshop on Recent Advances in Calorimetric Absorbed Dose Standards, Report CIRM 42. Teddington: National Physical Laboratory; 2000.
16. McEwen MR, Duane S. A Portable graphite calorimeter for measuring absorbed dose in the radiotherapy clinic ≅ In: Standards and Codes of Practice in Medical Radiation Dosimetry. Proceeding of the International Symposium Vienna, 2002, Vienna: IAEA; 2003.
17. Ross CK, Klassen NV. Water calorimetry for radiation dosimetry. Phys Med Biol 1996;41:1–29.
18. Williams AJ, Rosser KE, editors. Proceedings of the NPL Workshop on Recent Advances in Calorimetric Absorbed Dose Standards NPL Report CIRM 42. Teddington: National Physical Laboratory, 2000.
19. Seuntjens JP, DuSautoy AR. Review of calorimeter based absorbed dose to water standards. In: Standards and Codes of Practice in Medical Radiation Dosimetry. Proceeding of the International Symposium Vienna, 2002. Vienna: IAEA; 2003.
20. Fricke H, Morse S. The actions of x-rays on ferrous sulfate solutions. Phil Mag 1929;7(7):129.
21. Roos M, Hohlfeld K. Status of the primary standard of water absorbed dose for high energy photon and electron radiation at the PTB. Measurement Assurance in Dosimetry. Vienna: International Atomic Energy Agency; 1994. p 25–33.
22. International Commission on Radiation Units and Measurements (ICRU) Stopping powers for electrons and positrons (ICRU Report 37). Bethesda, MD: 1984.
23. Feist H, Muller U. Measurement of the total stopping power of 5.3 MeV electrons in polystyrene by means of electron beam absorption in ferrous sulphate solution. Phys Med Biol 1989;34:1863.
24. Faddegon BA, Ross CK, Rogers DWO. Measurement of collision stopping powers of graphite, aluminium and copper for 10 and 20 MeV electrons. Phys Med Biol 1992;37:1561–71.
25. MacPherson MS. Accurate measurements of the collision stopping powers for 5 to 30 MeV electrons Ph.D. dissertation. Ottawa: INMS, NRC; 1998. PIRS-0626.
26. Burns JE, Dale JWG. Conversion of absorbed-dose calibration from graphite to water NPL Report RSA(EXT)7. Teddington: NPL; 1990.
27. Pruitt JS, Loevinger R. The photon-fluence scaling theorem for Compton-scattered radiation. Med Phys 1982;9:176–179.
28. Nutbrown RF, Duane S, Shipley DR, Thomas RAS. Evaluation of factors to convert absorbed dose calibrations in graphite to water or mega-voltage photon beams NPL Report CIRM 37. Teddington: NPL; 2000.
29. Rogers DWO, et al. BEAM: A Monte Carlo code to simulate radiotherapy treatment units. Med Phys 1995;22:503–524.
30. Kawrakow I. Accurate condensed history Monte Carlo simulation of electron transport. I. EGSnrc, the new EGS4 version. Med Phys 2000;27:485–498.
31. Aird EGA, Farmer FT. The design of a thimble chamber for the Farmer dosemeter. Phys Med Biol 1972;17:169–174.
32. Mattsson LO, Johansson K-A, Svensson H. Calibration and use of parallel-plate ionization chambers for the determination of absorbed dose in electron beams. Acta Radiol Oncol 1981;20:385–399.
33. Williams AJ, McEwen MR, DuSautoy AR. A calculation of the water to graphite perturbation factor ratios for the NACP type 02 ionisation chamber using Monte Carlo techniques. NPL

Report CIRM(EXT)013. Teddington: National Physical Laboratory; 1998.

34. International Commission on Radiation Units and Measurements (ICRU) Average energy required to produce an ion pair (ICRU Report 31). Bethesda, MD: 1984.
35. Boag JW. Ionization measurements at very high intensities. I. Pulsed radiation beams. Brit J Radiol 1950;23:601–611.
36. Boag JW, Currant J. Current collection and ionic recombination in small cylindrical ionization chambers exposed to pulsed radiation. Brit J Radiol 1980;53:471–478.
37. International Commission on Radiation Units and Measurements (ICRU) The dosimetry of pulsed radiation, ICRU Report 34. Bethesda, MD: ICRU; 1982.
38. Derikum K, Roos M. Measurement of saturation correction factors of thimble-type ionization chambers in pulsed photon beams. Phys Med Biol 1993;38:755–763.
39. Burns DT, McEwen MR. Ion recombination for the NACP parallel-plate chamber in a pulsed electron beam. Phys Med Biol 1998;43:2033–2045.
40. DeBlois F, Zankowski C, Podgorsak EB. Saturation current and collection efficiency for ionization chambers in pulsed beams. Med Phys 2000;27:1146.
41. Boag JW. Ionization Chambers. In: Attix FH, Roesch WC, Tochilin E, editors. Radiation Dosimetry. Vol. II, New York: Academic; 1966. Chapt. 9, p 2–67.
42. Williams JA, Agarwal SK. Energy-dependent polarity correction factors for four commercial ionization chambers used in electron dosimetry. Med Phys 1997;24:785–790.
43. Nisbet A, Thwaites DI. Polarity and ion recombination correction factors for ionization chambers employed in electron beam dosimetry. Phys Med Biol 1998;43:435–443.
44. Pearce JAD. Characterisation of two new ionisation chamber types for use in reference electron dosimetry in the UK, NPL Report DQL-RD001. Teddington: National Physical Laboratory; 2004.
45. Bahar-Gogani J, Grindborg JE, Johansson BE, Wickman G. Long-term stability of liquid ionization chambers with regard to their qualification as local reference dosimeters for low dose-rate absorbed dose measurements in water. Phys Med Biol 2001;46:729–740.
46. Klassen NV, Shortt KV, Seuntjens J, Ross CK. Fricke dosimetry: the difference between G(Fe^{3+}) for ^{60}Co gamma-rays and high-energy X-rays. Phys Med Biol 1999;44:1609–1624.
47. Ross CK, Klassen NV, Shortt KR. The development of a standard based on water calorimetry for the absorbed dose to water. Proceeding of the NPL Calorimetry Workshop.Teddington: NPL; 1994.
48. Marre D, et al. Energy correction factors of LiF powder TLDs irradiated in high-energy electron beams and applied to mailed dosimetry for quality assurance networks. Phys Med Biol 2000;45:3657–3674.
49. Davis SD, et al. The response of LiF thermoluminescence dosemeters to photon beams in the energy range from 30 kV X rays to ^{60}Co gamma rays. Radiat Prot Dosimetry 2003;106:33–43.
50. Zeng GG, McEwen MR, Rogers DWO, Klassen NV. An experimental and Monte Carlo investigation of the energy dependence of alanine/EPR dosimetry: I. Clinical X-ray beams. Phys Med Biol 2004;49:257–270.
51. Izewska J, Bera P, Andreo P, Meghzifene A. Thirty Years of the IAEA/WHO TLD Postal Dose Quality Audits for Radiotherapy. Proceeding of the World Congress on Medical Physics. Chicago: AAPM; 2000.
52. Aguirre JF, et al. Thermoluminescence dosimetry as a tool for the remote verification of output for radiotherapy beams: 25 years of experience. In: Standards and Codes of Practice in Medical Radiation Dosimetry. Proceeding of the International Symposium Vienna; 2002, Vienna: IAEA; 2003.
53. Rosser KE, et al. The NPL absorbed dose to water calibration service for high energy photon beams. In: Flitton SP, editor Proceeding of the International Symposium on Measurement Assurance in Dosimetry (IAEA-SM-330/35) Vienna: IAEA; 1994. p 73.
54. Kalach NI, Rogers DWO. Which accelerator photon beams are ‘clinic-like’ for reference dosimetry purposes? Med Phys 2003;30:1546–1555.
55. Wessels BW, Paliwal BR, Parrot MJ, Choi MC. Characterization of Clinac-18 electron-beam energy using a magnetic analysis method. Med Phys 1979;6:45.
56. Deasy JO, Almond PR, McEllistrem MT. Measured electron energy and angular distributions from clinical accelerators. Med Phys 1996;23:675.
57. Almond PR. The physical measurements of electron beams from 6 to 8 MeV: absorbed dose and electrical calibration. Phys Med Biol 1967;12:13.
58. Klevenhagen SC. Physics and Dosimetry of Therapy Electron Beams. Madison: Medical Physics Publishing; 1993.
59. Institution of Physics and Engineering in Medicine and Biology (IPEMB) The IPEMB code of practice for electron dosimetry for radiotherapy beams of initial energy from 2 to 50 MeV based on an air kerma calibration. Phys Med Biol 1996;41:2557–2603.
60. Klevenhagen SC. Physics of electron beam therapy. Bristol: Adam Hilger; 1985. p 65.
61. Burns DT, Ding GX, Rogers DWO. R_{50} as a beam quality specifier for selecting stopping-power ratios and reference depths for electron dosimetry. Med Phys 1996;23:383.
62. Yin Z, Hugtenberg RP, Beddoe AH. Response corrections for solid-state detectors in megavoltage photon dosimetry. Phys Med Biol 2004;49:3691–3702.
63. Laub WU, Kaulich TW, Fridtjof N. A diamond detector in the dosimetry of high-energy electron and photon beams. Phys Med Biol 1999;44:2183–2192.
64. Planskoy B. Evaluation of diamond radiation dosemeters. Phys Med Biol 1980;25.
65. Mack A, et al. Precision dosimetry for narrow photon beams used in radiosurgery—Determination of Gamma Knife[sup [registered sign]] output factors. Med Phys 2002;29:2080–2089.
66. Ramani R, Russell S, O’Brien P. Clinical Dosimetry Using MOSFETS. Int J Rad Oncol Biol Phys 1997;37:956–964.
67. Childress NL, White RA, Rosen II. Dosimetric accuracy of Kodak EDR2 film for IMRT verifications. Med Phys 2005;32: 539–548.
68. McLaughlin WL, Desrosiers MF. Dosimetry systems for radiation processing. Radiat Phys Chem 1995;46:1163–1174.
69. Klassen NV, van der Zwan L, Cygler J. GafChromic MD-55: Investigated as a precision dosimeter. Med Phys 1997;24: 1924–1934.
70. Warkentin B, Steciw S, Rathee S, Fallone BG. Dosimetric IMRT verification with a flat-panel EPID. Med Phys 2003;30:3143–3155.
71. De Deene Y, Reynaert N, De Wagter C. On the accuracy of monomer/polymer gel dosimetry in the proximity of a high-dose-rate ^{192}Ir source. Phys Med Biol 2001;46: 2801–2825.
72. Baldock C. Radiotherapy gel dosimetry. In: Standards and Codes of Practice in Medical Radiation Dosimetry. Proceeding of the International Symposium Vienna, 2002. Vienna: IAEA; 2003.

Reading List

The following textbooks are recommended for further reading; they provide excellent coverage of the subject of radiation dosimetry.

Attix FH, Roesch WC, Tochilin E, editors. Radiation dosimetry. (Pts I,II,III) 2nd ed. New York: Academic Press; 1966–1969.

Greening JR. Fundamentals of radiation dosimetry. 2nd ed. Bristol [England]: A. Hilger in collaboration with the Hospital Physicists' Association; 1985.

Johns HE, Cunningham JR. The physics of radiology. 4th ed. Springfield, IL: Thomas.

Kase KR, Bjärngard B, Attix FH, editors. The Dosimetry of ionizing radiation. (Pts I, II, III) Orlando, FL: Academic; 1985.

Klevenhagen SC. Physics and Dosimetry of Therapy Electron Beams. Madison, WI: Medical Physics Publishing; 1993.

See also IONIZING RADIATION, BIOLOGICAL EFFECTS OF; RADIATION THERAPY SIMULATOR.

RADIATION DOSIMETRY, THREE-DIMENSIONAL

GEOFFREY S. IBBOTT
Anderson Cancer Center
Houston, Texas

INTRODUCTION

The goal of radiation therapy is to obtain the greatest possible local and regional tumor control, with the fewest complications. The response of many tissues to radiation can be characterized by a sigmoid curve. Relatively little response is seen until the dose reaches some threshold value, after which the response is quite rapid (1,2). In the region of steep response, relatively small variations in dose can yield significant differences in the response of both tumors and normal tissue (3). To minimize the variability of tissue response, the ICRU has recommended that the uncertainty in dose delivery be maintained below $\sim 5\%$ (4–6). Delivering a dose to a patient with a tolerance of 5% is not a simple matter (7). It has been estimated that the equipment used by most medical physicists to calibrate therapeutic radiation beams is itself calibrated with an overall uncertainty (expressed at the 95% confidence level) of $\sim 1.5\%$ (8). Uncertainties associated with the characterization of radiation beams, patient anatomy, and location of the target volume, as well as reproducibility of the treatment from day to day must be considered (9,10).

A comprehensive radiation therapy quality assurance program must address all sources of variability in the treatment of patients, in an effort to minimize variations. Technical aspects of quality assurance (QA) must address a wide array of issues, including the performance of simulation, treatment, and treatment planning equipment, the stability of measurement and test equipment, the accuracy and appropriateness of treatment planning calculations, and the accuracy and completeness of documentation. Technical quality assurance procedures should also address inventory, calibration, and treatment planning with brachytherapy sources. Recommendations for QA procedures can be found in a number of publications (11–23).

As the equipment used to deliver radiation therapy has evolved, methods of radiation dosimetry have also changed. Multifield, conformal radiation therapy (CRT), intensity-modulated radiation therapy (IMRT), stereotactic radiosurgery (SRS), and stereotactic radiation therapy (SRT) all produce dose distributions that can be highly irregular in three dimensions. Conventional two-dimensional (2D) planning and dosimetry systems are not adequate to simulate and measure such distributions. Instead, new dosimetry systems are required that can record and display these complex distributions (24). This article addresses recent developments in dosimetry systems, and their advantages and complications.

QUALITY ASSURANCE PROCEDURES REQUIRING DOSIMETRY SYSTEMS

External Beam Calibration Consistency-Basic Parameters

Detector systems are required for measurement of accelerator output, for compliance with published recommendations for quality assurance. Most published recommendations suggest that accelerator output constancy be monitored on a daily basis. Consequently, a dosimeter system that is rugged, reliable, and easy to operate is required. Most recommendations for daily output consistency suggest that deviations on the order 2–5% be detectable; therefore the dosimetry system does not need extremely high accuracy.

In addition, measurements of beam flatness and symmetry are recommended on a periodic basis, often weekly. Again, as these measurements are to demonstrate consistency of operation at the 2–5% level, high precision is not required. Several of the available array dosimeters systems are suitable for such frequent QA measurements of treatment unit performance.

External Beam Treatment Delivery, Planning Verification

Several treatment applications require the verification of delivered dose with relatively high accuracy. For example, IMRT requires the precise delivery of relatively small doses through a large number of fields. Even small errors in dose delivery can accumulate and result in a large error in the final dose. Monitoring of dose delivery during IMRT is done best using a real time measuring device, such as online portal imaging. Similarly, CRT delivery demands confirmation that the correct dose has been delivered. As CRT is generally delivered through static fields, point detectors may be used to measure the delivered dose in a suitable phantom. Several of the simpler point dosimeter systems described earlier are suitable for this purpose.

Likewise, SRS and SRT delivered with accelerators may need verification, particularly as SRS is delivered in single large fractions. Again, under most circumstances, point dosimeters are suitable here. However, the characterization of radiation beams for SRS–SRT requires a dosimeter with high spatial accuracy. Several of the detector systems described above would satisfy this requirement, although questions of electronic equilibrium must be addressed (25).

Total body treatments, such as photon TBI for systemic bone marrow ablation, or total skin electron therapy for cutaneous t-cell lymphoma, may require dosimetry to confirm the correct delivery of dose under these conditions of unusual field size and distances.

External Beam Treatment Delivery In Vivo

Modern external treatment delivery requires that doses be delivered with accuracy never before required. Procedures, such as IMRT, are delivered through many field segments, each delivering a small increment of dose. A systematic error in dose delivery can result in a significant error in the final dose received by the patient. Consequently, dosimetry devices for confirming correct dose delivery are necessary. These fall into three broad classes: surface dose measurements, transmission measurements, and true *In vivo* measurements.

Brachytherapy (LDR, HDR, IVBT)

Dosimeter systems are required for at least three purposes related to brachytherapy: source characterization, confirmation of dose distributions from arrangements of multiple sources, and *In vivo* dose measurements (26,27).

Imaging Procedures

Dosimetry measurements are required in cardiology, for procedures, such as cardiac ablation, in which patients can receive significant doses. A detector to be used in imaging must not be intrusive, meaning that it must be virtually transparent to the beam. It must measure dose over a large area, although accommodation needs to be made for the possibility that the beam may be moved during irradiation. Finally, a device attached to the source of radiation, such as a dose area product meter, may be used.

REQUIRED CHARACTERISTICS OF DOSIMETERS FOR QUALITY ASSURANCE

A dosimeter for modern CRT must possess a number of important characteristics. It must be tissue equivalent, as the dosimeter itself must not perturb the dose distribution. It must have a linear dose response over a clinically useful range. Ideally, its response would be independent of dose rate and of beam modality, making it useful for mapping dose distributions from isotope units, linear accelerators, or particle accelerator beams. Some dosimeters must be able to fill a volume, or conform to a surface. This will enable the dosimeter to either mimic any portion of human anatomy, or conform to a section of an anthropomorphic phantom.

The dosimeter must either provide immediate results or be stable for a sufficiently long period to enable irradiation and analysis. Under some circumstances, the delivery of the intended dose distribution may take some time, as is the case with brachytherapy. It is important that the dosimeter remain uniformly sensitive, and unaffected in response over the time required for irradiation. Further, the dosimeter must maintain the dose-deposition information throughout the time required for analysis. For some applications, it may be desirable to transport the dosimeter to another facility for analysis. The dosimeter must remain stable during shipment, unaffected by a variety of environmental conditions, throughout the analysis.

The accuracy and precision required of dosimeters for radiation therapy measurements depend on the intended use of the detector. Devices intended for reference calibration of treatment units should enable the determination of dose with an uncertainty of no more than 0.5%, expressed at the 95% confidence level ($k = 2$) (28,29). Dosimeters intended for verification of dose distributions should provide an uncertainty in dose measurement of no more than 2%, again expressed at the 95% confidence level ($k = 2$).

DETECTORS FOR THREE-DIMENSIONAL DOSIMETRY

Detector Arrays

A number of manufacturers have marketed arrays of conventional detectors using either ion chambers or diodes. These devices are not true three-dimensional (3D) dosimeters, but are included here because they provide 3D information through the use of one, or at most two, manipulations, such as translation across a beam. For example, linear diode arrays are available for the Scanditronix water phantom system, and for stand-alone QA device, such as the Sun Nuclear profiler. An array of ionization chambers has been described for verifying treatment planning for IMRT (30). The ion chambers are arranged in several parallel linear arrays, each one offset from the next. Twenty-four chambers, each 0.03 cm^3 in volume, are arranged in boreholes of a plastic-mounting frame. The assembly is positioned in a water phantom and maybe positioned in different orientations to allow measurements in different plains. Commercial ion chamber devices include the Thebes marketed by Nuclear Associates, an ion chamber array marketed by Wellhofer, the RBA-5 marketed by Gammex, and other devices. These devices range in number of detectors from few (four or five) to many (Fig. 1).

Plastic Scintillator

Some organic plastics fluoresce visible light when irradiated with ionizing radiation. Unlike the fluorescent screens used in imaging, organic scintillators have the

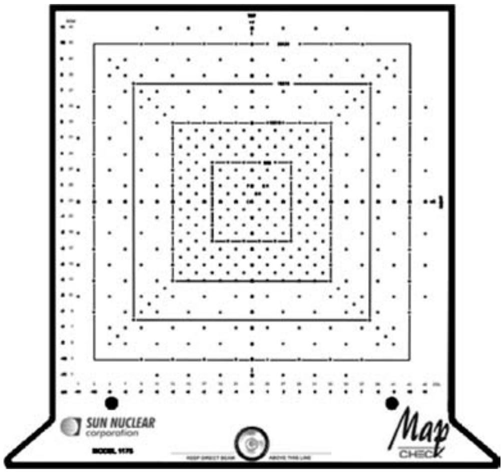


Figure 1. A diode array designed to display the intensity map of a therapy beam in real-time. (Courtesy of Sun Nuclear Corp.)

additional advantage of being approximately tissue-equivalent (31). However, this tissue-equivalence at present only exists at the energies conventionally used for megavoltage treatment. Most of the plastic scintillators currently available exhibit significant differences in the mass energy absorption coefficient relative to that of water. More recently, plastic scintillators have been developed for low energy photon dosimetry that are radiologically water-equivalent, have improved sensitivity over some others scintillators, and offer the potential for high spatial resolution (32,33).

Plastic scintillators may be used as point detectors, in which their potential for manufacturing into very small sizes yields the possibility for improved spatial resolution of measurements. Efforts also have been made to use plastic scintillators as 2D and 3D detectors (34). Two techniques have been used; the first being the use of plastic scintillators as a detector system themselves, using optical coupling through a light pipe assembly to a video detector. This method has been described previously (35–37). Significant difficulties still exist with the spatial resolution of these systems. Light emitted by the scintillator can travel some distance, in any direction, before reaching the light detector. Unless the plastic scintillator is thin, the resolution of the image will be degraded considerably. Some efforts have been to quench the light by adding dyes to the scintillator, to reduce the distance traveled by the light obliquely through the scintillator. Until this problem is resolved, the quality of the imaged dose distribution will not be adequate for radiation dosimetry.

A second technique involves the use of plastic scintillators to enhance the response of another detector, such as radiographic film (38). In this technique, radiographic film is sandwiched between sheets of organic plastic scintillator. Several investigators have noted that radiographic film has a tendency to overrespond to low energy photons (39,40). The use of an organic plastic scintillator has been proposed to enhance the response of radiographic film to higher energy photons, thus making the energy response of the film detector system more uniform.

Film

Radiographic film has long been used as a radiation detector, and as a QA device. Again, film itself is not a 3D dosimeter, but stacks of film have been used to measure dose distributions in 3D. The difficulties with film are well known; energy dependence, requirements for processing, variations from one batch to the next, dose rate dependence, positional dependence, and other issues have been discussed by a number of investigators (41). More recently, use of radiochromic film has been proposed. Radiochromic film requires no processing, has very little energy dependence, no known dose rate dependence, and requires minimal special handling techniques (42,43). The linearity of response of a recently developed model of film is shown in Fig. 2.

The use of film for verification of conformal and IMRT dose distributions has been recommended. At least one manufacturer has marketed a phantom intended for use with IMRT (see Fig. 3, for an example of such a device).

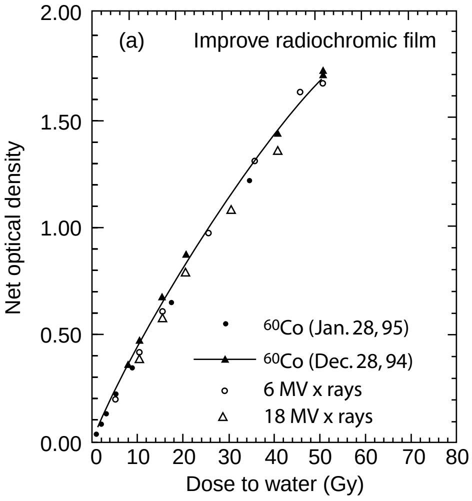


Figure 2. Energy dependence and linearity with dose of an improved radiochromic film. (Reproduced with permission from Ref. 42).

TLD Sheets and Plates

Lithium fluoride, a thermoluminescent material, has been used for many decades as a radiation detector (44). Its use has been limited principally to point measurements, because the dosimeter is provided either as extruded rods or chips, or as a powder that is encapsulated for use. Thermoluminescent dosimetry has a number of limitations, among them energy dependence, but most notably a requirement for delay between irradiation and processing. In addition, an expensive piece of equipment is required for readout of the material. The limitation of the device to point measurements has been addressed recently by the development of TLD sheets. In these, TL material is distributed in an array across a sheet of backing film. The film can be irradiated in much the same manner as conventional radiographic film, and may be immersed in a water phantom as necessary. As with film, 3D measurements can be made only by stacking multiple sheets of TL

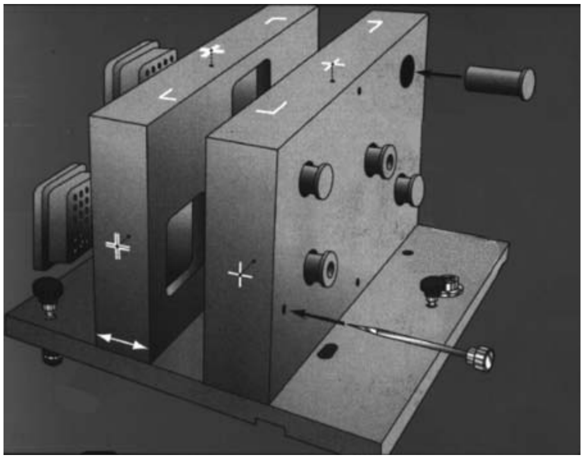


Figure 3. A phantom marketed for evaluating IMRT dose distributions. (Courtesy of Med-Tec Corp.)

material. After irradiation, and following the requisite delay, the film is inserted into a readout device that selectively heats the individual dosimeter regions using a laser. Light is collected from the heated regions using a photomultiplier tube. Through an automated operation, a matrix of data can be obtained quickly and efficiently. However, due to the cost of the reader, this dosimetry system is presently available only as a service (Inovision, Inc.)

Electronic Portal Imaging Devices

An important aspect of quality assurance in radiation therapy involves not just the dose delivered to the patient, but the correct positioning of the patient. For many years, positioning has been verified through the use of conventional radiographic film, or through the use of video imaging techniques (45,46). Video imaging permits only a check of the relative position of external landmarks. Radiographic film permits verification of the patient position through the visualization of internal boning anatomy, but requires a delay while the film is processed. The introduction of electronic portal imaging has brought to the clinic the possibility of immediate verification of patient position. With the introduction to clinical radiation therapy of modern techniques, such as IMRT, immediate verification of correct beam delivery is crucial. The failure or incorrect programming of a multileaf collimator can result in a completely unacceptable dose distribution. With on-line portal imaging, such errors may be detectable promptly, even during treatment (47–52). A further improvement has been the introduction of transmission flat-panel detectors up- and downstream from the patient. These allow the measurement of photon beam fluence entering and exiting the patient, and the estimation of dose within the patient. When combined with images of the patient made at multiple beam angles, as is done for multifield conformal treatment, or IMRT, it may be possible to reconstruct the dose distribution actually delivered to the patient in three dimensions.

GEL DOSIMETRY

Gel dosimetry has been examined as a clinical dosimeter since the 1950s (53,54). During the last two decades, however, the number of investigators has increased rapidly, and the body of knowledge regarding gel dosimetry has expanded considerably (55,56). Gel dosimetry is still considered by some to be a research project, and the introduction of this tool into clinical use is proceeding slowly. However, the interest in, and potential of, gel dosimetry for clinical use is demonstrated by the level of participation in three successful international workshops held to date on this subject (57–59). This section reviews the development of gel dosimetry, several of the formulations that have been investigated intensively, the characteristics of gel dosimetry that make it desirable for clinical use, the postulated and demonstrated applications of gel dosimetry, and some complications, setbacks, and failures that have contributed to the slow introduction into routine clinical use.

Fricke Gels

Nuclear magnetic resonance (NMR)-based gel dosimetry was introduced by Gore who recognized that the ferrous sulfate Fricke dosimeter (60,61) could be examined with magnetic resonance rather than spectrophotometry (55). The Fricke dosimeter is based on the radiation-induced and dose dependent transformation of ferrous (Fe^{2+}) ions into ferric (Fe^{3+}) ions. These two ions have different electron paramagnetic spin states and different ionic radii (60,61). Gore realized that the NMR spin–lattice and spin–spin relaxation rates ($1/T_1$ and $1/T_2$, respectively) of the water protons in the Fricke dosimeter are dependent on the amount of ferric ion present in the solution and that, because changes in these parameters produce the contrast of MR images, radiation induced changes in the solution should be visible by MRI (55). Soon afterward, other researchers began investigating the use of Fricke solutions incorporated into gel matrices (Fricke gels) to provide spatial stability of the dosimeter (62–65). The most common matrices investigated were gelatin, agarose, and sephadex. Each of these systems had its advantages and limitations, but agarose was probably used more than any other detector system. While agarose dosimeters are more sensitive to dose than gelatin-based systems, they are more difficult to produce because they must be bubbled with oxygen to ensure a uniform dose response.

Fricke gel dosimeters have a number of advantages; principle among them is the well-described understanding of the radiation chemistry of this system. In addition, the basic and NMR processes leading to the dosimetry response are well understood (66,67). Fricke gel dosimeters are tissue equivalent over a large range of photon energies. Like other gel dosimeters, they are prepared in a liquid form so that phantoms containing heterogeneities or conforming to anthropomorphic geometries can be constructed.

However, there are a number of significant problems associated with the use of Fricke gels for radiation dosimetry. The dosimeters require high doses, on the order of 10–40 Gy, for the radiation-induced changes to be observed by magnetic resonance imaging (MRI). The ferric ions produced by absorption of radiation diffuse readily through the gel or agarose matrix, leading to a decrease in signal intensity, and a loss of spatial information (64,66–69). Imaging must be performed within ~ 2 h of irradiation to avoid serious degradation of the dosimetric detail (70). The diffusion has been reduced by replacing the gelatin matrix with a poly (vinyl acohol) (PVA) matrix, which is less porous to the ferric ions (71). Other investigators have developed further methods to delay diffusion, although imaging must still be performed quite soon after irradiation (72). Some improvement in the diffusion of ions can be achieved by cooling the gel, but this is rarely practical in a clinical setting. Consequently, Fricke gel dosimetry has seen only limited clinical use.

Several improvements have been reported recently. For example, a Fricke gel dosimeter manufactured using a PVA cryogel technology has been described. The PVA is a common water-soluble polymer that can be cross-linked into its cryogel form by simply freezing and thawing. The cryogel is a rubber like material that holds its shape even at elevated temperatures. Preliminary reports of the PVA

Fricke gel dosimeter indicate that its ($1/T_1$) response has been found to be linear from 0 to 10 Gy, and the ion diffusion constant was found to be only 0.2–0.5 that of traditional preparations in gelatin or agarose (73,74). Representative ion diffusion constants are presented in Table 1 for several gel mixtures (68).

Some preliminary work using Fricke gel dosimetry in anthropomorphic phantoms has been reported (79). Several different gel compositions were investigated, including a lung equivalent gel that was developed with a density of 0.4 $g\cdot cm^{-3}$. This allows measurements of dose within the heterogeneity itself. However, diffusion of ions continues to be a problem with this dosimetry system.

Polymer Gels

Gels that replaced the Fricke solution with acrylic monomers were introduced in 1992 (80–82). Early work was conducted using a polyacrylamide gel based on the radiation-induced polymerization and cross-linking of bis and acrylamide. The formation of acrylic polymer chains largely resolved the problem of diffusion exhibited by Fricke gels, as the long polymer chains were too large to diffuse rapidly. The reciprocal of T_2, or R_2, the relaxation rate, was found to vary proportionally with dose, and MR imaging of polymer gels was shown to yield quantitative dose distributions (81). Subsequently, alternative gel formulations have been developed in which the bis and acrylamide are replaced with acrylic acid or methacrylic acid, which has yielded increased sensitivity of the gels, and reduced toxicity (83,84). However, the polymer gels continued to show another disadvantage; their response was inhibited by the presence of oxygen. This effect was addressed though the recent introduction of a class of polymer gel dosimeters containing oxygen scavengers (85,86). Several variations of these *normoxic* gel dosimeters (so-called because they can be prepared under normoxic conditions) have been characterized (87).

To avoid the disadvantages of the Fricke gel systems, a polymerizing gel dosimetry system was developed (MGS Research, Inc., Guilford, CT). A variety of polymerizing gels have been developed, many of which are based on acrylamide or acrylic acid, and are referred to as polyacrylamide gels (PAG). The dosimeters are based on

Table 1. Summary of Diffusion Measurements in the Literature[a]

Reference	Diffusion Coefficient, 10^{-3} $cm^2\cdot h^{-1}$	Gel Type and Concentration, %	Other Constituents, m*M*	Temperature, °C
(64)	18.3 ± 1.4	A 1	S 12.5, Fe^{3+} 1	
(64)	15.8 ± 1.1	A 1	S 25, Fe^{3+} 1	
(66)	19.1 ± 1.0	A 1.5	S 50, Fe^{2+} 1	25
(75)	10.9 ± 1.6[b]	A 1	S 50, Fe^{2+} 1, NaCl 1	15–17.5
(68)	9.7 ± 1.1	A 1	S 30, Fe^{2+} 1	22
(68)	11.9 ± 1.8	A 1	S 30, Fe^{2+} 1	22
(69)	12.5 ± 1.1	Agar	S 50, Fe^{2+} 1, NaCl 1	5
(69)	21.3 + 0.5	Agar	S 50, Fe^{2+} 1, NaCl 1	24
(76)	8.2 ± 0.1	G 4	S 26, Fe^{2+} 0.2, BE 5	10
(76)	9.1 ± 0.1	G 4	S 26, Fe^{2+} 0.2, BE 5, Fo 70	20
(76)	10.4 ± 0.1	G 4	S 26, Fe^{2+} 0.2, BE 5, P 0.6	10
(76)	4.4 ± 0.1	G 4	S 26, Fe^{2+} 0.2, BE 5, P 0.6	10
(76)	0.7 ± 0.1	G 8	S 26, Fe^{2+} 0.2, BE 5, Fo 46	20
(76)	1.0 ± 0.1	G 8	S 26, Fe^{2+} 0.2, BE 5, Fo 46, P 0.6	20
(76)	4.4 ± 0.1	G 4	S 26, Fe^{2+} 0.2, BE 5, XO 0.2	10
(76)	6.5 ± 0.1	G 4	S 26, Fe^{2+} 0.2, BE 5, BD 0.6	10
(76)	6.1 ± 0.1	G 4	S 26, Fe^{2+} 0.2, BE 5, Fo 46, XO 0.2	20
(76)	6.3 ± 0.1	G 4	S 26, Fe^{2+} 0.2, BE 5, AC 0.6	20
(76)	8.3 ± 0.1	G 4	S 26, Fe^{2+} 0.2, BE 5	10
(77)	14 ± 3	A 1.5	S 50, Fe^{2+} 0.5	22
(77)	20 ± 5	A 1.5	S 100, Fe^{2+} 0.5	22
(77)	22	A 1.5	S 200, Fe^{2+} 0.5	22
(77)	11	A 1.5	S 50, XO 0.25	22
(77)	5 ± 1	G 10	S 50 and 100, Fe^{2+} 0.5	22
(77)	9	A 1.5, G 3	S 50, Fe^{2+} 0.5	22
(77)	9	A 1, G 2	S 200, Fe^{2+} 0.5, XO 0.2	22
(77)	3 ± 1	A 1.5, G 3	S 50 and 100, Fe^{2+} 0.5, XO 0.1 & 0.25	22
(78)	14.6 ± 0.1	G	S 50, Fe^{2+} 1.5, XO 1.5	
(78)	8.1 ± 0.1	G	S 50, Fe^{2+} 1.5, XO 1.5	
(78)	8.2 ± 0.1	G + BA	S 50, Fe^{2+} 1.5, XO 1.5, BE 5.0	
(78)	17.8 ± 0.2	A 1.5	S 50, Fe^{2+} 1.5, XO 1.5	
(78)	16.3 ± 0.2	A 3	S 50, Fe^{2+} 1.5, XO 1.5	
(71)	1.4	PVA 20	S 50, Fe^{2+} 0.4, XO 0.4	20

[a] A - agarose, Agar= agar, g = gelatin, S= H_2SO_4, XO = xylenol orange, BE = benzoic acid, Fo = formaldehyde, P = phenanthroline, AC = acetylacetone, BD = bathophenanthroline disulfonic acid.

[b] Diffusion coefficient calculated in Ref. 76.

Table 2. Composition of BANG3 Polymer Gel Dosimeter

6% Methacrylic acid
1% Sodium hydroxide
5% Gelatin
88% Water

radiation-induced chain polymerization of acrylic monomers dispersed in a tissue-equivalent gel. The BANG polymer gel system is a proprietary PAG dosimeter made of a mixture of acrylic monomers in a tissue-equivalent gel. Early BANG gels were made from acrylic acid monomers and methylene-bis(acrylamide) cross-linker. More recently, the BANG3 dosimeter was introduce, which contains methacrylic acid monomer (see Table 2, from Ref. 84). Other proprietary response modifiers were added to adjust the dose range and sensitivity. Dissolved oxygen inhibits free radical polymerization reactions and is removed from the mixture by passing nitrogen through it while the gel remains above the gelling temperature, prior to sealing the vessel. Consequently, vessels of glass or other material not permeable to oxygen must be used for irradiating and imaging the gels.

The gelling agent in the BANG dosimeter is gelatin, which is used because the transverse NMR relaxation rate of water ($R_2 = 1/T_2$) in a gelatin gel is nearly an order of magnitude lower than that in agarose gels. Therefore the background R_2 in the gel is substantially reduced, which improves its dynamic range.

MR Imaging of Polymer Gels

Irradiation of the polymer gels induces polymerization and cross-linking of the acrylic monomers. As polymer microparticles are formed, they reduce the NMR relaxation times of neighboring water protons. Magnetic resonance imaging can be used to measure dose distributions in the gel (81,82,88). Water proton NMR (^{1}H NMR) transverse relaxation time T_2 can be τ determined from multiple spin–echo images. Images can be acquired using the Hahn spin–echo pulse sequence: $90^\circ\ -\tau-\ 180^\circ\ -\tau-$ *acquire* for four or more different values of τ. Typical pulse sequence parameters are TR = 2 s, TE = 11, 200, 400, and 600 ms. A field of view of 24 cm and a matrix of 128 × 256 can be used, with one acquisition and a 3 mm slice thickness.

More recently, it has been shown that spin–echo sequences other than the Hahn sequence described above can be used for gel imaging. Improved dose resolution can be achieved through the use of multiple spin–echo pulse sequences (89,90). Optimization of the imaging sequence is necessary, especially with regard to the number of echoes measured. The use of multiplanar imaging can reduce imaging times but can also lead to interference between image planes.

Once MR images have been obtained, they are most conveniently transferred via network to a computer for which a data analysis and display program has been written. One example of such a program has been described previously (82). The program calculates R_2 maps on the basis of multiple TE images, using a monoexponential nonlinear least-squares fit based on the Levenberg–Marquardt algorithm (91). The program also creates a dose-to-R_2 calibration function by fitting a polynomial to a set of dose and R_2 data points, obtained from gels irradiated in test tubes to known doses. This function can then be applied to any other R_2 map so that a dose map can be computed and displayed.

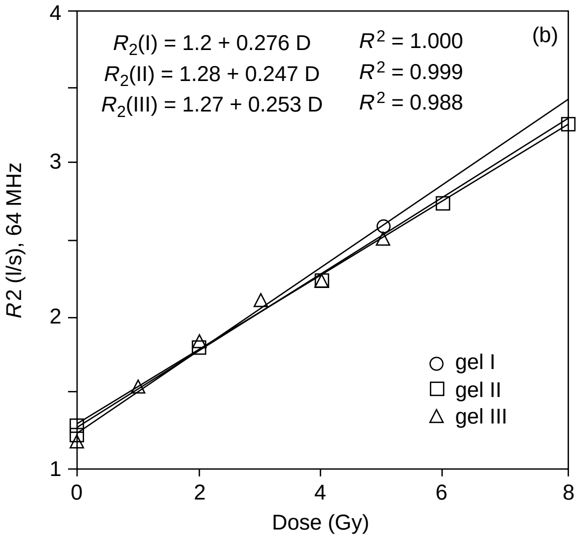


Figure 4. The dose dependence of the transverse relaxation rate (R_2) as a function of dose. Data from several experiments are shown indicating reproducibility over a wide range of doses (92).

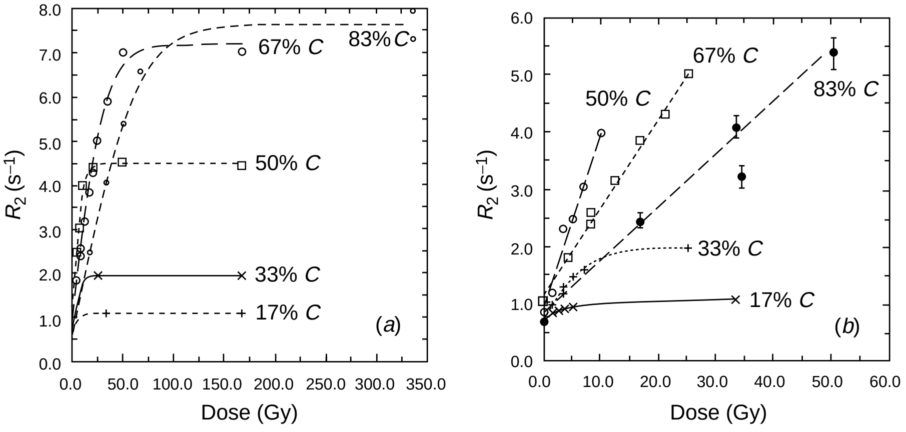
Figure 4 shows values of transverse relaxation rates (R_2) for the gels as a function of dose. The pooled data show that the dose response was highly reproducible over a wide range of doses. The dose response is well fitted by a straight line (92).

Additional experiments have shown that the response of the BANG gel can be adjusted by varying the concentration of cross-linker used per total amount of comonomer (93). Figure 5 demonstrates the relationship of R_2 to dose for five different values of the weight fraction of cross-linker per total comonomer. Figure 5b shows that, in the linear region of gel response, the greatest sensitivity of the gel was achieved at 50% cross-linker concentration. Similar data have been shown more recently for several different polymer gel mixtures (94).

The temperature of imaging has a large effect on both the gel sensitivity and its dynamic range (93,95). Dose sensitivity ($s^{-1}{\cdot}Gy^{-1}$) as a function of concentration of cross-linker is plotted in Fig. 6. Sensitivity is seen to reach a maximum at ~ 50% cross-linker (as described above), but sensitivity at all concentrations increases as the temperature at the time of imaging is reduced.

Figure 7 shows that the maximum R_2 achievable, and therefore the dynamic range of the gel, is dependent on the temperature at the time of imaging. While $R_2^{\max}$ increases with cross-linking, the dependence is enhanced by cooling the gel during NMR measurement.

For a number of gel compositions presently being evaluated, the fundamental chemistry and physics of response are well understood. Several gel compositions have been characterized in great detail (82,87,92,96–98). In polymer gels, for example, it is understood that the interaction of



(a) (b)

Figure 5. (a): Relationship between R_2 and dose for five different concentrations of cross-linker per total comonomer. (b) Lower dose region of the data from (a). (Reproduced with permission from Ref. 93.)

radiation with water produces free radicals, which trigger the cross-linking of monomers into polymer chains (81,99). The polymer chains bind water protons tightly causing a change in their paramagnetic properties that is detectable by magnetic resonance imaging (92,100). The relationship between dose and relaxation rate can be influenced by several additional factors, including accuracy of the calibration curve (101) and the aging characteristics of the gel (96,102,103).

The quality of the imaging process is affected by the homogeneity of the B_1 field (104) and the presence of eddy currents (105). Some additional complications due to the distortion of MR imaging systems have been identified (106).

Optical Scanning of Polymer Gels

Dosimetric results with MR imaging have been encouraging, but the need to use expensive and often inaccessible imaging systems renders this technique somewhat impractical. In most compositions, polymerization changes the optical characteristics, and measurements of optical density can be related to absorbed dose (85,107–111).

Optical computed tomography (OCT) of polymer gels can be conducted in a similar manner to X-ray CT. To date, OCT has been limited to transmission measurements, although the potential exists for measurements of attenuation, fluorescence, scatter, polarization and refractive index changes (112). Optical computed tomography has been performed by several investigators (107,109–113), but in general, the techniques all require the use of a cylindrical vessel to hold the gel, a tank filled with a medium matching the refractive index of the gel, and a monochromatic light source. Several of these systems use parallel-ray geometry and filtered back projection to reconstruct the image. At least one system uses a diffuse white light and cone-beam geometry (111).

An optical imaging system employing He–Ne laser CT scanning of the gel has been described (107). The scanner operates in a translate-rotate geometry and is capable of producing stacks of planar dose distributions with pixel size and slice thickness as small as 100 μm (114).

Optical scanning of several gels has been conducted using a modified version of a 3D optical CT laser scanner that was developed recently at MGS Research, Inc. (107,108,115,116). (see Fig. 8.)

The scanner, which is PC controlled, operates in a translate-rotate geometry and utilizes a single He–Ne laser

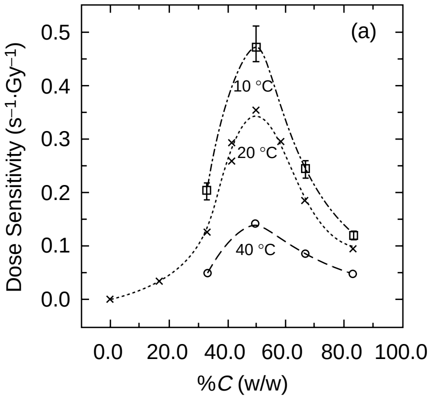


Figure 6. Dependence of the dose sensitivity on the cross-linker content, for three different temperatures. (Reproduced with permission from Ref. 93.)

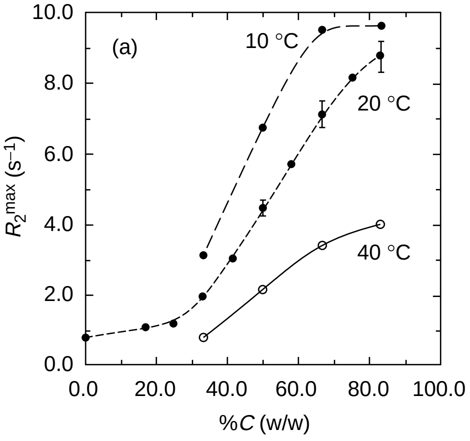


Figure 7. Dependence of the maximum R_2 on cross-linker fraction for three different temperatures. (Reproduced with permission from Ref. 93.)

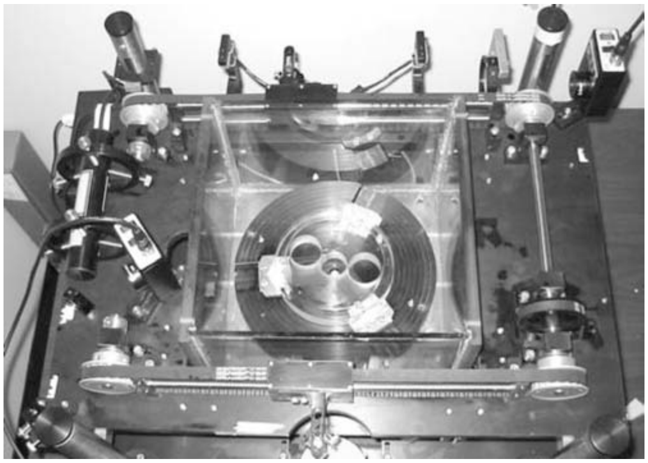


Figure 8. An optical scanner developed for use with polymer gels. (Photograph by M. Heard.)

light source and a photodiode, together with a reference photodiode to account for fluctuations of the laser output intensity. The gel is mounted on a central turntable and is immersed in a liquid that matches the gel's refractive index to minimize the loss of signal from projections at the edges of the gel. The platform on which the light source and the detector are mounted moves vertically. Isotropic resolution of 1 mm is achievable using this scanner, with scan times on the order of 8 min per plane. An image of a gel exposed to an ^{192}Ir high dose-rate (HDR) source appears in Fig. 9.

Further evaluation of an OCT system has been performed, to determine the stability and reproducibility of the system (118). In addition, characterization of gels has been performed to determine the optimum sensitivity consistent with the dynamic range of the scanner (119).

X-Ray CT Scanning of Polymer Gels

The formation of polymer chains increases the physical density of the gel, and the resulting change in attenuation coefficient can be measured by measurements of X-ray transmission, such as by computed tomography (120–125). While the change in density is small, it has been shown to vary proportionally with dose (122,126). This change in density leads to a small change in CT number when irradiated gels are examined with CT. Recent data show that this change can be as much as 0.2 kg·m^3·Gy^{-1} (127). An image of tubes of gel irradiated to different doses appears in Fig. 10. Methods for improving the quality of X-ray imaging have been developed, and include the acquisition of multiple images, background subtraction, and filtering (126,127).

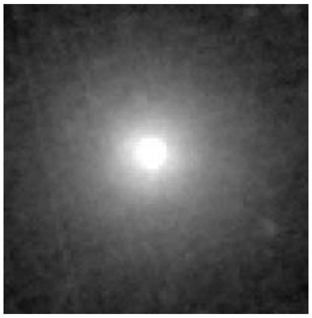


Figure 9. The OCT image of a polymer gel exposed to an ^{192}Ir HDR brachytherapy source. The central region was occupied by the source during irradiation and was replaced with irradiated gel for imaging (117).

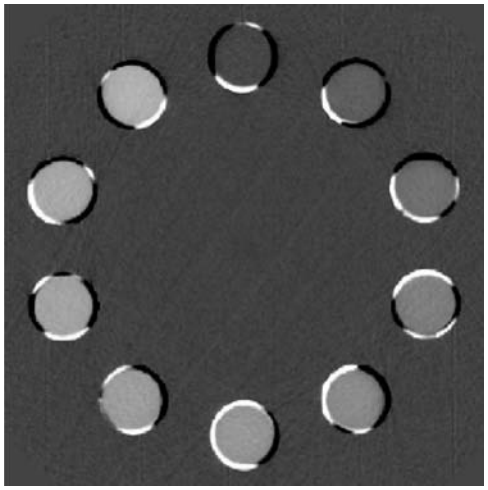


Figure 10. The CT image of several vials of polymer gel irradiated to different doses. (Reproduced with permission from Ref. 128.)

Ultrasound Imaging of Polymer Gels

Polymerization leads to changes in elasticity of the medium, and the corresponding changes in ultrasound absorption can be exploited (129–132). Ultrasound has been used to evaluate changes in density and elastic constant of a number of materials. Several different ultrasonic parameters can be measured and these can be used to characterize materials. The most commonly measured parameters attenuation and reflection coefficients, and the speed of propagation. A pulse-echo technique using one probe or a transmission technique using two probes is used to measure these parameters. These parameters can be related to structural properties of the sample including bulk density, elastic constants as well as sample inhomogeneities.

Vibrational Spectroscopic Imaging of Polymer Gels

Finally, vibrational spectroscopy can be used to demonstrate the conversion of monomers to polymer chains (133–136). Fourier transform (FT)–Raman vibration spectroscopy of polymer gel dosimeters has been investigated as a means by which the fundamental structure and properties of the dosimeters might be better understood. Raman spectroscopy has also been used to investigate the track structures of proton beams in polymer gel dosimeters (137). This study illustrated the difficulty in using polymer gel dosimeters to extract quantitative dose maps when exposed to proton radiation. Further studies will be required to determine whether Raman microscopy can be used routinely in the evaluation of polymer gel dosimeters.

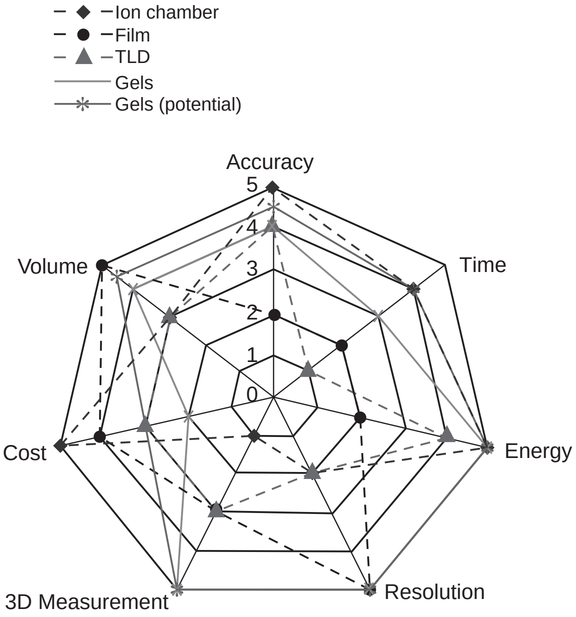


Figure 11. A spider plot, illustrating the capabilities of several common dosimetry systems, as well as gels, and the potential capabilities of gels. (Redrawn with permission from Ref. 110.)

CHARACTERISTICS OF GEL DOSIMETERS

Gel dosimeters have a number of characteristics that make them attractive for radiation dosimetry (138). A novel comparison of gel dosimeters with conventional dosimetry systems has been presented in the form of a spider plot (see Fig. 11, Ref. 110). This graphical presentation illustrates the relative performance of dosimeters, such as ion chambers, film, TLDs and gels by considering such parameters as accuracy, volume measured, cost, three-dimensionality, resolution, energy dependence, and time required for the measurement. Oldham has shown that gels compare favorably with the other detectors in most characteristics, including their relative accuracy, volumetric nature, inherent three-dimensionality, high resolution and lack of energy dependence over much of the important energy range (110). Methods for characterizing the response of gels have been found, and in particular, a technique for characterizing the dose resolution has been described (89,139,140).

However, today gels are still time-consuming and relatively expensive. Several dosimetric aspects have not yet been realized, including the absolute accuracy of measurement, and the ability to render a 3D dose distribution as opposed to multiple planes of data, although progress is being made rapidly on both aspects. In addition, the issues of cost and time required are being addressed. The availability of optical CT scanning and other imaging techniques are likely to drive down the cost of gel analysis, and improve the penetration of this modality into the clinic. At the same time, newer optical CT scanners equipped with more powerful computers are faster and can perform comprehensive imaging of gels in the time previously required for a single slice.

APPLICATIONS OF GEL DOSIMETRY

Potential applications of gel dosimetry have been summarized on several occasions in the recent past (97,138,141–143) although the field is developing rapidly. Today it is considered by many that gel dosimetry has useful characteristics that can facilitate radiation therapy dosimetry, especially in situations that are not handled well by conventional dosimeters. These characteristics include the ability to measure complex 3D dose distributions; to integrate dose accurately without dependence on dose rate, at least over a fairly wide range; tissue-equivalence; high spatial resolution; and lack of energy dependence over most of the kilovoltage and megavoltage range. With most gels, data are stored permanently, making gels suitable for performance of dosimetry at remote locations (144). They also are relatively safe to manufacture and handle, although some components such as acrylamide are toxic and must be handled with appropriate protection until mixed.

Demonstrated applications of gel dosimetry to date include basic dosimetry (depth dose, penumbra, wedge profiles) in photon, electron, neutron, and proton beams; dose distributions from imaging procedures; conformal therapy, stereotactic radiosurgery, and intensity-modulated radiation therapy (IMRT); dose distributions around brachytherapy sources (low and high dose rate, and intravascular sources); internal dosimetry (^{131}iodine doses); and evaluation of tissue heterogeneities. The advances made recently in these areas will be discussed.

Basic Dosimetry

Gel dosimeters have the capability to record and display the dose distribution throughout a 3D volume. This ability affords advantages over conventional dosimeters, even for basic dosimetry parameters such as percent depth dose in photon and electron beams (54,92,145). Gel dosimetry has been shown to be useful to validate simple multiple-field arrangements (146) and more complex anatomical situations including tangential breast treatment (147,148), conformal therapy (149) and scalp treatment with electron beams (150). Dynamic functions, such as a programmable wedge filter are difficult to measure with ionization chambers or diodes, and film is often used to provide data in a single plane. Gels have proven useful for capturing the dose distributions from programmable wedge filters, and allow distributions in multiple planes to be demonstrated from a single exposure (151).

Dose from Imaging Procedures

More recently, the use of gels to demonstrate dose distributions from imaging procedures has been explored (152,153). In a novel experiment, a high sensitivity gel was used to determine the dose from CT imaging. The benefit of this measurement is that the dose distribution throughout a patient volume can be estimated without requiring the use of numerous point dosimeters (e.g., TLD) and without averaging the dose along a line or throughout a volume (e.g., a pencil ionization chamber). These benefits may be most apparent in evaluating the dose distribution from helical CT scanners.

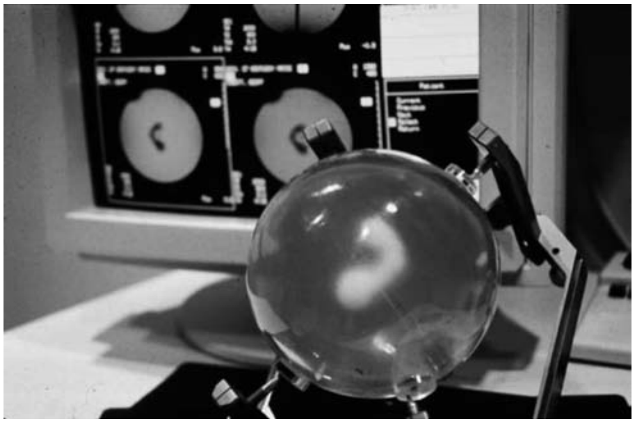


Figure 12. A BANG gel irradiated with a highly conformal dose distribution produced by a Gammaknife treatment unit. The distribution can be appreciated qualitatively without the need of imaging systems or processing. (Photograph by the author. See also Ref. 161.)

Evaluation of Conformal Dose Distributions

Stereotactic Radiosurgery. Gels have been used to demonstrate the dose distributions from stereotactic treatments both from dedicated multisource cobalt units and from linear accelerators (154–161). A clear benefit of gel dosimeters is that they can display a dose distribution, especially a highly conformal one as is produced by stereotactic radiosurgery techniques, so that it can be appreciated qualitatively in three dimensions without need of imaging systems or processing (see Fig. 12, Ref. 161).

In one series of measurements, gels were prepared in glass flasks chosen for their size and shape, which was comparable to that of a human head. Additional polymer gel material from the same batch was prepared in glass test tubes, for irradiation to selected doses, to generate a dose-response curve. The gels were prepared in Guilford, CT, and were shipped to Lexington, Kentucky for irradiation and analysis (161).

A gel prepared in a 16 cm diameter flask was fitted with a radiosurgical head frame (Leksell, Elekta Corporation, Atlanta, GA), as shown in Fig. 13. This flask was also equipped with a glass rod extending to near the center of the flask, to be used as a target. The MR images were obtained and were transferred to a Gammaknife treatment planning computer (Gammaplan; Elekta Corporation), where a complicated dose distribution was planned using multiple target points. Once the plan was completed, the coordinates of the individual target points were determined, and the gel was moved to the Gammaknife irradiation unit. Treatments were delivered to each of the target points, in accordance with the treatment plan. A dose of 10 Gy was delivered to the 100% isodose line.

Dosimetric imaging of the flask and test tubes containing gel was performed between 25 and 36 h after irradiation. The flask was placed in the head coil of the imager and the test tubes irradiated for calibration purposes were placed around the flask. The images were transferred via network to a Macintosh computer, and the DoseMap program was used to compute the maps of transverse relaxation rate (R_2).

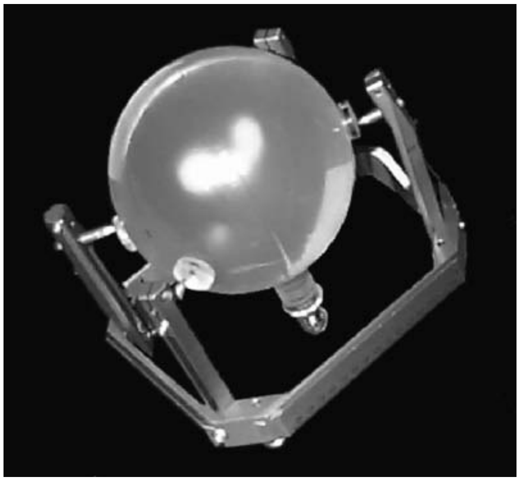


Figure 13. Photograph of a glass flask filled with the BANG Polymer Gel dosimeter. A glass rod was inserted into the gel to provide a target around which to localize the dose distribution. The flask was fitted with a Leksell stereotactic head frame. The gel is shown as it appeared following irradiation. (Photograph by the author).

A dose-response calibration curve was obtained as described earlier. Images of the gel-filled test tubes were obtained, and R_2 determined as a function of dose.

The calibration curve was then applied to R_2 maps of the flask irradiated with the Gammaknife unit. The result yielded an image of the dose distribution, as shown in Fig. 14a and 14b. As all scans were performed with the head ring and localizer box in place, the coordinates of the image plane could be determined. These image planes were located 1 mm from each of the corresponding treatment plans shown in Fig. 14a and 14b. Finally, isodose curves were drawn (by the DoseMap program) by interpolating within the measured dose distribution.

The measured dose distributions were compared with the treatment plans prepared prior to irradiation by superimposing the two data sets. The superimposed data are shown in Fig. 15a and 15b. The calculated and measured dose distributions were registered by aligning the point representing the tip of the glass rod.

The measured dose distributions compare favorably with the calculated dose distributions. In fact, the dose

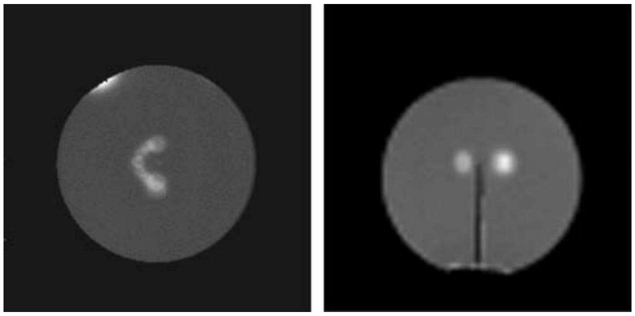


Figure 14. The R_2 maps obtained from the irradiated gel (a) Distribution in the axial plane. (b) Distribution in the sagittal plane. (Reproduced with permission from Ref. 161.)

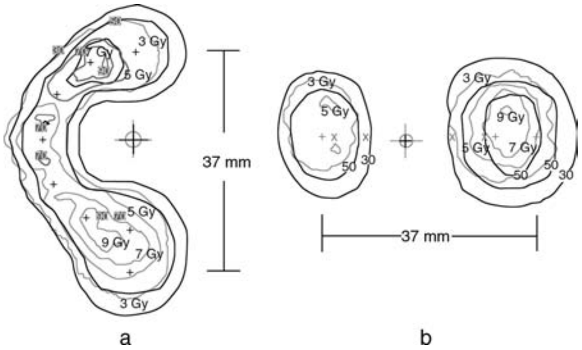


Figure 15. Composite figures showing both the treatment plan prepared using a Gammaplan treatment planning computer (drawn in black, labeled in percent of maximum dose) and isodose curves measured by the technique described in the text (drawn in gray, labeled in Gy). (a) The distribution in the axial plane containing the 8 isocenters. (b) The distribution in a perpendicular sagittal plane. (Reproduced with permission from Ref. 161.

map taken in the plane of the target points (Fig. 14a) indicates regions of overlap not demonstrated by the treatment planning system. As shown in Fig. 15a and 15b, the measured isodose lines conform in shape quite well with the calculated data, but seem to show a shift away from the glass target rod. The dose images were obtained in planes that were shifted 1 mm from the planes of dose calculation, and this shift might account partially for the difference in size and shape of the isodose curves. However, Fig. 15a shows a shift in the lateral (*X*) direction away from the glass target rod, which cannot be explained by a difference in the axial (*Z*) coordinates of the planes of calculation and measurement. Instead, it appears more likely that the dose distribution was placed $\sim$ 1 mm further from the glass target rod than intended.

Evaluation of Repeat-Fixation Stereotactic Radiotherapy. In recent years, fractionated stereotactic radiation therapy has been seen as a desirable method of delivering high dose radiation therapy to malignancies of the brain. Techniques developed for immobilizing the patient have also been applied more recently to intensity-modulated radiation therapy, in which conformal dose distributions are delivered through multiple fractions to one or more target volumes. In both techniques, reproducible positioning of the patient is critical, to ensure that the target volume receives the intended dose, and normal tissues are spared to the extent determined by treatment planning techniques. The BANG gel dosimeter has been used in a fractionated regimen to demonstrate the reproducibility of multiple setups under stereotactic position methods (158).

Intensity-Modulated Radiation Therapy (IMRT). Gels dosimeters have proven themselves to be valuable for evaluating and confirming IMRT dose distributions (146,162–169). Most investigations have been conducted in simple geometric phantoms (Fig. 16), but others have employed anthropomorphic phantoms in arrangements that allowed direct comparison with measurements using other techniques such as film and TLD (163,165,166).

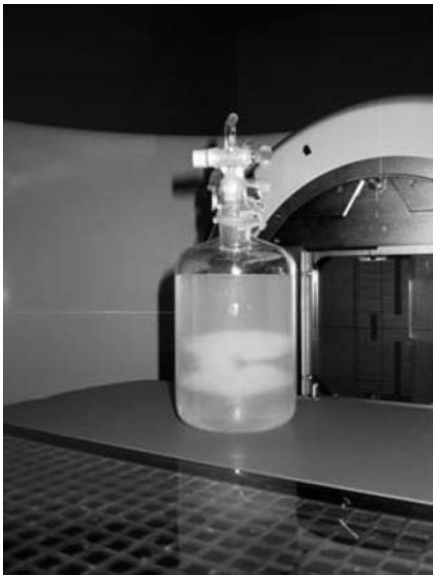


Figure 16. A cylindrical flask containing a normoxic gel shortly after irradiation with an IMRT treatment. The dose distribution is clearly visible, demonstrating the change in optical density with dose. (Reproduced from Ref. 167, with permission.)

Beach developed a gel insert for an existing anthropomorphic phantom that had been developed with film and TLD dosimeters (170). The phantom design revision included converting the existing imaging/dosimetry insert from a block-style design to a cylindrical design (Fig. 17). This insert contained embedded structures that simulated a primary and secondary target volume as well as an organ at risk (OAR). An additional insert was then constructed to house the polymer gel dosimeter. This insert was specially designed using Barex plastic. Both the imaging insert and the gel insert had an image registration system incorporated into their construction.

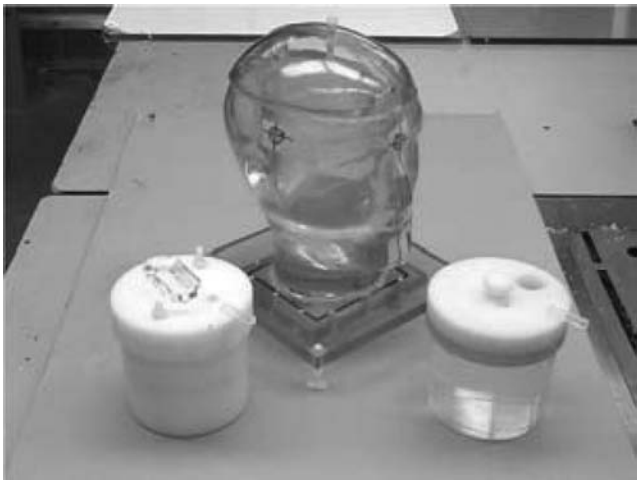


Figure 17. An anthropomorphic head-and-neck phantom developed by the Radiological Physics Center (170) showing the modifications made to accommodate a gel dosimeter.

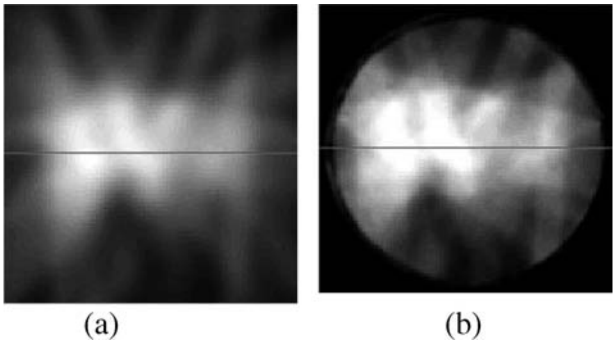


Figure 18. (a) A calculated dose distribution for an IMRT treatment, shown in a gray-scale format. (b) The measured dose distribution obtained from optical CT of a polymer gel, following irradiation with the treatment plan shown in (a). (From Ref. 165, with permission.)

X-ray CT images were obtained of the phantom with the imaging insert in place, and an IMRT treatment plan was developed. The phantom was then taken to the linear accelerator, the imaging insert was replaced with the gel insert, and the IMRT treatment was delivered.

A commercially available optical computed tomography (OCT) scanner (107) was commissioned for this project and future work with polymer gel dosimetry. The OCT scanner was used to image polymer gels before and after being irradiated. The preirradiation images were subtracted from the postirradiation images using a pixel-by-pixel subtraction method. The resultant images had net OD values that were directly proportional to the dose received by each given pixel. A comparison of the calculated dose distribution and the measured distribution is shown in Fig. 18.

Repeated measurements showed that a polymer gel imaged with optical CT was reproducible to within 1% (171). Repeated OCT imaging was shown to be consistent to within 1%. However, the results also showed that the techniques used to calibrate the gel (irradiation of a similar gel container with small-diameter beams delivering doses spanning the expected range) did not provide absolute dose measurements offering better agreement than 10% with the calculated data.

Duthoy compared the dose distribution measured with gels to the calculated distribution, for complex intensity-modulated arc therapy (IMAT) treatments in the abdomen (172). Vergote also examined IMAT with gels and observed a reproducible difference between calculations and measurements in low dose regions near steep dose gradients; a phenomenon also observed by Cadman et. al. and attributed to the failure of treatment planning systems to model the transmission of radiation through the rounded ends of multileaf collimator leaves (169,173).

Brachytherapy. Determining dose distributions and confirming the results of planning for brachytherapy treatment is historically difficult. No suitable methods of dosimetry have existed in the past to enable measurement and display of these 3D and complex distributions. Measurements around single sources have been possible only in a point-by-point fashion, such as with small ionization chambers or with thermoluminescence dosimeters (TLDs), (174) or in planar fashion with film (175). These methods are quite unsatisfactory for anything other than distributions around single sources, or very simple source arrangements. In contrast, the BANG polymer gel dosimetry system has the capability to measure and display complex dose distributions from complicated source arrangements. It is necessary to immerse the applicator containing the sources into the gel, or arrange for its introduction into a catheter already placed in the gel.

The ability of gels to record and display dose distributions around a high dose rate (HDR) source was first demonstrated over a decade ago (92,176,177). Maryanski et al. showed the dose distribution around a single catheter into which a high dose rate (HDR) remote afterloader source had been positioned (178). The HDR unit was programmed to dwell the source at several locations in the catheter, to deliver an elliptical dose distribution. After irradiation, the gel was imaged with MR, and a map of the dose distribution was computed. The map is shown in Fig. 19, where the color intensity is proportional to dose. Isodose lines, determined from the dose map data, are superimposed on the intensity map. Points at which the dose was computed by the treatment planning system also are shown. Excellent agreement between the position of the calculated dose points and the corresponding measured isodose lines indicates the agreement between doses measured by the gel and computed by the treatment planning system.

More recently, measurements have been made in close proximity to HDR ^{192}Ir sources (117,179) (see also Fig. 9). These measurements have shown that complications occur when measurements are made in the steep dose gradients close to an HDR source. Polymerization of the gel causes an increase in the gel density and a corresponding decrease in the volume filled by the gel. The change in density causes shrinkage of the gel in the vicinity of the source, distorting the resulting measured distribution. Changes to the composition of the gel to increase the concentration of gelatin

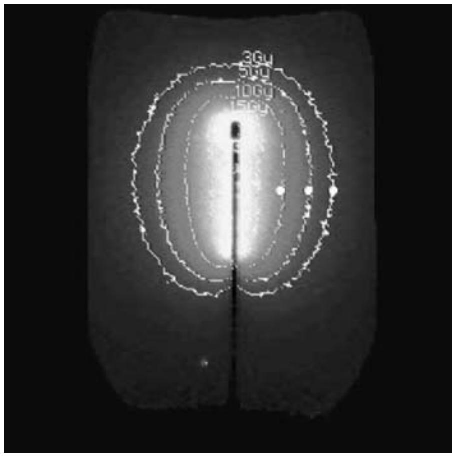


Figure 19. Use of the BANG gel to measure the dose distribution around an HDR source. The source was positioned in a catheter implanted in a BANG polymer gel. The figure illustrates a comparison between the dose distribution determined from a MRI image of the gel and the calculated dose distribution. (From Ref. 178.)

can mitigate the amount and effects of the density changes. Furthermore, there are suggestions that the high dose rates found near brachytherapy sources, particularly those of HDR afterloaders, can introduce temperature gradients that influences the polymerization of acrylamide monomer gels (87,93,180,181).

Efforts also have been made to characterize low dose rate (LDR) sources, such as prostate seeds (182–184), eye plaques (185), ^{137}Cs afterloading sources (186,187) and intravascular sources (188). Studies have indicated that the diffusion of monomers (or ferrous and ferric ions in Fricke gels) across steep dose gradients can introduce errors in measurement (92,189). As the use of gels to measure dose distributions from LDR sources requires long exposure times, diffusion of monomers or ions could introduce significant errors, and gels exhibiting high diffusion rates should be avoided for these measurements.

A further problem with gel dosimetry for LDR brachytherapy has been demonstrated by recent studies indicating energy dependence at lower energies. Data show that a polymer gel dosimeter under responds to radiation in the 20–60 keV range (190). Others have shown differences in gel response from one formulation to another, and suggest that the MAGAT gel is most water-equivalent over a wide range of energies (191). Changes in mass attenuation coefficient of polymer gels during irradiation can also introduce errors in the dose distributions measured around low energy sources.

Internal Dosimetry

Gel dosimetry has shown promise in the determination of dose distributions from administrations of unsealed radioactive sources (192). The authors embedded a vial of ^{131}I into a flask of polymer gel and observed a distance-dependent change in the T_2 signal. They also mixed ^{131}I into the gel and demonstrated a change in T_2 signal that was dependent on distance from the concentration of activity. No more recent investigations have been located.

Measurement of Neutron Dose Distributions

Some developments have been reported in characterizing fast and epithermal neutron beams with gel dosimetry (193–195). Thin layers of Fricke-xylenol orange gels have been irradiated in phantoms composed of insensitive gel. Adding ^{10}B or other nuclides with large cross-sections has increased the sensitivity of the gel dosimeter to neutrons. This technique has been used to determine the profiles of neutron beams used for boron neutron capture therapy. Some benefits of the use of gel dosimetry are the tissue-equivalence of the dosimeter to these energies, and the ability to separate the components of dose.

Measurement of Particle Dose Distributions

Several investigators have demonstrated the use of polymer gel dosimeters to record the dose distributions produced by proton beams (88,137,196–198). However, several authors have noticed disagreements between measurements with gels and conventional dosimeters such as diodes in the peak region of the distribution. Gustavsson has suggested that the response of gels, as they are based on the formation of free radicals, is dependent on the LET of the radiation (197,198). As the LET of the beam increases in the peak region, the local ionization density increases. As the distance between the radicals formed in the gel decreases, the likelihood of recombination of radicals increases. A decrease in the production of radicals with increasing LET has been described previously (199). Consequently, significant differences appear between depth dose measurements with gels and those with detectors such as diodes (see Fig. 20, Ref. 198).

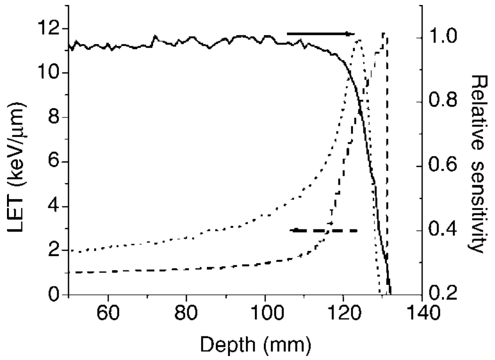


Figure 20. The variation in LET as a function of depth for a monoenergetic proton beam (dashed curve, left-hand scale) and the measured relative sensitivity for the gel dosimeter (full curve, right-hand scale). Also shown is the depth dose curve for the proton beam (dotted curve), normalized to 100% at the Bragg peak. (Reproduced with permission from Ref. 198.

Jirasek et al. performed track energy-deposition calculations and raman spectroscopy and reported agreement between these techniques and gel measurements (137). Their conclusion also was that the high density of delta-ray interactions close to the track of a proton resulted in high doses being delivered to the gel. These doses saturated the response of the gel by consuming the available monomer. This effect was greater near the end of the proton range, consistent with the results of other authors.

Gels have been used also to demonstrate the dose distribution produced by ^{12}C ions (200). Similar effects associated with decreased radical formation at high LET were observed in the carbon beam.

Evaluation of Tissue Heterogeneities

A valuable feature of gel dosimeters is that they are very nearly tissue-equivalent, particularly at photon beam energies above ~ 100 kV. Previous investigations have shown that the BANG gel, the MAGIC and MAGAS normoxic gels, as well as gels based on Fricke or vinyl solutions have electron densities within 1% of soft tissue, and effective atomic numbers in the range of 7.4 (190). However, several investigators have attempted to measure the effects of nonunit density tissues on external beam dose distributions. Early measurements were performed to estimate the dose distribution behind high atomic number heterogeneities, to simulate the presence of bone (201–204). More recently, measurements have been made behind or adjacent to cavities filled with air or with lung-equivalent plastic (168). To attempt a measurement in lung-equivalent gel, Olberg produced a foam of gel with the approximate density of lung tissue (205). Other investigators have evaluated the promise of gel dosimeters to simulate lung tissue, by introducing polystyrene foam beads into a gel mixture (206). While these measurements showed promise, there were several sources of error. First, the introduction of air, or air-containing polystyrene beads introduced the possibility of oxygen contamination. Purging the polystyrene beads with nitrogen, or using nitrogen rather than air to foam the gel addressed this problem. The introduction of air or polystyrene eliminated the possibility of evaluating the measured dose distribution by optical scanning, and MR imaging must be used. The presence of air may lead to partial volume imaging effects that could introduce errors into the measurement.

COMPLICATIONS TO BE ADDRESSED

As was suggested earlier in this article, there are a number of complications associated with gel dosimetry that remain to be addressed, and that are inhibiting the routine use of gels in the clinic. Some of these are listed below, with short descriptions of the causes of the problems, and possibilities for correcting them.

Imaging Artifacts

This article has discussed several methods of generating images of dose distributions using gels. Principal methods are MRI, OCT, and X-ray CT. Each of these imaging methods is prone to imaging artifacts, although the type of artifact and its causes are different with the different modalities. In MRI, for example, susceptibility artifacts can result from variations in the conductivity of the volume being imaged, and interference is likely when multiplanar imaging of adjacent planes is attempted. The presence of air or low-density structures can lead to partial volume effects or susceptibility artifacts.

In OCT, any structure that blocks the light beam is likely to cause a streak artifact, similar to those produced by high densities in X-ray CT images. In addition, the refraction of the light at interfaces between the gel and other materials can cause ring artifacts or distortion of the image. The artifacts found in OCT images have been described (110). An example of the artifact caused by high optical densities is shown in Fig. 21.

When X-ray CT is used, artifacts can result from the low signal to noise ratio that occurs because of the very small density differences present in the gel. These artifacts have been investigated in some detail previously (121).

Temperature Dependence

The existence of a dependence on temperature during irradiation of polymer gels was not recognized immediately, but it has since been shown that this dependence exists. Furthermore, the temperature dependence can be more pronounced for some polymer gel formulations than others. The polymerization that occurs as a result of irradiation of the gel is exothermic, and consequently can lead to a temperature rise that influences further polymerization of gel in response to continuing exposure. In extreme cases, this temperature rise can exceed several degrees Celcius (207).

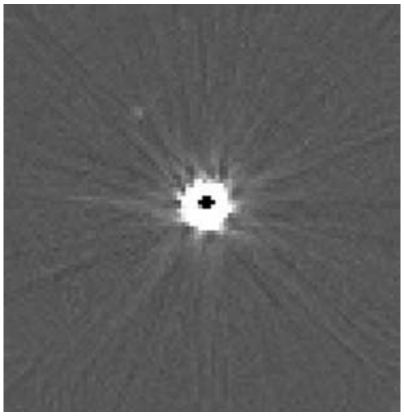


Figure 21. An optical CT scan of a normoxic gel irradiated with a low dose rate ^{125}I brachytherapy source. The high optical densities close to the source completely attenuate the laser, resulting in a streak artifact.

Oxygen Sensitivity

The sensitivity of polymer gels to oxygen has been discussed extensively, and several investigators have responded by developing gels that contain oxygen scavengers, such as the MAGIC gel (86). The oxygen scavenger removes oxygen present in the gel at the time of manufacture, even if this is done in normoxic conditions. It can remove additional small amounts of oxygen, but ultimately will be overwhelmed if exposure to normal atmosphere is ongoing. While this problem has been addressed, it still creates minor inconvenience that might limit the successful introduction of gels into routine clinical use. The characteristics of several normoxic gel dosimeters have been investigated in detail (90,208).

Tissue Equivalence and Energy Dependence

Gels, both Fricke and polymer types, compete well when compared to other dosimeters in terms of their tissue equivalence and energy dependence. In comparison to thermoluminescence dosimeters (TLD), radiographic film, and even ionization chambers, for example, gels are considerably less energy dependent and are much more tissue equivalent (209). However, under extreme conditions of photon energies below 60 keV, and LET values greater than $\sim 2.5\ \text{keV}\cdot\mu^{-1}$, gels show a dependence that remains to be fully characterized (190).

Simulation of Nonunit Densities

The benefits of gels discussed in the previous paragraph lead to the inability of gels to easily simulate nonunit density tissues. To date, limited efforts have been described to create low density gel mixtures, to simulate lung tissue. No attempts have been described to date to create high density mixtures and are unlikely to be with today's emphasis on the use of gelatin-based formulations.

Diffusion of Monomer in Steep Gradients

The diffusion of monomer, and the shrinkage of gel proportional to the creation of long polymer chains, can be addressed through the development of better gel mixtures. Avoiding small monomers such as acrylamide can reduce the rate of diffusion in regions of steep dose gradient, such as the penumbra of radiation beams (210). Employing different concentrations of gelatin might reduce or eliminate problems associated with shrinkage of gels in high dose regions. Some normoxic gels may demonstrate decreased diffusion in regions of steep dose gradient (90,208).

SUMMARY AND CONCLUSIONS

The importance of quality assurance in radiation therapy is well documented. High quality, safe, and effective radiation therapy is dependent upon the proper operation of equipment, the accuracy of alignment devices, and the dependability of dosimetry procedures. The accurate delivery of the prescribed dose depends on procedures involving dosimetry systems. Properly functioning dosimeters are necessary to assure that the equipment is properly calibrated and that treatment planning procedures are conducted correctly.

A wide variety of dosimetry systems are available to medical physicists today. Choosing the correct dosimetry system for any given application requires an understanding of the operation of the device and its appropriateness for the intended circumstances. This presentation has reviewed a number of dosimetry systems presently available, their design and operation, and some of the uses for which they are valuable.

Gel dosimetry offers the promise of accurate and convenient dosimetry under a variety of circumstances. In most of the examples discussed above, gel dosimeters offer a number of advantages over conventional dosimeters. Chief among these is the ability to measure a complex dose distribution throughout a volume with a single radiation exposure. Additional advantages include tissue equivalence, high spatial accuracy, good dose precision, and reasonable convenience.

However, gel dosimetry continues to experience little acceptance in the clinic, largely because some aspects of promise have not been achieved, and because of a perceived lack of convenience. Members of the radiation physics community are apparently not convinced that the benefits of gels sufficiently outweigh conventional dosimeters such as film and TLD. It is incumbent on those of us working with gels to encourage more widespread use, by taking every opportunity to display the results of measurements with gels.

BIBLIOGRAPHY

1. Cunningham JR. Development of computer algorithms for radiation treatment planning. Int J Radiat Oncol Biol Phys 1989;16:1367.
2. Fischer JJ, Moulder JE. The steepness of the dose-response curve in radiation therapy. Radiology 1975;117:179–184.
3. Hendrickson FR. Precision in radiation oncology. Int J Radiat Oncol Biol Phys 1981;8:311–312.
4. ICRU Rep No. 24: Determination of absorbed dose in a patient irradiated by beams of X or gamma rays in radiotherapy procedures, Washington (DC), 1976, International Commission on Radiation Units and Measurements.
5. ICRU report No. 42, Use of computers in external beam radiotherapy procedure with high energy photons and electrons, Washington (DC), 1988. International Commission on Radiation Units and Measurements.
6. ICRU report No. 50, Prescribing, recording, and reporting photon beam therapy, Washington (DC), 1993, International Commission on Radiation Units and Measurements.
7. Leunens G, et al. Assessment of dose inhomogeneity at target level by in vivo dosimetry; Can the recommended 5% accuracy in the dose delivered to the target volume be fulfilled in daily practice? Radiother Oncol 1992;25:245–250.
8. Ibbott GS, et al. Uncertainty of calibrations at the accredited dosimetry calibration laboratories. Med Phys 1997;24(8):1249–1254.
9. Kartha PKI, Chung-Bin A, Hendrickson FR. Accuracy in clinical dosimetry. Br J Radiol 1973;46:1083–1084.
10. Kartha PKI, et al. Accuracy in patient setup and its consequence in dosimetry. Med Phys 1975;2:331–332.
11. Ahuja SD. Physical and technological aspects of quality assurance in radiation oncology. Radiol Tech 1980;51(6): 759–774.
12. American Association of Physicists in Medicine. Radiation treatment planning dosimetry verification, AAPM Task Group 23 Test Package, 1984. AAPM Rep No. 55, 1995.
13. American College of Radiology: ACR Standard for Radiation Oncology Physics for External Beam Therapy Res. 15-1994, American College of Radiology Standards, adopted 1995 by the American College of Radiology, 1891 Preston White Drive, Reston, VA 22091.
14. American College of Radiology: ACR Standard for Radiation Oncology, Res. 38-1995, American College of Radiology Standards, adopted 1995 by the American College of Radiology, 18941 Preston White Drive, Reston, VA 22091.
15. American College of Radiology, ACR Standards for the performance of Brachytherapy Physics: Manually-Loaded Sources. Res. 25-1995, American College of Radiology Standards, adopted 1995 by the American College of Radiology, 1891 Preston White Drive, Reston, VA 22091.
16. Annett CH. Program for periodic maintenance and calibration of radiation therapy linear accelerators. Appl Radiol 1979;6: 77–80.
17. Earp KA, Gates L. Quality assurance: A model QA program in radiation oncology. Radiol Technol 1990;61(4):297–304.
18. Ibbott GS, et al. Quality Assurance Workshops for Radiation Therapy Technologists. Appl Radiol, March–April 1977.
19. International Electrotechnical Commission Rep No. 976, Medical electrical equipment, Geneva, Switzerland, 1993, Bureau Central de la Commission Electrotechnical Internationale.
20. International Electrotechnical Commission Rep No. 977, Medical electrical equipment: Medical electron accelerators in the range 1 MeV to 50 MeV. Guidelines for functional performance characteristics, Geneva, Switzerland, 1993, Bureau Central de la Commission Electrotechnique Internationale.
21. JCAHO Accreditation Manual for Hospitals, 1995, Oak Brook Terrace (IL): Joint Commission on Accreditation of Healthcare Organizations; 1995.
22. Kutcher GJ, et al. Comprehensive QA for radiation oncology; Report of AAPM Radiation Therapy Committee Task Group 40. Med Phys 1994;21(4):581–618.
23. Van Dyk J, editor. The Modern Technology of Radiation Oncology: A Compendium for Medical Physicists and Radiation Oncologists. Madison (WI): Medical Physics Publishing; 1999.

24. CIRMS 2004 - Council on Ionizing Radiation Measurements and Standards: Fourth Report on Needs in Ionizing Radiation Measurements and Standards, Dec 2004. CIRMS, P. O. Box 1238, Duluth, GA 30096. Available at www.cirms.org. 2004.
25. Rice RK, Hansen JL, Svensson GK, Siddon RL. Measurements of dose distributions in small beams of 6 MV X-rays. Phys Med Biol 1987;32:1087–1099.
26. Alecu R, Alecu M. *In-vivo* rectal dose measurements with diodes to avoid misadministrations during intracavitary high dose rate brachytherapy for carcinoma of the cervix. Med Phys 1999;26(5):768–770.
27. Alecu R, Loomis T, Alecu J, Ochran T. Guidelines on the implentation of diode in vivo dosimetry programs for photon and electron external beam therapy. Med Dosimetry 1999;24(1): 5–12.
28. Mijnheer B, et al. Quality assurance of treatment planning systems: Practical examples for non-IMRT photon beams. 2004 ESTRO, Mounierlaan 83/12 - 1200 Brussels (Belgium)
29. International Atomic Energy Agency Technical Report 430. Commissioning and quality assurance of computerized planning systems for radiation treatment of cancer. Vienna: IAEA; 2004.
30. Karger CP, Heeg P. A system for 3-dimensional dosimetric verification of treatment plans in intensity-modulated radiotherapy with heavy ions. Med Phys Oct 1999;26(10).
31. Knoll GF. Radiation Detection and Measurement, New York: Wiley; 1989, p 216–227.
32. Kirov AS, et al. Towards two-dimensional brachytherapy dosimetry using plastic scintillator: New highly efficient water equivalent plastic scintillator materials. Med Phys August 1999;26(8).
33. Beddar AS, Mackie TR, Attix FH. Water-equivalent plastic scintillation detectors for high energy beam dosimetry: II. Properties and measurements. Phys Med Biol 1992;37: 1901–1913.
34. Olde GL, Brannon E. Three dimensional scintillation dosimeter. Rev Sci Instrum 1959;30:1014–1016.
35. Perera H, et al. Rapid 2-dimensional dose measurement in brachytherapy using plastic scintilaator sheet: Linearity, signal-to-noise ratio, and energy response characteristics. Int J Radiat Oncol Biol Phys 1992;23:1059–1069.
36. Kirov AS, et al. Two-dimensional dosimeter using plastic scintillator: Localization of the scintillation process. Med Phys 1997;24:1005.
37. Kirov AS, et al. New highly efficient water equivalent plastic scintillator materials for radiation dosimetry. Med Phys 1998;25:A153.
38. Yeo IJ, Wang CKC, Burch SE. A new approach to film dosimetry for high-energy photon beams using organic plastic scintillators. Phys Med Biol 1999;44:3055–3069.
39. Yeo IJ, Wang C, Burch SE. A scintillation method for improving X-ray film dosimetry in photon radiotherapy (abstract). Med Phys 1996;23:1161.
40. Burch SE, Yeo IJ, Wang CK. A new approach to film dosimetry for high-energy photon beams: lateral scatter filtering. Med Phys 1997;24:775–783.
41. Seamon JM, Ibbott GS. Errors introduced in electron beam film dosimetry. Med Dosimet 1987;12(2):35–37.
42. Meigooni AS, Sanders MI, Ibbott GS, Szeglin SR. Dosimetric characteristics of an improved radiochromic film. Med Phys 1996;23(11):1883–1888.
43. McLaughlin WL, Miller A, Fiban S, Pejtersen K. Radiochromic plastic film for accurate measurement of radiation absorbed dose and dose distributions. Radiat Phys Chem 1977;9:737–474.
44. Cameron JR, Suntharalingam N, Kenney GN. Thermoluminescent Dosimetry. Madison (WI): The University of Wisconsin Press; 1968.
45. Conner WG, et al. Patient repositioning and motion detection using a video cancellation system. Int J Radiat Oncol Biol Phys 1975;1:147–153.
46. Rogus RD, Stern RL, Kubo HD. Accuracy of a photogrammetry-based patient positioning and monitoring system for radiation therapy. Med Phys 1999;26(5):721–728.
47. Curtin-Savard AJ, Podgorsak EB. Verification of segmented beam delivery using a commercial electronic portal imaging device. Med Phys 1999;26(5):737–742.
48. Petrascu O, et al. Automatic on-line electronic portal image analysis with a wavelet-based edge detector. Med Phys 2000; 27(2):321–329.
49. Gilhuijs KGA, Van Herk M. Automatic on-line inspection of patient setup in radiation therapy using digital portal images. Med Phys 1993;20:667–677.
50. Van Herk M, Bel A, Gilhuijs KGA, Vijlbrief RE. A comprehensive system for the analysis of portal images. Radiother Oncol 1993;29:221–229.
51. Fritsch D, et al. Core-based portal image registration for automatic radiotherapy treatment verification. Int J Radiat Oncol Biol Phys 1995;33:1287–1300.
52. Dong L, Boyer AL. An image correlation procedure for digitally reconstructed radiographs and electronic portal images. Int J Radiat Oncol Biol Phys 1995;33:1053–1060.
53. Day MJ, Stein G. Chemical effects of ionizing radiation in some gels. Nature(London) 1950;166:146–147.
54. Andrews HL, Murphy RE, LeBrun EJ. Gel dosimeter for depth dose measurements. Rev Sci Instr 1957;28:329–332.
55. Gore JC, Kang YS, Schulz RJ. Measurement of radiation dose distributions by nuclear magnetic resonance (NMR) imaging. Phys Med Biol 1984;29:1189–1197.
56. Schreiner LJ. Gel dosimetry: Motivation and historical foundation. In DOSGEL 1999: Proc 1st Int Workshop Radiation Therapy Gel Dosimetry (Canadian Organization of Medical Physicists, Edmonton) Schreiner LJ, Audet C, editors.
57. DOSGEL 1999. Proc 1st Int Workshop Radiation Therapy Gel Dosimetry (Lexington, KY). In: Schreiner L J, Audet C, editors. Ottawa, Ontario, Canada: Canadian Organization of Medical Physicists; 1999.
58. DOSGEL 2001. Proc 2nd Int Conf Radiotherapy Gel Dosimetry. In: Baldock C, De Deene Y editors. Brisbane, Queensland, Australia: Queensland University of Technology; 2001.
59. DOSGEL 2001. Proc 3rd Int Conf Radiotherapy Gel Dosimetry. In: Baldock C, De Deene Y, editors. Gent University, Gent, Belgium. J Phys Conf Ser 2004; **3**.
60. Fricke H, Morse S. The chemical action of Roetgen rays on dilute ferrosulphate solutions as a measure of dose. Am J Roent Radium Ther Nul Med 1927;18:430–432.
61. Fricke H, Hart EJ. Chemical Dosimetry, Vol. 2. In: Attix FH, Roesch WC, editors. Radiation Dosimetry. New York: Academic Press; 1966.
62. Appleby A, Christman EA, Leghrouz A. Imaging of spatial radiation dose distribution in agarose gels using magnetic resonance. Med Phys 1987;14:382–384.
63. Olsson LE, Petersson S, Ahlgren L, Mattsson S. Ferrous sulphate gels for determination of absorbed dose distributions using MRI technique: basic studies. Phys Med Biol 1989;34: 43–52.
64. Schulz RJ, deGuzman AF, Nguyen DB, Gore JC. Dose-response curves for Fricke-infused agarose gels as obtained by nuclear magnetic resonance. Phys Med Biol 1990;35:1611–1622.
65. Olsson LE, Appleby A, Sommer JA. A new dosimeter based on ferrous sulphate solution and agarose gel. Appl Radiat Isot 1991;42:1081.
66. Olsson LE, Westrin BA, Fransson A, Nordell B. Diffusion of ferric ions in agarose dosimeter gel. Phys Med Biol 1992a; 37:2243–2252.

67. Baldock C, Harris PJ, Piercy AR, Healy B. Experimental determination of the diffusion coefficient in two-dimensions in ferrous sulphate gels using the finite element method. Aust Phys Eng Sci Med 2001; 24:19–30.
68. Balcolm BJ, et al. Diffusion in Fe(II/III) radiation dosimetry gels measured by MRI. Phys Med Biol 1995;40:1665–1676.
69. Harris PJ, Piercy A, Baldock C. A method for determining the diffusion coefficient in Fe(II/III) radiation dosimetry gels using finite elements. Phys Med Biol 1996;41:1745–1753. Baldock C, et al. Temperature dependence of diffusion in Fricke gel MRI dosimetry. Med Phys 1995;22:1540.
70. Schreiner LJ. Fricke gel dosimetry. Proc 2nd Int Conf Gel Dosimetry, DOSGEL 2001, 15–22.
71. Chu KC, et al. Polyvinyl alcohol Fricke hydrogel and cryogel: two new gel dosimetry systems with low Fe^{3+} diffusion. Phys Med Biol 2000;45:955–969.
72. Kelly RU, Jordan KJ, Battista J. Optical CT reconstruction of 3D dose distributions using the ferrous benzoic-xylenol (FBX) gel dosimeter. Med Phys 1998;25:1741–1750.
73. Chu KC, et al. A Novel Fricke Dosimeter using PVA Cryogel, DOSGEL'99, Proceedings of the 1st International Workshop on Radiation Therapy Gel Dosimetry. In: Schreiner LJ, Audet C, editors. Ottawa, Ontario, Canada: Canadian Organization of Medical Physicists, 1999.
74. Chu K, Rutt BK. Polyvinyl alcohol cryogel: an ideal phantom material for MR studies of arterial flow and elasticity. Magn Reson Med 1997;37:314–319.
75. Gambarini G, et al. Dose-response curve slope improvement and result reproductibility of ferrous-sulphate-doped gels analyzed by NMR imaging. Phys Med Biol 1994;39:703–717.
76. Rae WID, et al. Chelator effect on ion diffusion in ferrous-sulfate-doped gelatin gel dosimeters as analyzed by MRI. Med Phys 1996;23:15–23.
77. Kron T, Jonas D, Pope JM. Fast T-1 imaging of dual gel samples for diffusion measurements in NMR dosimetry gels. Magn Reson Imaging 1997;15:211–221.
78. Pedersen TV, Olsen DR, Skretting A. Measurement of the ferric diffusion coefficient in agarose and gelating gels by utilizatino of the evolution of a radiation enduced edge as reflected in relaxation rate images. Phys Med Biol 1997;42: 1575–1585.
79. Scherer J, et al. 3D Fricke gel dosimetry in antropomorphic phantoms. Proc 1st Int Workshop Radiation Therapy Gel Dosimetry, Lexington(ky) July 21–23, 1999; p 211–213.
80. Maryanski MJ, Gore JC, Schulz RJ. 3D Radiation Dosimetry by MRI: Solvent Proton Relaxation Enhancement by Radiation-Controlled Polymerization and Crosslinking in Gels. 11th Annu Sci Meet Soc Magnetic Resonance in Medicine, Berlin, (Poster No. 1325). 1992.
81. Maryanski MJ, Gore JC, Kennan RP, Schulz RJ. NMR relaxation enhancement in gels polymerized and cross-linked by ionizing radiation: a new approach to 3D dosimetry by MRI. Magn Reson Imaging 1993;11:253–258.
82. Maryanski MJ, et al. Radiation therapy dosimetry using magnetic resonance imaging of polymer gels. Med Phys 1996;23:699–705.
83. Baldock C, et al. Experimental procedure for the manufacture of polyacrylamide gel (PAG) for magnetic resonance imaging (MRI) radiation dosimetry. Phys Med Biol 1998; 43:695–702.
84. Maryanski MJ. Radiation-sensitive polymer-gels: properties and manufacturing. Proc 1st Int Conf Gel Dosimetry, DOSGEL 1999. Queens University Printing Service, Kingston, Ontario, Canada. 63–73.
85. Maryanski MJ, Gore JC, Schulz RJ. Three-Dimensional Detection, Dosimetry and Imaging of an Energy Field by Formation of a Polymer in a Gel. US Patent 5,321,357.
86. Fong PM, Keil DC, Does MD, Gore JC. Polymer gels for magnetic resonance imaging of radiation dose distributions at normal room atmosphere. Phys Med Biol 2001;46:3105–3113.
87. De Deene Y, et al. A basic study of some normoxic polymer gel dosimeters. Phys Med Biol 2002;47:3441–3463.
88. Maryanski MJ, et al. Three dimensional dose distributions for 160 MeV protons using MRI of the tissue-equivalent BANG Polymer-gel dosimeter. Particles (PTCOG Newsletter) Jan 10–11 1994a.
89. Baldock C, et al. Dose resolution in radiotherapy polymer gel dosimetry: effect of echo spacing in MRI pulse sequence. Phys Med Biol 2001;46:449–460.
90. De Deene Y, Baldock C. Optimization of multiple spin-echo sequences for 3D polymer gel dosimetry. Phys Med Biol 2002;47:3117–3141.
91. Marquardt DW. An algorithm for least-squares estimation of nonlinear parameters. J Soc Ind Appl Math 1963;11:431–441.
92. Maryanski MJ, et al. Magnetic resonance imaging of radiation dose distributions using a polymer-gel dosimeter. Phys Med Biol 1994;39:1437–1455.
93. Maryanski MJ, Audet C, Gore JC. Effects of crosslinking and temperature on the dose response of a BANG polymer gel dosimeter. Phys Med Biol 1997;42:303–311.
94. Hrbacek J, Spevacek V, Novotny J Jr, Cechak T. A comparative study of four polymer gel dosimeters. J Phys Conf Ser 2004;3:150–154.
95. De Deene Y, De Wagter C. Artefacts in multi-echo T2 imaging for high-precision gel dosimetry. III. Effects of temperature drift during scanning. Phys Med Biol 2001;46:2697–2711.
96. De Deene Y, et al. An investigation of the chemical stability of a monomer/polymer gel dosimeter. Phys Med Biol 2000; 45:859–478.
97. MacDougall ND, Pitchford WG, Smith MA. A systematic review of the precision and accuracy of dose measurements in photon radiotherapy using polymer and Fricke MRI gel dosimetry. Phys Med Biol 2002;47:R107–R121.
98. Jirasek AI, Duzenli C, Audet C, Eldridge J. Characterization of monomer/crosslinker consumption and polymer formation observed in FT-Raman spectra of irradiated polyacrylamide gels. Phys Med Biol, 2001;46:151–165.
99. Spinks JWT, Woods RJ. An Introduction to Radiation Chemistry. New York, London, Sydney: Wiley; 1964.
100. Kennan RP, et al. The effects of cross-link density and chemical change on magnetization transfer in polyacrylamide gets. J Magn Res B 1996;100:267–277.
101. Oldham M, et al. Improving calibration accuracy in gel dosimetry. Phys Med Biol 1998; 43:2709–2720.
102. Baldock C, et al. Investigation of polymerisation of radiation dosimetry polymer gels Proceedings. 1st International Workshop on Radiation Therapy Gel Dosimetry (Lexington, KY). Schreiner L J, Audet C, editors. Canadian Organisation of Medical Physics 1999. p 99–105.
103. McJury M, Oldham M, Leach MO, Webb S. Dynamics of polymerization in polyacrylamide gel (PAG) dosimeters I. Ageing and long-term stability Phys Med Biol 1999;44: 1863–1873.
104. De Deene Y, et al. Artefacts in multi-echo T2 imaging for high-precision gel dosimetry. II. Analysis of B1 field inhomogeneity. Phys Med Biol 2000;45:1825–1839.
105. De Deene Y, et al. Artefacts in multi-echo T2 imaging for high-precision gel dosimetry. I. Analysis and compensation of eddy currents. Phys Med Biol 2000;45:1807–1823.
106. Watanabe Y, Perera GM, Mooij RB. Image distortion in MRI-based polymer gel dosimetry of Gamma Knife stereotactic radiosurgery systems. Med Phys 2002;29:797–802.
107. Maryanski MJ, Zastavker YZ, Gore JC. Radiation dose distributions in three dimensions from tomographic optical

density scanning of polymer gels: II. Optical properties of the BANG polymer gel. Phys Med Biol 1996;41:2705–2717.

108. Gore JC, Ranade M, Maryanski MJ, Schulz RJ. Radiation dose distributions in three dimensions from tomographic optical density scanning of polymer gels: I. Development of an optical scanner. Phys Med Biol 1996;41:2695–2704.
109. Oldham M, Siewerdsen JH, Shetty A, Jaffray DA. 1998b; High resolution gel-dosimetry by optical-CT and MR scanning. Med Phys 2001;28:1436–1445.
110. Oldham M et al. Optical-CT gel dosimetry I: Basic investigations. Med Phys 2003;30:623–634.
111. Wolodzko JG, Marsden C, Appleby A. CCD imaging for optical tomography of gel radiation dosimeters. Med Phys 1999;26:2508–2513.
112. Jordan K. Advances in optical CT scanning for gel dosimetry. J Phys Conf Ser 2004;3:115–121.
113. Oldham M. Optical CT scanning of polymer gels. J Phys: Conf Ser 2004;3:122–135.
114. Maryanski MJ. High-resolution 3D dosimetry for endovascular brachytherapy using optical laser CT microimaging of BANG polymer gels. Med Phys 1998;25:A107.
115. Knisely JPS et al. Three-dimensional dosimetry for complex stereotactic radiosurgery using a tomographic optical density scanner and BANG polymer gels. In: Radiosurgery 1997, Dondziolka D, editors Vol 2. Basel : Karger; 1998. p 251–260.
116. Islam KTS, et al. Initial evaluation of commercial optical CT-based 3D gel dosimeter. Med Phys 2003;30:2159–2168.
117. Heard MP and Ibbott GS. Measurement of brachytherapy sources using MAGIC gel. J Phy: Conf Ser 2004;3:221–223.
118. Xu YS, Wuu C-S, Maryanski MJ. Performance of a commercial optical CT scanner and polymer gel dosemeters for 3-D dose verification. Med Phys 2004;31:3024.
119. Xu Y, Wuu C-S, Maryanski MJ. Determining optical gel sensetivity in optima CT scanning of gel dosemeters. Med Phys 2003;30:2257.
120. Hilts M, Audet C, Duzenli C, Jirasek A. Polymer gel dosimetry using X-ray computed tomography: A feasibility study. Phys Med Biol 2000;45:2559–2571.
121. Trapp JV, et al. An experimental study of the dose response of polymer gel dosimeters imaged with x-ray computed tomography. Phys Med Biol 2001;46:2939–2951.
122. Trapp JV, Michael G, De Deene Y, Baldock C. Attenuation of diagnostic energy photons by polymer gel dosimeters. Phys Med Biol 2002;47:4247–4258.
123. Audet C, Hilts M, Jirasek A, Duzenli C. CT gel dosimetry technique: comparison of a planned and measured 3D stereotactic dose volume. J Appl Clin Med Phys 2002;3; 110–118.
124. Brindha S, Venning A, Hill B, Baldock C. Experimental investigation of the attenuation properties of normoxic polymer gel dosimeters. Med Phys 2004;31:1886.
125. Hilts M, Audet C, Duzenli C, Jirasek A. Polymer gel dosimetry using X-ray computer tomography: feasibility and potential application to stereotactic radiosurgery Proc. 1st Int. Workshop on Radiation Therapy Gel Dosimetry (Lexington, KY USA) Schreiner L J Audet C editors. 1999.
126. Hilts M, Duzenli C. Image filtering for improved dose resolution in CT polymer gel dosimetry. Med Phys 2004a;31:39–49.
127. Hilts M, Jirasek A, Duzenli C. Effects of gel composition on the radiation induced density change in PAG polymer gel dosimeters: a model and experimental investigations. Phys Med Biol 2004;49:2477–2490.
128. Hilts M, Jirasek A, Duzenli C. The response of PAG density to dose: a model and experimental investigations. J Phy: Conf Ser. 2004c;3:163–167
129. Mather ML, Whittaker AK, Baldock C. Ultrasound evaluation of polymer gel dosimeters. Phys Med Biol 2002;47: 1449–1458.
130. Mather ML et al. Investigation of ultrasonic properties of PAG and MAGIC polymer gel dosimeters. Phys Med Biol 2002;47:4397–4409.
131. Mather ML, et al. Acoustic evaluation of polymer gel dosimeters. Proc Int Symp Standards and Codes of Practice in Medical Radiation Dosimetry. International Atomic Energy Agency, Vienna; 2002c. p 234–235.
132. Mather ML, Baldock C. Ultrasound tomography imaging of radiation dose distributions in polymer gel dosimeters. Med Phys 2003;30:2140–2148.
133. Baldock C et al. Fourier transform Raman spectroscopy of polyacrylamide gels (PAGs) for radiation dosimetry. Phys Med Biol 1998; 43:3617–3627.
134. Baldock C. X-ray computer tomography, ultrasound and vibrational spectroscopic evaluation techniques of polymer gel dosimeters. J Phy Conf Ser, 2004;3:136–141.
135. Lepage M, Whittaker AK, Rintoul L, Baldock C. ^{13}C-NMR, ^{1}H-NMR and FT-Raman study of radiation-induced modifications in radiation dosimetry polymer gels. J Appl Polym Sci 2001;79:1572–1581.
136. Rintoul L, Lepage M, Baldock C. Radiation dose distributions in polymer gels by Raman spectroscopy. Appl Spectrosc 2003;57:51–57.
137. Jirasek A, Duzenli C. Relative effectiveness of polyacrylamide gel dosimeters applied to proton beams: Fourier transform Raman observations and track structure calculations. Med Phys 2002;29:569–577.
138. McJury M et al. Radiation dosimetry using polymer gels: methods and applications. Br J Radiol 2000;73:919–929.
139. Lepage M, Jayasakera PM, Back SAJ, Baldock C. Dose resolution optimization of polymer gel dosimeters using different monomers. Phys Med Biol 2001;46:2665–2680.
140. Trapp JV et al. Dose resolution in gel dosimetry: effect of uncertainty in the calibration function. Phys Med Biol 2004;49:N139–N146.
141. Day MJ. Radiation dosimetry using nuclear magnetic resonance an introductory review. Phys Med Biol 1990;35:1605.
142. Bonnett D. A review of application of polymer gel dosimetry. DOS GEL 2001. Proc 2nd Int Conf Radiotherapy Gel Dosimetry. In: Baldock C, DeDeene Y, editors. Queensland University of Technology, Brislane, Queensland, Australia. p 40–48
143. Ibbott GS. Applications of Gel Dosimetry. J Phys Conf Ser 2004;3:58–77.
144. Ibbott GS, Maryanski MJ, Avison RG, Gore JC. Investigation of a BANG polymer gel dosimeter for use as a mailed QA device. Med Phys 1995;22:951.
145. Haraldsson P, Back SA, Magnusson P, Olsson LE. Dose response characteristics and basic dose distribution data for a polymerization-based dosimeter gel evaluated using MR. Br J Radiol 2000;73:919–929.
146. Oldham M et al. An investigation into the dosimetry of a nine-field tomotherapy irradiation using BANG-gel dosimetry Phys Med Biol 1998;43:1113–1132.
147. Baldock C et al. A dosimetry phantom for external beam radiation therapy of the breast using radiation-sensitive polymer gels and MRI. Med Phys 1996;23:1490.
148. Love PA, Evans PM, Leach MO, Webb S. Polymer gel measurement of dose homogeneity in the breast: comparing MLC intensity modulation with standard wedged delivery. Phys Med Biol 2003;48:1065–1074.
149. De Deene Y et al. Three-dimensional dosimetry using polymer gel and magnetic resonance imaging applied to the verification of conformal radiation therapy in head-and-neck cancer. Radiother Oncol 1998;48:283–291.
150. Trapp JV, et al. The use of gel dosimetry for verification of electron and photon treatment plans in carcinoma of the scalp. Phys Med Biol 2004;49:1625–1635.

151. Bengtsson M et al. Measurement of dynamic wedge angles and beam profiles by means of MRI ferrous sulphate gel dosimetry. Phys Med Biol 1996;41:269–277.
152. Hill B, Venning C, Baldock C. Acceptance testing of a computer tomography scanner using normoxic polymer gel dosimetry. Med Phys 2004;31:1786.
153. Hill B, Venning C, Baldock C. X-ray computer tomography dose response of normoxic polymer gel dosimeters. Br J Radiol. In press. 2005.
154. Olsson LE, Arndt J, Fransson A, Nordell B. Three-dimensional dose mapping from gamma knife treatment using a dosimeter gel and MR-imaging. Radiother Oncol 1992; 24:82–86.
155. Schulz RJ, Maryanski MJ, Ibbott GS, Bond JE. Assessment of the Accuracy of Stereotactic Radiosurgery Using Fricke-Infused Gels and MRI. Med Phys 1993;20:1731–1735.
156. Ibbott GS et al. Stereotactic radiosurgery simulation using MRI and a polymer-gel dosimeter. Med Phys May/June 1993;20(3) .
157. Ibbott GS, et al. Use of BANG polymer gel dosimeter to evaluate repeat-fixation stereotactic radiation therapy. Med Phys 1996;23:1070.
158. Meeks SL et al. Image registration of BANG gel dose maps for quantitative dosimetry verification. Int J Radiat Oncol Biol Phys 1999;43:1135–1151.
159. Novotny J Jr et al. Quality control of the stereotactic radiosurgery procedure with the polymer-gel dosimetry. Radiother Oncol 2002;63:223–230.
160. Scheib SG, Gianolini S. Three-dimensional dose verification using BANG gel: a clinical example. J Neurosurg 2002;97: 582–587
161. Ibbott GS, et al. Three dimensional visualization and measurement of conformal dose distributions using magnetic resonance imaging of BANG polymer gel dosimeters. Int J Radiat Oncol Biol Phys 1997;38:1097–1103.
162. Low DA et al. Evaluation of polymer gels and MRI as a 3D dosimeter for intensity-modulated radiation therapy. Int J Radiat Oncol Biol Phys 1999;26:154.
163. Beach M et al. Implementation of a Polymer Gel Dosimetry Insert for An Anthropomorphic Phantom Used to Evaluate Head and Neck Intensity-Modulated Radiation Therapy. desseitation, UT M. D. Anderson Cancer Center. 2003.
164. De Neve W. Clinical delivery of intensity modulated conformal radiotherapy for relapsed or second-primary head and neck cancer using a multileaf collimator with dynamic control. Radiother Oncol 1999;50:301–314.
165. Ibbott G, Beach M, Maryanski M. An anthropomorphic head phantom with a BANG® polymer gel insert for dosimetric evaluation of IMRT treatment delivery, Standards and Codes of Practice in Medical Radiation Dosimetry, Proc Int Symp. Vienna 2002;2:361–368.
166. Ibbott GS, Beach ML, Maryanski MJ. IMRT QA with an Anthropomorphic Phantom Employing a Polymer Gel Dosimeter. Int Organization Med Phys Proc. Vol. 1 2003.
167. Gustavsson H et al. MAGIC-type polymer gel for three-dimensional dosimetry: Intensity-modulated radiation therapy verification. Med Phys 2003;30:1264–1271.
168. Vergote K, et al. Application of monomer/polymer gel dosimetry to study the effects of tissue inhomogeneities on intensity-modulated radiation therapy (IMRT) dose distributions. Radiother Oncol 2003;67:119–128.
169. Vergote K, et al. Validation and application of polymer gel dosimetry for the dose verification of an intensity-modulated arc therapy (IMAT) treatment. Phys Med Biol 2004;49:287–305.
170. Molineu A et al. Design and implementation of an anthropomorphic quality assurance phantom for intensity modulated radiation therapy. Int J Radiat Oncol Biol Phys 2005; (In press)
171. Heard MP. Characterizing Dose Distributions of Brachytherapy Sources using Normoxic Gel. MS dissertation M. D. Anderson Cancer Center, Houston (TX) 2005.
172. Duthoy W, et al. Whole abdominopelvic radiotherapy (WAPRT) using intensity-modulated arc therapy (IMAT): first clinical experience. Int J Rad Oncol Biol Phys 2003;57:1019–1032.
173. Cadman P et al. Dosimetric considerations for validation of a sequential IMRT process with a commercial treatment planning system. Phys Med Biol 2002;47; 3001–3010.
174. Nath R, Melillo A. Dosimetric characteristics of a double wall ^{125}I source for interstitial brachytherapy. Med Phys 1993;20:1475–1483.
175. Muench PJ, Meigooni AS, Nath R, McLaughlin WL. Photon energy dependence of the sensitivity of radiochromic film and comparison with silver halide film and LiF TLDs used for brachytherapy dosimetry. Med Phys 1991;18:769–775.
176. Schreiner LJ, et al. Imaging of HDR brachytherapy dose distributions using NRM Fricke-gelatin dosimetry. Magn Reson Imaging 1994;12:901–907.
177. Olsen DR, Hellesnes J. Absorbed dose distribution measurements in brachytherapy using ferrous sulphate gel and magnetic resonance imaging. Br J Radiol 1994;67:1121–1126.
178. Maryanski MJ, et al. Magnetic Resonance Imaging of Dose Distributions from Brachytherapy Sources Embedded in Tissue Equivalent BANG Polymer Gel Dosimeters. Med Phys 1994;21:919 (abstract).
179. De Deene Y et al. On the accuracy of monomer/polymer gel dosimetry in the proximity of a high-dose-rate ^{192}Ir source. Phys Med Biol 2001;46:2801–2825.
180. Gelfi C, Righetti P G. Polymerization kinetics of polyacrylamide gels: II. Effect of temperature. Electrophoresis 1981;2: 220–228.
181. Omidian H, Hashemi SA, Sammes PG, Meldrum IG. Modified acrylic-based superabsorbent polymers. Effect of temperature and initiator concentration. Polymer 1988;39:3459–3466.
182. Ibbott G, et al. Characterization of a New Brachytherapy Source by BANG® Gel Dosimetry. DosGel 99: Proc 1st Int Workshop Radiation Therapy Gel Dosimetry. Canadian Organisation of Medical Physicists and the Canadian College of Physicists in Medicine. 196–198. 1999.
183. Ibbott GS et al. Characteristics of a new brachytherapy source by BANG® gel dosimetry. Int J Rad Oncol Biol Phys. 1999;45(35):417.
184. Heard M, Ibbott G, Followill D. Characterizing Dose Distributions of Brachytherapy Sources Using Normoxic Gel (WIP), AAPM Annual Meeting 2003.
185. Chan M et al. The measurement of three dimensional dose distribution of a ruthenium-106 ophthalmological applicator using magnetic resonance imaging of BANG polymer gels. J Appl Clin Med Phys 2001;2:85–89.
186. Gifford K et al. Verification of Monte Carlo calculations around a fletcher suit delclos ovoid with radiochromic film and normoxic polymer gel dosimetry. Med Phys 2004;31.
187. Gifford K et al. Verification of monte carlo calculations around a Fletcher suit delclos ovoid with normoxic polymer gel dosimetry. J Phys: Conf Ser, 2004;3:217–220.
188. Wuu C-S, et al. Dosimetry study of Re-188 liquid balloon for intravascular brachytherapy using polymer gel dosimeters and laser-beam optical CT scanner. Med Phys 2003;30: 132–137.
189. Vergote K, et al. On the relation between the spatial dose integrity and the temporal instability of polymer gel dosimeters. Phys Med Biol 2004;49:4507–4522.
190. Pantelis E, et al. Polymer gel water equivalence and relative energy response with emphasis on low photon energy dosimetry in brachytherapy. Phys Med Biol 2004;49, 3495–3514.

191. Venning AJ, Brindha S, Hill B, Baldock C. Preliminary study of a nomoxic PAG gel dosemeter with tetrabis (hydroxymethyl phosphonium chloride as an antioxidant. J Phip: Conf Ser 3 2004; 155–158.
192. Courbon F et al. Internal dosimetry using magnetic resonance imaging of polymer gel irradiated with iodine-131. Preliminary results. Proc 1st Int Workshop Radiation Therapy Gel Dosimetry (Lexington, KY). In: Schreiner L J Audet C, editors. Canadian Ottawa, Ontario, Canada: Organization of Medical Physicists, 1999.
193. Gambarini G et al. Three-dimensional determination of absorbed dose by spectrophotometric analysis of Ferrous-Sulphate Agarose gel. Nucl Instrum and Meth 1999; A 422:643–648.
194. Gambarini G. Gel dosimetry in neutron capture therapy. Proc 2nd Int Conf Radiotherapy Gel Dosimetry (Brisbane, Australia). 2001. p 89–91.
195. Gambarini G. et al. Fricke-gel dosimetry in boron neutron capture therapy. Radiat Prot Dosim 2002;101:419–422.
196. Bäck S A et al. Ferrous sulphate gel dosimetry and MRI for proton beam dose measurements. Phys Med Biol 1999; 44:1983–1996.
197. Gustavsson H, Karlsson A, Back S, Olsson LE. Dose response characteristics of a new normoxic polymer gel dosimeter. Ph.D. dissertation Department of Medical Radiation Physics, Lund University, Malmo University Hospital. 2004.
198. Gustavsson H et al. Linear energy transfer dependence of a normoxic polymer gel dosimeter investigated using proton beam absorbed dose measurements. Phys Med Biol 2004;49: 3847–3855.
199. Swallow AJ. Radiation Chemistry: an Introduction London: Longman Group limited; 1973.
200. Ramm U, et al. Three-dimensional BANG™gel dosimetry in conformal carbon ion radiotherapy. Phys Med Biol 2000;45: N95–N102.
201. Vergote Ket al. Comparisons between monomer/polymer gel dosemetry and dose computations for an IMRT treatment of a thorox phantom. In: Baldock- C, De Deene Y editor DOSGEL 2001, Proc 2nd Int Conf Radiotherapy Gel Dosemetry. Queensland University of Technology, Brisbane, Queensland, Australia.
202. Hepworth SJ et al. Dose mapping of inhomogeneities positioned in radiosentive polymer gels. Nucl. Instrum. Methods Phys Res A 1999;422:756–760.
203. Gum F, et al. Preliminary study on the use of an inhomogeneous anthropomorphic Fricke gel phantom and 3D magnetic resonance dosimetry for verification of IMRT treatment plans, Phys Med Biol 2002;47; N67–N77.
204. Watanabe Y, Mooij RB, Perera GM, Maryanski MJ. Heterogeneity phantoms for visualization of 3D dose distributions by MRI-based polymer gel dosimetry. Med Phys 2004;31: 975–984
205. Olberg S, Skretting A, Bruland O, Olsen DR. Dose distribution measurements by MRI of a phantom containing lung tissue equivalent compartments made of ferrous sulphate gel. Phys Med Biol 2000;45:2761–2770.
206. Borges JA, BenComo J, Ibbott GS. A 3 Dimensional Gel Dosimetry Lung Equivalent (WIP), AAPM annual meeting, 10–14 Aug 2003, San Diego(CA); 2003.
207. Salomons GJ, Park YS, McAuley KB, Schreiner LJ. Temperature increases associated with polymerization of irradiated PAG dosimeters. Phys Med Biol 2002;47:1435–1448.
208. De Deene Y et al. Dose-response stability and integrity of the dose distribution of various polymer gel dosimeters. Phys Med Biol 2002;47:2459–2470.
209. Keall PJ, Baldock C. A theoretical study of the radiological properties and water equivalence of Fricke and polymer gels used for radiation dosimetry. Aust Phys Eng Sci Med 1999;22: 85–91.
210. McAuley KB. The chemistry and physics of polyacrylamide gel dosimeters: why they do and don't work. J. Phys Conf Ser 2004;3; 29–33.

See also PHANTOM MATERIALS IN RADIOLOGY; RADIATION DOSIMETRY FOR ONCOLOGY; RADIATION THERAPY, INTENSITY MODULATED; RADIATION THERAPY SIMULATOR; RADIOSURGERY, STEREOTACTIC.

RADIATION, EFFECTS OF. See IONIZING RADIATION, BIOLOGICAL EFFECTS OF; NONIONIZING RADIATION, BIOLOGICAL EFFECTS OF.

RADIATION PROTECTION INSTRUMENTATION

GLENN P. GLASGOW
Loyola University of Chicago
Maywood, Illinois

RADIATION PROTECTION INSTRUMENTATION

Radioactive materials and equipment that generate radiation are prevalent in industry, military, education, science and medical facilities, and even in the home. Many scientific instruments perform dedicated radiation measurement tasks; the nuclear power industry employs possibly the greatest number of instruments of different designs and degrees of sophistication. This article describes similar instruments commonly used for radiation protection in medicine. Instruments used for radiation dosimetry for medical treatments (e.g., radiotherapy ionization chambers,) and those used for medical treatments (e.g., nuclear medicine well-ionization chambers) are excluded. Included are instruments used for the general tasks of detecting radiation, determining the types of radiation or species of radionuclides present, determining quantities of radionuclides, and measuring radiation levels around materials and equipment. The focus is how instruments detect radiation, not their electronic circuity, which is described only briefly in a few instances. Before choosing an instrument, the user must know about the availability and choice of instruments, types and sources of radiation, special terms that describe quantities of radiation, and measures of biological dose equivalency that individuals receives in the presence of radiation. A science discipline, *Health Physics*, and a scientific society, the *Health Physics Society*, are devoted to these topics (1). Since the1988 first edition of this article, major changes in medical radiation protection instrumentation include the development of the Internet for dissemination, by manufacturers and vendors, of information about instrument designs, operating parameters, and performances; improved performance and electronic circuitry using chips with complementary metal oxide semiconductors (CMOS) microprocessor technology of various types; miniaturization of computer components that reduce the weight and size of instruments; and new

Table 1. Some Distributors and Manufacturers of Radiation Protection Instruments

Company (Product Lines)	Internet Address	Electronic Mail Address
Berkeley Nucleonics	http://www.berkeleynucleonics.com	info@berkeleynucleonics.com
Berthold Technologies GmbH & Co.	http://www.bertholdtech.com	info@BertholdTech.com
Canberra Industries (Packard)	http://www.canberra.com	customersupport@canberra.com
Capintec, Inc	http://www.capintec.com	getinfo@capintec.com
Cardinal Health Nuclear Pharmacy Services (Inovision, Victoreen)	http://www.nps.cardinal.com	npsinfo@cardinal.com
Durridge Company, Inc.	http://www.durridge.com	sales@durridge.com
Far West Technology & Health Physics Instruments	http://www.fwt.com	info@fwt.com
Global Dosimetry Solutions	http://www.dosimetry.com	info@dosimetry.com
International Specialty Products	http://www.ispcorp.com	customerservicecenter@ispcorp.com
Landauer, Inc.	http://www.landauerinc.com	custserv@landauerinc.com
LAURUS Systems, Inc	http://www.laurussystems.com	sales@laurussystems.com
Ludlum Measurements, Inc.	http://www.ludlums.com	ludlum@ludlums.com
Ortec	http://www.ortec-online.com	info@ortec-online.com
Perkin Elmer Life and Analytical Sciences	http://www.las.perkinelmer.com	products@perkinelmer.com
Technical Associates	http://www.tech-associates.com	tagold@nwc.com
Thermo Electron Corporation	http://www.thermo.com	enviromental.radiation@thermo.com

definitions and terms used to describe radiation quantities and units [*Note*: In addition to the common prefixes of kilo- (k), mega- (M), giga- (G), milli- (m), micro- (μ), nano- (n), pico- (p), note the use of the somewhat less familiar femto- (f; 10^{-15}), atto- (a; 10^{-18}), zepto- (z; 10^{-21}), and yocto- (y; 10^{-24})] (2). Manufacturers market smaller, compact survey meters, personnel dosimeters, and specialized detectors and monitors with improved performance. We review common features of instruments, such as ionization chambers, gas-proportional counters, Geiger–Müeller (GM) tubes, scintillation and solid-state detectors, other less-common detectors, and photographic films.

AVAILABILITY OF INSTRUMENTS AND INFORMATION

Table 1 lists some major companies and manufacturers of radiation protection instruments, their worldwide web Internet addresses, and their electronic mail addresses. Commercial product catalogues, usually now available on the Internet, contain a wealth of specific information on the theory and operation of instruments. This company list represents no endorsement by the author; these companies were selected because their worldwide web Internet sites provide details about common radiation protection instruments advertised for research, laboratory, environmental, security, medical, and health physics (safety and protection) applications. Basic instruments require only modest modifications for specific field applications. Table 2 contains a typical product list of medical radiation protection instruments. Instruments are regularly reviewed in Technology Monitor articles in *Health Physics* (3). One general interest group shares information regarding procedures, selection, testing, and standardization of instruments (4). Basic radiation detection principles and instrument designs are described in university level science textbooks in Health Physics curricula; comprehensive descriptions appear in Knoll (5), Shapiro (6), Turner (7), and Gollnick (8).

CHOICE OF INSTRUMENTS

A radiation field often consists of multiple types of radiation. Instruments usually must have the capability to detect particular types of radiation and produce relative or absolute measures of their magnitudes, while discriminating against other types of radiation. Often the radiation energies must be measured. Common medical uses, Table 3, include equipment radiation surveys, area monitoring, area and personnel contamination surveys, personnel dosimetry, finding misplaced radioactive materials (radioactive seeds or sources), surveying radioactive packages, air sampling, and emergency response tasks. Individuals choosing radiation protection instruments for measurements preferably should know about the radiation environment under investigation. Is it predominantly photon

Table 2. A List of Some Medical Radiation Protection Products

- Air Monitors
- Alarm Ratemeters
- Alpha Detectors
- Alpha/Beta/Gamma Detectors
- Alpha/Beta Detectors
- Area Monitors
- Beta Detectors
- Beta/Gamma Detectors
- Connectors
- Counters
- Detector Accessories
- Dosimeters
- Gamma Detectors
- Geiger Counters
- Geiger-Müller Probes
- Ion Chambers
- MicroR Meters
- Neutron Detectors
- Neutron Meters
- Package Monitors
- Pocket Dosimeters
- Pocket Survey Meters
- Portable Accessories
- Portable Scaler/Ratemeters
- Proportional Probes
- Response Kits
- Sample Counters
- Sample Holders
- Scalers and Accessories
- Scintillation Well Counting and Detection Systems
- Specialized Monitors
- Specialized Portable Meters
- Survey Meters
- Test Equipment
- Wipe Counters
- X-ray Monitors

Table 3. Some Typical Radiation Protection Instruments and Their Major Features

Type	Generic Name	Characteristics	Uses
		Portable Survey Meters	
Alpha, Beta, Gamma	Ion Chamber	Air ionization chambers to detect alpha, beta, gamma, and X rays from 50 $nSv{\cdot}h^{-1}$ to 20 $Sv{\cdot}h^{-1}$; sliding alpha and beta shield	General purpose survey meters with large range features
Beta, gamma	Geiger Counter Rate Meters	Multiple ranges up to 60 kcpm and 5 $nSv{\cdot}h^{-1}$; uses halogen quenched GM tube; multiple attachable probes	General purpose survey meters for lower level (5 $nSv{\cdot}h^{-1}$) surveys
Gamma	Gamma Survey Meters	Multiple ranges to 10 $\mu Sv{\cdot}h^{-1}$; halogen quenched GM tube with energy compensation to 40 keV	General purpose survey meter for gamma ray surveys
Alpha, (H-3), beta	Gas Proportional Survey Meter	Measures alphas, low energy beta to 500 kcpm	For measurements in presence of volatile vapor, high γ-ray fields, and for surface contamination
Gamma	Micro "R" Meter	Uses 1 × 1 in. NaI(Tl) scintillator to measure 0.1 $\mu Sv{\cdot}h^{-1}$	For sensitive low level surveys of $\mu Sv{\cdot}h^{-1}$ levels
Alpha, gamma	Alpha-Gamma Scintillation Counter	Measures alpha to 2×10^6 cpm using scintillator; measures gammas to 20 $mSv{\cdot}h^{-1}$ using GM tubes	For simultaneous measurements of alpha–gamma contamination
Alpha, beta, gamma, neutrons, X rays	Scaler, Ratemeter, Single Channel Analyzer	Multiple ranges (1 $nSv{\cdot}h^{-1}$–10 $Sv{\cdot}h^{-1}$; 1 cpm–500 kcpm) with single-channel analyzer with selected windows	For use with multiple probes of many types; measures identifies type of radiation or radionuclide
Neutrons	Neutron rem Meter	Measures equivalent dose neutrons using BF_3 tube in a cadmium loaded polyethylene moderator	General purpose neutron detection for thermal to high energy neutrons
		Personnel Electronic Dosimeters	
Gamma	Alarming Dose Rate Meter	Scintillation detector sensitive to 0.1 $\mu Sv{\cdot}h^{-1}$ to 20 $\mu Sv{\cdot}h^{-1}$ with multiple preset alarm levels	Medical personnel monitoring
Gamma, beta, neutron	Alarming Dose Rate Meter	Silicon semiconductor detectors; 10 μSv–1 Sv	Medical personnel monitoring
		Area Monitors	
Gamma	Area Radiation Monitors	Alarming counter rate meters with adjustable alarm that sounds when exposure rates exceed preset levels. Usually have GM tube detectors	Used to monitor areas where personnel prepare and use sources; used to determine that remote control sources have retracted to a safe.
		Air Monitors and Samplers	
Beta, Gamma	Beta–Gamma Air Particulate Monitor	Measure airborne particulate beta emitting particles using pancake-type GM tubes; ^{133}Xe monitors	Alarm monitor for laboratories using radioactive gasses emitting beta particles
		Well Counting Systems	
Gamma, beta	Liquid Scintillation Counting system	Counting of wipe tests from labs, sources to identify type and amount of radionuclide	General radiation control and containment
		Spectroscopy	
Gamma, beta	Multichannel Analyzer with NaI(Tl) or Ge detector	Identification of radonuclides by characteristic spectral analysis	General radiation control and containment, nuclear medicine labs, research labs, and so on.

(X ray or γ ray) radiation, charged particle (proton, beta particle) radiation, neutron radiation, or mixtures thereof? Spectroscopy measurements can determine the types and energy distributions, but often measurements require simpler detection or measurement devices. The radiation environment may be characterized by the maximum energy of radiation, whether the radiation source is continuous, as with an X-ray unit, rapidly pulsed as with some linear accelerators, or is random decay from a radioisotope. Is the measurement made in the primary direct beam, or in scattered radiation beam filtered by radiation barriers? Choice of instruments depends on why the radiation is being measured. It is desirable to know the approximate magnitude of radiation, expressed in some appropriate units, and the approximate energy of the radiation. It is then possible to estimate personnel equivalent dose rates in the radiation field. Multiple instruments with different features may be required to properly characterize and measure a radiation environment.

TYPES OF RADIATION

Radiation is a general term used to describe the emission and propagation of energy through space or a material. Texts describing instruments describe types of radiation (5–8). Mohr and Taylor, on behalf of The Committee on Data for Science and Technology, updated, through 2002, the numerical (*Note*: In the interest of better science, we present exact values as they are not widely published nor readily available!) parameters of common radiation types (9). Radiation may be classified as directly ionizing, indirectly ionizing, or nonionizing (e.g., microwaves, laser lights, ultrasound, not further discussed here). Particulate forms of radiation (protons, electrons) possess one unit of electrical charge (160.217653 zC is the unit of electrical charge of an electron) and directly ionize atoms and molecules, as do particulate radiations with multiple charges, such as alpha particles.

Indirecty ionizing forms of radiation (X rays, γ rays, neutrons) lack electrical charge, but interact with matter and produce secondary charged particles (electrons, positrons) that ionize atoms. The magnitude of the energy possessed by the radiation, frequently expressed in millions of electron volts (MeV) and the mass of particulate radiation, expressed in atomic mass units (1 amu is defined as 1/12 of the mass of the carbon atom, and equals 0.00166053886 ykg) are important physical parameters that arise in describing the properties of radiation.

Protons have a mass of 1.007276446 amu, possess one unit of positive charge and are one of the core particles of the nucleus. Protons are heavy charged particles, that lose energy mostly by ionization and excitation of atoms as they exert electromagnetic forces on the orbital electrons surrounding the nucleus. Loosing only a small fraction of energy during each interaction, protons move through matter mostly in a straight-line path leaving in their wake ionized or excited atoms. Heavy charged particles require great energy to penetrate tissue. A 10 MeV proton has a range of ~ 0.11 cm, while a 100 MeV proton has a range of ~ 7.5 cm. Protons with several tens of million electronvolts of energy are used at research facilities with particle accelerators as probes to study nuclear structure. Neutrons with a mass of 1.008664915 amu are slightly more massive than protons, but lack charge and interact in matter primarily by collisions with protons to which they impart a portion of their energy during the collision. Neutrons are generally classified as thermal if their energies are < 0.5 eV, intermediate if their energies are > 0.5 eV, but < 0.2 MeV, and fast if it is > 0.2 MeV, but < 20 MeV. Lacking charge, neutrons are generally more penetrating in tissue than protons of the same energy and interact with atoms by elastic and inelastic collisions. Numerous radioactive sources serve as neutron generators; an alpha particle source, such as ^{241}Am, may be mixed with a light metal, such as beryllium to produce neutrons by a (α, n) reaction. Nuclear reactors are prolific neutron generators and numerous research facilities have accelerators capable of producing neutrons. Alpha particles, usually with several million electronvolts of energy, consist of two protons and two neutrons, and appear when certain nuclides decay into more stable nuclides, such as the decay of ^{238}U to ^{234}Th or the decay of ^{226}Ra and certain daughters. While large mass and double charge prevents even the most energetic alpha particles from penetrating much beyond the most superficial layer of external tissue, alpha particles are hazardous when ingested into the sensitive epithelium of the lungs.

Electrons have a mass of 0.00054857990945 amu, a small fraction of the mass of a proton, but carry an equal quantity of negative charge. Electrons with several tens of million electronvolts of energy can be generated with electron linear accelerators and many other pieces of equipment are capable of generating less energetic, but still hazardous electrons. Electrons interacting in a material can also produce a spectrum of bremsstrahlung X rays with the maximum X-ray energy identical to the maximum energy of the electrons. These X rays are far more penetrating that the electrons. Beta particles with several million electronvolts of energy arise from the nucleus during certain radioactive decay processes. Negative beta particles, negatrons, or ordinary electrons, possess a spectrum of energies below their maximum energy. Many nuclear transformations yield multiple beta particles; a few, for example, the decay of ^{32}P to ^{32}S, yield a single negative beta particle. Positrons have the same mass as electrons, but carrying a positive charge and arise in certain radionuclide transformations, such as the decay of ^{22}Na to ^{22}Ne. Beta particles and positrons with several million electronvolts of energy are more penetrating than alpha particles and even minute quantities of radioisotopes producing these particles, such as ^{32}P, can potentially produce damaging skin burns if spilled on the skin and left unattended.

Gamma rays frequently arise when a daughter radioactive nuclide in an excited state, decays by beta particle decay, or by other modes of decay to form a more stable nuclide, such as the decay of ^{60}Co to ^{60}Ni, yielding 1.17 and 1.33 MeV γ rays. These electromagnetic rays can possess several million electronvolts of energy and, lacking charge and mass, the more energetic γ rays can penetrate deeply into tissue and other materials. Following their interaction in a medium, such as tissue, they generate ionizing secondary electrons that actually produce the damage to cells.

X rays are a form of electromagnetic radiation arising from changes in the arrangements in the orbital electrons surrounding the nucleus, yielding characteristic X rays of several tens of kiloelectronvolts (keV) of energy. Another form of X rays, bremsstrahlung, are produced when energetic electrons with energy of several million electronvolts strike high Z targets yielding X rays with very high energies. Hence, X rays can span a broad energy range, from a few fractions of million electronvolts to several tens of million electronvolts depending on how they are produced. Like γ rays, the most energetic X rays have great potential to deeply penetrate matter. The term photon is used to describe X rays, γ rays, or other form of electromagnetic energy without referring to the method of production or source of the radiation.

Nuclei of atoms, such as a deuterium, ^{2}H, may be accelerated in highly energetic linear accelerators as a probe to study the properties of the nucleus of various elements. Because of the great mass and charge, heavy charged particles must possess several tens of million electronvolts of energy to penetrate tissue.

In addition to this limited list of types of radiation, numerous radioactive isotopes of the elements, such as ^{60}Co, ^{137}Cs, ^{131}I, and ^{125}I, and many more are widely used in medicine for a host of applications. Radioisotopes, through decay, can produce alpha particles, γ rays, beta particles, positrons, and X rays and each radioisotope has a unique spectrum of radiation that allows it to be identified even in the presence of other radioisotopes. While many forms of radioactive materials are encapsulated solids, others are unsealed and as liquids or as gases, are more readily dispersed during accidental releases. Hence, radiation is a term used to refer to many different forms of particulate and nonparticulate radiations with energies from fractions of a million electronvolt to tens and hundreds of million electronvolts. Obviously, the means of detecting radiation must be specific for the types of radiations present in a specific locale.

SOURCES OF RADIATION

The most energetic X rays, γ rays, and heavily charged particles are found almost exclusively in government or university sponsored scientific accelerator research facilities. Photons (X and γ rays) with energies as high as several million electronvolts are the most prevalent forms of radiation as equipment yielding X and γ ray are widely used in medical facilities. X-ray imaging in hospitals is probably the single most common medical use of radiation, with diagnostic use of radiopharmaceuticals next in importance. Beta particle sources and electron producing equipment are the next most prevalent sources of radiation. Neutron producing sources and equipment are frequently found in university research laboratories but are less common in medical facilities.

RADIATION QUANTITIES AND UNITS

Radiation protection definitions and terms often lack clear meanings, as noted by Storm and Watson (10). International commissions make recommendations, but national councils and regulatory bodies in different countries (or even within the same country) adopt or apply the recommendations differently (11). As radiation quantities and units, Table 4, are used on instrument displays, users must understand both historical and newly adopted radiation units. Gollnick offers a useful review (8). We limit this discussion to popular historical quantities and units and provided brief, albeit, limited descriptions of current quantities and units recommended by the International Commission on Radiological Protection, Systeme International

Table 4. International Radiation Concepts and Units[a]

Concept	Quantity	Symbol	SI unit	Numerical Value	Relationship to Other Concepts
Ionization of air by X and γ rays	Exposure	X	None	1 R = 0.000258 C of charge released per kilogram of air	
Kinetic energy released per unit mass of material	kerma (collisional)	$K_{\mathrm{air}}^{\mathrm{col}}$	Sv	1 Gy of energy transfered per kilogram of material	$K_{\mathrm{air}}^{\mathrm{col}} = XW/e$
Absorption of energy in a material	Absorbed dose	D	Gy	1 Gy = 1 J of energy absorbed per kilogram of material	$D = X f_{\mathrm{med}}$
Risk of biological energy for different forms of radiation	Radiation weighting factor[b,c]	W_{R}	None	X rays, γ rays, beta particles = 1; Thermal neutrons, high energy particles = 5; Alpha particles, fast neutrons = 20	
Equivalent biological effect in humans	Equivalent dose[d]	H_{T}	Sv		$H_{\mathrm{T}} = \Sigma W_{\mathrm{R}} D$
Total (50 y) cummulative dose to an organ for internal radiation	Committed[d] equivalent dose	$H_{\mathrm{T}}(50)$	Sv		
Reduced risk of partial body exposure to radiation	Tissue weighting factor[c]	W_{T}	None	Skin, bone surface = 0.01; bladder, liver = 0.05 colon; stomach = 0.12 gonads = 0.20	
Sum of weighted equivalent doses of partial body exposures	Effective dose[d]	E	Sv		$E = \Sigma W_{\mathrm{T}} H_{\mathrm{T}}$
Sum of weighted total (50 y) cummulative doses to organs from internal radiation	Equivalent dose	E (50)	Sv		$E(50) = \Sigma W_{\mathrm{T}} H_{\mathrm{T}}(50)$

[a] For complete concepts, definitions, and descriptions, see Refs. 11, 13 and 14.

[b] The equivalent concept, Q, albiet with different values, is used in the United States by the National Council on Radiation Protection Units and the Nuclear Regulatory Commission.

[c] For complete list of W_{R}, W_{T}, see Refs. 11–13.

[d] Similar, but different nomenclature is used in the United States by the National Council on Radiation Protection and Units and The Nuclear Regulatory Commission.

d' Unites, and the International Commission on Radiological Units and Measurements (11–13).

Counts (events) or count rates (events per unit time) are denoted on instruments that detect the presence and relative magnitudes of radiation. Count rates of a few counts per minute (cpm) to millions of cpm are possible, depending on the radiation field intensity.

Exposure, denoted by the symbol, X, is the measure of the ability of X and γ rays of energies 10 keV to < 3 MeV to ionize air and is the quotient of $\Delta Q/\Delta m$, where ΔQ is the sum of all charges of one sign produced in air when all of the electrons liberated by photons in a mass Δm of air are completely stopped. Exposure is expressed in a special unit, the Roentgen (R), equal to 258 μC of charge per kilogram of air at standard temperature and pressure. As with count rates, exposure rates generally vary from $\sim 5\ \mu\mathrm{R{\cdot}h^{-1}}$ associated with natural background radiation to nearly $1\ \mathrm{MR{\cdot}h^{-1}}$ in a linear accelerator X-ray beam. Because of its historical use in science and medicine, exposure remains popular even though its continued use is not recommended in the Systeme International d' Unites (14).

Kinetic energy released per unit mass of material, denoted in lower case, by the acronym, kerma, is a measure for indirectly ionizing radiations (photon, neutron) interacting in a material, of the total kinetic energy of all charged particles released per unit mass of material. Kerma possesses a collision component from the kinetic energy imparted by inelastic collisions with electrons and a radiation component (usually much smaller) from interactions with nuclei. For X rays absorbed in air, collision kerma in air is the product of exposure and the average energy, $\overline{W}$, required to produce an ion pair per unit electrical charge. Kerma is measured in gray (Gy), which is one joule (J) of energy released per kilogram (kg) of the medium. Turner offers a complete description of kerma and its relationship to other quantities (7).

Absorbed dose, denoted by the symbol D, is a measure of the amount of the released energy that is absorbed in the medium per unit mass of the medium. Under conditions of charged-particle equilibrium, with negligible energy loss, absorbed dose equals kerma. One joule of energy absorbed per kilogram of the medium, the gray, is widely used in medicine as a measure of absorbed dose, as is the submultiple, the centigray (cGy), 1/100th of a gray. The older special unit (now abandoned) of absorbed dose, rad, an acronym for roentgen absorbed dose, represents 0.01 J of energy absorbed per kilogram of the medium (1 Gy = 100 rad).

The absorbed dose in a medium may be determined from the exposure in air at the same point in the medium by multiplying the exposure by a conversion factor, f_{med} that converts the exposure in air to dose in the medium. The factor, f_{med}, is slightly < 1 for most biological materials, except bone, where values as high as 3 occur for the soft X rays in the diagnostic energy range.

Different types of radiation produce different degrees of biological damage when the same amount of energy per unit mass is deposited in the biological system. The radiation weighting factor, w_{R}, which replaced an older similar concept, quality factor, Q (now abandoned) is a measure of this phenomena (11). Radiation weighting factors of unity, Table 4, are assigned to most electrons, X and γ rays, while factors as high as 20 are assigned to alpha particles and fast neutrons. Hence, the biologically equivalent effect in tissue of the absorption of 1 Gy of alpha particles is 20 times more severe than the absorption of 1 Gy of 1 MeV γ rays.

Equivalent dose, usually denoted by the symbol H_{T}, is the term used in radiation control programs to monitor and record the biological equivalency of exposure to amounts of radiations of different energies that an individual has received. The equivalent dose in sievert (Sv), the product of the absorbed dose, D, in gray and the radiation weighting factor, w_{R}, is commonly used to express the biologically equivalency of absorbed doses of particular types and energies of radiation. Historically, this equivalency was expressed in the special unit, rem (now abandoned) an acronym for roentgen equivalent human. As the f_{med} factor and radiation weighting factors, w_{R}, ~ 1 for most photon energies commonly encountered in many situations, the units for kerma (Gy), absorbed dose (Gy), and equivalent dose (Sv) are nearly numerically equivalent and are often used interchangeably, as were the older abandoned special units, R, rad, and rem.Table 4 summarizes these relationships.

Committed equivalent dose, usually denoted by the symbol $H_{\mathrm{T}}(50)$, in sieverts, is employed to consider exposure within the body from internally deposited radioactivity. It represents the total cumulative dose delivered over 50 years to an organ system by ingested radioactivity.

Effective dose, usually denoted by the symbol E, considers the consequences of partial body radiation exposure (11). Tissue weighting factors, w_{T}, account for the reduced effects that occur when only a portion (or organ system) of the body is irradiated. Values for w_{T} (Table 4) range from 0.01 for bone surfaces to 0.2 for the gonads. The effective dose, E, in sieverts, is the sum, over the body, of the products, $w_{\mathrm{T}} H_{\mathrm{T}}$ for each partially irradiated portion of the body.

Committed effective dose, usually denoted by the symbol, $E(50)$, in sieverts, is employed to consider exposure from internally deposited radioactivity. The committed effective dose, $E(50)$, is the sum, over the body of the products, $w_{\mathrm{T}}\ H_{\mathrm{T}}(50)$, for each partially irradiated portion of the body > 50 years.

For individuals experiencing both external and internal radiation exposure, methodologies exist (omitted here) to combine effective dose and committed effective dose to estimate cumulative risks from both types of exposures (8).

Activity, denoted by the symbol A, describes an amount of radioactivity, expressed in decays per second (dps). One becquerel (Bq) equals one decay per second. The curie (Ci), the original term used to describe an amount of activity, equals 37 Gdps. Trace amounts of radioactivity are generally expressed in the microcuries (μCi) quantities and laboratory cleanliness standards are often expressed in picocuries (pCi) or smaller amounts. One becquerel equals ~ 27 pCi. Activities of millions of curies are commonly found in power reactors while curie and millicurie (mCi) quantities of materials are commonly used in medical applications.

COMMON FEATURES OF INSTRUMENTS

Radiation protection instruments in a facility can be broadly categorized as those (e.g., area monitors, personnel scanners) used in fixed locations and those (e.g., GM detectors, survey meters) moved for use in multiple locations. Fixed instruments usually are large, heavy, and use permanent electrically power. By design, they may have more features and offer more sensitive detection or measurement features than their portable counterparts. Portable instruments generally are small, lighter, and battery powered. Some feature tripod stands for temporary uses of long duration. Instruments may be categorized as those that detect or measure a quantity of radiation and those that grossly or specifically identify types of radiation or radionuclides. Instruments of all types usually feature multiple signal ranges because of the wide variations in the signal (counts, exposure, dose, etc.) monitored. The number of photons or particulate forms of radiation from a sizable radioactive source or large piece of equipment frequently varies inversely (or approximately so) as the square of the distance of the finite size detector from the radiation source and multiple signal ranges, usually in multiples of 3 or 10, are required to map the spatial variation of the radiation around the source or equipment that will vary from the highest signals measured very near the source to the smallest signal measured at distances far from the source of radiation. Current devices often feature auto-zeroing and auto-ranging of scales. Instrument scales indicate the magnitude or measure of the radiation detected. Often, more than one unit will appear on the scales, or users may electronically select from multiple choices of units. Some instruments feature a rate mode, usually per second, minute, or hour or integrate mode that sums the signals over some predetermined time periods. Instrument scales may only be correct for specific radiation conditions under which the instrument was calibrated. Instruments can yield incorrect readings when used under noncalibrated radiation conditions.

Efficiency of a device is a measure of the number of output parameters (counts, pulses, etc.) produced relative to the number of input parameters (γ rays, particles, etc.) producing the output. Efficiencies (absolute, intrinsic, relative, etc.) have technical definitions beyond the scope of this article. Different applications require instruments with different degrees of efficiency, but, generally, high efficiency is desirable.

Signal accuracy is highly variable and depends on the type of radiation monitored and the design of the instrument. The more intense and hazardous the radiation field the greater the necessity for more accurate measurements. Conversely, measurements in very low level radiation fields need not be as accurate as the risk presented to personnel is roughly proportionately less. For example, ionization chambers used as survey meters will be calibrated to be accurate to 10% at the one-third and two-thirds of full-scale deflection while GM counters may only be 50% accurate over their full-scale range. An instrument should provide precise reproducible readings. Instruments need to reproducibly repeat measurements at the same locations when used repeatedly in the same radiation fields. Portable instruments often feature rugged weatherproof designs with lightweight features, such as ergonomic antifatigue handles to facilitate outdoor use for long times. Individuals with various skill levels frequently use instruments; some personnel use instruments infrequently. In both situations, instruments subsequently suffer some misuse and abuse. Hence, simplicity of use is a highly desirable feature. Instruments should be designed with a minimum of controls or knobs to be adjusted, On, Off, and Battery Check switch positions must be clearly indicated and any scale selection switches should be labeled in an unambiguous manner. Some instruments feature audible signals whose intensity is proportional to the magnitude of the radiation signals. Such a feature often is useful on the most sensitive scale, but may be undesirable on the higher ranges; hence, usually the audible signal is switch selectable and may be turned off when desired. Potentiometer controls necessary during calibration adjustments and voltage setting control should not be so accessible that they can be easily changed. Such controls often are recessed or located on a rear panel so that they can only be changed in a deliberate manner. While every instrument will not possess all of the features discussed here, this discussion has included those found most commonly.

Compact, lightweight designs are easily achieved using microprocessors, liquid-crystal displays, modern CMOS electronics, and phenol, or acrylonitrile–butadiene–styrene (ABS) plastic cases. Some devices feature automatically backlight displays in low ambient light conditions. Current instruments frequently are available with either digital scales or analog scales; some offer both displaying a digital reading and an analogue bar graph emulating analogue meter movement. While digital scales are usually easily read in a constant radiation environment, they are inappropriate in rapidly changing radiation fields as the signal changes rapidly and rapidly changing digital readouts are difficult to read and interpret. For these situations, a freeze mode indicates a peak reading.

The radiation detector may physically be in the instrument with the associated electronics necessary to process the signals, connected to the counting electronics by a cable, or feature a remote reading capability, allowing the observer to stay in an area where radiation exposure is minimal. A cable connection is common in applications where the observer is physically in the radiation field monitored. Cable connectors, with instrument displays of the probe connected, allow the use of multiple probes or detectors (GM tubes, neutron probes, proportional counters, scintillation probes) with different features with a single count rate meter. Instruments with built-in detectors are free of cable problems, such as the loss of charge by poor cable insulation. Many devices feature an RS-232 interface or a universal serial bus (USB) cable that can connect to a computer; data software packages allow data retrieval, time–date stamps, or use parameter selections, such as programmable flashing displays and audible alarms, or measurements for specific applications. Some instruments feature data logging, the sequential capturing of hundreds or thousands of data points under different measurement conditions. Data captured by the instrument can be downloaded, by cable connection or by via infrared (IR) communication, to a personal computer for processing with a numerical spreadsheet (3).

Generally, instruments feature a battery check (a known scale deflection on an analog device or a brief audible tone or indicator lamp on a digital device) that allows the user to determine that the battery possesses enough charge to operate the instrument successfully. Voltage stability is important; many instruments feature two power supplies, one for the counting electronics and another for the constant voltage required across the detecting volume. Currently, multiple 9 V alkaline batteries are widely used, providing 100–500 h of operation. The voltage across the detecting volume is usually required to be the most stable. A small variation in this voltage can lead to large changes in the observed signal, depending on the design and mode of operation of the instrument. Voltage stability of 0.1% is usually required for the voltage across the detector in most radiation detection instruments.

A zero check allows the zero point on the scale, previously set to zero in a radiation free environment, to be checked in the presence of radiation. Some devices feature auto-zeroing scales. Some instruments have an attached constancy source, a minute quantity of radioactive material, such as 0.06 μCi of ^{238}U or 10 μCi of ^{137}Cs, which, when placed on or near the detector in a predetermined geometry, yields a predetermined signal on the most sensitive range. Proper instrument use requires all three items, battery function, zero point, and known response, to be checked prior to each use. A reduced signal or complete loss of signal from the radiation protection instruments is particularly dangerous because the user falsely concludes that little or no radiation is present.

Radiation detectors generally are designed either to monitor individual events (counts) or pulses or to integrate (sum) counts or pulses that occur in such a short time interval that they cannot be electrically separated. In pulse mode, individual events or signals are resolved in 1 μs, 1 ns, or even smaller time intervals. In integrate mode, the quantity measured is the average of many individual events in some very short time period. Some devices feature signal integration when the device is used in a rate mode.

Response time of an instrument measures how rapidly an instrument responds to the radiation detected. The response time is short (fractions of a second) on the higher multiple scales and becomes longer (several seconds) as the scale multiple decreases with the longest resolving times occurring on the most sensitive scale. The response time (*T*) is called the time constant, and is proportional to the product of the resistance (*R*) of the electronic counting circuitry and its capacitance (*C*). (Fig. 1). Many instruments feature a slow response switch that allows electronic averaging of a rapidly varying scale signals.

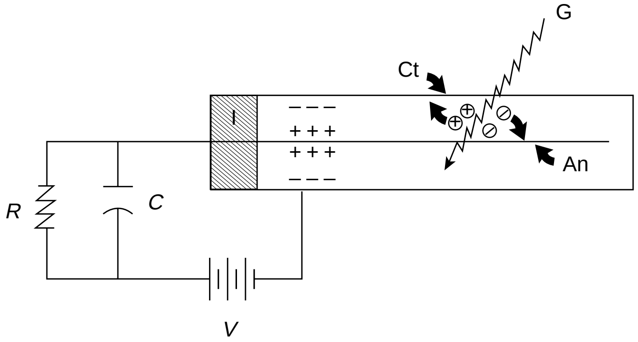


Figure 1. Simple schematic of a gas-filled ion chamber. A voltage (*V*) is maintained across the central wire anode (An) and chamber wall cathode (Ct). An incident (γ-ray (*G*) produces ion pairs; they move to the anode and cathode producing a pulse in the circuit containing a resistor (*R*) and capacitor (*C*).

Energy independence is desired for most radiation survey instruments, such as ionization chambers and GM counters; the signal is independent of the energy of radiation detected, but is proportional to the magnitude (counts, exposure, etc.) of the radiation field being monitored. However, many instruments exhibited a marked energy dependency at lower X or γ ray energies; the signal varies as the energy of the radiation varies even a constant magnitude radiation field (Fig. 2). Knowledge of the energy dependency of an instrument and of the approximate energy of the radiation field to be monitored is essential in properly using a radiation detector. Whether the signal from the instrument is higher or lower than it should be relative to the signal observed at its calibration energy depends on many parameters. Calibration of radiation detectors is required annually for some regulatory agencies; instruments usually display a calibration sticker indicating the most recent calibration date, the calibration source, or sources if several were used to obtain energy response of the instrument, the scale readings obtained (often in mSv·h^{-1}, mR·h^{-1}, or other multiples, or cpm), and the accuracy and precision of those readings, expressed as a percentage of the scale reading, any necessary scale correction factors to be used specific scales, and instrument response to a reference source containing a minute quantity of radioactive material. By performing a battery check

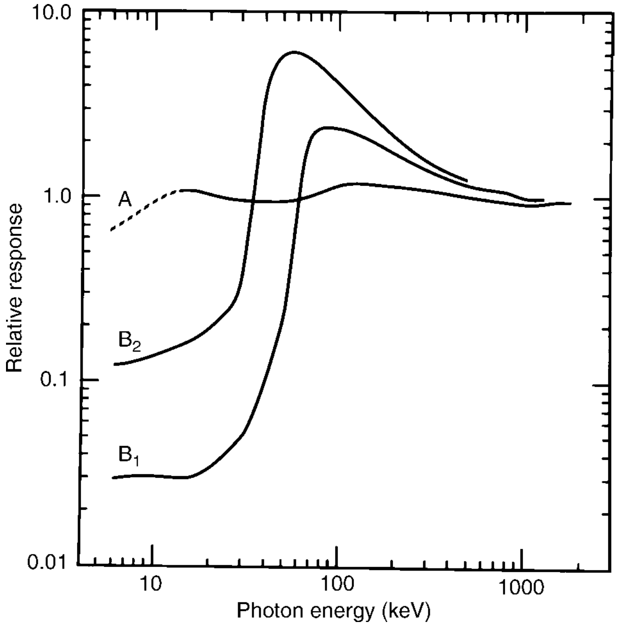


Figure 2. Typical relative response versus incident photon energy (kV): (A) is the ionization chamber; (B$_1$) is a Geiger counter with thin window shield closed. (B$_2$) is a Geiger counter with thin window shield open.

or by observing that the battery condition is acceptable by the absence of a low battery indicator, a zero-scale reading check, and meter response check using a reference source, the user can determine that the instrument is operating correctly before use. Moreover, the average energy of radiation at a particular location in an area is highly dependent on the relative amounts of primary and scattered radiation present and frequently varies within an area. Energy dependency is, of course, advantageous when it is desirable to measure the energy spectrum in addition to determine the intensity of radiation.

Some instruments are environmentally sensitive; graphs or tables providing correction factors as a function of temperature, pressure, and humidity indicate the degree to which the signal is altered by environmental conditions. For instruments with the detector volume open to the air, corrections based on thermodynamic gas laws for the mass of air present in the detector are employed.

Excessive humidity can cause incorrect instrument readings. Humidity can cause leakage of current in cables, at electrical contacts, and at other locations in the electronic circuitry. Usually, an instrument will require a warm-up time of 1–2 min or longer. Proper warm-up allows electronic circuitry to stabilize and yields more stable and reproducible signal readings.

Strong radio frequency (RF) fields associated with some equipment generating radiation can cause improper signals in some radiation measurement instruments. The susceptibility of the instrument to strong rf fields will usually be discussed in the user's manual.

Many radiation protection instruments exhibit geotropism, orientation (gravitational) dependency, or angular dependency because radiation incident from the sides and rear of the instrument are attenuated more by metal casing surrounding the counting electronics than radiation incident on the sensitive detecting volume. The proper orientation of the instrument for measurement in a radiation field and the degree of angular response will be indicated in the users manual or on the calibration certificate.

As previously noted, some devices are designed to identify different types of radiation or to identify species of radionuclides. Many detectors will feature a thin detection window of only a few milligrams per square centimeters of thickness protected by a thicker filter, a sliding, or rotating shield, that allows the least energetic forms of radiation, such as soft X rays and alpha and beta particles to be detected through the thin window when the thicker filter is removed. Conversely, with the shield in place alpha and beta particles are discriminated against and only higher energy radiations are detected. Other windowless detectors are designed to detect the low energy radiation by flowing a radioactive gas through the detector.

Radionuclide identification requires spectroscopy, the identification of the characteristic radiation spectra of multiple radionuclides each in the presence of others. Resolution is the measure of the abilities of devices to distinguish a single energy in multiple energy radiation spectra. Different applications require instruments with different degrees of resolution. Spectroscopy formerly was limited to using heavy fixed laboratory-based NaI(Tl) detectors or Ge(Li) detectors often with a multichannel analyzer to quantify and identify radionuclides in test samples. Currently, in-field spectroscopy can be performed with small handheld or portable NaI(Tl)-based detectors or with portable high purity germanium (HPGe) detectors that allow radionuclide identification of the most common radionuclides.

IONIZATION CHAMBERS

The ionization chamber (Fig. 1) consists of a cavity, frequently cylindrical, with a positively charged central electrode (anode) insulated from the chamber walls (cathode) at negative potential. The direct reading pocket dosimeter, with an external dosimeter charger (Fig. 3) is a simple ionization chamber. When fully charged, an internal quartz fiber, visible under a magnification lens, is deflected to a "zero" reading. As the dosimeter is irradiated, the charge is reduced proportionally to the amount of radiation received. Older pocket chambers are being replaced with personnel detectors or monitors with more electronic versatility. As a survey meter, an external power source (Fig. 1) provides the voltage potential; a resistor and capacitor in parallel (or equivalent electronic circuitry) are used to collect the charge produced when ionization occurs in the chamber. The ionization chamber electrode polarity may be reversed for special applications. Incident X or γ rays interact in air or a tissue equivalent gas, producing positive and negative ions in the chamber. If the voltage is sufficiently high to prevent recombination, that is, the positive and negative ions rejoining before they reach the charged surfaces, the negative ions will be attracted to the central electrode and the positive ions will be collected on the chamber wall. The collected charge flows to the capacitor and one electronic pulse is detected in the counting circuitry. In open air chambers, the filling gas is air at ambient temperature and pressure and appropriate corrections to the charge collected may be required as previously discussed.

Historically, ionization chambers were designed to measure exposure (*R*); newer instruments may offer equivalent dose readings (Sv or their submultiples) (Fig. 4). The walls of the chamber must be sufficiently thick for electronic equilibrium to be established, that is, the number of electrons entering and leaving the cavity is the same and the chamber walls are sufficiently thick to stop any electrons arising from the interaction of the radiations with the gas or in the chamber walls. Moreover, the chamber walls are usually designed of air equivalent materials. Sealed ionization chambers may be filled with a tissue equivalent gas and usually are designed to measure collision air kerma. The chamber size must be small relative to the dimensions of the irradiating beam so the chamber is uniformly irradiated. Typical ionizing voltages required across the sensitive detecting volume are $\sim$ 150–300 V (Fig. 5), sufficiently high to present recombination of the positive and negative ions, but not high enough to cause additional ionizations that amplify the signal. Ion chamber currents are low, usually 1 pA or 1 fA. A 10 $\mathrm{mSv{\cdot}h^{-1}}$ γ ray field yields $\sim$ 1 pA; extraneous currents must be minimized in order to

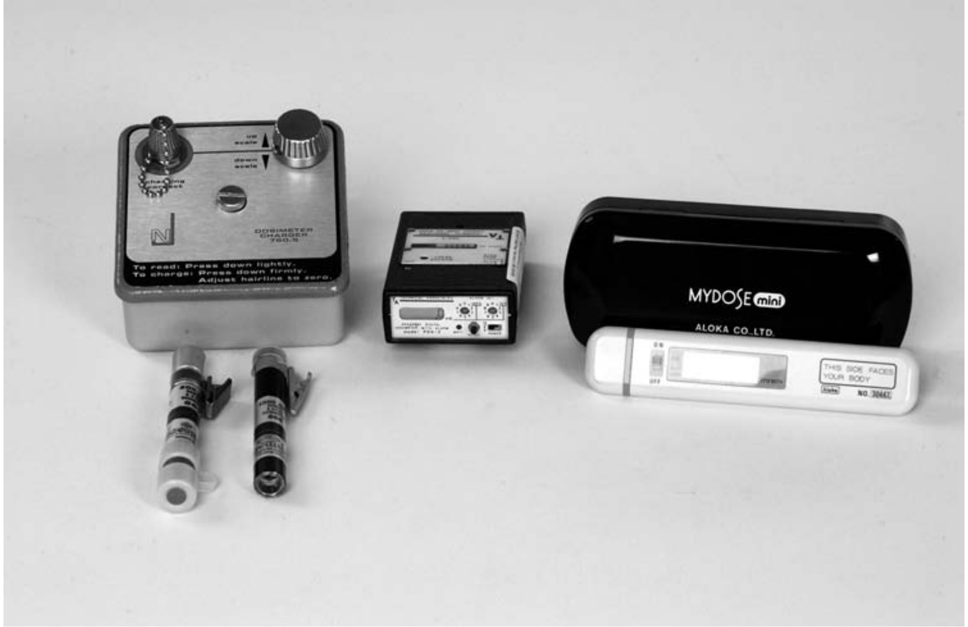

DOSIMETER CHARGER 750-5
To read: Press down lightly.
To charge: Press down firmly.
Adjust hairline to zero.
MYDOSE mini
ALOKA CO.,LTD.
THIS SIDE FACES YOUR BODY

Figure 3. Personnel dosimeters. Low dose (bottom left; Dosimeter Corp., Model 862;) and high dose (bottom right; Dosimeter Corp., Model 866) γ- and X-ray pocket dosimeters, with charger (top left; Jordan Nuclear Co, Model 750-5). An alarming personal digital dose meter (center; Technical Associates, Model PDA-2) and a miniature pocket digital dosimeter (right; Aloka Co, LTD., Model MYDOSE-mini).

measure such a small current. Electrical leakages can occur across lint, dust, or loose conductive materials between interior conducting surfaces. Cable connections can exhibit greater leakage current in high humidity. A guard ring design in some chambers minimizes leakage and polarization currents that arise after the collection potential is initially turned on. The insulator between the outer and inner electrodes is divided into two segments with the conductive guard ring in between; any leakage current through the insulator is collected and prevented from contributing to the true current. The currents from ionization chambers are normally measured using a potential drop across a high resistor or a rate of charge method. The small currents from the ion chamber are amplified by a vibrating reed electrometer. The amplified

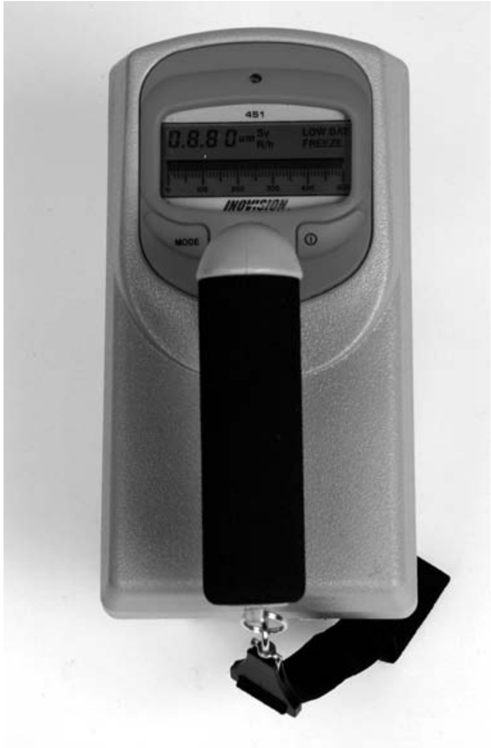


Figure 4. A portable ionization chamber survey meter (Cardinal Health; Inovision, Model 451). The display shows the features during the initial check phase immediately after turning on the instrument.

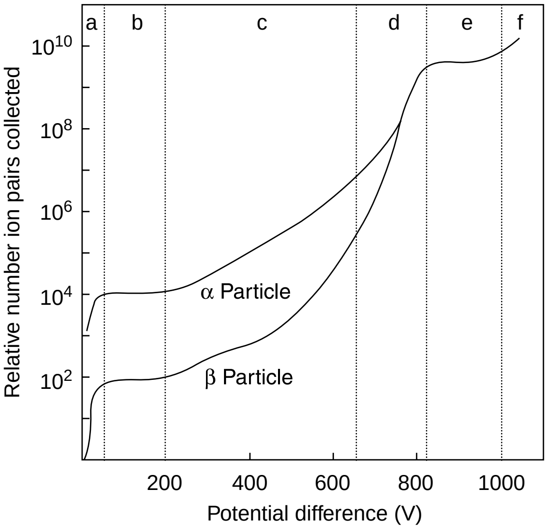


Figure 5. Voltage dependence of a gas-filled cylindrical ionization chamber: (a) Voltage is insufficient to prevent ion recombination. (b) Ionization chamber voltages are sufficient to prevent recombination. (c) Proportional counter voltages, the number of secondary ion pairs is proportional to the number of primary ion pairs. (d) Limited proportionality region. (e) Geiger voltages produce maximum number of ion pairs from a single primary ion pair. (f)Continuous discharge region.

current passes through a precision resister and the voltage drop across the resistor is proportional to the current. If collected on a capacitor, the rate of charge collected on the capacitor is proportional to the current. This latter method is used for smaller current measurements while the former is used in ionization chambers designed for more rugged use.

Ionization chambers normally exhibit good energy independence (Fig. 2) over a large energy range, and this makes them useful for measurement of X or γ rays with energies from $\sim$ 7–1000 keV and for measuring low (1 μSv$\cdot$h^{-1}) to high (10 mSv$\cdot$h^{-1} or higher multiples) dose rates. Open-air ionization chambers are less useful for low dose rates of < 1 μSv$\cdot$h^{-1}. Pressurized (up to 8 atm) ionization chambers allow accurate measurements < 1 μSv$\cdot$h^{-1}. Properly modified ionization chambers, with sliding shields, may be used to monitor alpha, beta, and neutron radiation. For example, an ion chamber with boron on its interior chamber wall or containing boron gas may utilize the high cross-section of boron for neutrons and the subsequent ^{10}B, ^{7}Li reaction, and the chamber will detect neutrons using the subsequent alpha particles from this reaction.

GAS PROPORTIONAL COUNTERS

Gas proportional counters (Fig. 6) have similar design features as ionization chambers, but employ higher voltages between the central electrode and the chamber walls. The typical operating voltages of 300 up to 1000 V (Fig. 5) are sufficiently high that, following an ionizing event in the chamber, the positive and negative ions generate additional ionizations so that the number of ions from the initial ionizing events are multiplied $\sim$ 1 thousand to 1 million times. The resulting signal is proportional to the energy deposited by the initial number of ionizing events. Proportional chambers can be used in either the pulse or integrate mode, but the pulse mode is used most commonly. They are capable of detecting individual ionizing events. Because of amplification, the current from proportional chambers is much higher than those from ionization chambers. As the signal from a proportional current is dependent on the operating voltage, a highly stable power supply is required. The choice of detector gas in thin-window proportional counters depends on the type of radiation to be detected. For counting alpha particles, helium or argon gas frequently is used. For counting beta particles, a high multiplication gas is required, such as methane (CH_4) or a mixture of a polyatomic gas and a rare gas, such as argon. The gasses also help make the proportionally of the chamber more independent of operating voltage. Proportional counters are generally cylindrical in shape and the central electrode is a very fine wire of uniform diameter as any variation in the electrode's diameter causes small variations in the resulting signal. Gas (often a mixture of 90% argon and 10% methane) flow proportional counters usually have a sample of the radioactive material flow through the chamber. Either 2π (180°) or 4π (360°) solid geometries are used, and these systems are very useful for counting low energy beta particles, alpha particles, or very low energy photons. Proportional counters are useful for spectroscopy (energy determination) measurements.

Proportional counters have the ability to discriminate between alpha and beta particles by discriminating between the magnitudes of the signals produced. Gas proportional counters may be used to measure fluence or absorbed dose.

When neutron spectra are poorly known, neutron rem meters are used to estimate the equivalent dose for fast neutrons. Older style neutron detectors consisted of a proportional counter either lined with boron or filled with boron trifluoride gas; the boron has a high capture cross-section thermal neutron detector. The subsequent charged

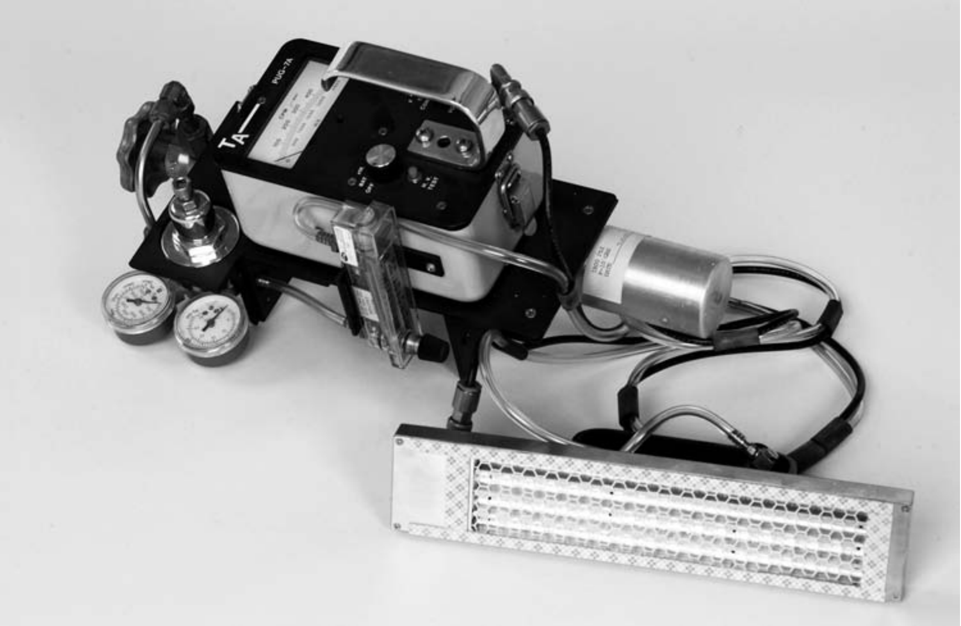


Figure 6. Portable gas proportional counter with alpha probe (Technical Associates, Model PUG-7A).

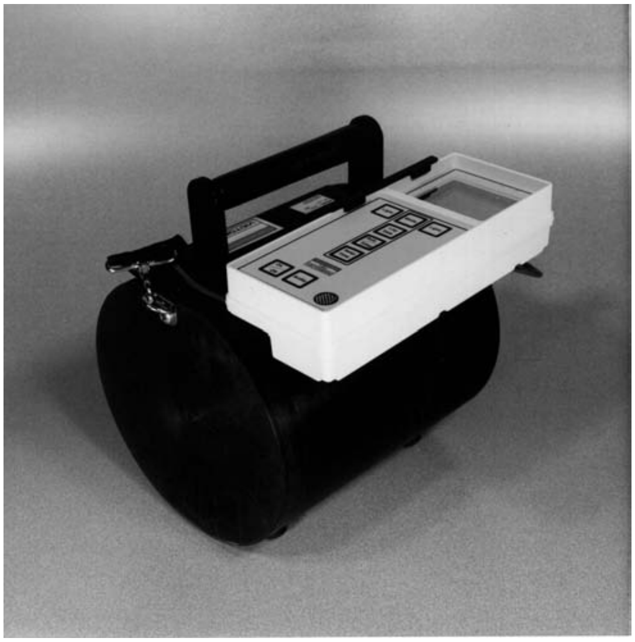


Figure 7. Portable neutron survey meter (Cardinal Health, Inovision, Model190N). (Courtesy Cardinal Health.)

particles (alpha particles) from this reaction are readily counted. Current neutron detectors use ^{3}He as the fill gas and detect both the proton and tritium, ^{3}H, from the subsequent reaction (Fig. 7). Fast neutrons can be moderated in several centimeters of high density polyethylene to thermalize the neutrons for detection by the methods described. Olsher et al. (15) described recent improvements in neutron rem meter instrumentation.

GEIGER–MÜELLER COUNTERS

If the voltage on an ion chamber is increased to $\sim$ 900–1000 V (Fig. 5), the proportionately exhibited at lower voltages is lost. Each initial radiation interaction in the walls or gas of the detector results in complete ionization of the gas in the detector. Interactions in the detector are spatially dependent, but generally, the following sequence occurs. Electrons produced following the initial ionizing event lose energy as they drift toward the anode. They lack enough energy to produce secondary ionization until they approach the anode when secondary ionization begins to occur. This secondary ionization builds up rapidly producing an avalanche of electrical charge in the detector. These processes reduce the potential difference between the central electrode and the chamber walls and the avalanche terminates. Once the necessary ionizing potential is reestablished, the detector is ready again. One undesirable aspect of the movement of the positive ions to the cathode and their resulting collisions with the cathode causes additional electrons to be ejected from the cathode. These additional electrons are undesired and may be controlled by manufacturing a tube containing a quenching and a filling gas. Organic quenching gases, such as ethanol or ethyl formate, are depleted by this process. An inorganic filling gas, such as chlorine, recombines to provide a continuous supply (8). The energy of the undesired electrons dissociates these organic molecules rather than starting new avalanches in the tube. The number of organic molecules available for quenching limits this method of quenching. An alternative method uses halogen gases, usually bromide or chlorine. The extra energy of these electrons is used to disassociate these halogen molecules. As opposed to the organic molecules, halogen reassociates so that the same atoms are available again to continue the process.

Geiger–Müeller tubes generally are used as pulse-type detectors of radiation. Their response is a function of the intensity of the radiation field. The movement of the positive ions to the cathode, previously described, requires from 100 to 200 μs and during this time interval the GM tube is unable to respond to other radiation interactions, have not recovered (reassociated) and are unable to resolve additional events (7). In very intense radiation fields, the relative long resolving times of GM tubes creates periods in which the tube is insensitive to radiation events; the GM tube may not respond to radiation, giving a false zero or low reading when an intense field is present. However, GM tubes are excellent as very sensitive detectors of X and γ rays in low level radiation fields. Commercial manufacturers offer at least three different GM tubes designs (Figs. 8,9) for specific applications. Pancake probes, with covers only a few milligrams per centimeter squared thick, allow the detection of alpha particles $>$ 3.5 MeV, beta particles $>$ 35 keV, and γ-rays $>$ 6 keV, while thin end window probes detect alpha particles $>$ 4 MeV, beta particles $>$ 70 keV, and γ rays $>$ 6 keV (Fig. 8). Some pancake-type probes (Fig. 9) feature removable tin and copper filters $\sim$ 3 mm thick that allow energy compensated exposure rates measurements. Energy-compensated GM probes feature a design that reduces response energy dependency so that it responds more like an ionization chamber (Fig. 5). Geiger–Müeller instruments are basically count rate meters, but may be calibrated in exposure rate for a specified energy of photons. Use of the meter in other energy spectrums different from the calibration spectrum invalidates the meter reading in μSv/h and mSv/h, but still the instrument allows the detection of radiation in the count rate mode.

SCINTILLATION DETECTORS

Luminescence is a physical process in which a substance, a scintillator, absorbs energy and then reemits the energy in the visible or near visible energy range. Prompt scintillators that deexcite in 10 ns following luminescence exhibit many useful properties as radiation detectors. For every photon or particle detected, a single pulse is normally counted and the size of the pulse generated is related to the energy deposited by the radiation interacting in the scintillator. Scintillators exhibit great sensitivity and yield high count rates. They can measure fluence, exposure, or absorbed dose if calibrated for the energy range of interest. Moreover, their exceptional sensitivity allows measurement of radiation rates at or near background levels, such as 1 $\text{nSv}\cdot\text{h}^{-1}$. Solid inorganic scintillators includes sodium

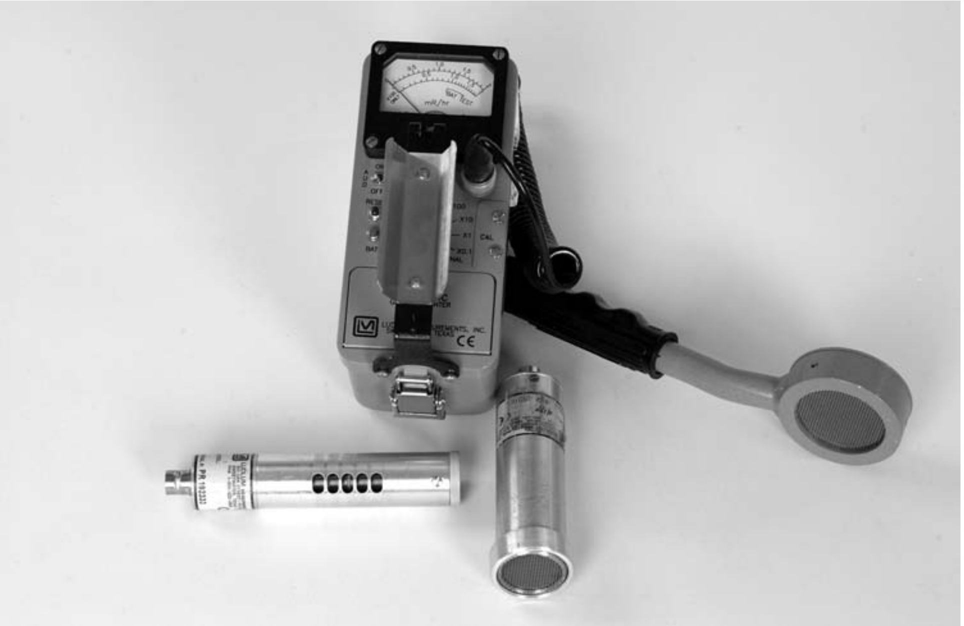


Figure 8. Portable GM counter (top; Ludlum Measurements, Inc. Model 14C) with an open side-window probe (left; Ludlum Measurements, Inc. Model 44-38), an end thin-window probe (center; Ludlum Measurements, Inc., Model 44-7), and a pancake probe (right; Ludlum Measurements, Inc., Model 44-9).

iodide crystals with trace amounts of thallium, NaI(Tl); cesium iodide with thallium, Csl(Tl); cesium fluoride, CsF; zinc sulfide with silver, ZnS(Ag); and bismuth germanium oxide, BiGeO, also known as BGO. The trace amounts of impurities in these inorganic salt crystals serves as luminescent process activators that promote the efficient conversion of the incident radiation energy into light. The scintillator crystal is connected to a photomultiplier by direct contact or through a light pipe. The crystal and photomultiplier must be encased in a light tight case to prevent light leaks. Typical crystals are cylindrical, ~ 1 in. (2.54 cm) diameter by 1 in. (2.54 cm) thick, but larger sizes (Fig. 10) are available for more sensitive measurements. The resulting photomultiplier signal is amplified by the associated counting electronics. Detectors with thin windows are available for the detection of low energy X rays and energetic beta particles. Inorganic solid crystals are relatively dense and reasonably efficient for detecting higher energy photons (Fig. 11). However, they are also hydroscopic and to protect them from absorbing moisture are encased in light reflecting cases that promote good efficiency. Organic crystal scintillators produce their light by a molecular process. Anthracene and transtiblene are the most widely used organic crystal scintillators. Incoming radiation excites electrons to higher energy levels of vibrational states; the electrons subsequentially decay with a release of energy. Organic liquid scintillators are formed by dissolving organic scintillators in liquid organic

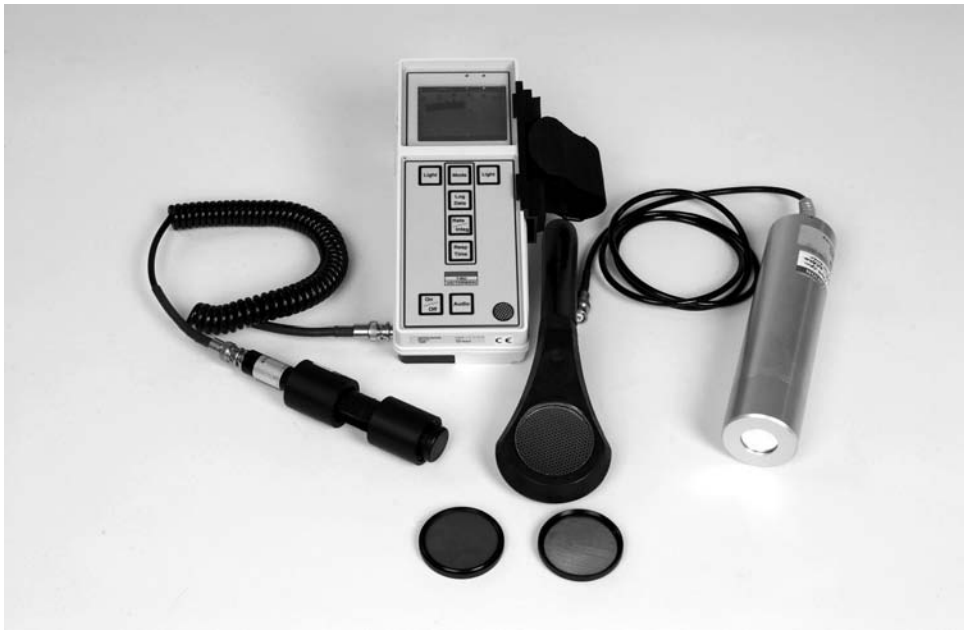


Figure 9. A counter (center; Cardinal Health; Victoreen, Model 190) with an energy-compensated sliding window GM probe (left; Cardinal Health; Inovision, Model 90-12), a pancake detector with filters (center; Cardinal Health; Inovision, Model 489-118FS), and a 1 × 1 in. NaI(Tl) detector (right; Cardinal Health; Inovision, Model 425-110).

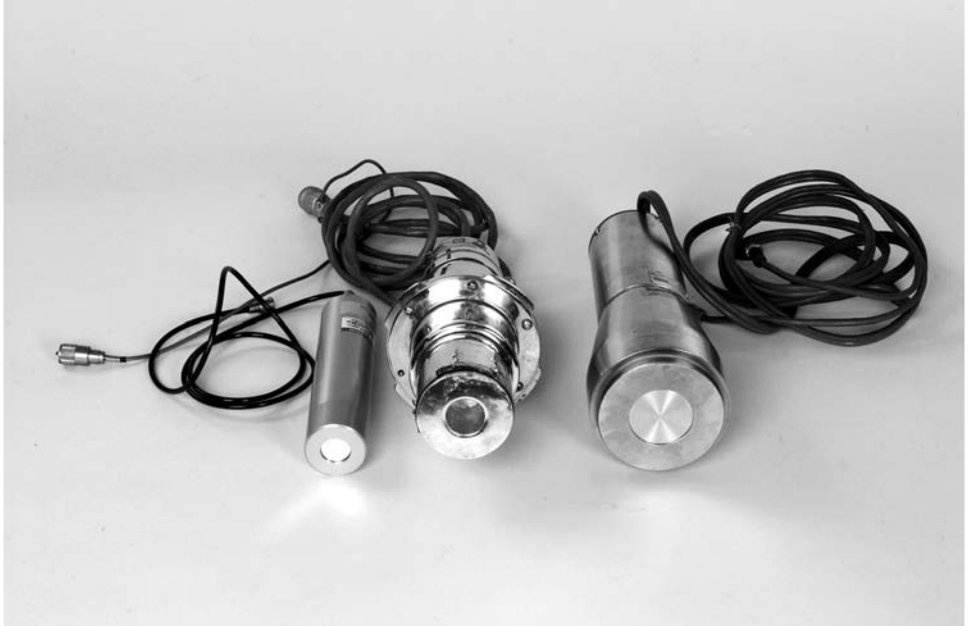


Figure 10. The NaI(Tl) detectors: 1 × 1 in. (left; Cardinal Health; Inovision, Model 425-110); 2 × 2 in. (5.08 cm) with center well (center; Nuclear Chicago, Model 321330), and 3 × 3 in. (2.62 cm) (right; Picker Nuclear Omniprobe, Model 2830-A).

solvents, such as xylene, toluene, and phenylcyclohexane. A wave shifter fluorescent material shifts the wavelength of the light from the main solute to a longer wavelength and lower energy, so that the wavelength more closely matches the spectral response of the photocathode. Liquid organic scintillators are widely used because the sources of ionizing radiation can be dissolved into the solvent and made a part of the scintillator solution. Low energy beta emitters' tritium, ^{3}H (19 keV), and ^{14}C (156 keV), are counted with high efficiency by these methods (Fig. 12). Modern pulse processing methods allow separation of alpha and beta events.

Plastic scintillators consist of organic scintillators that have been dissolved in a solvent and the existing solvent polymerized to form plastic scintillators. As plastics can be made ultrathin, they can be useful for detecting low energy particles of radiation in the presence of gamma rays or for

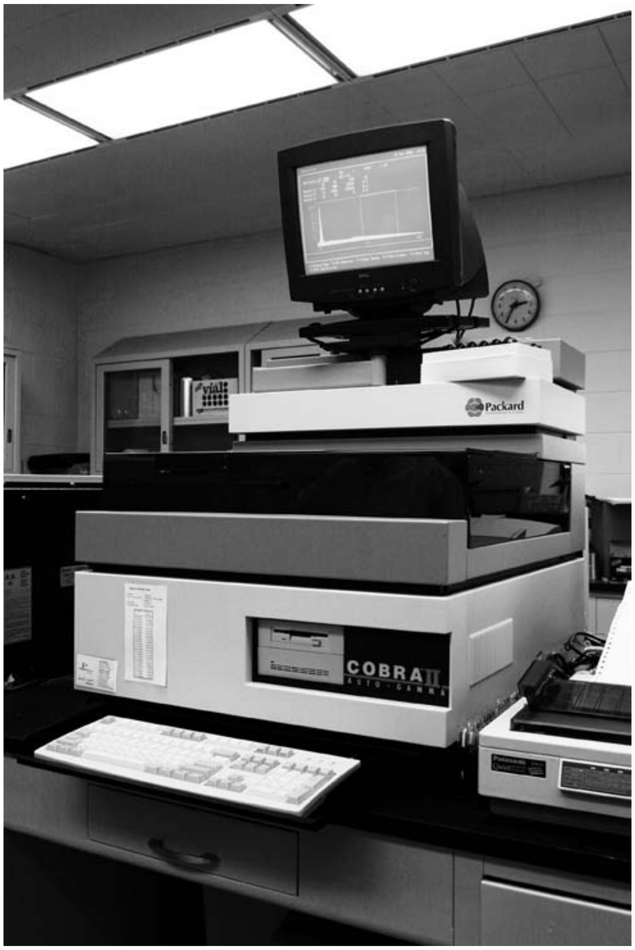


Figure 11. A γ-counter system with a 1 × 1 in. (2.54 cm) NaI(Tl) detector to identify and measure γ rays. (Canberra; Packard, Model Cobra II Auto-Gamma.)

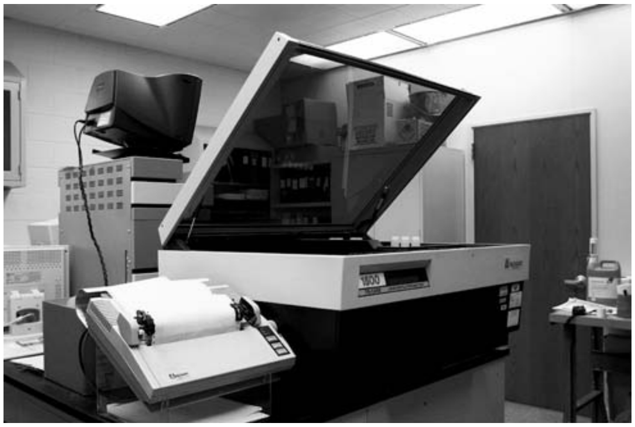


Figure 12. A liquid scintillation counter system for beta particles (Canberra; Packard, Model B1500 Tri-Carb).

detecting heavy particles. Plastic scintillators are available in many physical configurations. Neutrons can be defected by incorporating a neutron sensitive material, such as lithium, into the solvent and the subsequent plastic scintillator. Nobel gas scintillators consist of high purity concentration of helium or xenon, which have the property that, following radiation interaction in the gas, both visible and ultraviolet (UV) light is emitted. While these materials exhibit a very short deexcitation time of 1 ns, they yield little light and the conversion of light is reasonably inefficient; nevertheless, they do have numerous applications when a fast response time is required.

As with GM probes, scintillation detectors are available as rectangular pancake probes to detect beta particles > 100 keV and γ rays > 25 keV, thin scintillators to detect γ and X rays > 10 keV, flashlight-like probes to detect alpha particles > 350 keV and beta particles > 14 keV, and conventional thick-crystal cylindrical probes for γ and X rays > 60 keV.

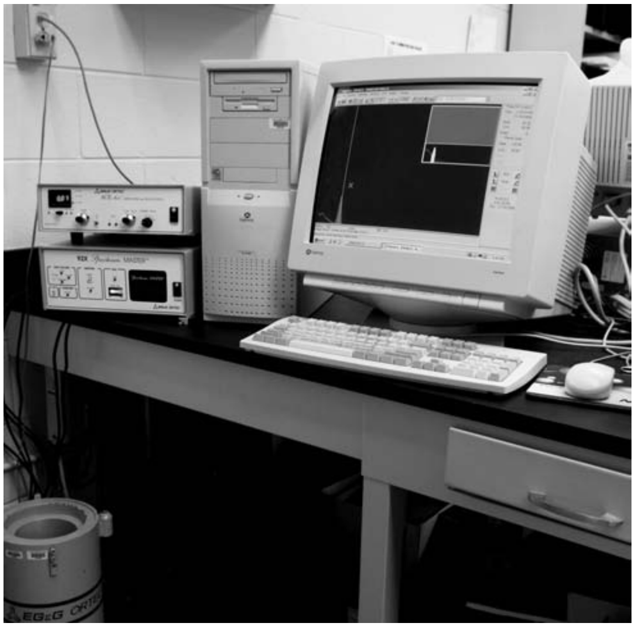


Figure 13. A multiple component NaI(Tl) spectroscopy system. Detector shield (lower left; Ortec) with a NaI(Tl) detector (not shown); amplifier/bias supply(upper center; Ortec, Model Acemate); spectral analyzer (lower center; Ortec, Model 92X Spectrum Master), and computer display (left).

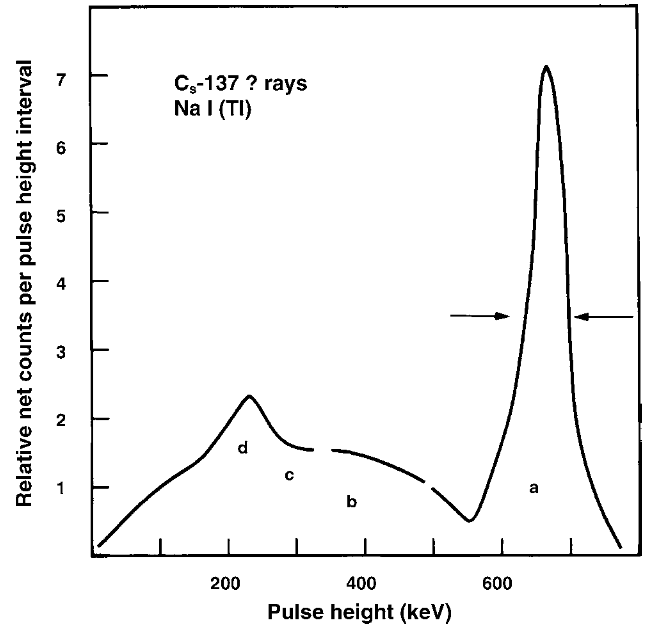


Figure 14. A spectrum measured with a 1×1 in. (2.54 cm) NaI(Tl) dectector: (a) ^{137}Cs full-energy (0.66 MeV) peak; (b) Compton edge (0.48 MeV); (c) Compton distribution; (d) backscatter peak (0.18 MeV).

Photocathodes have many applications in devices that measure or detect radiation, such as image intensifiers, vidicon tubes, and other detectors. Usually, the electronics required to amplify the initial signal does not have as short a resolving time as the detector proper, but with modern solid-state electronics, the resolving times have been shortened to less than microseconds. Because of their extreme sensitivity, scintillator detectors are useful when detection and subsequent identification, by spectral analysis, of a type of radioactivity is required (Fig. 13). While the spectral peaks associated with scintillators are reasonably broad, their energy resolution is sufficient to allow rapid identification of minute quantities of various radioisotopes (Fig. 14). Formerly limited to laboratory analysis, NaI(Tl) spectroscopy is now available in portable handheld units (Fig. 15) that, using quadratic compression conversion (QCC), can identify 128 radionuclides in real-time (1 s intervals). Quadratic compression conversion creates spectral energy peaks whose widths vary proportionally to the NaI(Tl) crystal's energy resolution. All energy peaks are displays with the same peak width that allows radionuclides' distinct spectra to be more readily identified. The electronics associated with scintillators must be extremely stable, but often variations are introduced by environmental factors. High permeability magnetic shielding materials, for example, Mu-metal, often are used to shield against stray magnetic fields. Temperature and humidity variation can produce undesirable electronic noise. As previously noted, some crystals are hydroscopic and the moisture absorbed can reduce the efficiency of the process. Pulse discrimination techniques are often used to distinguish one type of radiation from another.

SOLID-STATE RADIATION DETECTORS

Thermoluminescence (TL) is the emission of visible light released by heating previously irradiated solid-state crystals. The light emitted from a thermoluminescent crystal is proportional to the amount of radiation to which the crystal has been exposed, and this proportionality holds over a large range (10^2–10^5) of exposures. At very high exposures, nonlinearity is exhibited and the amount of visible light released is no longer proportional to the amount of radiation detected. Thermoluminescent dosimetry (TLD) materials (Table 5) commonly used in medicine, include lithium fluoride (LiF), available in three forms (TLD-100, -600, -700,), and lithium borate manganese (Li_2B4O_7:Mn) (TLD-800).

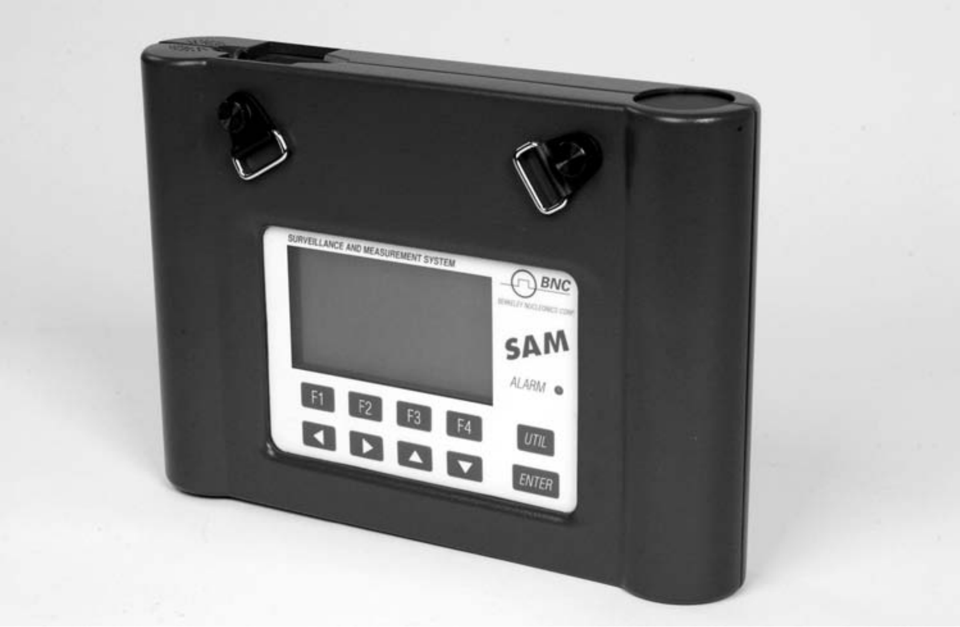


Figure 15. A portable NaI(Tl) surveillance and measurement (spectroscopy) system. (Berkeley Nucleonics Corp., SAM Model 935).

Other TLDs used in environmental dosimetry applications, calcium fluoride manganese (CaF:Mn) (TLD-400), calcium fluoride dysprosium (CaF_2:Dy) (TLD-200), calcium sulfate dysprosium ($CaS0_4$:Dy) (TLD-900), and aluminum oxide (Al_2O_3:C) (TLD-500), are not further described here. X rays, γ rays, and neutrons are easily detected with different TLD materials; the detection of higher energy beta particles is possible, but quantification of the amount of beta radiation and calibration of the solid-state detectors for beta radiation is often more difficult than for X and γ rays. The TL materials exhibit an enhanced response to lower energy ($<$ 200 keV) X or γ rays as compared to higher energy (1 MeV) X or γ rays. For LiF, the over response is only a factor of 1.2, but for lithium borate manganese ($Li_2B_4O_7$:Mn) (TLD-800), there is an under response of $\sim$ 0.9. By adding filters, the energy response of a given type of crystal can be made more uniform and this is commonly done with TLD detectors used as personnel radiation monitors (Table 5).

Table 5. Properties of Some Thermoluminescent Materials

Property/type	LiF:Mg, Ti (TLD-100)	^{6}LiF:Mg, Ti(TLD-600)	^{7}LiF:Mg, Ti (TLD-700)	LI$_2$B$_4$O$_7$:Mn (TLD-800)
Applications	Health and medical dosimetry	Neutron dosimetry	Gamma dosimetry	Neutron dosimetry
Relative concentrations	^{6}Li (7.5%) ^{7}Li (92.5%)	^{6}Li (95.6%) ^{7}Li (4.4%)	^{6}Li (0.007%) ^{7}Li (99.993%)	NA[a]
Density (g mL^{-1})	$\sim$ 2.6 (ribbons) $\sim$ 1.3 (powder)	2.64	2.64	$\sim$ 2.4 (ribbons) $\sim$ 1.2 (powder)
Effective Z for photoelectric absorption	8.2	8.2	8.2	7.4
TL Emission spectra	350–600 nm (400 nm max)	250–600 nm	350–600 nm (400 nm max)	530–630 nm (605 nm max)
Temperature of main TL glow peak	195 °C	195 °C	195 °C	200 °C
Efficiency at ^{60}Co relative to LiF	1.0	1.0	1.0	0.15
Energy response 30 keV/^{60}Co	1.25	1.25	1.25	0.9
Useful range	mR–3 $\times$ 10^5 R	mR–10^5 R	mR–3 $\times$ 10^5 R	50mR–10^6 R
Fading	Negligible* 5%/years at 20 °C	5%/year	5%/year	$<$ 5% in 3 months
Preirradiation anneal	400 °C at 1 h + (100 °C 2 h or 80 °C at 16 h)	400 °C at 1 h + (100 °C at 2 h or 80 °C at 16 h)	400 °C at 1 h + (100 °C at 2 h or 80 °C at 16 h)	300 °C at 15 min
Postirradiation anneal	100 °C at 10 min	100 °C at 10 min	100 °C at 10 min	
Special feature	Low dose rate dependence	Highly sensitive to thermal neutrons	Insensitive to neutrons	High dose dosimetry

[a] Not available = NA.

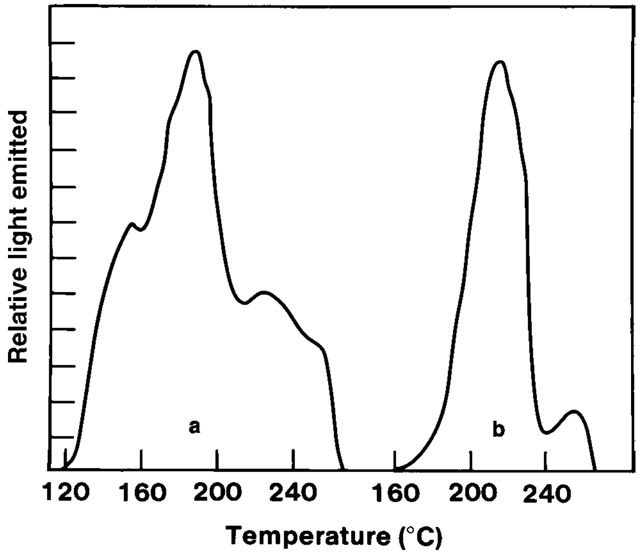


Figure 16. Glow cures for TLD-100 (natural LiF) rods exposed to 10 R. (a) Without proper preparation annealing, multiple natural peaks occur; (b) Using a 1 h annealing at 400 °C and a 2 h annealing at 100 °C a smoother curve is obtained.

Physically, TLD materials consist of loose powder contained or embedded in plastic holders, compressed crystals (chips), extruded rods, and chips on a card in a configuration that allows reproducible heating of the detector materials to selected temperatures. The TL materials exhibit some undesirable features as radiation detectors. At room temperature, fading or loss of signals occurs, from $< 0.5\%$ per month for LiF to 5% in 3 months for $Li_2 B_4 O_7$:Mn. The degree of fading can be controlled, to some extent, by proper preparation (annealing) procedures.

The optical readers generally consist of a heating pan or device that allows the TL material to be uniformly heated in a controlled manner, at a specified temperature for a given period of time. The heating device and material holder are directly below a photomultiplier tube that usually has some filters to remove any IR radiations and transmits light in the blue-green region of the visible spectrum. The signal from the photomultiplier is amplified and used to prepare a glow curve (Fig. 16) a plot of the intensity of light versus the heating cycle of heating elements. Different TL materials exhibit different glow peaks. Numerous peaks occur in the curve and either the area under the curve or the height of the major peak is chosen to be proportional to the amount of radiation to which the material was exposed. Proper preannealing (extended heating at a controlled temperature) and postannealing of the material will remove some smaller undesirable peaks (Fig. 16) leaving the main peak that is used for measurement purposes.

Current TL readers offer automatic features for glow curve analysis and processing of large numbers of samples (chips, rods, or cards).

Lithium flouride is the most commonly used TL material for personnel dosimetry and consists of natural lithium. Lithium-6 is preferentially sensitive to thermal neutrons and ^{7}Li is insensitive to thermal neutrons, but sensitive to γ-rays. Hence, by using paired ^{6}Li and ^{7}Li materials, it is possible to measure thermal neutrons in the presence of γ rays. Use of natural LiF detectors in radiation environments that contain low levels of thermal neutrons will yield incorrect dose equivalents for personnel, as the thermal neutrons will cause an apparent over response of LiF calibrated only to detect and measure γ rays (16).

Optically stimulated luminescence (OSL) is the release of light by a phosphor following its irradiation by a laser. Aluminum oxide (Al_2O_3) containing carbon impurities exhibit OSL releasing a blue light when excited by a green laser light. Some personnel radiation monitors employ OSL technology that offers some improvement over TLD-based personnel radiation dosimeters (17). The OSL dosimeter offers greater sensitivity, stability, and accuracy than TLD dosimeters. Aluminum oxide is highly linear from 1 mSv to 10 Sv; there is little signal fading > 1 year. It does exhibit an energy dependency below ~ 100 keV. However, unlike TLD chips, the aluminum oxide element can be reread multiple times to confirm an initial reading, an advantage for personnel dosimeter applications. The Luxel badge (Fig. 17; Table 6) contains the Al_2O_3 phosphor element, with 20 mg·cm^{-2} open filter (paper wrapper), 182 mg·cm^{-2} copper filter, and 372 mg·cm^{-2} tin filter in a heat-sealed, light-tight hexagonal plastic badge (18). The combination allows detection of beta particles > 150 keV with a 100 μSv threshold and X and γ rays > 5 keV with a 10 μSv threshold.

Photoluminescence (PL) occurs when the irradiated crystal emits visible light when exposed to UV light instead of heat. Silver activated glass encapsulated PL detectors are available in numerous shapes, sizes, and radiation levels as low as 100 μSv are detectable, but the detectors are commonly used to detect higher exposures. Appropriate filters can be used to make the energy response more linear, but at exposures of 0.1 Sv or higher the response of these detectors is nonlinear. These materials exhibit some signal fading that depends on the composition of the glass. Heating the glass detectors post postirradiation for 30–60 min at 150 °C yields maximum luminescence. Reannealing requires 40 °C for 1 h, which restores the material to its preirradiation state. For low level exposure measurements, care is required to keep glass detectors free of dirt, dust, and other materials that would reduce the amount of light transmitted through the detector.

Semiconductor materials used for radiation detection, Si, Ge, CdTe, HgI_2, and GaAs, have band gaps, the region between the valence and conducting band, of < 2.2 eV. Electrons migrate from the valance to the conduction band, leaving holes, that act like positively charged electrons, in the valence band. In a p-type semiconductor, the current is carried by the positively charged holes; in an n-type semiconductor, current is carried by the electrons. Usually, a potential difference is maintained across the solid-state semiconductor such that the depletion layer is devoid of electrons and holes. The interaction of X rays, γ rays, alpha particles, or beta particles generates additional electrons in the depletion layer; these are then swept away by the potential difference across the material, yielding a small current whose magnitude is proportional to the intensity of the incident radiation Energy required to generate

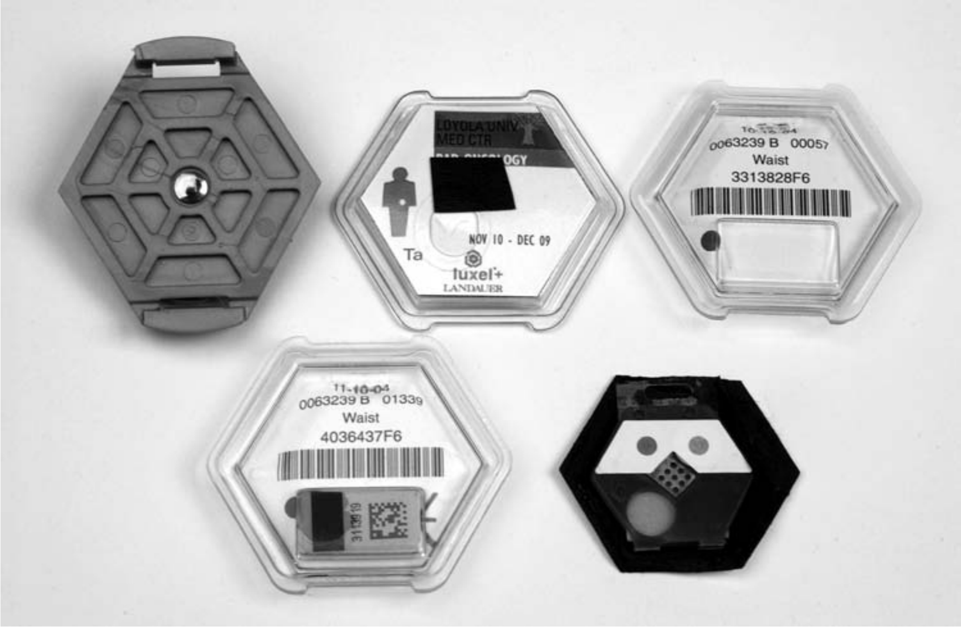


Figure 17. The Luxel Dosimeter: Holder (top left); front view (top center); rear view (top right); rear view with attached Neutron 144 detector (bottom left); detector element showing Cu filter (right), the Al filter (left), and plastic filter (circle, lower center). (Courtesy Landauer, Inc.)

electron–hole pairs range from ~ 3 to 6.5 eV. Semiconductors exhibit excellent linear response over a large energy range and greater efficiency than gas detectors. (For a full description of semiconductor physics, see Refs. 7 and 8.)

Surface-barrier detectors essentially consist of an n-type and p-type layers that function as anode and cathode, with an intrinsic I layer (depletion region), without electrons, in between, in which radiation interactions occur. Depletion regions range from 0.1 to 1 mm. The composite is commonly called P-I-N structure. With a moderate reverse bias applied, radiation interactions create electron–hole pairs with each electron and hole leaving, or depleting, the intrinsic layer, and, with appropriate circuitry, creating a detection and counting system.

Silicon surface-barrier diode counters are used for charged-particle (alpha and beta particle) detectors. Alpha resolution ranges from 12 to 35 keV; beta resolution ranges from 6 to 30 keV. Passivated Implanted Planar Silicon (PIPS) detectors use implanted contacts more rugged than conventional surface-barrier contacts. They can be costumed-designed for specific applications.

Germanium (Ge) detectors are used for γ- X-ray spectroscopy to identify not only the amount of radiation present, but also to identify the type of radioactivity present. These detectors have excellent energy resolution and are used to identify the individual X and γ rays from a radioisotope, and with peak fitting programs have the ability to resolve closely lying peaks (Fig. 18). Lithium drifted geranium, Ge(Li), detectors have been replaced by high purity pure geranium (HPGe) detectors that only required liquid-nitrogen cooling (Fig. 19) or electrically refrigerated cryostats during measurements or use; otherwise they are kept at room temperature. Lithium drifted silicon, Si(Li), detectors are still used. With microprocessors, these instruments are now used for on-site identification of samples that may contain several different radioisotopes, a process previously limited to the laboratory. They are important tools in research facilities where minute quantities of many

Table 6. Parameters of Some Commercial Personnel Dosimeters[a]

Monitor Designation	Sensitive Element	Primary Filter and Thickness, mg cm^{-2}	Radiation Detected and Detection Threshold, mSv
Gardray	Film or	Film wrapper (35)	β(0.4)
	4-chip TLD	Plastic (325)	X, γ(0.1)
		Aluminum (375)	Thermal neutron (0.1)
		Lead (1600)	
		Cadmium (1600)[b]	
T	2 TLD chips	Plastic (75)	β(0.4)
		Plastic (200)	X, γ(0.1)
Neutrak 144	TLD-600	Cadmium (660)	Thermal neutrons (0.1)
	TLD-700		Fast neutrons (0.2)
	CR-39		
Luxel	AI_2O_3	Open (20)	β (0.1)
		Copper (182)	X,γ (0.01)
		Tin (372)	

[a] Courtesy of Landauer, Inc. Parameters quoted are those listed in the company's advertisement.
[b] Cadmium filter provided with film systems only.

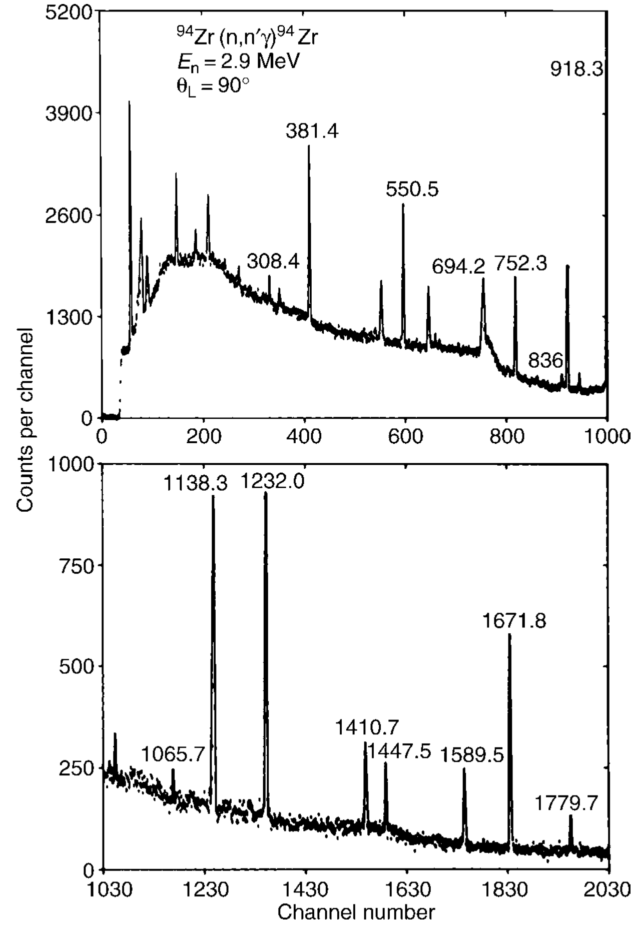


Figure 18. A γ-ray spectrum measured with a 35 cm Ge(Li) detector with good energy resolution.

different radioisotopes may potentially accumulate and radioisotope identification is required.

Cadmium telluridide (CdTe) detectors are popular as they are hygroscopic, do not require a photomultiplier, require only a 50 V bias, and operate at room temperatures. They exhibit high sensitivity, but their energy resolution is intermediate between that of NaI(Tl) and Ge detectors. They are available in a multitude of small sizes for special applications.

OTHER DETECTORS

Other physical changes that arise in materials as a result of irradiation include coloration and nuclear activation. Color changes occurs in some materials following their irradiation. Thin films with a cyanide emulsion (GAF-CHROMIC EBT) will exhibit a deep orange color following their irradiation by γ rays to a dose of 10–100 Gy (19). These materials exhibit excellent linearity > 1–800 cGy of radiation dose. While developed for high dose radiation dosimetry studies on linear accelerators, they can be used for high dose radiation protection studies.

Neutrons may be detected by numerous nuclear reactions or by the process of counting the number of recoil proton tracks produced in certain neutron sensitive materials, materials, such as boron (20). A polycarbonate material CR-39 (allyl diglycol carbonate) is insensitive to X rays, γ rays, and beta particles. However, incident neutrons collide with the protons, which produced changed particle tracks; the tracks are enhanced by chemically etching the polycarbonate, so they will be more visible under a microscope. This technology is used as one component in a composite personnel dosimeter used to monitor radiation therapy personnel working around linear accelerators with X-ray energies > 10 MeV (Table 6) (18).

Superheated Drop Detectors (SDD) consists of a small container of gel holding superheated drops ~ 0.1 mm diameter. Neutrons produce recoil protons that strike the drops, causing them to vaporize, generating an audible pop or sound that can be counted. The detector is insensitive to γ-rays, and is independent of neutron energy to ~ 14 MeV. Sensitivity is ~ 80 bubbles per 10 μSv, with a minimum threshold of ~ 1 μSv; there is a linear relationship between the number of bubbles and the neutron dose. The SDD technology has been incorporated into equivalent dose neutron survey meters with replaceable SSD cartridges that must be changed after exposure to certain maximum doses. Alternately, bubbles in samples can be visually counted to determine neutron dose (8).

PHOTOGRAPHIC DETECTORS

Photographic emulsions that darken in proportion to the amount of radiation they receive represent one of the earliest methods of detecting radiation. Most films consist of a thin plastic sheet with ~ 0.2 mm emulsions one or both sides; the emulsion, usually silver halide granules in a gelatin mixture is covered with a thin protective plastic coat. Modern films have emulsions specific for optimal detection of certain energy and intensity of radiation. Radiation incident on the emulsion changes the clear silver halide ions, forming a latent image. During processing additional silver is deposited, the darkness of the film is determined by where silver ions are deposited on the film and the amount of silver deposited. While most films are limited to dynamic ranges of ~ 10^3, multiple film packets can be used in combination to extend the dynamic ranges to 10^5 or higher. Fast films are sensitive to the lowest levels of radiation, while slow films require greater exposure to darken the films. Usually, the film is used in a protective cover, which can be a cassette, commonly used for imaging and including metal screens to enhance the image. Rapid processing film is wrapped in a thin light tight paper wrapping that prevent light leaks and may be used without a cassette.

Personnel monitors frequently use film as the radiation detector (Table 6). The small film packet in its light wrapper is carried in a plastic holder, but the film wrapping is sufficiently thin to allow the transmission of the low energies X rays, γ rays, or beta particles. Films usually exhibit an enhanced sensitivity of several factors of 10, to lowest energies of radiation, below ~ 100–150 keV, relative to their response to γ rays with energies of 150–1.5 MeV. By using filters of copper, aluminum, and lead of varying thickness in the plastic holder in front of the film, estimates

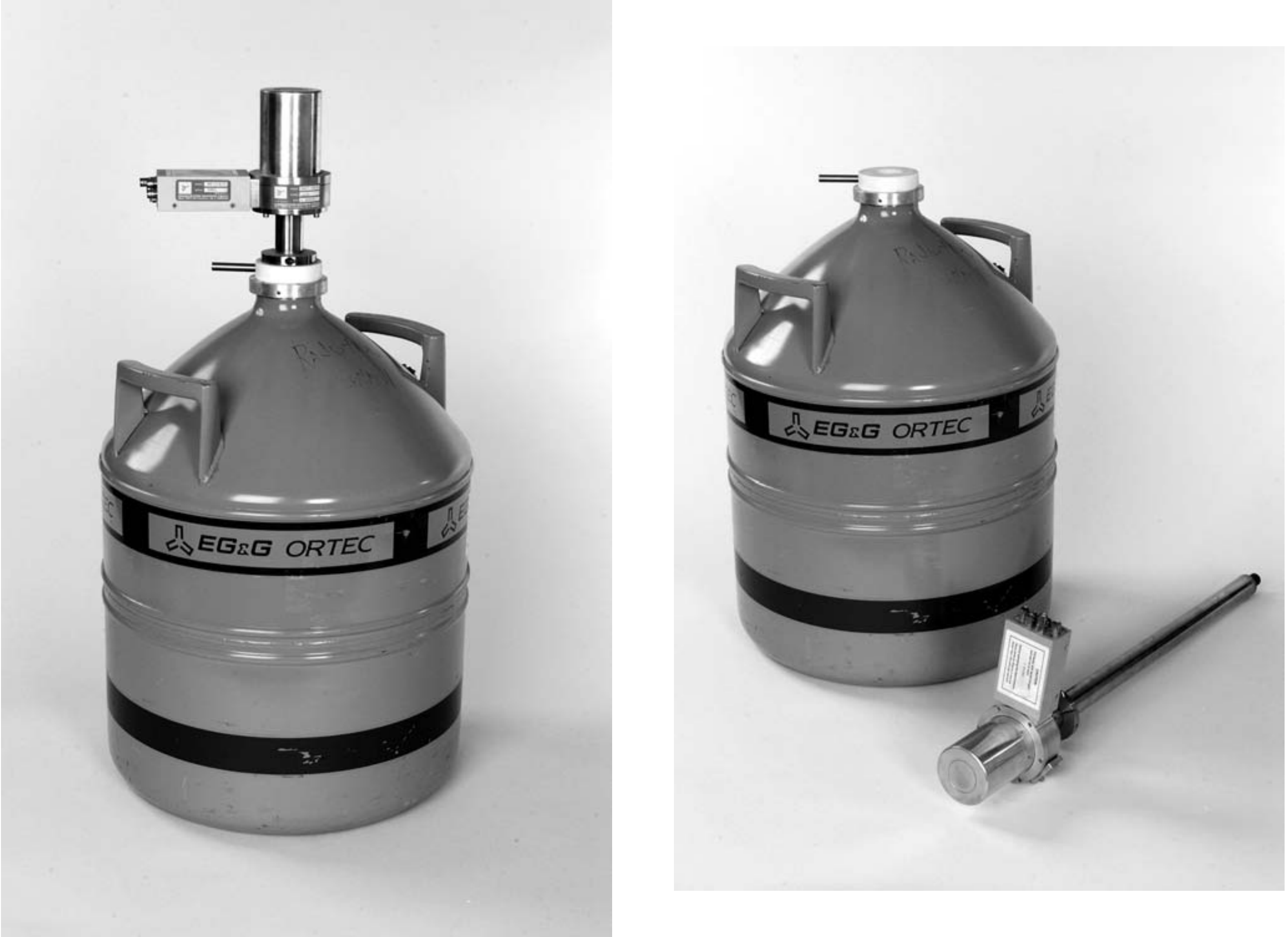


Figure 19. Left Panel: liquid nitrogen dewar (left; Ortec, Model unknown) for cooling a germanium detector (right; Princeton Gamma Tech, Model RG-11B/C;); Right Panel: The assembled detector ready for use.

of the energies of the radiations darkening the film can be made. One distinct advantage of film dosimeters is the permanent record generated. A disadvantage is that radiation incidence at an angle to the filter appears to have passed through a thicker filter than actually available. Film badges or monitors are available in numerous configurations for the body, head, wrist, hand, and fingers. Normally film badges are exchanged monthly if radiation levels are low; individuals working in a higher level radiation environment may be monitored more frequently.

Numerous environmental factors can fog photographic film producing erroneous personnel exposure readings. While films are free of electromagnetic interference, excessive humidity can influence results, as can excessive heat. Images in films can fade with time so prompt collection and processing of personnel monitors is important in obtaining accurate results. Special metal filters, such as cadmium, may be used in a film holder to produce a neutron sensitive film by the (n, α) reaction in cadmium. Film dosimeter are widely used as personnel monitors and the overall accuracy of a film dosimetry system is usually at least 50% or better depending on the energy range of use. Lower energy radiation present in small amounts yields the greater uncertainty in the accuracy of monitor readings. With film dosimeters, the method of film processing is very important, as inconsistent methods of developing film will lead to substantial errors in the final results. Generally, film processors have their developer chemistry optimized for a particular type of film. Monitoring the temperature of the developing chemistry is very important and frequent use of calibration films, films previously given known exposures to radiation, is required to maintain the integrity of a film dosimetry system. Proper care and maintenance and rigorously scheduling of chemical developer replenishers are necessary.

ACKNOWLEDGMENTS

Equipment photographs courtesy of Todd Senglaub, Loyola Radiation Control, and Alvin Hayashi, Loyola Medical Media. I wish to thank Josephine Davis, Loyola Dept. Radiation Oncology, for her patience and excellent assistance preparing this manuscript.

BIBLIOGRAPHY

1. Health Physics Society. (No Date). Home Page. [Online] Health Physics Society. Available at http://www.health-physics.com. [2004, Nov. 16]. 2004.

2. Glasgow GP. Radiation protection instruments. Webster JG, editor. Wiley Encyclopedia of Medical Devices and Instrumentation. New York: Wiley; 1988.
3. Kasper K. Ludlum Model 2360. Health Phys 2003;84:1.
4. Health Physics Instrumentation Committee (1999, Sept. 14) [Online] Department of Energy. Available at http://www.llnl.gov/HPIC/HPICHP.HTML. [2004, Nov. 16]. 2004.
5. Knoll GF. Radiation Detection and Measurement. 3rd ed. New York: Wiley; 2000.
6. Shapiro J. Radiation Protection: A Guide for Scientists, Regulators, and Physicians. 4th ed. Cambridge (MA): Harvard University Press; 2002.
7. Turner JE. Atoms. Radiation, and Radiation Protection. 2nd ed. New York: Wiley; 1995.
8. Gollnick DA. Basic Radiation Protection Technology. 4th ed. Altadena (CA): Pacific Radiation Corporation; 2001.
9. Mohr PJ, Taylor BN. The Fundamental Physical Constants, [2004, Aug] Physics Today. [Online]. Available at http://www.physicstoday.org/guide/fundconst.pdf. [2004, Nov. 16]. 2004.
10. Storm DJ, Watson CR. On being understood: Clarity and jargon in radiation protection. Health Phys 2002;82:373–386.
11. 1990 Recommendations of the ICRP. Publication 60, International Commission on Radiological Protection. New York: Pergamon Press; 1990.
12. International Commission on Radiation Units and Measurements. Quantities and units in radiation protection dosimetry. Bethesda (MD): ICRU Publications; ICRU Report No. 51: 1993.
13. International Commission on Radiation Units and Measurements. Fundamental quantities and units for ionizing radiation. Bethesda (MD): ICRU Publications; 1998.
14. Bureau International des Poids et Measures. Le Systeme International d'Unites (SI). French and English Texts. Servres, France: Bureau International des Poids et Measures; 1991.
15. Olsher RH, et al.. WENDI: An improved neutron rem meter. Health Phys 2000;79:170–181.
16. Glasgow GP, Eichling J, Yoder RC. Observations on personnel dosimetry for radiotherapy personnel operating high energy linacs. Health Phys 1986;50:789.
17. Kasper K. Optically stimulated luminescence dosimeters. Health Phys 2001;81:1108–1109.
18. Products and.Services. (No Date). Home Page. [Online]. Landauer, Inc. Available at http://www.landauerinc.com/luxelosl.htm. [2004, Nov. 16]. 2004.
19. GAFCHROMIC EBT Product Brief. (No Date). Home Page. [Online]. International Specialty Products. Available at http://www.ispcorp.com/products/dosimetry/index.html. [2004, Nov. 19]. 2004.
20. Kumamoto Y, Noda Y. Measurements of effective doses of natural background levels of neutrons with etched-tracked detectors. Health Phys 2002;83:553–557.

See also CODES AND REGULATIONS: RADIATION; EQUIPMENT MAINTENANCE, BIOMEDICAL; SAFETY PROGRAM, HOSPITAL; X-RAY EQUIPMENT DESIGN.

RADIATION THERAPY, INTENSITY MODULATED

WALTER GRANT III
Baylor College of Medicine
Houston, Texas

INTRODUCTION

The past 10 years has seen the rapid emergence of a new radiation therapy process that has become known as intensity modulated radiotherapy, or IMRT. This new process has expanded so rapidly that there are many misconceptions of its origin as well as its identity. The concept of modifying the standard radiation pattern emitted from an external radiotherapy unit has been employed for decades, for example, in the form of physical wedges or compensators. To begin to understand this new technology, it is important to recognize what differentiates the non-uniform intensity patterns associated with IMRT from those accomplished with other methods. For example, The National Cancer Institute Collaborative Working Group for IMRT stated that IMRT is, "An advanced form of image-guided 3dCRT that utilizes variable beam intensities across the irradiated volume that have been determined using computer optimization techniques". (1). While this definition is correct, it is a generalization that adds to the lack of clarity regarding IMRT and a review of the literature should allow a better understanding of the basic technologies required to produce an IMRT treatment.

In 1982, Brahme, Roos, and Lax (the BRL paper) published an article (2) describing the exact solution for the beam intensities required for a new nonlinear wedge shape that could be used to create improved dose uniformity for targets in a cylindrical phantom that were on or near the axes of symmetry and rotation of a problem in arc therapy. The authors also discussed the similarities to the imaging processes used in computed tomography (CT). This technique was used to treat 25 patients.

In 1987, Cormack (3) extended the BRL concept to noncircular symmetry. These results have nonexact solutions, but the author discussed logical approaches to selecting the best intensities.

The major step came just a year later in 1988 when Brahme published an article (4) that proposed the use of inverse treatment planning using filtered backprojections to solve both stationary and rotational problems. He believed that this approach would have a large impact and stated that it, "...largely avoids the trial and error approach often applied in treatment planning of today". His planning scheme was to have the physician place constraints on the doses to tumors and normal tissues and allow a computer to determine the location and intensities for the beams that achieved these results, as opposed to the traditional method where the planner places the beams and then evaluates the resulting dose distribution.

In diagnostic CT, one uses filtered backprojections to eliminate the artifacts that occur during image reconstruction. These filters mathematically convert a uniform beam from an X-ray tube into a nonuniform beam in order to achieve the proper images. They can have fine spatial resolution as well as negative values. It is this reality that creates the fingerprints of the process that is known as IMRT. In order to perform the reverse process in radiation therapy, we would have to use small beams and be able to produce a negative radiation source. This means that we likely will not be able to produce an exact solution, but only a "best" solution based on the clinical constraints of the patient and the source of radiation. For this reason, an optimization algorithm must be employed. We now have the unique markers for IMRT, the ability to plan and delivery small beamlets of varying intensities that have been determined by computer optimization.

This article also cleared the path for the two delivery approaches used today in IMRT, multiple stationary gantry positions (fixed field) with dynamic intensity map creation and arc therapy delivery (tomotherapy) of dynamic intensity maps. Each of these techniques used different technologies for delivery of the dose pattern and had unique problems to overcome. In order to appreciate the complexities of the technologies, each will be addressed separately, beginning with the fixed field technique.

The limit for the effectiveness of radiation therapy as a treatment modality always has been the volume of normal tissue being irradiated. While early patients had this volume reduced by the use a library of lead blocks that were hand placed, the introduction of low melting point alloys (5) to create patient specific blocking introduced a major improvement in accomplishing the goal of the reduction of dose to normal tissue. The disadvantages of this system include the time to create the blocking as well as the weight of the finished product. To overcome these problems, multileaf collimators (MLC) were introduced (6). These devices were not created to increase blocking effectiveness, but to provide a more efficient blocking system. It took over a decade for these systems to mature to their potential, but today they are a common feature on linear accelerators. Figure 1 shows the exit port of an accelerator with the MLC leaves fully retracted. This would deliver a square radiation field. Using the MLC, one can shape the radiation field easily and Fig. 2 shows a field in the shape of a diamond. This, and any other shape, are set quickly by computer control.

The designs of the commercial MLCs have similar characteristics and some differences. The best place to find particular information is in the American Association of Physicists in Medicine (AAPM) Report No. 72 (7). There are some distinctions that will be made here as they affect the ability of these devices to create the required intensity maps for IMRT, but one should read Report No. 72 for more detail.

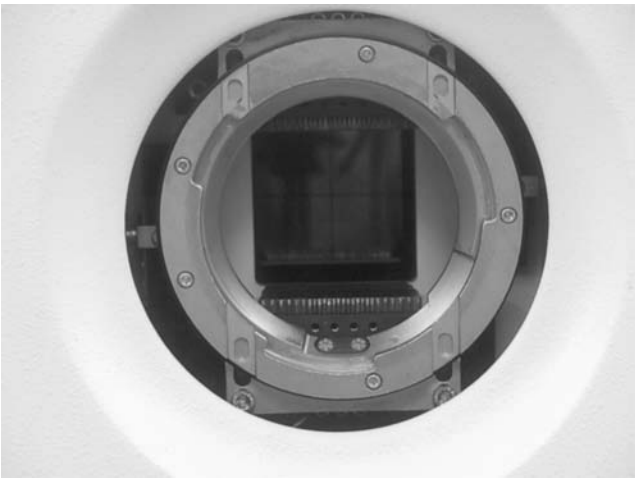


Figure 1. The exit port of a linear accelerator with the MLC parked completely under the primary collimating jaws.

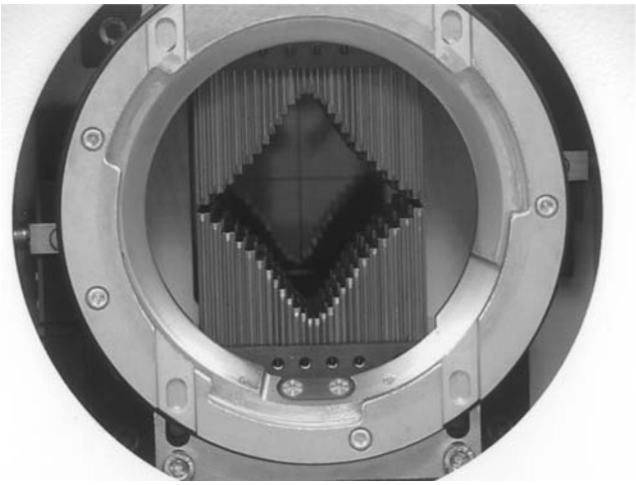


Figure 2. The MLC positioned to deliver a radiation pattern in the shape of a diamond.

Most vendors now have leaves that project to a 1.0 cm width at the treatment distance for the machine, although there are recent advances to 0.5 cm for a general MLC and 0.3 cm for a special purpose MLC. The MLC will either replace one set of primary collimators in the head of the machine or be attached to the machine as a tertiary collimator. The former allows for more clearance between the patient and the exit port of the linear accelerator, but will have greater accuracy requirement for leaf position as inaccuracy is magnified by distance increased distance from the patient. An additional advantage is that these leaves can be double focused, allowing for a beam penumbra that is constant as a leaf moves across the field. Tertiary collimators of today have directly opposite characteristics.

There are additional characteristics that one needs to evaluate for the use of MLCs in general and IMRT in particular. One characteristic is the maximum distance a leaf may travel across the center of the field. This will determine the maximum field that can be treated with IMRT unless some additional facility of the collimator is introduced.

Another is the capability for a leaf moving from one side to pass its two neighboring leaves from the opposite side. The capacity to interdigitate leaves is important for some intensity modulation segmentation techniques, as well as create stand alone or “island” blocks in the central portion of a treatment field. Figure 3 shows an interdigitated set of leaves. The leaves on the edges of the pattern are set to the centerline of the radiation field and one can see alternating leaves extending over the center line from each side.

A final characteristic is that no MLC has adjacent leaves with smooth surfaces. In order to prevent leakage radiation from streaming through the interface of two leaves, leaves are keyed or notched to eliminate a direct path for the radiation. During the creation of an intensity map, adjacent leaves may get far enough apart that there can be a significant increase in unwanted leakage radiation because the leaf edge is not as thick as the leaf center. This is called the tongue and groove effect and will be part

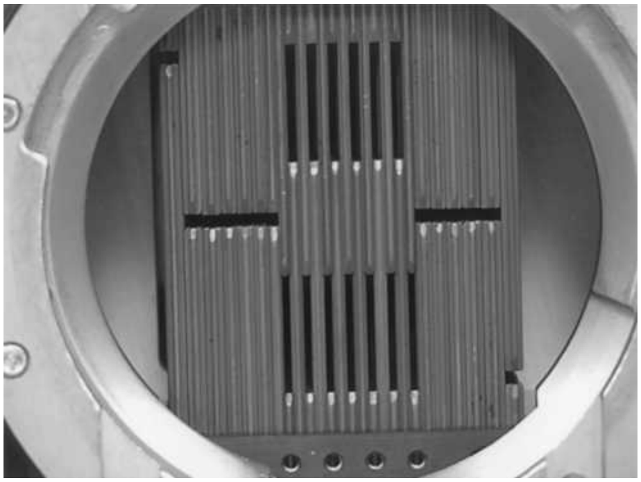


Figure 3. An example of leaves that are interdigitated. The closed leaves are set at the middle of the field.

of a latter discussion regarding the creation of the intensity maps for IMRT.

Although there was a decade of development for the MLC, it was obvious that these devices could be used to create intensity maps and investigators worked on two methods, both of which are clinically available today although not from every vendor. The first involves the creation of a final intensity map by the accumulation of multiple static fields, or segments. This technique was investigated theoretically by Bortfeld et al. (8) and Webb (9) while Bortfeld et al. (10) conducted the first experiment of such delivery in 1994. As a single segment is completed, the beam is turned off and the MLC instructed to change to the next position. Then the beam is turned on and the next segment is delivered. The NCI–CWG–IMRT defines this process as Static MLC IMRT or sMLC. As an example, the MLC in Fig. 4 has the leaves positioned at the boundaries of a radiation field to be treated. If one begins by positioning all the leaves to the left-side boundary and moves them independently across the field at specified dose intervals, one can deliver a non-uniform dose pattern as shown in Fig. 5 with the darker colors indicating higher dose.

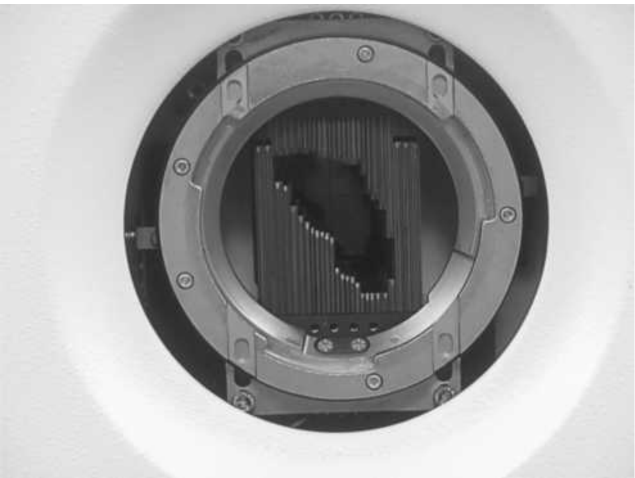


Figure 4. A MLC with the leaves positioned at the boundary of a shaped radiation field. The leaves on the right will move to meet the leaves on the left and then all leaves will move across the radiation field to create a nonuniform intensity map.

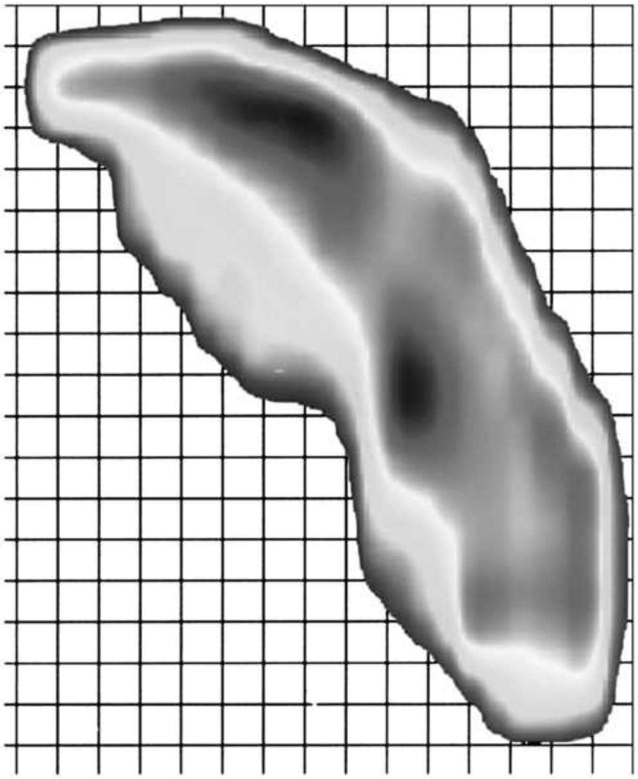


Figure 5. The planar intensity map created by the field from Fig. 4 (black is the highest radiation exposure).

All vendors currently deliver this methodology although with some meaningful differences. As mentioned earlier, the location of the MLC plays a major role in the positional accuracy of any leaf at the treatment distance and there are differences in the time each vendor requires to assure accuracy before initiating the beam on sequence of the accelerator. This time ranges from milliseconds to multiple seconds and will discourage the use of intensity maps with many segments in order to achieve clinically realistic delivery times. The use of this technique may result in segments that have the beam on for a fraction of a second, thereby requiring the accelerator to reach its steady-state performance in that same short time. The accelerator manufacturers have developed this capability on newer accelerators, but one should be careful when attempting IMRT on older accelerators. The second method involves the creation of intensity maps by varying the dose rate while leaves are in motion. The NCI–CWG–IMRT defines this process as Dynamic MLC IMRT or dMLC. Currently, it is available from only one vendor. The technique was postulated by Kallman et al. (11) and the optimal trajectory equations developed by Stein et al. (12), Svennson et al. (13), and Spirou and Chui (14). To successfully deliver such a treatment, the accelerator needs to monitor leaf positions and alter the dose rate as required to allow leaves to reach the next positions at the proper time. While this is the faster means of creating an intensity map, it also has the potential for more discrepancies in leaf position and dose than the static approach. These discrepancies are likely to be small and, with many ports being delivered, are likely to produce small degradation in the desired pattern. It is also possible to identify potential large errors and modify them to be small. This process is usually done with software programs that optimize the segment delivery. By using this segmental optimization, one can control the magnitude of such things as the number of segments, the tongue and groove effect, and minimize the potential for one beamlet to have a much higher intensity than any of the others,

Delivery of IMRT with fixed field techniques is the most common method used because there are so many MLCs in use and the MLC can be used for conventional beam shaping without intensity modulation of the beam. However, the methodology has some drawbacks. Linear accelerators are expensive (~ $1.8 M) and it is important economically to treat large numbers of patients on each accelerator. For this reason, the number of ports to be treated must be limited to a few, optimal orientations. For some disease sites, such as prostate, these orientations can be predetermined and applied to all patients. In other disease sites, such as head and neck tumors, the disease presentation may have unique problems that require additional planning time to seek a satisfactory number of ports and orientations. There is no single answer to this problem as situations at each institution vary. For example, a small center might only have one or two machines while large institutions are likely to have more and can identify one that can treat fewer patients.

The alternative to fixed field IMRT is to deliver the treatment with arc therapy or tomotherapy. Since IMRT has such a strong similarity to CT, this is a logical approach. However, with the proliferation of high energy accelerators over the past 20 years, the use of arc therapy has diminished dramatically and is employed mainly in special procedures such as stereotactic radiosurgery (SRS). There are three implementations of tomotherapy and each has an analogy to current CT technology. In 1992 a neurosurgeon, Dr. Mark P. Carol, introduced a commercial system called Peacock (15) to be used as a SRS tool. The system consisted of a unique multileaf collimator (MIMiC) that consisted of two rows of 20 vanes each, with each vane projecting to a nominal $1 \times 1\,\text{cm}^2$ beam at the normal isocenter of an accelerator (100 cm). The MIMiC is operated pneumatically and is a binary collimator meaning that the vane is either open (and permitting radiation to pass) or closed (blocking radiation). Figure 6 shows a patient's view of the MIMiC vanes in an alternating open/close pattern. One can see the two independent rows. The MIMiC is a removable tertiary collimator, so the accelerator could also be used for traditional clinical treatments. Figure 7 shows the MIMiC and it's associated hardware in place on an accelerator. Either a 5 or 10° sector of the arc delivery is considered a fixed field and a beamlet's intensity is based around the center of the arc sector. For a 10° sector, that means a beam with 100% intensity is open for the entire 10°, while a beam with 50% intensity is closed the first 2.5° of the sector, open for the next 5° and closed the last 2.5°. Because of the MIMiCs design, one treats a 2 cm length of the patient in one arc. It is then necessary to index the patient precisely and deliver another arc. The process is repeated until the total length of the tumor is treated. A CT analogy is a dual slice axial scanner and has the NCI–CWG–IMRT designation as Sequential Tomotherapy. This system also has historical importance as it was used to treat the first patient with IMRT (16) as defined by the NCI–CWG–IMRT and is still popular in 2004.

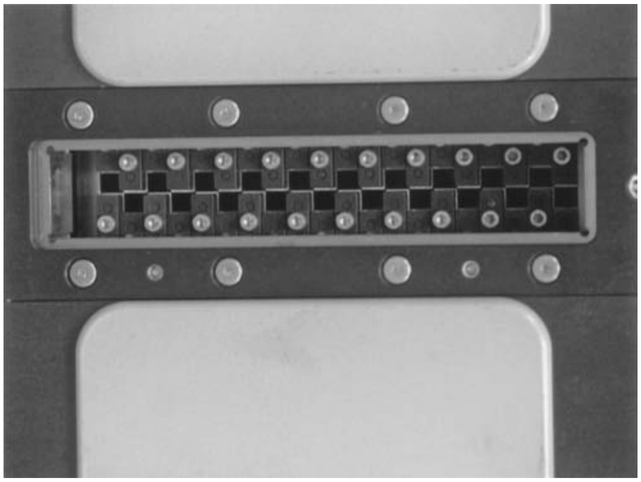


Figure 6. The MIMiC collimator as viewed by the patient. Alternating leaves are open and closed.

In 1995, Yu (17) introduced the concept of intensity modulated arc therapy (IMAT) that uses a traditional MLC to deliver tomotherapy. It basically is a combination of Dynamic Arc, a conformal delivery where the MLC changes its borders as a function of gantry angle but does not modulate the beam in the field, and the creation of intensity maps by using arcs with different shapes multiple times. This technique has limitations in that there are no planning systems that automatically create the fields, so the planner has to develop the skills to do this and that

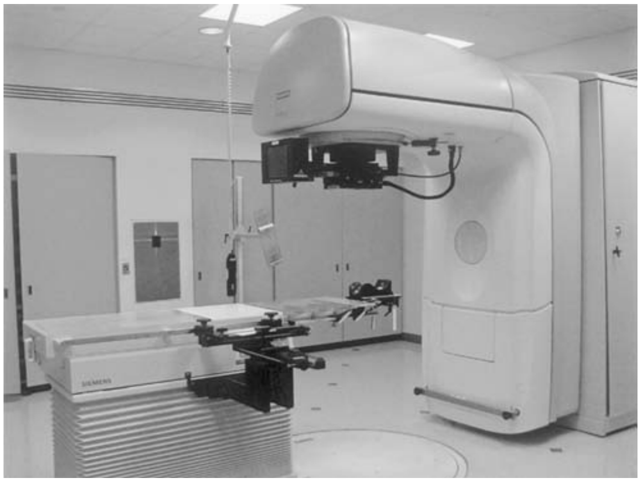


Figure 7. A sequential tomotherapy system mounted on a linear accelerator.

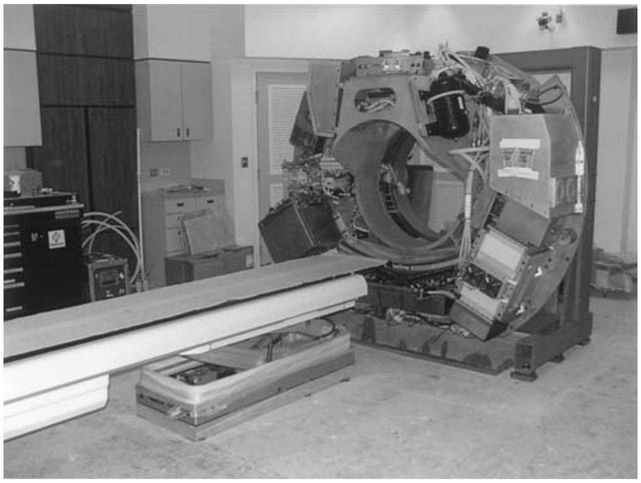


Figure 8. A helical tomotherapy machine at installation. One can see the gantry ring holding the accelerator components.

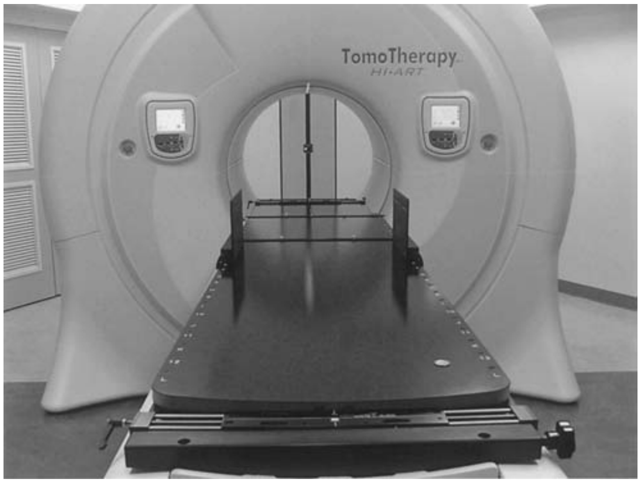


Figure 9. A helical tomotherapy machine in an operational configuration and looking much like a diagnostic CT scanner.

introduces a problem because, given N possible intensities, there are $(N!)^2$ combinations that can produce that pattern. As an example, an intensity map with only 3 intensities can be created using 36 possible patterns. However, with the use of multiple gantry angles, one can create clinically acceptable plans for many simple disease presentations. This technique is extremely useful on machines that would require seconds to verify leaf positions if one were doing sMLC. Since it treats the entire volume in each arc, a CT analogy would be a multislice axial scan. More information on using this technique is found in a more recent article by Yu and Shepard (18).

In 1993, Mackie et al. (19) described a novel machine designed to treat IMRT only with a helical delivery system just like a helical CT scanner. For this reason, the system has been defined as Helical Tomotherapy by the NCI–CWG–IMRT. This machine has the physical appearance of a CT scanner, with a 6 MV waveguide and associated electronics mounted on a rotating annulus. A unit at the time of installation is shown in Fig. 8. The patient is treated with 6 MV X-rays by having the treatment couch moving at a constant speed through the bore of the rotating system. In order to modulate the beam, this machine has a pneumatically driven binary collimator. In addition, this machine had a series of Xenon detectors mounted on the rotating annulus opposite the waveguide so that the radiation exiting from the patient can be captured and analyzed. This machine became available commercially in 2001, with the first clinical installations occurring in 2003. Figure 9 shows a helical tomotherapy machine ready for patient treatment. Because it is the first new design of an external beam treatment machine since the linear accelerator was introduced at Stanford in 1955, it deserves some additional scrutiny.

As with helical CT, the operator can control the pitch (the distance the table travels per revolution) as well as the slice width (the machine has moveable jaws that can be set from 0.5 to 5.0 cm prior to the initiation of treatment). Using a pitch >1 and a large slice width, one can create a megavoltage CT image (the waveguide is tuned for 3.5 MV X-rays) that can be used to verify or correct patient position prior to treatment. During treatment the exit dose can be collected in the xenon detectors and mapped back into the patient to determine the magnitude of any discrepancies between planned and actual delivery, whether they be caused by mechanical errors or patient–organ motion. These types of tools have long been desired to help insure the accuracy of these more conformal dose distributions.

Similar tools are finding their way into conventional linear accelerators as vendors are adding kilovoltage X-ray tubes as well as megavoltage cone CT to their equipment. There are questions as to how one facilitates the use of fixed-field delivery with the need to do arc rotations to gather the necessary information to create the images that are part of the desired schema.

The last issue to address is the use of optimized treatment planning as a necessary part of the IMRT concept. It was stated earlier that perfect shapes in CT often require the application of mathematical filters with negative values and, since we cannot deliver negative radiation, optimization tools are required to achieve the best result possible give the planning constraints. The evaluation of optimization techniques continues today but the application of optimization techniques in use today is very straightforward. They are either a gradient descent or stochastic algorithm. In the late 1980s, Webb (20) as well as Mohan et al. (21) investigated the stochastic algorithm, Simulated Anneling, as a possible algorithm. This was appropriate because the algorithm was powerful and no one was sure just how complicated the optimization need be to create a clinically useful plan. Because of this, Carol designed his system based on Webb's work with this stochastic optimization algorithm.

By their nature, stochastic optimizations are slower than gradient descent algorithms because they permit the solution to get worse for a number of iterations as an attempt to avoid being trapped in a local minimum. Over time, it became clear that gradient descent algorithms are useable and today all planning systems now use gradient descent algorithms and some use both. There is a product that uses no optimization at all and was given an IMRT moniker by the vendor, but this product does not meet the NCI–CWG–IMRT definition of IMRT. Research interest in the subject continues to grow. A literature search of "inverse treatment planning" found 39 publications in 1998, 48 in 2000, and 108 in 2002, so we can expect optimization algorithms to continue to be tested and improved.

The rapid acceptance of IMRT as a delivery technique is unprecedented in radiation therapy. A mere 6 years after Brahme postulated the concept in 1988, Carol had produced a commercial system that was used clinically. Within a decade after that event, not only do vendors supply additional technology for their accelerators to treat IMRT, but also a new vendor has emerged with a dramatically new machine to treat only IMRT. Image guided systems are being adapted to take advantage of the power to shape radiation fields easily and precisely with IMRT. This technology should continue to expand.

BIBLIOGRAPHY

1. NCI–CWG–IMRT, Intensity-modulated radiotherapy: Current status and issues of interest. Int J Radiat Oncol Biol Phys 2001;51:880–914.
2. Brahme A, Roos JE, Lax I. Solution of an integral equation encountered in rotation therapy. Phys Med Biol 1982;27: 1221–1229.
3. Cormack AM, Cormack RA. A problem in rotation therapy with x-rays: Dose distributions with an axis of symmetry. Int J Radiat Oncol Biol Phys 1987;13:1921–1925.
4. Brahme A. Optimization of stationary and moving beam radiation therapy techniques. Radiother Oncol 1988;12:129–140.
5. Powers WE et al. A new system of field shaping for external-beam radiation therapy. Radiology 1973;108:407–411.
6. Sofia JW. Computer controlled, multileaf collimator for rotational radiation therapy. AJR Am J Roentgenol 1979;133: 956–957.
7. Boyer A et al. Basic applications of multileaf collimator; AAPM Report No. 72. Madison (WI): Medical Physics Publishing; 2001.
8. Bortfeld T, Burkelbach J, Boesecke R, Schlegel W. Methods of image reconstruction from projections applied to conformation radiotherapy. Phys Med Biol 1990;35:1423–1434.
9. Webb S. Optimization by simulated annealing of three-dimensional conformal treatment planning for radiation fields defined by a multileaf collimator. Phys Med Biol 1991;36: 1201–1226.
10. Bortfeld T et al. Realization and verification of three-dimensional conformal radiotherapy with modulated fields. Int J Radiat Oncol Biol Phys 1994;30:899–908.
11. Kallman P, Lind B, Eklof A, Brahme A. Shaping of arbitrary dose distributions by dynamic multileaf collimation. Phys Med Biol 1988;33:1291–1300.
12. Stein J, Bortfeld T, Dorschel B, Schlegel W. Dynamic x-ray compensation for conformal radiotherapy by means of multi-leaf collimation. Radiother Oncol 1994;32:163–173.
13. Svensson R, Kallman P, Brahme A. An analytical solution for the dynamic control of multileaf collimators. Phys Med Biol 1994;39:37–61.
14. Spirou SV, Chui CS. Generation of arbitrary intensity profiles by dynamic jaws or multileaf collimators. Med Phys 1994;21: 1031–1041.
15. Carol MP. Integrated 3-D conformal multivane intensity modulation delivery system for radiotherapy. Hounsell AR, Wilkinson JM, Williams PC, editors. Proceedings of the 11th International Conference on the Use of Computers in Radiation Therapy. Madison (WI): Medical Physics Publishing; 1994.
16. Butler EB, Woo SY, Grant 3rd W, Nizin PS. Clinical realization of 3d conformal intensity modulated radiotherapy. Int J Radiat Oncol Biol Phys 1995;32:1547–1548.
17. Yu CX. Intensity-modulated arc therapy with dynamic multileaf collimation: An alternative to tomotherapy. Phys Med Biol 1995;40:1435–1449.
18. Yu C, Shepard D. Treatment planning for stereotactic radiosurgery with photon beams. Technol Cancer Res Treat 2003;2:93–104.
19. Mackie TR, et al. Tomotherapy: A new concept for the delivery of dynamic conformal radiotherapy. Med Phys 1993;20:1709–1719.
20. Webb S. Optimisation of conformal radiotherapy dose distributions by simulated annealing. Phys Med Biol 1989;34: 1349–1370.
21. Mohan R, et al. Clinically relevant optimization of 3-d conformal treatments. Med Phys 1992;19:933–944.

See also COMPUTED TOMOGRAPHY; RADIOTHERAPY, THREE-DIMENSIONAL CONFORMAL.

RADIATION THERAPY SIMULATOR

DANIEL A. LOW
SASA MUTIC
Washington University School of Medicine
St. Louis, Missouri

INTRODUCTION

Radiation therapy, or radiation oncology, is one of the primary modalities (along with surgery and chemotherapy) for the treatment of cancer patients. Its origins can be traced to the early 1900s and today radiation therapy facilities can be found in most major medical centers and many free standing practices. This medical specialty and its success depend strongly on the technology used for cancer diagnosis and planning, delivery, and verification of patient treatments. Therefore, the amount of efforts and resources invested in the improvement of radiation therapy related technologies is significant. One of the cornerstones of modern radiation therapy practices are volumetric patient images, computed tomography (CT), magnetic resonance (MR) imaging, nuclear medicine imaging (positron emission tomography (PET) and single positron emission tomography (SPECT)), and ultrasound (US). Medical images are used for cancer detection, disease staging, treatment planning, for verification of treatment delivery, and for evaluation of treatment outcomes and patient follow up. Imaging devices that are used to image cancer patients for radiation therapy treatment planning are called Radiation Therapy Simulators. The distinguishing characteristics of radiation therapy simulators, in addition to their basic imaging properties, are that these devices need to have the following characteristics; (*1*) the modality allows patients to be imaged in their treatment position, (*2*) that the acquired images have high spatial accuracy, and (*3*) that the dataset be able to provide image datasets that are of sufficient quality to be used for validating the radiation beam shape and anatomic location. While these may seem like relatively straightforward requirements, the design and implementation of radiation therapy simulators can be technically challenging. The main source of technical difficulties stems from the fact that the design of the medical imaging devices on which the simulators are based has historically been driven by the needs of diagnostic radiology that have less concern for patient positioning or for the spatial accuracy of the image datasets. For diagnostic scanning, patients often assume a comfortable position with their arms on the side or on abdomen–chest. Diagnostic physicians typically need to determine the presence of anatomic or functional anomalies, so a quantified determination of the size, shape, or location of internal organs or tumors relative to the imaging modality hardware is not a primary consideration.

For radiation therapy imaging, the patient extremities (arms and legs) are often positioned away from the torso to provide access by the radiation treatment beams. Patients are also imaged in immobilization devices that are subsequently used during treatment. Additionally, image spatial accuracy and the geometry of images is extremely important in order to precisely deliver the radiation to the tumors while avoiding radiation sensitive organs. Radiation therapy simulator design is based on an imaging device that was originally developed for diagnostic imaging, and then the device is modified to accommodate patient imaging in the radiotherapy treatment position and to improve image spatial accuracy and geometry to satisfy the needs of radiation therapy treatment planning. This approach is slowly changing and more devices are being designed exclusively for radiation therapy or major features of diagnostic imaging equipment are designed with radiation therapy in mind. This change in manufacturer attitude is reflected in description of radiation therapy simulators in the rest of this article.

The majority of simulation history in radiation therapy is based on conventional simulators (1–6). However, the modern practice of radiation therapy is dominated by CT simulators. Shortly after the introduction of clinical CT scanners in the early 1970s, it was realized that this imaging modality has much to offer in a radiation oncology setting. The CT images provide volumetric information not only about target volumes, but about critical structures as well. Using CT images for radiation therapy treatment, planning has improved dose delivery to target volumes while reducing dose to critical organs. The CT images also provide relative electron density information for heterogeneity-based dose calculations. A major weakness of CT imaging is a relatively limited soft-tissue contrast. This limitation can be overcome by using CT images in conjunction with MR studies for treatment planning. The PET images can be used to add physiological information. Ultrasound has also been useful for imaging in brachytherapy. Multimodality imaging-based treatment planning and target and normal structure delineation offer an opportunity to better define the anatomic extent of target volumes and to define their biologic properties.

Tatcher (7) proposed treatment simulation with CT scanners. This short article described the feasibility of CT simulator and indicated potential economical benefits. In 1983, Goitein and Abrams (8,9) further described multidimensional treatment planning based on CT images. Sherouse et al. (10,11) went on to describe a CT image based virtual simulation process that they referred to as a "software analog to a conventional simulation". This series of manuscripts described software tools and addressed technical issues that affect today's CT-simulation process. The manuscripts pointed out the need for fast computers, specialized software, but also for improved patient immobilization and setup reproducibility.

The radiation oncology community eagerly embraced the concept of virtual simulation and in early 1990s commercial software packages became available. These systems consisted of a diagnostic CT scanner, external laser positioning system, and a virtual simulation software work station. One of the early commercial CT simulation packages is shown in Fig. 1.

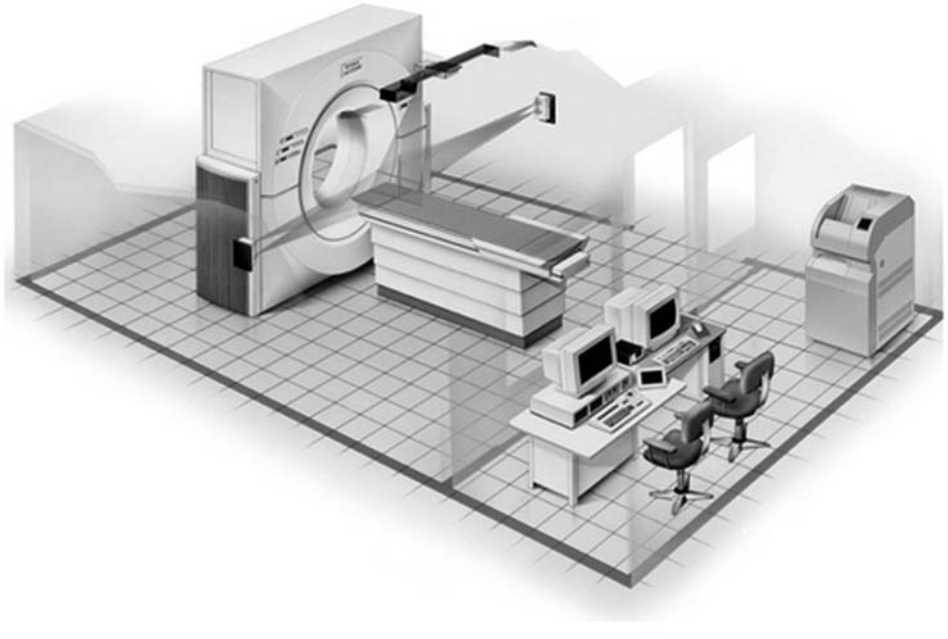


Figure 1. The CT-simulator room layout. (Image courtesy of Philips Medical Systems, Cleveland, Ohio.)

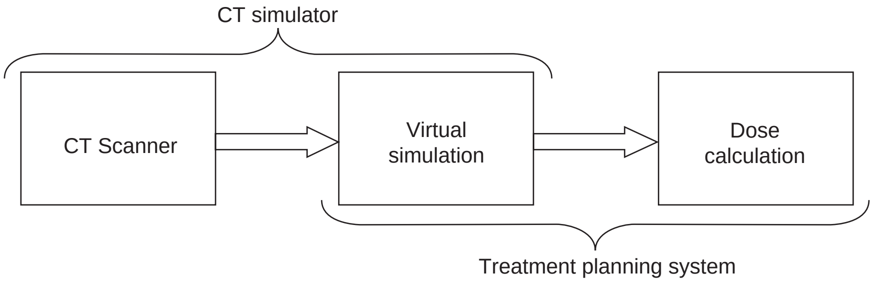


Figure 2. Place of CT simulation in radiotherapy treatment planning process. (Reprinted with permission from Ref. 12.)

The CT simulators have matured to a point where they are one of the cornerstones of modern radiation oncology facilities. Today's systems incorporate specially designed large bore CT scanners, multislice CT scanners, high quality laser positioning systems, and sophisticated virtual simulation packages. Many systems incorporate dose calculation capabilities and treatment plan analysis and evaluation tools.

Additional virtual simulation software features and functions along with increased efficiency and flexibility have enabled CT simulators to replace conventional simulators in many facilities. This trend seems to be further fueled by the increased demand for imaging studies for conformal three-dimensional (3D) and intensity modulated radiation therapy (IMRT) treatment planning where conventional simulators are of limited value. Figure 2 shows the place of CT simulation in the treatment planning process.

Both MR and PET–CT simulators are recent developments in radiation therapy simulation and are designed to complement CT simulation process and shortcomings of CT imaging. Magnetic resonance imaging offers superior soft tissue contrast and PET provides information about biological tissue properties. Computed tomography has relatively poor soft tissue contrast and provides rather limited information about functional tissue characteristics. Both MR and PET imaging in radiation therapy imaging greatly enhance our ability to accurately define anatomical and biological properties of tumors and normal tissues.

The implementation of simulation and treatment planning process varies greatly between radiation oncology departments. This diversity is in part driven by significant technical differences between simulation and treatment planning systems offered by different manufacturers. The discussion of radiotherapy simulators provided here describes general characteristic of processes and technology used for radiation therapy treatment planning. For more specific details, readers are referred to suggested readings list.

TECHNOLOGY OVERVIEW

In the late 1990s, the imaging equipment manufacturers began designing major devices (CT, MR, and PET scanners) specifically for radiation therapy or with radiation therapy needs in mind. This paradigm change resulted in a multitude of imaging devices available for radiation therapy simulation. Not only are there new devices (CT, MR, PET), but conventional simulators are being improved as well in order to be able to compete with other imaging modalities.

CONVENTIONAL SIMULATOR

The radiation therapy simulator has been an integral component of the treatment planning process for > 30 years. Conventional simulators are a combination of diagnostic X-ray machine and certain components of a radiation therapy linear accelerator (1–6). A conventional simulator, as seen in Fig. 3, consists of a diagnostic X-ray unit and fluoroscopic imaging system (X-ray tube, filters, collimation, image intensifier, video camera, generator, etc. (13), patient support assembly (a model of the treatment table), laser patient positioning and marking system, and simulation and connectivity software. The treatment table and the gantry are designed to mimic the geometric functions of a linear accelerator. The gantry head is designed to accommodate the common beam modification devices (blocks, wedges, compensating filters), in a geometry that mimics the linear accelerator. The simulator provides transmission radiographs with radiation portal-defining collimator settings outlined by delineator wires. By using primarily bony landmarks, the physician delineates the radiation portal outlines.

Imaging chain: One of the major improvements in conventional simulator design was the replacement of image intensifiers and video camera systems by amorphous silicon detectors. The new imagers produce high spatial and contrast resolution images which approach film quality, Fig. 4. More importantly, these images are distortion-free, a feature that is important for accurate geometric representation of patient anatomy. The introduction of high quality digital imagers in conventional simulation further facilitates the concept of film-less radiation oncology departments.

Simulation software: Conventional simulation software has also undergone many improvements. Modern simulators use the Digital Image Communications in Medicine (DICOM) standard (14) for data import capabilities. Treatment field parameters can be imported directly from the treatment planning computer for verification on the simulator. The software can then automatically set the simulator parameters according to the treatment plan. This facilitates efficient and accurate verification of patient

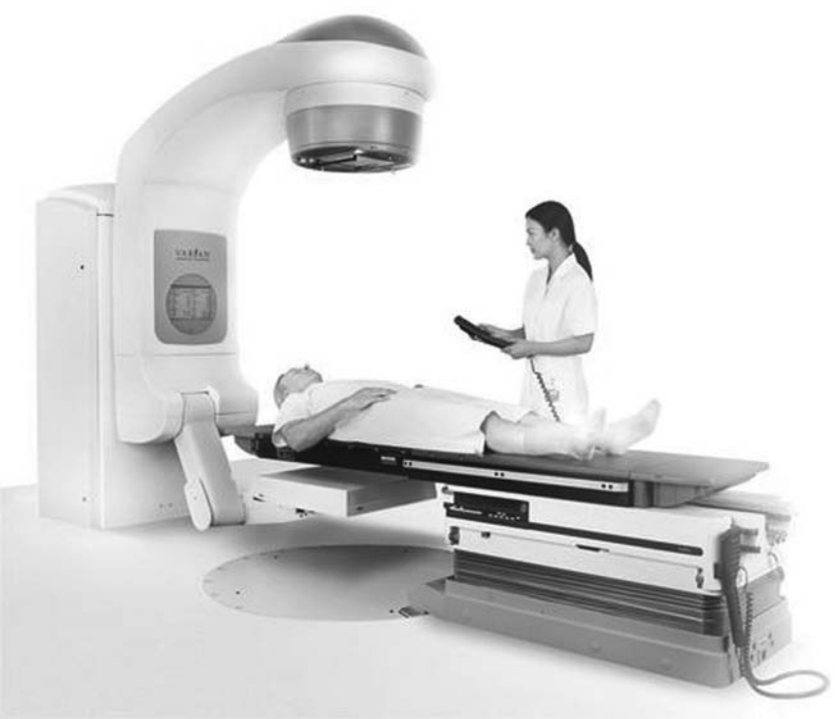


Figure 3. Modern version of a conventional simulator. (Image courtesy of Varian Medical Systems, Palo Alto, California, copyright © 2002.)

treatment setup on the conventional simulator. These simulators also have DICOM export capabilities that improves the reliability of treatment setup parameter transfer directly to a record and verify system or to a treatment planning computer. The ability to import and capture digital images enables conventional simulators to have tools for automatic correlation of treatment planning and verification fields.

Another potential improvement for conventional simulators is the capability of providing cone-beam CT (15,16). Because newer simulators are equipped with digital imaging hardware, two-dimensional (2D) projection image datasets are acquired while the gantry is rotating. The projection data is used to reconstruct a high spatial resolution CT image dataset. This capability will significantly improve imaging capabilities and usefulness of these devices. Figure 5 shows a cone beam CT image from a conventional simulator.

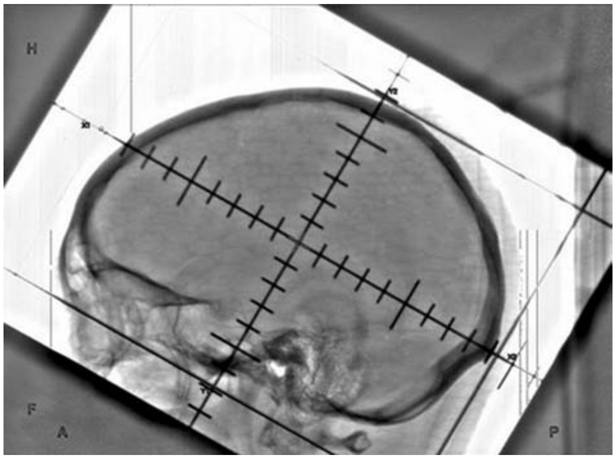


Figure 4. Digital image of a head from a modern conventional simulator equipped with an amorphous silicon imager. (Image courtesy of Varian Medical Systems, Palo Alto, California, copyright © 2002.)

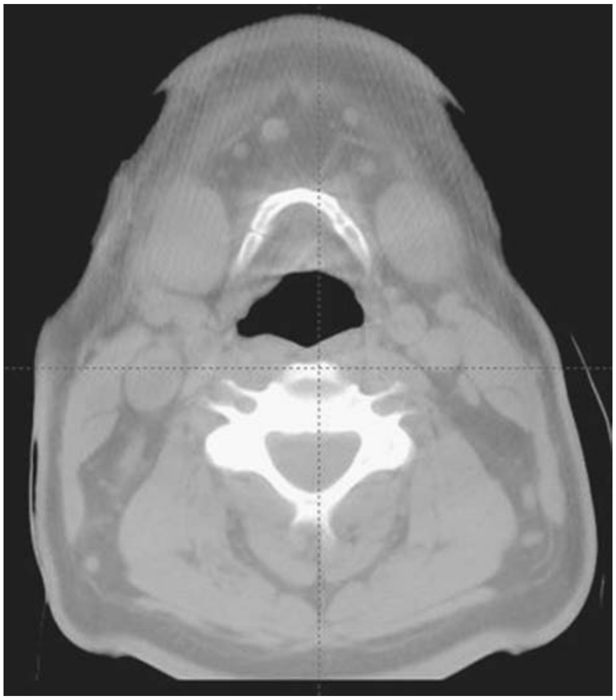


Figure 5. Cone beam CT image of a head acquired on a conventional simulator. (Image courtesy of Varian Medical Systems, Palo Alto, California, copyright © 2002.)

While it is often mentioned that conventional simulators can be completely replaced with CT simulators, new features and usefulness of conventional simulators are slowing down this process. Conventional simulator continues to be an important component of radiotherapy process even though its use for treatment planning of many tumor sites has been significantly reduced.

CT Simulator

Computed tomography simulator (10,11,17–26) consists of a CT scanner, laser patient positioning–marking system, virtual simulation–3D treatment planning software, and different hardcopy output devices, Fig. 1.

The CT scanner is used to acquire volumetric CT scan of a patient that represents the virtual patient and the simulation software recreates the functions of a conventional simulator. In recent years, the three most significant changes in CT-simulation technology have been the introduction of a larger gantry bore opening (Large Bore CT) (27), multislice image acquisition (Multislice CT) (28), and addition of CT-simulation software directly on the CT scanner control console. These innovations improve the efficiency and accuracy of the CT-simulation process. They also improve the patient experience by allowing patients to be positioned in more comfortable positions while reducing the simulation procedure time.

Large Bore CT: Large bore CT scanners (defined here as having > 70 cm diameter bores) were specifically designed with radiation therapy needs in mind. One of the requirements in treatment of several cancer sites (breast, lung, vulva, etc.) is for extremities to be positioned away from the torso. When acquiring a CT scan with a patient in such treatment position, extremities often cannot fit through a conventional 70 cm diameter scanner bore opening. In such situations, patient positioning needs to be modified to acquire the scan. This can result in less than optimal treatment position (patient may be less comfortable and therefore the daily setup reproducibility may be compromised) or in a mismatch between the imaging and treatment positions. The first large bore CT simulator was introduced in 2000, and several additional models with enlarged bore opening have been introduced since then.

Large bore scanners also have increased the available scan field of view (SFOV), which determines the largest dimension of an object that can be fully included in the CT image. It is typically 48–50 cm in diameter on most conventional 70 cm bore opening scanners. For treatment planning purposes it is necessary to have the full extent of the patient's skin on the CT image. Lateral patient separation can often be $>$ 48–50 cm and the skin is then not visible on CT images. Increased SFOV available on large bore scanners solves this problem. There are, however, differences in implementation of extended SFOV and validity of quantitative CT values (quantitative CT) at larger image sizes. The CT numbers for some scanners are accurate only for smaller SFOVs and the values toward the periphery of large SFOV images are not reliable. This can be a concern for some dose calculation algorithms because inaccurate CT numbers can lead to dose calculation errors. The impact of CT number accuracy for increased SFOV images on dose calculation accuracy should be evaluated during scanner commissioning.

Multislice CT: In 1992, Elscint introduced a scanner that had a dual row of detectors and could simultaneously acquire two images (slices). Since then, multislice CT has gained wide spread acceptance and scanners that can acquire up to 64 slices are now available from all major vendors. The basic design behind the multislice CT technology is that multiple rows of detectors are used to create several images for one rotation of the X-ray tube around the patient.

One of the obstacles for radiation therapy scanning with single-slice scanners is the limited tube heat loading capability. Often, fewer images are taken, slice thickness is increased, the image quality is decreased (reduced mAs), or scan pitch is increased to reduce the amount of heat produced during the scan and to allow for the entire scan to be acquired in a single acquisition. Due to the longer length of imaged volume per tube rotation (multiple slices acquired simultaneously), the tube heat loading for a particular patient volume is reduced when using a multislice scanner relative to a single-slice scanner and multislice scanners are generally not associated with tube heat loading concerns. Faster acquisition times and decreased tube loading of multislice scanners, which allow longer volumes to be scanned in a single acquisition, can provide an advantage over single-slice systems for treatment planning purposes. Multislice technology can be especially beneficial for imaging of the thorax where breathing artifacts can be minimized with faster scanning. Multislice technology also facilitates dynamic CT scanning, often referred to as 4D CT (29,30). This application of multislice CT in radiation therapy is yet to be fully explored.

Multislice scanners are also capable of acquiring thinner slices that can result in better quality digitally reconstructed radiographs, used for treatment portal validation, and more accurate target delineation because of the improved spatial resolution with thinner slices, (Fig. 6).

CT simulator tabletop: This section and discussion about simulator tabletops applies equally to all simulators used in radiation therapy (conventional, MRI, CT, and PET) and treatment machines. Tabletops used for patient support in radiation therapy during imaging or treatment should facilitate easy, efficient, reproducible, and accurate patient. It is not only important that a tabletop improves patient positioning on a single device (i.e., treatment machine), but the repositioning of a patient from one imaging or treatment device to another also has to be considered. A great improvement in this process is if all tabletops involved in patient simulation and treatment have a common design. They do not have to be identical, but they should have the same dimensions (primarily width), flex and sag under patient weight, and they should allow registration (indexing) of patient immobilization devices to the tabletop. Figure 7 demonstrates this concept. The CT simulator tabletop has the same width as the linear accelerator used for patient treatment and both allow registration of patient immobilization system to the treatment couch. The ability to register the immobilization device and the patient to a treatment table is extremely important and improves immobilization, set-up

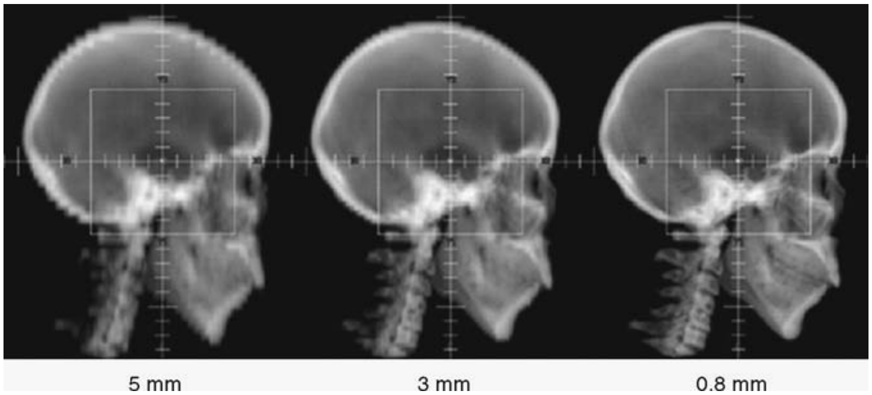


Figure 6. The CT slice thickness DRRs are 5, 3, and 0.8 mm. Thinner slice thickness images reveal much more relevant anatomical detail.

reproducibility, accuracy, and efficiency. The patient is always positioned in the same place on the treatment machine and patient daily setup can be facilitated using the treatment couch positions. If the patient is registered to the treatment couch, the coordinates of the couch used for patient treatment can become a part of parameters that are set and tracked in the linear accelerator record and verify system.

Patient marking lasers: A laser system is necessary to provide reference marks on patient skin or on the immobilization device. Figure 1 shows a laser system for a CT simulator:

Wall lasers: Vertical and horizontal, mounted to the side of the gantry. These lasers can be fixed or movable.

Sagittal laser: Ceiling or wall mounted single laser, preferably movable. Scanner couch can move up/down and in/out, but cannot move left/right, therefore the sagittal laser should move left–right to allow marking away from patient mid line.

Scanner lasers: Internally mounted, vertical and horizontal lasers on either side of the gantry and an overhead sagittal laser.

MR Simulator

The MR images for radiotherapy treatment planning are usually acquired in diagnostic radiology departments because few radiation oncology departments have a dedicated MR scanner. Furthermore, currently the majority of radiotherapy MR studies are limited to brain imaging. The MR scanner has a superior soft tissue contrast compared to CT imaging and there are several benefits that MR can offer for target delineation based on this advantage. There have been several reports describing use of MR scanners for imaging and treatment simulation in radiotherapy (31–35). Some of these reports have suggested that MR studies can be used without a corresponding CT scan for radiotherapy treatment planning. Indeed, if spatial distortions (the geometry of imaged objects is not always reproduced correctly), which is the largest concern with MR imaging, can be removed or minimized MR studies can be used as the primary imaging modality for several treatment sites. Superior soft tissue contrast provided by MR can also be an advantage for treatment planning of certain extracranial tumor sites like prostate, for example (36,37).

Conventional MR scanners are not well suited for extracranial imaging for treatment planning. The main difficulty is placement of patient in treatment position with immobilization device in the scanner. The small diameter and long length of conventional MR scanner openings significantly limits patient positioning options for imaging. Open MR scanners do not share these difficulties and patients can be scanned in conventional

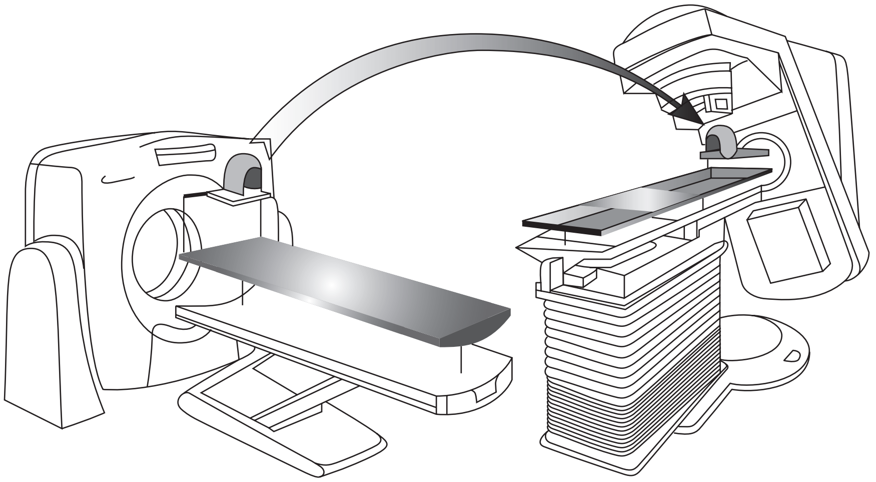


Figure 7. Similarity in design of simulator and treatment machine tabletops allows efficient and accurate reproducibility of patient positioning. (Image courtesy of MED-TEC, Inc, Orange City, Iowa.)

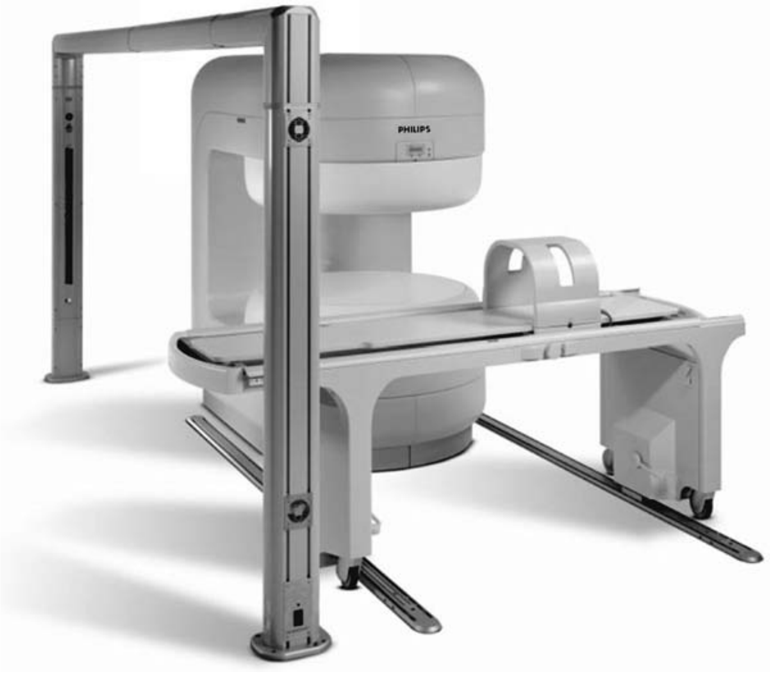


Figure 8. A MR simulator. (Image courtesy of Philips Medical Systems, Cleveland, Ohio.)

treatment positions. At least one manufacturer offers an open MR scanner that has been modified to serve as a radiotherapy simulator, Fig. 8. The scanner table is equipped with a flat top and external patient alignment lasers. The geometry of the scanner is then similar to the CT simulator. Another manufacturer offers a 70 cm diameter gantry opening conventional MRI scanner. The depth of the scanner opening is 125 cm. The dimensions of this scanner are very similar to a conventional CT scanner and in fact the scanner could be mistaken for a CT scanner. The ergonomics of this scanner are also well suited for radiotherapy simulation.

One of the major challenges with MR imaging for radiotherapy treatment planning are geometric distortions in acquired images. The MR scanners are often equipped with correction algorithms that will minimize geometrical distortions. These corrections do not affect the entire image and only the center portion of the image (center 20–35 cm diameter) is adequately correct (within 2 mm). Therefore, the representation of patient's skin and peripheral anatomy for larger body sections may be inaccurate. The effect of these inaccuracies must be evaluated if dose distributions will be calculated directly on MR images.

PET–CT Simulator

The PET images for radiotherapy planning can come from a standalone PET scanner or a combined PET–CT unit. Combined PET–CT scanners are being installed in radiation oncology departments and are used for PET scanning, but also these machines can be used for CT scanning only without PET acquisition. Due to this purpose, these scanners can be classified as CT simulators, though PET–CT simulator term may be more appropriate. Combined PET–CT scanners offer several advantages for radiotherapy imaging and are generally preferred over stand-alone units.

The first combined PET–CT prototype was introduced in 1998 at the University of Pittsburgh (38), since then all major manufacturers have produced several commercial models. The key description of PET–CT scanners is that a PET and a CT scanner are *combined* in the same housing. Meaning that there are two gantries (PET and CT) combined in one housing sharing a common couch. Image reconstruction and scanner operation is increasingly performed from one control console.

Combined PET–CT scanner design varies among different vendors with respect to PET detectors, image quality and resolution, speed, image field of view; number of CT slices, scanner couch design, gantry bore opening, and other considerations. Currently, the commercially available scanners have a 70 cm gantry opening for the CT portion, though large bore CT scanners will likely become part of PET–CT scanners in the future. The PET gantry opening ranges in diameter from 60 to 70 cm, meaning that some of the commercial scanners have a nonuniform gantry opening as the patient travels from the CT portion of the scanner to the PET side. More importantly, the scanners with the smaller gantry opening on the PET side will pose the same difficulties for radiotherapy scanning as stand-alone PET scanners. Again, the size of patient immobilization devices and patient scan–treatment position will have to be adapted to the size of the gantry opening.

The combined PET/CT technology offers two major benefits for radiotherapy planning. First, because the images are acquired on the same scanner, providing that the patient does not move between the two studies, the patient anatomy will have the same coordinates in both studies. These images have been registered using hardware rather than software registration. The second benefit of the combined PET–CT units is that CT images are used to measure attenuation correction factors (ACFs) for the PET emission data, obviating the need for a time-consuming PET transmission scan (39,40). The use of CT images to generate PET ACFs reduces the scan time up to 40% and also provides essentially noiseless ACFs compared with those from standard PET transmission measurements (41). Shorter scan times can benefit radiotherapy patients who are scanned in treatment position that often can be uncomfortable and difficult to tolerate for prolonged amounts of time. One of the concerns with ACFs generated from CT images is mismatch or misalignment between CT and PET images due to respiration motion. The PET images are acquired during many cycles of free breathing and CT images are acquired as a snapshot in time at full inspiration, partial inspiration, or some form of shallow breathing. The breathing motion will cause mismatch in anatomy between PET and CT images in the base of lung and through the diaphragm region. This mismatch can result in artifacts in these areas that may influence diagnosis and radiotherapy target definition in this region. There are various gating methods that can be used during image acquisition to minimize the motion component and essentially acquire true, motionless, images of patient anatomy. Gated or 4D CT (with time being the fourth dimension) can be used to generate more reliable ACFs and also for radiotherapy treatment planning where gated delivery methods are being used.

Virtual Simulation Software

As with all software programs, user-friendly, fast, and well functioning virtual (CT) simulation software with useful features and tools will be a determining factor for success of a virtual simulation program. Commercially available programs far surpass in-house written software and are the most efficient approach to virtual simulation. Several features are very important when considering virtual simulation/3D treatment planning software:

Contouring and localization of structures: Contouring and localization of structures is often mentioned as one of the most time consuming tasks in the treatment planning process. The virtual simulation software should allow fast user-friendly contouring process with help of semiautomatic or automatic contouring tools. An array of editing tools (erase, rotate, translate, stretch, undo) should be available. An ability to add margins in three dimensions and to automatically draw treatment portals around target volumes should be available. An underlining emphasis should be functionality and efficiency.

Image processing and display: Virtual simulation workstation must be capable of processing large volumetric sets of images and displaying them in different views as quickly as possible (near real-time image manipulation and display is desired). The quality of reconstructed images is just as important as the quality of the original study set. The reconstructed images (DRRs and multiplanar reconstruction) are used for target volume definition and treatment verification and have a direct impact on accuracy of patient treatments.

Simulator geometry: A prerequisite of virtual simulation software is the ability to mimic functions of a conventional simulator and of a medical linear accelerator. The software has to be able to show gantry, couch, collimator, and jaw motion, SSD changes, beam divergence, and so on. The software should facilitate design of treatment portals with blocks and multileaf collimators.

DISCUSSION

As radiotherapy treatment planning and delivery technology and techniques change, so does the treatment simulation. The most significant change in the recent past has been the wide adoption of CT simulation to support conformal radiotherapy and 3D treatment planning. A CT simulation has gone from a concept practiced at few academic centers to several available sophisticated commercial systems located in hundreds of radiation oncology departments around the world. The concept has been embraced by the radiation community as a whole. The acceptance of virtual simulation comes from improved outcomes and increased efficiency associated with conformal radiation therapy. Image-based treatment planning is necessary to properly treat a multitude of cancers and CT simulation is a key component in this process. Due to demand for CT images, CT scanners are commonly found in radiation oncology departments. As CT technology and computer power continue to improve so will the simulation process, and it may no longer be based on CT alone. The PET–CT combined units are commercially available and could prove to be very useful for radiation oncology needs. Several authors have described MR simulators where the MR scanner has taken the place of the CT scanner. It is difficult to predict what will happen over the next 10 years, but it is safe to say that image based treatment planning will continue to evolve.

One great opportunity for an overall improvement of radiation oncology is the better understanding of tumors through biological imaging. Biological imaging has been shown to better characterize the extent of disease than anatomical imaging and also to better characterize individual tumor properties. Enhanced understanding of individual tumors can improve selection of the most appropriate therapy and better definition of target volumes. Improved target volumes can utilize the full potential of IMRT delivery. Biological imaging can also allow evaluation of tumor response and possibly modifications in therapy plan if the initial therapy is deemed not effective.

Future developments in radiotherapy treatment planning simulation process will involve the integration of biological imaging. It is likely that this process will be similar to the way that CT scanning was implemented in radiotherapy. The imaging equipment is initially located in diagnostic radiology facilities and as the demand increases the imaging is gradually moved directly to radiation oncology.

BIBLIOGRAPHY

1. Bomford CK, et al. Treatment Simulators. BJR 1989; (Suppl. 23).

2. Connors SG, Battista JJ, Bertin RJ., On the technical specifications of radiotherapy simulators. Med Phys 1984;11:341–343.
3. Greene D, Nelson KA, Gibb R., The use of a linear accelerator "simulator" in radiotherapy. BJR 1964;37:394–397.
4. McCullough EC., Radiotherapy treatment simulators, in Advances in radiation oncology physics: dosimetry, treatment planning, and brachytherapy. In: Purdy JA, ed. Woodbury (NY): American Institute of Physics; 1992. pp 491–499.
5. McCullough EC, Earle JD., Selection, acceptance testing and quality control of radiotherapy simulators. Radiology 1979; 131:221–230.
6. Van Dyk J, Munro PN. Simulators. In: Van Dyk J, Editor. The modern technology in radiation oncology. Wisconsin Medical Physics Publishing; 1999. pp 95–129.
7. Tatcher M., Treatment simulators and computer assisted tomography. BJR 1977;50:294.
8. Goitein M, Abrams M. Multi-dimensional treatment planning: I. Delineation of anatomy. Int J Rad Oncol, Biol, Phys 1983;9(6): 777–787.
9. Goitein M, et al. Multi-dimensional treatment planning: II. Beam's eye-view, back projection, and projection through CT sections. Inter J Rad Oncol, Biol, Phys 1983;9(6):789–797.
10. Sherouse G, et al. Virtual Simulation: Concept and Implementation. In: Bruinvis IAD, et al. ed. Ninth Int Conf Use of Computers in Radiation Therapy. North-Holland Publishing Co.; 1987. pp 433–436.
11. Sherouse GW, Bourland JD, Reynolds K. Virtual simulation in the clinical setting: some practical considerations. Int J Radiat Oncol Biol Phys 1990;19:1059–1065.
12. Mutic S, et al. Quality assurance for CT simulators and the CT simulation process: Report of the AAPM Radiation Therapy Committee Task Group No. 66. Med Phys 2003;30: 2762–2792.
13. Bushberg JT, et al. Fluoroscopy, in The Essential Physics of Medical-Imaging. 2nd ed. Baltimore: Lippincott Williams & Wilkins; 2002.
14. (NEMA), N.E.M.A. Digital Imaging Communications in Medicine (DICOM). 1998.
15. Jaffray DA, et al. A radiographic and tomographic imaging system integrated into a medical linear accelerator for localization of bone and soft-tissue targets. Int J Rad Oncol, Biol, Phys 1999;45:773–789.
16. Jaffray DA, et al. Flat-panel cone-beam computed tomography for image-guided radiation therapy. Int J Rad Oncol, Biol, Phys 2002;53:1337–1349.
17. Kushima T, Kono M. New development of integrated CT simulation system for radiation therapy planning. Kobe J Med Sci 1993;39(5–6):197–213.
18. Nagata Y, et al. CT simulator: a new 3-D planning and simulating system for radiotherapy: Part 2. Clinical application [see comments]. Int J Rad Oncol, Biol, Phys 1990;18(3): 505–513.
19. Nishidai T, et al. CT simulator: a new 3-D planning and simulating system for radiotherapy: Part 1. Description of system [see comments]. Int J Rad Oncol, Biol, Phys 1990;18(3): 499–504.
20. Butker EK, et al. Practical Implementation of CT-Simulation: The Emory Experience. In: Purdy JA, Starkschall G, eds. A Practical Guide to 3-D Planning and Conformal Radiation Therapy. Middleton (WI): Advanced Medical Publishing; 1999. pp 58–59.
21. Coia LR, Schultheiss TE, Hanks G, eds. A Practical Guide to CT Simulation. Madison (WI): Advanced Medical Publishing; 1995.
22. Conway J, Robinson MH. CT virtual simulation. Br J Radiol 1997;70:S106–S118.
23. Galvin JM. Is CT simulation the wave of the future? [letter; comment]. Med Phys 1993;20(5):1565–1567.
24. Heidtman CM. Clinical applications of a CT-simulator: precision treatment planning and portal marking in breast cancer. Med Dosimetry 1990;15(3):113–117.
25. Jani SK, ed. CT Simulation for Radiotherapy. Madison (WI): Medical Physics Publishing; 1993.
26. Van Dyk J, Taylor JS. CT-Simulators. In: Van Dyk J, ed. The Modern Technology for Radiation Oncology: A Compendium for Medical Physicist and Radiation Oncologists. Madison (WI): Medical Physics Publishing; 1999. pp 131–168.
27. Garcia-Ramirez JL, et al. Performance evaluation of an 85 cm bore X-ray computed tomography scanner designed for radiation oncology and comparison with current diagnostic CT scanners. Int J Rad Oncol, Biol, Phys 2002;52:1123–1131.
28. Klingenbeck_Regn K, et al. Subsecond multi-slice computed tomography: basics and applications. Eur J Radiol 1999;31(2): 110–124.
29. Keall P. 4-dimensional computed tomography imaging and treatment planning. Sem Rad Oncol 2004;14:81–90.
30. Low DA, et al. A method for the reconstruction of 4-dimensional synchronized CT-scans acquired during free breathing. Med Phys 2003;30:1254–1263.
31. Mah D, Steckner M, Palacio E. Characteristics and quality assurance of dedicated open 0.23 T MRI for radiation therapy simulation. Med Phys 2002;29:2541–2547.
32. Potter R, Heil B, Schneider L. Sagittal and coronal planes from MRI for treatment planning in tumors of brain, head and neck: MRI assisted simulation. Radiother Oncol 1992;23: 127–130.
33. Okamoto Y, Kodama A, Kono M. Development and clinical application of MR simulation system for radiotherapy planning with reference to intracranial and head and neck regions. Nippon Igaku hoshasen Gakkai Zasshi 1997;57(4): 203–210.
34. Schubert K, et al. Possibility of an open magnetic resonance scanner integration in therapy simulation and three-dimensional radiotherapy planning. Strahlenther Onkol 1999; 175(5):255–231.
35. Beavis AW, Gibbs P, Dealey RA. Radiotherapy treatment planning of brain tumors using MRI alone. BJR 1998; 71:544–548.
36. Chen L, et al. MRI based treatment planning for radiotherapy: dosimetric verification of prostate IMRT. Int J Rad Oncol, Biol, Phys 2004;60:636–647.
37. Lee YK, Bollet M, Charles-Edwards G. Radiotherapy treatment planning of prostate cancer using magnetic resonance imaging alone. Radiother Oncol 2003;66(2):203–216.
38. Beyer T, et al. A combined PET/CT scanner for clinical oncology. J Nuclear Med 2000;41:1369–1379.
39. Bailey DL. Data acquisition and performance characterization in PET. In: Valk PE, et al. eds. Positron emission tomography: Basic science and clinical practice. London: Springer-Verlag; 2003. pp 69–90.
40. Bailey DL, Karp JS, Surti S. Physics and Instrumentation in PET. In: Valk PE, et al. eds. Positron Emission Tomography: Basic Science and Clinical Practice. London: Springer-Verlag; 2003. pp 41–67.
41. Townsend DW, et al. PET/CT today and tomorrow. J Nucl Med 2004;45(Supl 1):4s–14s.

See also RADIATION DOSIMETRY FOR ONCOLOGY; RADIOTHERAPY TREATMENT PLANNING, OPTIMIZATION OF; X-RAY EQUIPMENT DESIGN; X-RAY QUALITY CONTROL PROGRAM.

RADIATION THERAPY TREATMENT PLANNING, MONTE CARLO CALCULATIONS IN

Chen-Shou Chui
Ellen Yorke
Ren-Dih Sheu
Memorial Sloan-Kettering
Cancer Center
New York City, New York

INTRODUCTION

Boltzmann Transport Equation

The distribution of radiation in a three-dimensional (3D) heterogeneous medium is governed by the Boltzmann transport equation, which is a partial differential integral equation in six dimensions [3D for position, two (2D) for direction, and one (1D) for energy].

Due to the complex nature of the Boltzmann transport equation, analytic solutions are generally not available except for very simple, idealized cases. For realistic problems, numerical methods are needed. The Monte Carlo method is probably the most accurate and widely used method for solving radiation transport problems in 3D, heterogeneous geometry.

Random Sampling, Law of Large Numbers, Central Limit Theorem

The basic idea of the Monte Carlo method is to simulate physical events by random sampling from known probability distributions. For example, the step size a photon particle travels before the next interaction is sampled from the exponential distribution. The particular interaction event is sampled from the relative probabilities of competing interaction types. If a Compton event occurs, the energy and direction of the outgoing photon and electron are sampled from the Klein–Nishina distribution.

There are two mathematical principles underlying the Monte Carlo method: (1) The law of large numbers, and (2) the central limit theorem. The law of large numbers says that as the sample size increases to infinity, the sample average approaches the mean of the probability distribution from which the samples were drawn. Since in practice the sample size is finite, the error of the sample average needs to be estimated. For this, we make use of the central limit theorem, which states that the distribution of the sample averages approaches a normal distribution if the sample size is sufficiently large. Moreover, the standard deviation of the mean is inversely proportional to the square root of the total number of histories. Thus, in order to reduce the standard deviation of the mean by a factor of 2, the number of histories needs to be increased by a factor of 4. To estimate the statistical uncertainty, it is common practice to divide the total number of histories into separate groups, with each group containing sufficiently large number of histories. The sample average of each group, according to the central limit theorem, is normally distributed. The standard deviation of the mean is then calculated from these group averages. Alternatively, a history-by-history method can be used to estimate the standard deviation of the mean. In this method, both the quantity of interest and its square are tallied on the fly for each history. After the simulation is completed, the standard deviation of the mean is then calculated using the sum and the sum of squares of all histories. This method tends to have smaller uncertainty in the uncertainty estimate than the multiple group method.

Applications of Monte Carlo Method in Medical Physics

The Monte Carlo method has been applied to a variety of medical physics problems (1). For radiation therapy, these include the simulation of the machine head (2–22); dose calculation for external photon beams (23–38); electron beams (39–49); proton beams (50–54); and brachytherapy (55–86). For nuclear medicine, it has been used to calculate organ doses due to internal emitters (87–97). For diagnostic radiology, it is used for calculating beam characteristics and dosimetry (98–111). For radiation measurement, various correction factors for ion chambers have been calculated with Monte Carlo (112–129).

A number of Monte Carlo codes have been developed and applied to problems in medical physics (39,43,130–141). For the purpose of radiation therapy treatment planning, the EGS4 system and its user codes (43,130,133,135) are probably the most widely used in North America.

MONTE CARLO SIMULATION

Overview

A Monte Carlo simulation consists of two major components: transport and interaction. Transport moves a particle from one position to another while interaction determines the outcome of a particular interaction event. If an interaction results in multiple outgoing particles, as in the case of Compton scattering or pair production, then each outgoing particle forms its own transport-and-interaction track. This repetitive transport-and-interaction continues until either the particle travels out of the geometry, for example, exiting the patient body, or its energy falls below a cutoff energy.

Geometry Specification

For a given Monte Carlo simulation, the 3D geometry needs to be described by a collection of mathematical objects. One such package, called combinatorial geometry (142) has been used by a number of Monte Carlo codes. The combinatorial geometry package provides a set of primitive objects, such as sphere, cylinder, box, cone, and so on. An object in question is then modeled by logical combinations of these primitives. For example, a hemisphere can be formed as an intersection of a sphere and a box. The physical properties of the object are also assigned, including the material composition and the physical density. The combinatorial geometry package is a powerful tool that can be used to define very complex geometric objects such as the treatment machine head.

If a 3D image set, such as computed tomography (CT), is used for dose calculation, then the entire 3D voxel array defines the geometry. Since the CT image is typically acquired with kilovoltage X rays, when it is used for dose calculation for megavoltage photons or electrons, the CT Hounsfield number needs to be converted into electron density ratios relative to water. For radiation therapy dose calculations, the typical energy range is up to 20 MeV and the physical properties of the voxel can often be considered as water equivalent but with varying densities. If, however, consideration of material composition is important, such as a metal implant, then different materials can be assigned to the corresponding voxels. This can be done by creating a lookup table for material (air, lung, fat, muscle, water, bone, soft bone, metal) as a function of Hounsfield number and then subdividing each material as a function of density.

Transport

During particle transport, the step size is sampled from the exponential distribution. The exponent in the exponential distribution is related to the mean free path that depends on the particle energy and the material in which the particle travels. For neutral particles (photons or neutrons), the step size is relatively large. For example, the mean free path of a ^{60}Co photon in water is $\sim$ 16 cm. For charged particles (electrons, positrons, protons), the step size is very short due to the Coulomb force. As a result, direct simulation would be very inefficient. To overcome this problem, multiple steps are condensed into a single step, called the condensed-history step (143). Energy deposition due to continuous slowing down is calculated along this step and angular deflection is sampled at the end of the step based on multiple scattering theory.

When transporting a particle, it may cross the boundary of one geometric object to another. When this happens, the original step has to be truncated at the boundary. This is because the next geometric object may have different physical properties from the current one. The remaining step size will have to be adjusted based on the new object and resampled if necessary.

For charged particles, there is continuous energy loss due to ionization and excitation. The energy can be deposited uniformly along the step or at a point randomly selected within the step. For neutral particles, there is no energy deposition during a step. However, if the KERMA (Kinetic Energy Released per unit Mass) approximation is made, then energy can also be deposited using the energy absorption coefficient and the step length, which is an estimate of the photon fluence.

Interaction Types

For radiation therapy dose calculation, we are mostly interested in photons and electrons, so the discussion of interaction types here is limited to these radiation particles only.

For photons, the interaction types are photoelectric, Compton scattering, and pair production. Coherent scattering is relatively unimportant as it involves no energy loss and only small angular deflection. In a photoelectric interaction, the incoming photon collides with an atom and ejects one of the bound electrons (typically K shell). The accompanying fluorescent X rays are low energy photons and are usually ignored in radiation therapy dose calculations. In a Compton scattering event, the incoming photon knocks off a loosely bound electron, considered as a free electron, from the atom. The photon itself is deflected with a lower energy. This is the dominant event for interaction of megavoltage photons with matter. The angular distribution of the outgoing particle is governed by the Klein–Nishina formula, and the angle of the outgoing particle uniquely defines its energy. In a pair production, the incoming photon is absorbed in the field of the nucleus and a positron–electron pair is produced. For this interaction to occur, the photon energy must be $>$ 1.022 MeV, the sum of the rest mass energies of the positron–electron pair.

For electrons and positrons, the discrete interaction types are bremsstrahlung production, delta-ray production, and positron annihilation. Bremsstrahlung production is caused by the deceleration of charged particles (electrons and positrons) passing by the atomic nuclei. This is the mechanism by which photon beams are produced in a linear accelerator. The bremsstrahlung energy spectrum is continuous with the maximum energy equal to the kinetic energy of the incoming electron. The angular distribution is largely forward. A delta-ray is the secondary electron ejected from the atom resulting from a large energy transfer from the incoming electron or positron. If the incident particle is an electron, then the energy of the delta-ray cannot exceed one-half of the incident electron energy, for by definition, the outgoing electron with the lesser energy is the delta-ray. If the incident particle is a positron, then it can give up all its energy to the delta-ray. Positron annihilation is the process that occurs when a positron and an electron collide. If they are approximately at rest relative to each other, they destroy each other upon contact, and produce two photons of 511 keV each that are emitted in opposite directions. If they are moving at different relative speeds, the energies of the photons emitted will be higher.

APPLICATIONS IN RADIATION THERAPY DOSE CALCULATION

Beam Characteristics

In order to perform dose calculation using the Monte Carlo method, it is necessary to have accurate information about the radiation field incident upon the patient, that is, the phase-space data. This data is difficult to obtain by empirical means, therefore in practice it is obtained by Monte Carlo simulation of the machine head. Figure 1 shows a typical configuration of the machine head for a medical linear accelerator that produces clinical photon beams. The components directly in the beam are the target, the flattening filter, and the monitor chamber. The components that collimate the beam are the primary collimator, and the upper and lower collimating jaws. The phase space data can be collected on two scoring planes above and below the collimating jaws, respectively. For clinical electron beams, the machine head is similar to that of photon beams except that the target is removed, the flattening filter is replaced by a scattering foil system, and an additional applicator is used for further collimation of the electron beam (Fig. 2).

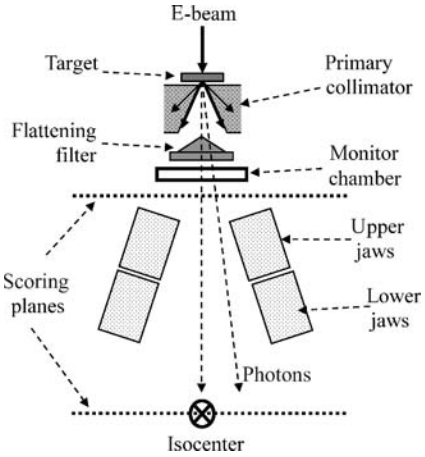


Figure 1. Production of a clinical photon beam (drawing not to scale).

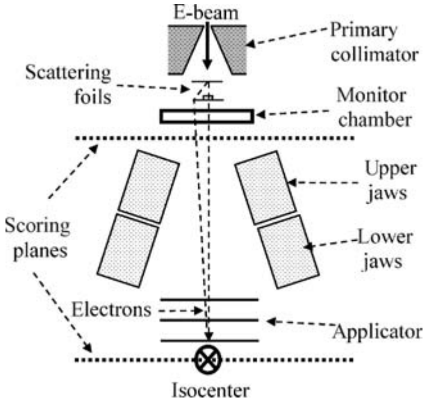


Figure 2. Production of a clinical electron beam (drawing not to scale).

Figure 3 shows the energy spectra of a 15 MV photon beam. Note that the low energy photons have been filtered out by the flattening filter. Moreover, the spectrum near the center of the beam, say, within 3 cm of the central axis, is harder than that away from the axis, say, 10–15 cm from the center. The reason is that the flattening filter is thicker in the middle, thus absorbing more low energy photons. Figure 4 shows the energy spectrum of a clinical 9 MeV electron beam. It is clear there are two peaks corresponding to the thin part and the thick part of the scattering foil system. Figure 5 shows the angular distribution of the electrons at the isocenter plane (~100 cm from the entrance to the primary collimator). Due to the significant scattering of electrons in the scattering foils as well as in the air space above the isocenter, the angular spread is diffused and approximates a normal distribution.

This information about beam characteristics such as the energy and angular distributions is difficult to measure, but can be calculated by Monte Carlo with relative ease.

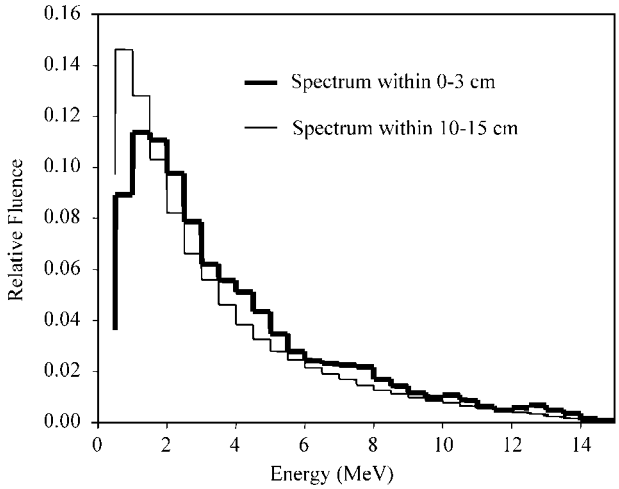


Figure 3. Energy spectra of a 15 MV photon beam.

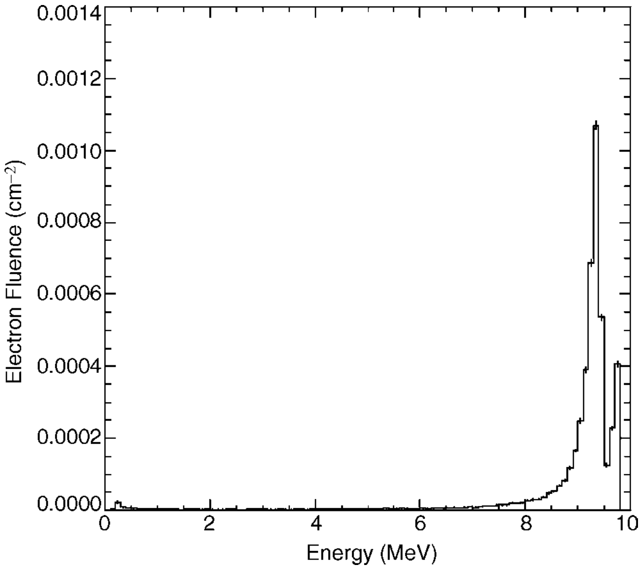


Figure 4. The energy spectrum of a clinical 9 MeV electron beam.

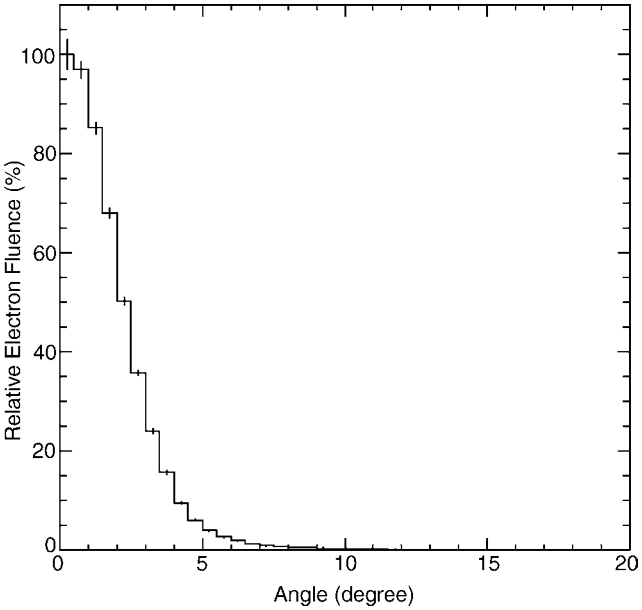


Figure 5. The angular distribution of a clinical 9 MeV electron beam.

Treatment Planning Dose Calculation

Many modern dose calculation algorithms other than Monte Carlo are quite accurate for megavoltage (^{60}Co–20 MV) external photon beam radiation therapy for sites that are composed of soft tissue (density ~ 1 g·mL^{-1}) and bone (brain, pelvis, limbs) (144). However, these algorithms are less accurate when electronic equilibrium is lost due to more severe tissue inhomogeneities. This may be a clinical concern in the lung, where soft tissue tumors are surrounded by low density ($\sim$ 0.2–0.3 g·mL^{-1}) lung and in the head and neck (H&N) due to the presence of air cavities. Although Monte Carlo is currently impractical for routine clinical use, Monte Carlo calculations based on patient CT scans and inhomogeneous phantoms provide clinically valuable information, especially when combined with high resolution phantom measurements (film and/or TLD). For a summary of the status of the field (145).

Lung Cancer

Since lung has lower electron density than soft tissue, there is reduced attenuation of the primary photons of a beam traversing lung compared with the same path length in soft tissue. Most inhomogeneity correction algorithms can account for this effect (144). However, other, more subtle effects are described with reasonable approximation only by superposition-convolution algorithms (146,147) and most accurately by Monte Carlo. The cause of these effects is the long range of the secondary electrons in lung compared to soft tissue (the range is approximately inversely proportional to the ratio of lung to soft tissue density). Energy is thus transported outside the beam's geometric edge, resulting in a broader beam penumbra and reduced dose within the beam. Also, especially for a small soft tissue target embedded in a very low density medium and irradiated with a tight, high energy beam, there is a build-down (low dose region) at the entrance surface and sides of the target. All these effects are more pronounced for higher energy beams and lower density lungs (longer electron ranges) and smaller fields. The clinical concern is that treatment plans developed with algorithms that do not account for these effects can result in target underdose and/or overdose to normal tissues in penumbral regions. Figure 6 shows characteristic differences between the dose distribution of a single 6 MV photon beam predicted by a measurement-based pencil beam calculation that accounts only for changes in primary attenuation and that predicted by a Monte Carlo calculation. Lung radiation treatments usually consist of two or more beams, incident on the tumor in a cross-fire technique. Figure 7 shows that even in this patient's full four-field plan, these characteristic differences between the Monte Carlo and pencil beam calculations persist.

The degree to which the target underdose and broader penumbra in lung may compromise complications-free tumor control has been addressed in several studies (24,33,34,148–155). References 148–151 used measurements only to investigate penumbra broadening and build-down effects. A recent study (154) combined film dosimetry and EGSnrc and DOSXYZnrc Monte Carlo calculations to study the dose distribution in a 2 × 2 × 2 cm acrylic ($\sim$ tissue density) cube embedded in cork, simulating a small lesion in lung irradiated with a single and with parallel opposed photon beams from 4 to 18 MV. The parallel opposed geometry is a common field arrangement for treatment of lung tumors. Cork density, field size, and depth of the lesion in cork were varied. For the entire target cube to receive at least 95% of the dose to its center required field edges of the parallel opposed fields to be at least 2 cm from the cube even for the most favorable case (4 MV photons).

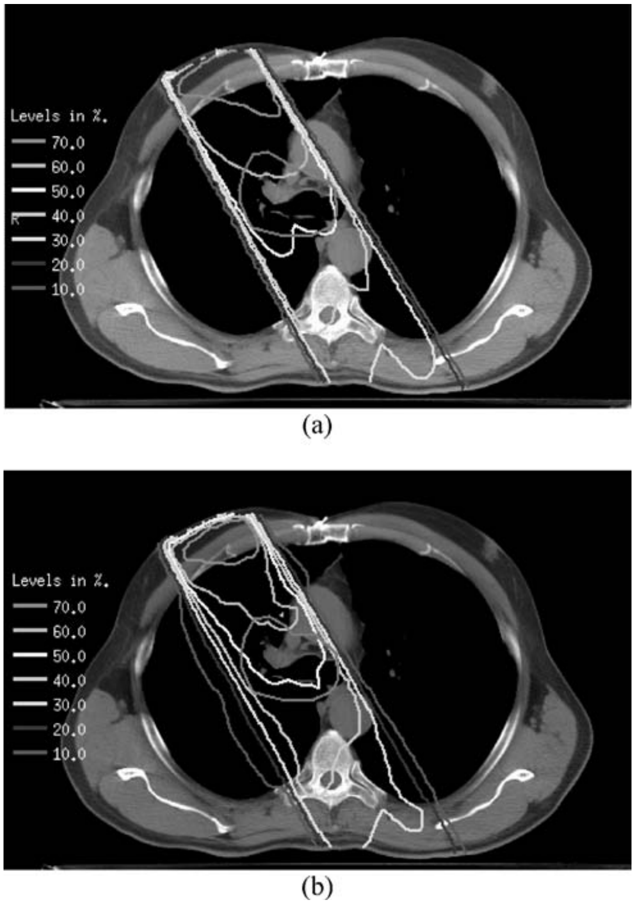
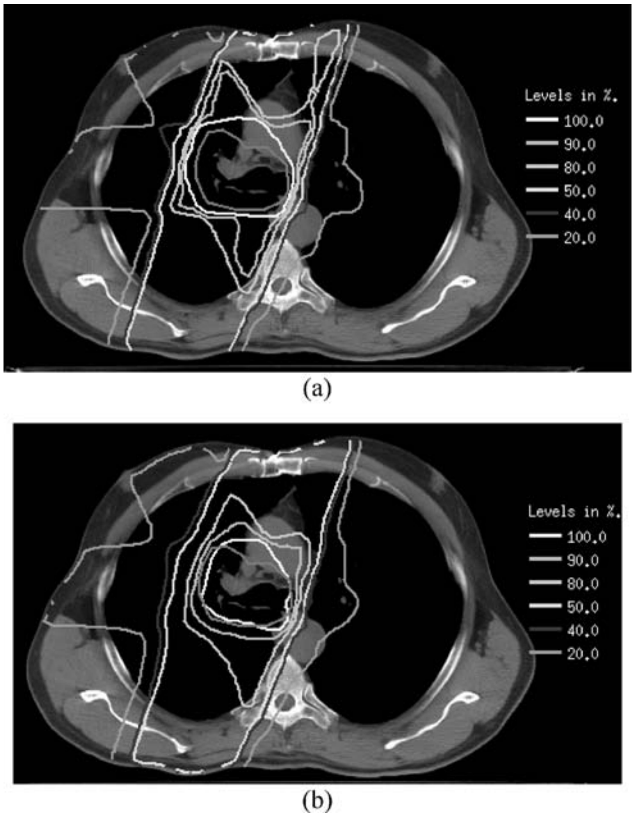


Figure 6. Dose distribution of a single 6 MV photon beam (a) predicted by a measurement-based pencil beam calculation that accounts only for changes in primary attenuation and (b) that predicted by a Monte Carlo calculation. The red contour indicates the target. Please use online version for color figure.

Other recent studies from different institutions have compared more complex treatment plans designed on anatomical phantoms or patient CT image sets and calculated with Monte Carlo versus the local treatment planning system calculation algorithm (24,34,152,153,155). In these studies, as in routine clinical practice, the beam is shaped to cover the planning target volume (PTV), which is larger than the grossly visible tumor gross target volume (GTV). The margin is intended to account for microscopic disease, setup error and breathing motion. Based on these studies, it is expected that (a) Results depend on the treatment planning system algorithm (34,153,155); (b) For the same planning system, results are patient (phantom geometry)



(a)

(b)

Figure 7. Dose distributions of a four-field plan in the lung. Figure (a) and (b) show dose distributions on a transverse plane predicted by a measurement-based pencil beam calculation and by a Monte Carlo calculation, respectively. The red contour indicates the target. Please see online version for color figure.

dependent as well as dependent on beam energy (24,33,34,153,155); (c) Changes quoted depend on the dosimetric coverage factor being evaluated. Mean target dose and dose encompassing 95% of the target volume are relatively insensitive indices; minimum dose (a single point) and dose-volume points on rapidly changing portions of the dose-volume histogram are more sensitive; (d) The PTV is usually underdosed relative to expectations from the treatment planning system. The degree of underdose varies from 1 to 20%, for 6 MV photons, depending on the dosimetric coverage factor, the lung density and the tumor location. For most of the cases reported, the underdosage is $< 10\%$; (e) The planning system results for coverage of the GTV are more similar to Monte Carlo, as is expected because the margin results in a larger distance from geometric beam edge to the GTV border than to the PTV border; (f) The greatest differences are for tumors surrounded by very low density lungs (33,34); (g) There are greater differences for high (e.g., 15 MV) than low (6 MV) energy beams (153); (h) Normal organ doses (primarily lung and spinal cord) are only slightly affected.

Head and Neck Cancer

H&N cancer radiation therapy usually includes photon irradiation with low megavoltage beams (^{60}Co—6 MV). Target tissue often borders on naturally occurring or surgical air cavities. Experiments demonstrate that build-down accompanying the loss of electronic equilibrium in air cavities in tissue-equivalent phantoms can cause up to a 25% underdose within the first millimeter of tissue (156–159), with particularly pronounced effects for small ($\leq 5 \times 5$ cm^2) fields, such as are used for treatment of larynx cancer. Whether this impacts on local control of larynx cancers treated with 6 MV beams versus ^{60}Co has not been resolved (160). The penumbra broadening and loss of dose within the beam that are noted in lung also occur in air cavities but the small size of these cavities, compared to the size of a lung, prevent these effects from posing a serious clinical problem.

Monte Carlo calculations compared well with parallel plate ion chamber measurements for single field and parallel opposed field irradiation (4, 6, and 8 MV photons) of a $4 \times 4 \times 4$ cm^3 cavity centered in a $30 \times 30 \times 16$ cm^3 phantom (161) though neither method had the spatial resolution to probe the build-down region in detail. A few studies have compared dose distributions on patient CT image sets for clinical beam arrangements as calculated with the local planning system and with Monte Carlo calculations for the same beams (31,33,162). Differences between the two calculation methods are more noticeable for individual beams than when all the beams (from two to seven, depending on the plan) are combined for the overall treatment plan. Monte Carlo predicts inferior target coverage compared to the planning system, but the differences, which depend on dosimetric index and tumor geometry, are less than in lung. Spinal cord maximum dose differences of < 1 Gy were reported in (31) (with the Monte Carlo calculation sometimes higher, sometimes lower) and 3 Gy higher as calculated by Monte Carlo in (162).

DISCUSSION

For treatment planning dose calculations, Monte Carlo is potentially the most accurate method. Monte Carlo dose calculation for electron beams has recently become available on a commercial treatment planning system (48). For photon beams, however, it has not been practical for routine clinical use due to its long running time. To improve the computation efficiency, there are variance reduction techniques available. The most common techniques are splitting and Russian roulette (136). In splitting, a particle is artificially split into multiple particles in important regions to produce more histories. In Russian roulette, particles are artificially terminated in relatively unimportant regions to reduce the number of histories. In both techniques, the particle weight, of course, needs to be adjusted to reflect the artificial increase or decrease of histories.

In addition to dose calculation, perhaps a more important application of Monte Carlo is to provide information that cannot be easily obtained by measurement. For example, in the simulation of the machine head, the phase-space data provide information on the primary and scattered radiation from various components in the machine head. These data provide important information in understanding the beam characteristics and may be used for other dose calculation methods.

BIBLIOGRAPHY

1. Andreo P. Monte Carlo techniques in medical radiation physics. Phys Med Biol 1991;36(7):861–920.
2. Mohan R, Chui C, Lidofsky L. Energy and angular distributions of photons from medical linear accelerators. Med Phys 1985;12(5):592–597.
3. Han K, et al. Monte Carlo simulation of a cobalt-60 beam. Med Phys 1987;14(3):414–419.
4. Chaney EL, Cullip TJ, Gabriel TA. A Monte Carlo study of accelerator head scatter. Med Phys 1994;21(9):1383–1390.
5. Lovelock DM, Chui CS, Mohan R. A Monte Carlo model of photon beams used in radiation therapy. Med Phys 1995 ;22(9):1387–1394.
6. Lee PC. Monte Carlo simulations of the differential beam hardening effect of a flattening filter on a therapeutic X-ray beam. Med Phys 1997;24(9):1485–1489.
7. Bhat M, et al. Off-axis X-ray spectra: a comparison of Monte Carlo simulated and computed X-ray spectra with measured spectra. Med Phys 1999;26(2):303–309.
8. Libby B, Siebers J, Mohan R. Validation of Monte Carlo generated phase-space descriptions of medical linear accelerators. Med Phys 1999;26(8):1476–1483.
9. Ma CM, Jiang SB. Monte Carlo modelling of electron beams from medical accelerators. Phys Med Biol 1999;44(12):R157–R189.
10. Siebers JV, et al. Comparison of EGS4 and MCNP4b Monte Carlo codes for generation of photon phase space distributions for a Varian 2100C. Phys Med Biol 1999; 44(12):3009–3026.
11. van der Zee W, Welleweerd J. Calculating photon beam characteristics with Monte Carlo techniques. Med Phys 1999;26(9): 1883–1892.
12. Deng J, et al. Photon beam characterization and modelling for Monte Carlo treatment planning. Phys Med Biol 2000;45(2): 411–427.
13. Bieda MR, Antolak JA, Hogstrom KR. The effect of scattering foil parameters on electron-beam Monte Carlo calculations. Med Phys 2001;28(12):2527–2534.
14. Antolak JA, Bieda MR, Hogstrom KR. Using Monte Carlo methods to commission electron beams: a feasibility study. Med Phys 2002;29(5):771–786.
15. Ding GX. Energy spectra, angular spread, fluence profiles and dose distributions of 6 and 18 MV photon beams: results of monte carlo simulations for a varian 2100EX accelerator. Phys Med Biol 2002;47(7):1025–1046.
16. Sheikh-Bagheri D, Rogers DW. Monte Carlo calculation of nine megavoltage photon beam spectra using the BEAM code. Med Phys 2002;29(3):391–402.
17. van der Zee W, Welleweerd J. A Monte Carlo study on internal wedges using BEAM. Med Phys 2002;29(5):876–885.
18. Ding GX. Using Monte Carlo simulations to commission photon beam output factors–a feasibility study. Phys Med Biol 2003;48(23):3865–3874.
19. Van de Walle J, et al. Monte Carlo model of the Elekta SLiplus accelerator: validation of a new MLC component module in BEAM for a 6 MV beam. Phys Med Biol 2003; 48(3):371–385.
20. Verhaegen F, Seuntjens J. Monte Carlo modelling of external radiotherapy photon beams. Phys Med Biol 2003;48(21):R107–R164.
21. Fix MK, et al. Monte Carlo source model for photon beam radiotherapy: photon source characteristics. Med Phys 2004; 31(11):3106–3121.
22. Pena J, et al. Commissioning of a medical accelerator photon beam Monte Carlo simulation using wide-field profiles. Phys Med Biol 2004;49(21):4929–4942.
23. DeMarco JJ, Solberg TD, Smathers JB. A CT-based Monte Carlo simulation tool for dosimetry planning and analysis. Med Phys 1998;25(1):1–11.
24. Wang L, Chui CS, Lovelock M. A patient-specific Monte Carlo dose-calculation method for photon beams. Med Phys 1998;25(6):867–878.
25. Jeraj R, Keall P. Monte Carlo-based inverse treatment planning. Phys Med Biol 1999;44(8):1885–1896.
26. Laub W, et al. Monte Carlo dose computation for IMRT optimization. Phys Med Biol 2000;45(7):1741–1754.
27. Lewis RD, et al. Use of Monte Carlo computation in benchmarking radiotherapy treatment planning system algorithms. Phys Med Biol 2000;45(7):1755–1764.
28. Keall PJ, et al. Monte Carlo dose calculations for dynamic IMRT treatments. Phys Med Biol 2001;46(4):929–941.
29. Li XA, et al. Monte Carlo dose verification for intensity-modulated arc therapy. Phys Med Biol 2001;46(9):2269–2282.
30. Shih R, Lj XA, Hsu WL. Dosimetric characteristics of dynamic wedged fields: a Monte Carlo study. Phys Med Biol 2001;46(12):N281–N292.
31. Wang L, Yorke E, Chui CS. Monte Carlo evaluation of tissue inhomogeneity effects in the treatment of the head and neck. Int J Radiat Oncol Biol Phys 2001;50(5):1339–1349.
32. Ma CM, et al. A Monte Carlo dose calculation tool for radiotherapy treatment planning. Phys Med Biol 2002;47(10): 1671–1689.
33. Wang L, Yorke E, Chui CS. Monte Carlo evaluation of 6 MV intensity modulated radiotherapy plans for head and neck and lung treatments. Med Phys 2002;29(11):2705–2717.
34. Yorke ED, et al. Evaluation of deep inspiration breath-hold lung treatment plans with Monte Carlo dose calculation. Int J Radiat Oncol Biol Phys 2002;53(4):1058–1070.
35. Leal A, et al. Routine IMRT verification by means of an automated Monte Carlo simulation system. Int J Radiat Oncol Biol Phys 2003;56(1):58–68.
36. Wieslander E, Knoos T. Dose perturbation in the presence of metallic implants: treatment planning system versus Monte Carlo simulations. Phys Med Biol 2003;48(20):3295–3305.
37. Heath E, Seuntjens J, Sheikh-Bagheri D. Dosimetric evaluation of the clinical implementation of the first commercial IMRT Monte Carlo treatment planning system at 6 MV. Med Phys 2004;31(10):2771–2779.
38. Yang J, et al. Modelling of electron contamination in clinical photon beams for Monte Carlo dose calculation. Phys Med Biol 2004;49(12):2657–2673.
39. Kawrakow I, Fippel M, Friedrich K. 3D electron dose calculation using a Voxel based Monte Carlo algorithm (VMC). Med Phys 1996;23(4):445–457.
40. Keall PJ, Hoban PW. Super-Monte Carlo: a 3-D electron beam dose calculation algorithm. Med Phys 1996;23(12): 2023–2034.
41. Scora D, Faddegon BA. Monte Carlo based phase-space evolution for electron dose calculation. Med Phys 1997;24(2): 177–187.
42. Jiang SB, Kapur A, Ma CM. Electron beam modeling and commissioning for Monte Carlo treatment planning. Med Phys 2000;27(1):180–191.
43. Kawrakow I. Accurate condensed history Monte Carlo simulation of electron transport. I. EGSnrc, the new EGS4 version. Med Phys 2000;27(3):485–498.
44. Lee MC, et al. Monte Carlo based treatment planning for modulated electron beam radiation therapy. Phys Med Biol 2001;46(8):2177–2199.
45. Bjork P, Knoos T, Nilsson P. Influence of initial electron beam characteristics on monte carlo calculated absorbed dose distributions for linear accelerator electron beams. Phys Med Biol 2002;47(22):4019–4041.

46. Deng J, Lee MC, Ma CM. A Monte Carlo investigation of fluence profiles collimated by an electron specific MLC during beam delivery for modulated electron radiation therapy. Med Phys 2002;29(11):2472–2483.
47. Doucet R, et al. Comparison of measured and Monte Carlo calculated dose distributions in inhomogeneous phantoms in clinical electron beams. Phys Med Biol 2003;48(15):2339–2354.
48. Cygler JE, et al. Evaluation of the first commercial Monte Carlo dose calculation engine for electron beam treatment planning. Med Phys 2004;31(1):142–153.
49. Coleman J, et al. A comparison of Monte Carlo and Fermi-Eyges-Hogstrom estimates of heart and lung dose from breast electron boost treatment. Int J Radiat Oncol Biol Phys 2005;61(2):621–628.
50. Carlsson AK, Andreo P, Brahme A. Monte Carlo and analytical calculation of proton pencil beams for computerized treatment plan optimization. Phys Med Biol 1997;42(6): 1033–1053.
51. Paganetti H. Monte Carlo method to study the proton fluence for treatment planning. Med Phys 1998;25(12):2370–2375.
52. Fippel M, Soukup M. A Monte Carlo dose calculation algorithm for proton therapy. Med Phys 2004;31(8):2263–2273.
53. Jiang H, Paganetti H. Adaptation of GEANT4 to Monte Carlo dose calculations based on CT data. Med Phys 2004;31(10): 2811–2818.
54. Paganetti H. Four-dimensional Monte Carlo simulation of time-dependent geometries. Phys Med Biol 2004;49(6):N75–N81.
55. Williamson JF, Morin RL, Khan FM. Monte Carlo evaluation of the Sievert integral for brachytherapy dosimetry. Phys Med Biol 1983;28(9):1021–1032.
56. Burns GS, Raeside DE. Monte Carlo simulation of the dose distribution around 125I seeds. Med Phys 1987;14(3):420–424.
57. Williamson JF. Monte Carlo evaluation of specific dose constants in water for 125I seeds. Med PhysM 1988;15(5):686–694.
58. Angelopoulos A, et al. Accurate Monte Carlo calculations of the combined attenuation and build-up factors, for energies (20–1500 keV) and distances (0–10 cm) relevant in brachytherapy. Phys Med Biol 1991;36(6):763–778.
59. Williamson JF, Li Z. Monte Carlo aided dosimetry of the microselectron pulsed and high dose-rate 192Ir sources. Med Phys 1995;22(6):809–819.
60. Weaver K, et al. A source model for efficient brachytherapy computations with Monte Carlo. Med Phys 1996;23(12): 2079–2084.
61. Cheung YC, et al. The dose distribution close to an 192Ir wire source: EGS4 Monte Carlo calculations. Phys Med Biol 1997;42(2):401–406.
62. Baltas D, et al. Application of the Monte Carlo integration (MCI) method for calculation of the anisotropy of 192Ir brachytherapy sources. Phys Med Biol 1998;43(6):1783–1801.
63. Daskalov GM, Loffler E, Williamson JF. Monte Carlo-aided dosimetry of a new high dose-rate brachytherapy source. Med Phys 1998;25(11):2200–2208.
64. Mainegra E, Capote R, Lopez E. Dose rate constants for 125I, 103Pd, 192Ir and 169Yb brachytherapy sources: an EGS4 Monte Carlo study. Phys Med Biol 1998;43(6):1557–1566.
65. Wang R, Sloboda RS. Monte Carlo dosimetry of the VariSource high dose rate 192Ir source. Med Phys 1998;25(4): 415–423.
66. Karaiskos P, et al. A Monte Carlo investigation of the dosimetric characteristics of the VariSource 192Ir high dose rate brachytherapy source. Med Phys 1999;26(8):1498–1502.
67. Reynaert N, et al. Monte Carlo calculations of dose distributions around 32P and 198Au stents for intravascular brachytherapy. Med Phys 1999;26(8):1484–1491.
68. Casal E, et al. Monte Carlo calculations of dose rate distributions around the Amersham CDCS-M-type 137Cs source. Med Phys 2000;27(1):132–140.
69. Hedtjarn H, Carlsson GA, Williamson JF. Monte Carlo-aided dosimetry of the Symmetra model I25.S06 125I, interstitial brachytherapy seed. Med Phys 2000;27(5):1076–1085.
70. Li Z, Palta JR, Fan JJ. Monte Carlo calculations and experimental measurements of dosimetry parameters of a new 103Pd source. Med Phys 2000;27(5):1108–1112.
71. Mainegra E, Capote R, Lopez E. Radial dose functions for 103Pd, 125I, 169Yb and 192Ir brachytherapy sources: an EGS4 Monte Carlo study. Phys Med Biol 2000;45(3):703–717.
72. Mainegra E, Capote R, Lopez E. Anisotropy functions for 169Yb brachytherapy seed models 5, 8 and X1267. An EGS4 Monte Carlo study. Phys Med Biol 2000;45(12):3693–3705.
73. Williamson JF. Monte Carlo modeling of the transverse-axis dose distribution of the model 200 103Pd interstitial brachytherapy source. Med Phys 2000;27(4):643–654.
74. Ballester F, et al. Technical note: Monte-Carlo dosimetry of the HDR 12i and Plus 192Ir sources. Med Phys 2001;28(12): 2586–2591.
75. Capote R, Mainegra E, Lopez E. Anisotropy function for 192Ir low-dose-rate brachytherapy sources: an EGS4 Monte Carlo study. Phys Med Biol 2001;46(5):1487–1499.
76. Rivard MJ. Monte Carlo calculations of AAPM Task Group Report No. 43 dosimetry parameters for the MED3631-A/M125I source. Med Phys 2001;28(4):629–637.
77. Chan GH, Prestwich WV. Monte carlo investigation of the dosimetric properties of the new 103Pd BrachySeedPd-103 Model Pd-1 source. Med Phys 2002;29(9):1984–1990.
78. Hedtjarn H, Carlsson GA, Williamson JF. Accelerated Monte Carlo based dose calculations for brachytherapy planning using correlated sampling. Phys Med Biol 2002;47(3):351–376.
79. Ibbott GS, Meigooni AS, Gearheart DM. Monte Carlo determination of dose rate constant. Med Phys 2002;29(7):1637–1638.
80. Bohm TD, DeLuca Jr PM, DeWerd LA. Brachytherapy dosimetry of 125I and 103Pd sources using an updated cross section library for the MCNP Monte Carlo transport code. Med Phys 2003;30(4):701–711.
81. Medich DC, Munro JJ. 3rd, Monte Carlo calculated TG-43 dosimetry parameters for the SeedLink 125Iodine brachytherapy system. Med Phys 2003;30(9):2503–2508.
82. Anagnostopoulos G, et al. The effect of patient inhomogeneities in oesophageal 192Ir HDR brachytherapy: a Monte Carlo and analytical dosimetry study. Phys Med Biol 2004; 49(12):2675–2685.
83. Ballester F, et al. Monte carlo dosimetric study of best industries and Alpha Omega Ir-192 brachytherapy seeds. Med Phys 2004;31(12):3298–3305.
84. Lymperopoulou G, et al. A monte carlo dosimetry study of vaginal 192Ir brachytherapy applications with a shielded cylindrical applicator set. Med Phys 2004;31(11):3080–3086.
85. Reniers B, Verhaegen F, Vynckier S. The radial dose function of low-energy brachytherapy seeds in different solid phantoms: comparison between calculations with the EGSnrc and MCNP4C Monte Carlo codes and measurements. Phys Med Biol 2004;49(8):1569–1582.
86. Perez-Calatayud J, et al. Monte carlo and experimental derivation of TG43 dosimetric parameters for CSM-type Cs-137 sources. Med Phys 2005;32(1):28–36.
87. Furhang EE, Chui CS, Sgouros G. A Monte Carlo approach to patient-specific dosimetry. Med Phys 1996;23(9):1523–1529.

88. Tagesson M, Ljungberg M, Strand SE. A Monte-Carlo program converting activity distributions to absorbed dose distributions in a radionuclide treatment planning system. Acta Oncol 1996;35(3):367–372.
89. Liu A, et al. Monte Carlo-assisted voxel source kernel method (MAVSK) for internal beta dosimetry. Nucl Med Biol 1998; 25(4):423–433.
90. Clairand I, et al. DOSE3D: EGS4 Monte Carlo code-based software for internal radionuclide dosimetry. J Nucl Med 1999;40(9):1517–1523.
91. Zaidi H. Relevance of accurate Monte Carlo modeling in nuclear medical imaging. Med Phys 1999;26(4):574–608.
92. Chao TC, Xu XG. Specific absorbed fractions from the image-based VIP-Man body model and EGS4-VLSI Monte Carlo code: internal electron emitters. Phys Med Biol 2001;46(4): 901–927.
93. Kvinnsland Y, Skretting A, Bruland OS. Radionuclide therapy with bone-seeking compounds: Monte Carlo calculations of dose-volume histograms for bone marrow in trabecular bone. Phys Med Biol 2001;46(4):1149–1161.
94. Yoriyaz H, Stabin MG, dos Santos A. Monte Carlo MCNP-4B-based absorbed dose distribution estimates for patient-specific dosimetry. J Nucl Med 2001;42(4):662–669.
95. Ljungberg M, et al. A 3-dimensional absorbed dose calculation method based on quantitative SPECT for radionuclide therapy: evaluation for (131)I using monte carlo simulation. J Nucl Med 2002;43(8):1101–1109.
96. Kinase S, et al. Evaluation of specific absorbed fractions in voxel phantoms using Monte Carlo simulation. Radiat Prot Dosimet 2003;105(1–4):557–563.
97. Wolf I, et al. Determination of Individual S-Values for (131)I Using Segmented CT Data and the EGS4 Monte Carlo Code. Cancer Biother Radiopharm 2005;20(1):98–102.
98. Chan HP, Doi K. Radiation dose in diagnostic radiology: Monte Carlo simulation studies. Med Phys 1984;11(4):480–490.
99. Dance DR, Day GJ. The computation of scatter in mammography by Monte Carlo methods. Phys Med Biol 1984;29(3): 237–247.
100. Kulkarni RN, Supe SJ. Radiation dose to the breast during mammography: a comprehensive, realistic Monte Carlo calculation. Phys Med Biol 1984;29(10):1257–1264.
101. Kulkarni RN, Supe SJ. Monte Carlo calculations of mammographic X-ray spectra. Phys Med Biol 1984;29(2):185–190.
102. Boone JM, Seibert JA. Monte Carlo simulation of the scattered radiation distribution in diagnostic radiology. Med Phys 1988;15(5):713–720.
103. Papin PJ, Rielly PS. Monte Carlo simulation of diagnostic X-ray scatter. Med Phys 1988;15(6):909–914.
104. Gao W, Raeside DE. Orthovoltage radiation therapy treatment planning using Monte Carlo simulation: treatment of neuroendocrine carcinoma of the maxillary sinus. Phys Med Biol 1997;42(12):2421–2433.
105. Verhaegen F, et al. Monte Carlo modelling of radiotherapy kV X-ray units. Phys Med Biol 1999;44(7):1767–1789.
106. Boone JM, Cooper 3rd VN. Scatter/primary in mammography: Monte Carlo validation. Med Phys 2000;27(8):1818–1831.
107. Boone JM, Buonocore MH, Cooper 3rd VN. Monte Carlo validation in diagnostic radiological imaging. Med Phys 2000;27(6):1294–1304.
108. Ng KP, Kwok CS, Tang FH. Monte Carlo simulation of X-ray spectra in mammography. Phys Med Biol 2000;45(5):1309–1318.
109. Peplow DE, Verghese K. Digital mammography image simulation using Monte Carlo. Med Phys 2000;27(3):568–579.
110. Kramer R, et al. Backscatter factors for mammography calculated with Monte Carlo methods. Phys Med Biol 2001; 46(3):771–781.
111. Ay MR, et al. Monte carlo simulation of X-ray spectra in diagnostic radiology and mammography using MCNP4C. Phys Med Biol 2004;49(21):4897–4917.
112. Andreo P, Nahum A, Brahme A. Chamber-dependent wall correction factors in dosimetry. Phys Med Biol 1986;31(11): 1189–1199.
113. Rogers DW. Calibration of parallel-plate chambers: resolution of several problems by using Monte Carlo calculations. Med Phys 1992;19(4):889–899.
114. Ma CM, Nahum AE. Calculations of ion chamber displacement effect corrections for medium-energy X-ray dosimetry. Phys Med Biol 1995;40(1):45–62.
115. Ma CM, Nahum AE. Monte Carlo calculated stem effect correction for NE2561 and NE2571 chambers in medium-energy X-ray beams. Phys Med Biol 1995;40(1):63–72.
116. Mobit PN, Nahum AE, Mayles P. An EGS4 Monte Carlo examination of general cavity theory. Phys Med Biol 1997; 42(7):1319–1334.
117. Ferreira IH, et al. Perturbation corrections for flat and thimble-type cylindrical standard ionization chambers for 60Co gamma rays: Monte Carlo calculations. Phys Med Biol 1998;43(10):2721–2727.
118. Ferreira IH, et al. Monte Carlo calculations of the ionization chamber wall correction factors for 192Ir and 60Co gamma rays and 250 kV X-rays for use in calibration of 192Ir HDR brachytherapy sources. Phys Med Biol 1999;44(8):1897–1904.
119. Borg J, et al. Monte Carlo study of correction factors for Spencer-Attix cavity theory at photon energies at or above 100 keV. Med Phys 2000;27(8):1804–1813.
120. Seuntjens JP, et al. Absorbed-dose beam quality conversion factors for cylindrical chambers in high energy photon beams. Med Phys 2000;27(12):2763–2779.
121. Mazurier J, et al. Calculation of perturbation correction factors for some reference dosimeters in high-energy photon beams with the Monte Carlo code PENELOPE. Phys Med Biol 2001;46(6):1707–1717.
122. Fu Y, Luo Z. Application of Monte Carlo simulation to cavity theory based on the virtual electron source concept. Phys Med Biol 2002;47(17):3263–3274.
123. Mainegra-Hing E, Kawrakow I, Rogers DW. Calculations for plane-parallel ion chambers in 60Co beams using the EGSnrc Monte Carlo code. Med Phys 2003;30(2):179–189.
124. Piermattei A, et al. The wall correction factor for a spherical ionization chamber used in brachytherapy source calibration. Phys Med Biol 2003;48(24):4091–4103.
125. Rogers DW, Kawrakow I. Monte Carlo calculated correction factors for primary standards of air kerma. Med Phys 2003;30(4):521–532.
126. Siegbahn EA, et al. Calculations of electron fluence correction factors using the Monte Carlo code PENELOPE. Phys Med Biol 2003;48(10):1263–1275.
127. Capote R, et al. An EGSnrc Monte Carlo study of the micro-ionization chamber for reference dosimetry of narrow irregular IMRT beamlets. Med Phys 2004;31(9):2416–2422.
128. McCaffrey JP, et al. Evidence for using Monte Carlo calculated wall attenuation and scatter correction factors for three styles of graphite-walled ion chamber. Phys Med Biol 2004;49(12):2491–2501.
129. Sempau J, et al. Electron beam quality correction factors for plane-parallel ionization chambers: Monte Carlo calculations using the PENELOPE system. Phys Med Biol 2004;49(18): 4427–4444.
130. Nelson WR, Hirayama H, Rogers DWO. The EGS4 Code System. 1985; Stanford Linear Accelerator Center.
131. GEANT team, GEANT version 315. 1992, CERN-data handling division, report DD/EE/84-1 revision.

132. Baro J, et al. PENELOPE: an algorithm for Monte Carlo simulation of the penetration and energy loss of electrons and positrons in matter. Nucl Instrum Methods B 1995;100:31–46.
133. Ma C-M, et al. DOSXYZ users manual. Ottawa: NRCC. 1995.
134. Neuenschwander H, Mackie TR, Reckwerdt PJ. MC–a high-performance Monte Carlo code for electron beam treatment planning. Phys Med Biol 1995;40(4):543–574.
135. Rogers DW, et al. BEAM: a Monte Carlo code to simulate radiotherapy treatment units. Med Phys 1995;22(5):503–524.
136. Briesmeister JF. MCNP-A general Monte Carlo N-Particle transport code, version 4B. 1997; Los Alamos National Laboratory report LA-12625-M.
137. Sempau J, Wilderman SJ, Bielajew AF. DPM, a fast, accurate Monte Carlo code optimized for photon and electron radiotherapy treatment planning dose calculations. Phys Med Biol 2000;45(8):2263–2291.
138. VMC++, electron and photon Monte Carlo calculations optimized for Radiation Treatment Planning, in Advanced Monte Carlo for Radiation Physics, Particle Transport Simulation and Applications: Proceedings of the Monte Carlo 2000. In: Meeting Kling A, et al. editors. Berlin; Lisbon: Springer, 2001; p 229–236.
139. Hartmann Siantar CL, et al. Description and dosimetric verification of the PEREGRINE Monte Carlo dose calculation system for photon beams incident on a water phantom. Med Phys 2001;28(7):1322–1337.
140. Salvat F, Fernandez-Varea JM, DSempau J. PENELOPE-A code system for Monte Carlo simulation of Electron and Photon Transport. 2003; Issy-les-Moulineaux, France: OECD Nuclear Energy Agency.
141. van der Zee W, Hogenbirk A, van der Marck SC. ORANGE: a Monte Carlo dose engine for radiotherapy. Phys Med Biol 2005;50(4):625–641.
142. Guber W, et al. A geometric description technique suitable for computer analysis of both the nuclear and conventional vulnerability of armored military vehicles. Washington (DC): 1967.
143. Berger M. Monte Carlo calculation of the penetration and diffusion of fast charged particles, in Methods in Computational Physics. In: Alder B, Fernbach S, Rotenberg M, editors. New York: Academic; 1963. p 135–215.
144. Papanikolaou N, et al. Tissue inhomogeneity corrections for megavoltage photon beams. Medical Physics Publishing; 2004.
145. Fraass BA, Smathers J, Deye J. Summary and recommendations of a National Cancer Institute workshop on issues limiting the clinical use of Monte Carlo dose calculation algorithms for megavoltage external beam radiation therapy. Med Phys 2003;30(12):3206–3216.
146. Ahnesjo A. Collapsed cone convolution of radiant energy for photon dose calculation in heterogeneous media. Med Phys 1989;16(4):577–592.
147. Mackie TR, Scrimger JW, Battista JJ. A convolution method of calculating dose for 15-MV X-ray. Med Phys 1985;12(2): 188–196.
148. Ekstrand KE, Barnes WH. Pitfalls in the use of high energy X-ray to treat tumors in the lung. Int J Radiat Oncol Biol Phys 1990;18(1):249–252.
149. White PJ, Zwicker RD, Huang DT. Comparison of dose homogeneity effects due to electron equilibrium loss in lung for 6 MV and 18 MV photons. Int J Radiat Oncol Biol Phys 1996;34(5):1141–1146.
150. Yorke E, et al. Dosimetric considerations in radiation therapy of coin lesions of the lung. Int J Radiat Oncol Biol Phys 1996;34(2):481–487.
151. Klein EE, et al. A volumetric study of measurements and calculations of lung density corrections for 6 and 18 MV photons. Int J Radiat Oncol Biol Phys 1997;37(5):1163–1170.
152. Miften M, et al. Comparison of RTP dose distributions in heterogeneous phantoms with the BEAM Monte Carlo simulation system. J Appl Clin Med Phys 2001;2(1):21–31.
153. Wang L, et al. Dosimetric advantage of using 6 MV over 15 MV photons in conformal therapy of lung cancer: Monte Carlo studies in patient geometries. J Appl Clin Med Phys 2002;3(1):51–59.
154. Osei EK, et al. EGSNRC Monte Carlo study of the effect of photon energy and field margin in phantoms simulating small lung lesions. Med Phys 2003;30(10):2706–2714.
155. Chetty I, et al. The influence of beam model differences in the comparison of dose calculation algorithms for lung cancer treatment planning. Phys Med Biol 2005;50:801–815.
156. Epp ER, Boyer AL, Doppke KP. Underdosing of lesions resulting from lack of electronic equilibrium in upper respiratory air cavities irradiated by 10MV X-ray beams. Int J Radiat Oncol Biol Phys 1977;2(7–8):613–619.
157. Beach JL, Mendiondo MS, Mendiondo OA. A comparison of air-cavity inhomogeneity effects for cobalt-60, 6-, and 10-MV X-ray beams. Med Phys 1987;14(1):140–144.
158. Niroomand-Rad A, et al. Air cavity effects on the radiation dose to the larynx using Co-60, 6 MV, and 10 MV photon beams. Int J Radiat Oncol Biol Phys 1994;29(5):1139–1146.
159. Ostwald PM, Kron T, Hamilton CS. Assessment of mucosal underdosing in larynx irradiation. Int J Radiat Oncol Biol Phys 1996;36(1):181–187.
160. Parsons JT, et al. Treatment of early and moderately advanced vocal cord carcinoma with 6-MV X-rays. Int J Radiat Oncol Biol Phys 2001;50(4):953–959.
161. Kan WK, et al. The effect of the nasopharyngeal air cavity on X-ray interface doses. Phys Med Biol 1998;43(3):529–537.
162. Seco J, et al. Head-and-neck IMRT treatments assessed with a Monte Carlo dose calculation engine. Phys Med Biol 2005;50: 817–830.

See also RADIATION DOSE PLANNING, COMPUTER-AIDED; RADIOTHERAPY TREATMENT PLANNING, OPTIMIZATION OF; STATISTICAL METHODS.

RADIATION THERAPY, QUALITY ASSURANCE IN

GLEN GEJERMAN
JOSEPH HANLEY
Hackensack University Medical
Hackensack, New Jersey

INTRODUCTION

The curative goal in radiation therapy is to deliver sufficient doses of tumoricidal radiation to a target volume while protecting the contiguous normal tissues. As radiation dose and accuracy of treatment delivery correlate with improved disease-free survival and avoidance of toxicity, quality control must be maintained throughout the planning and delivery of radiotherapy. The International Commission on Radiation Units and Measurements (ICRU) has recommended that treatment should be delivered to within 5% of the prescribed dose. Treatment planning and delivery is a multistep process that includes clinical decision making, patient immobilization, simulation, delineation of target and avoidance structures, determination of beam number and orientation, dose calculation, dosimetric scrutiny, patient set up, and treatment administration. In order to meet the ICRU's stringent recommendation, each of these steps must achieve better than 3% accuracy, and yet several studies have shown that geometric uncertainty and dosimetric inaccuracy impact the clinical reality of radiotherapy. Patient misidentification, set-up variability, organ motion, block or multileaf collimation placement errors, and dosimetric miscalculation can lead to underdosing or overdosing the target volume and unintentional irradiation of normal surrounding tissue. Studies have shown that these types of errors occur at various stages in the treatment planning and delivery process, are often due to inadequate communication and mistakes in data transfer, and quality assurance procedures facilitate early detection of and reduction in their occurrence (1,2). A recent review of radiation therapy errors over a 5 year period at a major tertiary cancer center found that 44% of errors were due to field deviations, 38% due to incorrect use of beam modifiers, and 18% due to deviations from the prescribed dose. Once a quality improvement intervention that addressed several technological issues such as electronic charting was initiated, a significant impact in error reduction was noted (3). The numerous important quality assurance duties coupled with the increasing sophistication of radiation treatment planning and delivery systems, calls for an integrated comprehensive program that validates, verifies, and maintains accuracy throughout the entire process of radiation therapy delivery (4).

Treatment inaccuracies can be divided into systematic or treatment preparation variations (which include positioning errors, organ motion during treatment planning simulation, contouring errors, field shaping errors, and machine calibration errors) and random or execution variations (which include day-to-day patient misalignment and organ motion during treatment). A comprehensive quality assurance program must encompass both systematic and random uncertainty in treatment planning and delivery in order to minimize their occurrence. Although both types of errors lead to geometrical deviations of the target volume, they have different effects on the delivered dose. Systematic errors cause a displacement of the dose distribution away from the intended target and random errors lead to blurring of the dose distribution. The impact of systematic errors on target dose and the tumor control probability is therefore much greater than the impact of random execution variations (5,6). Identifying and correcting planning preparation errors early in the treatment process is critical in order to mitigate their impact on the treatment outcome.

To minimize treatment inaccuracies, it is essential that each radiation oncology department establish a “quality system” or quality assurance program to provide the organizational structure, responsibilities, procedures, processes, and resources for assuring the quality of patient management. A series of checklists to monitor compliance with the objectives of the program should be developed and applied. Due to the ever-changing nature of radiation oncology, this quality assurance program should be reviewed annually.

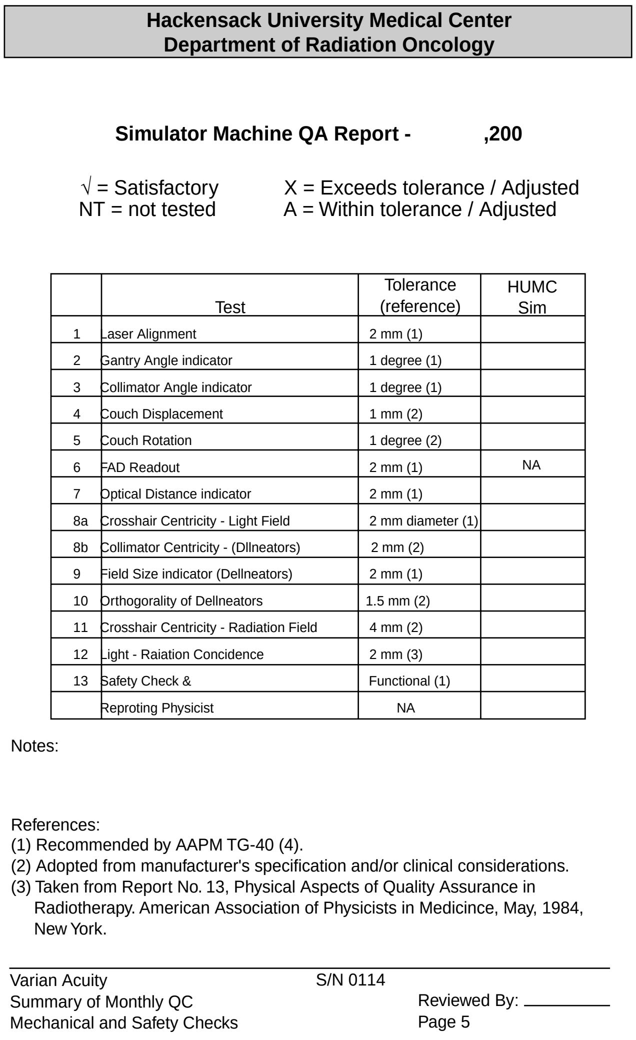
SIMULATION QA

The foundation of treatment planning is the simulation process. After the radiation prescription has been filled out to include the intended target, the organs to avoid, and the total and fractional doses, the patient undergoes treatment planning simulation, during which anatomical data is acquired, patient topography is measured, and the target volume and avoidance structures are delineated. Radiation therapy simulators use fluoroscopic, X-ray and computed tomography (CT) techniques to visualize internal anatomy in relation to external landmarks and can be divided into two broad categories: conventional or fluoroscopic simulators and CT simulators. These simulators can replicate the treatment machine geometry either physically, in the case of a conventional simulator, or virtually on a computer for CT simulation. The quality assurance program for the physical simulators must be parallel to that of the treatment machines, so that the geometric relationship between the treatment unit and the target volume can be accurately and consistently reproduced. To ensure the same accuracy in the case of virtual simulation, the treatment unit must be precisely modeled in the simulation computer.

To minimize intratreatment movement, and to ensure accurate daily positioning, patients are simulated (conventional or CT) in the treatment position with the use of special immobilization devices. These devices extend beyond the treatment site and rigidly immobilize the patient while providing them with support to enhance relaxation and minimize movement. Studies comparing set-up variations in immobilized versus free set up of patients note a significant reduction in positioning errors. In patients without custom immobilization, the percentage of fractions with set-up errors greater than 5 mm ranged from 17–57% and errors greater than 10 mm occurred in 15% of fractions (7,8). A randomized trial analyzing patients in the prone position receiving pelvic radiotherapy found a statistically significant benefit when using rigid immobilization. In the group treated without immobilization, 31% of port films had isocenter deviations greater than 10 mm compared with 11% in the immobilized patients. Average set-up deviations in the anteroposterior, right-left, and superior-inferior directions were 5.2 mm, 3.2 mm, and 4.3 mm in the patients treated without immobilization versus 2.9 mm, 2.1 mm, and 3.9 mm in those treated with rigid immobilization, respectively (9). Patient-related uncertainties also include organ motion. The patient’s treatment position can impact the extent of both inter- and intrafractional movement. A randomized trial analyzing organ motion during treatment demonstrated less prostate motion in the supine treatment position. The mean anterior-posterior organ motion was 0.1 mm for patients treated in the supine position as opposed to 0.7 mm in those treated prone (10). These data demonstrate why proper immobilization is such a vital part of the quality assurance program. It should be noted that the integrity of immobilization devices should be checked on a daily basis during the treatment course.

Quality assurance of the conventional simulator is necessary to avoid inaccuracies that could lead to target-beam misalignment. After installation, and prior to clinical use, a detailed customer acceptance procedure is often performed and can act as a baseline for ongoing testing. A complete QA program for a simulator should follow the guidelines detailed in the American Association of Medical Physicists (AAPM) Task Group 40 (TG40) report (4) and be augmented by any city, state, or federal guidelines. Table 3 of TG40 specifies the tests to be performed and at what frequency and to what tolerance to perform them. As a minimum, the lasers and the optical distance indicator should be checked daily. On a monthly basis, evaluation of mechanical uncertainties includes a check of field size settings and rotational settings, light-field radiation field congruence, treatment couch movement, laser positioning, and cross-hair centering. An example of a monthly QA checklist is shown in Fig. 1. A more thorough series of mechanical checks should be performed on an annual basis. These checks include determination of true axes of rotation of the simulator and the coincidence of these axes. Annual tests of the X-ray generator are essential to ensure that the exposure to the patient is minimized. It is important to check the imaging chain as image quality can directly affect patient care. Specialized phantoms can be used to determine the spatial and contrast resolution. Conventional simulators are equipped with X-ray and fluoroscopic modes, and both modes should be tested.

During conventional simulation, fluoroscopy is used to determine the treatment portals to cover the appropriate target volume. When using fluoroscopic simulation, one must be certain that the field size and shape adequately encompasses the target volume. Without detailed knowledge of the true extent of the tumor volume, fluoroscopically determined treatment portals may result in inadequate coverage. In an analysis of patients with cervical cancer, reconstruction of CT-defined tumor volumes on the simulation films demonstrated inadequate anterior and posterior borders in a significant proportion of patients (11). When the simulation and target localization is completed, the position of the treatment field's isocenter is marked on the patient's skin or immobilization apparatus. In order to be able to consistently reproduce the treatment set up established at the simulation, three or more laser beams are used to establish fiducial marks often called triangulation points on the patient. A detailed recording of these simulation parameters, including gantry and collimator settings and the source-skin distance (SSD) for each treatment field, in the treatment chart will allow for accurate repositioning in the treatment room (12). The data acquired at the time of simulation can also be directly captured into a Record & Verify system (R&V), obviating the need for manual entry with the potential of transcription error. At completion of the simulation, a set of simulation films that show the field size, field isocenter, and projected anatomy from the chosen beam direction and distance is obtained. These films are used as the standard by which the future port films in the treatment room will be measured for set-up accuracy and to assess patient movement.



Hackensack University Medical Center
Department of Radiation Oncology

Simulator Machine QA Report - ,200

√ = Satisfactory X = Exceeds tolerance / Adjusted
NT = not tested A = Within tolerance / Adjusted

	Test	Tolerance (reference)	HUMC Sim
1	Laser Alignment	2 mm (1)	
2	Gantry Angle indicator	1 degree (1)	
3	Collimator Angle indicator	1 degree (1)	
4	Couch Displacement	1 mm (2)	
5	Couch Rotation	1 degree (2)	
6	FAD Readout	2 mm (1)	NA
7	Optical Distance indicator	2 mm (1)	
8a	Crosshair Centricity - Light Field	2 mm diameter (1)	
8b	Collimator Centricity - (Dllneators)	2 mm (2)	
9	Field Size indicator (Dellneators)	2 mm (1)	
10	Orthogorality of Dellneators	1.5 mm (2)	
11	Crosshair Centricity - Radiation Field	4 mm (2)	
12	Light - Raiation Concidence	2 mm (3)	
13	Safety Check &	Functional (1)	
	Reproting Physicist	NA	

Notes:

References:
(1) Recommended by AAPM TG-40 (4).
(2) Adopted from manufacturer's specification and/or clinical considerations.
(3) Taken from Report No. 13, Physical Aspects of Quality Assurance in Radiotherapy. American Association of Physicists in Medicince, May, 1984, New York.

Varian Acuity S/N 0114
Summary of Monthly QC Reviewed By: ________
Mechanical and Safety Checks Page 5

Figure 1. Monthly checklist for simulator machine QA.

Over the past several years, CT simulators have been replacing or used in conjunction with conventional simulators. The CT simulator acquires CT images of a patient and sends them to a computer workstation on which a virtual simulation can be performed. The patient's anatomy can be reconstructed in 3D allowing for a display in a beam's eye view (BEV) perspective of the target and its relation to the normal surrounding tissues in different treatment angles. The BEV is used to create beam apertures that geometrically conform to the projections of the target and normal anatomy through different treatment angles. The computer software can be used to define the treatment isocenter, and a CT-generated virtual image of the patient is then used to complete the simulation with a digitally reconstructed radiograph (DRR) or a digitally composited radiograph (DCR). DRRs are computed radiographs that use the CT simulation data to provide planar reference images with the target volume, organs at risk, isocenter, and field edges shown. The DCR is created by computer enhancement or suppression of the CT numbers that allow for better visualization of the targeted organs. Although the principles of patient immobilization, treatment field delineation, and patient coordinate marking are similar to those of conventional simulation, the digital nature of CT simulation requires additional quality controls. During the acceptance testing, a CT dataset of a humanoid phantom is used for treatment planning to assess contouring capabilities, isocenter calculation, target localization, and DRR reconstruction and data transfer from the CT to the treatment planning system. The phantom is then used to test field size accuracy, virtual gantry and collimator rotation, and the ability to accurately shift the isocenter. In addition to standard QA procedures for conventional CT scanners, CT simulators require interval testing of the laser system and of the data link to the virtual simulation computer system that allows tumor contouring, isocenter and field size definition, transfer of coordinates to the patient's skin, as well as construction of the DRR (13). Upon completion of virtual simulation, the patient's images, contours, and treatment beams are electronically sent to the treatment planning system.

TREATMENT PLANNING QA

Modern treatment planning systems consist of complex software run on sophisticated platforms with multiple peripheral devices. Recognizing the challenge of ensuring proper maintenance and use of these increasingly complicated systems, the AAPM Task group 53 (TG53) published a comprehensive set of quality assurance guidelines that can be applied to clinical radiotherapy planning (14). Acceptance testing and commissioning of a treatment planning system provides the benchmark by which the system will be evaluated during the periodic quality assurance testing. Acceptance testing is performed after installation but prior to clinical use of the system. The process entails testing that the system's hardware, software, and peripheral devices function according to manufacturer's specifications. These tests ensure that the system can properly acquire patient data, process anatomical contouring, orient beam direction, perform dose calculation, display the resultant isodose plots, and print hard copies of the approved treatment plan's parameters. The ability to properly transfer imaging data is confirmed by scanning phantoms of known geometry with internal markers and transferring the imaging data to the treatment planning system. The transferred data is then compared with film images to validate orientation, measurement, and fiducial positioning. System commissioning involves extensive testing of the dosimetric algorithms for a variety of clinical scenarios. The physical properties of each treatment unit have to be entered into the system and checked for consistency with manufacturer's specifications. Data such as percent depth-dose tables, off-axis profiles, and output factors are acquired using a computer-controlled water phantom for each treatment beam on each treatment unit to be used in the planning system. Phantoms with known geometric target volumes are used to simulate common clinical scenarios and treatment plans are evaluated to verify calculated dose distributions. Although anthropomorphic phantoms (that are shaped like the human body) are well-suited to test clinical treatment techniques, geometric phantoms (that are cylindrical or cubic) have more reliable ionization chamber positioning (15). If dosimetric calculations that account for inhomogeneities within the patient are to be performed, a CT number to electron density calibration curve must be established, which is performed for each CT acquisition unit that sends images to the treatment planning system. A phantom containing plugs of known electron density is scanned on the CT and the corresponding CT number is determined. These numbers are plotted versus electron density to derive the calibration curve for that scanner. As an incorrect conversion of CT number to electron density can lead to significant dosimetric miscalculations, the American College of Radiology (ACR) recommends testing this calibration curve monthly. Although the most accurate form of dose calculations are Monte-Carlo-based, these calculations are computationally intensive and cannot currently be used for routine planning. All other dose calculation algorithms used in treatment planning systems have limitations, and it is essential to understand where these limitations manifest, for example, in areas where electronic equilibrium does not exist, such as lung-tissue interfaces. Routine periodic quality assurance testing of the planning system consists of daily, monthly, and annual tests. Daily tests validate the performance of input devices such as point digitizers and the accuracy of output devices such as printers. Monthly tests can involve calculating computer checksums for the treatment planning software executables and machine data, to ensure the program and data has not been modified. Annual tests are more involved and should include a subset of standard treatment plans that cover a wide range of clinical scenarios ranging from point-dose calculations, 2D, 3D conformal radiation therapy (3DCRT), and intensity-modulated radiation therapy (IMRT) plans. This set of standard plans are used for testing whenever software upgrades, either patches or version changes, are applied.

Once the imaging data has been transferred to the treatment planning system, the target volume and avoidance structures must be delineated if not already defined at the CT simulation. This delineation can be the major contributor to overall uncertainty in the treatment planning chain, as many factors exist that contribute to this uncertainty. It is imperative that the treating physician and the radiation treatment planner share a common vocabulary regarding the tumor volume and the additional margins necessary to account for organ motion and set-up inaccuracies. Prescribing and designing a treatment plan to a target without correcting for geometric uncertainties will result in a substantially different delivered dose than the intended one. In order to address these issues, the ICRU Report 50 (16) recommended using specific definitions regarding margins and volumes. The gross tumor volume (GTV) represents the visible tumor. The clinical target volume (CTV) denotes the GTV with an additional region encompassing suspected microscopic spread. The planning target volume (PTV) contains the CTV with margins added to account for geometric uncertainties. These margins are determined based on the extent of uncertainty caused by patient and tumor movement as well as the inaccuracies in beam and patient setup. Several margin recipes based on geometrical uncertainties and coverage probabilities have been published; however, their clinical impact remains to be proven (17). The organs at risk (OAR) are the normal tissues that are contiguous with the CTV (such as small bowel, rectum, and spinal cord) whose radiation tolerance can affect the maximum deliverable dose and treatment technique. The ICRU report 62 (18) refined the definition of the PTV with the concepts of internal margin and set-up margin. Internal margin uncertainty that is caused by physiological changes such as respiratory movement cannot be easily modified without using respiratory gating techniques. In contrast, set-up margin uncertainty can be more readily minimized by proper immobilization and improved machine accuracy. The report also addressed the issue of OAR mobility by introducing the planning organ at risk volume (PRV) in which additional margins are added to account for the geometric uncertainty of these organs. In order to avoid significant radiation toxicity and to maintain post treatment quality of life, the planning physician must be vigilant when considering avoidance structures. In a Radiation Therapy and Oncology Group (RTOG) analysis of the impact of dose escalation in prostate cancer, a lack of physician awareness leading to unnecessary exposure of the penile bulb to high radiation doses lead to treatment-induced impotence (19).

Even with a common terminology and attention to detail when delineating the anatomical structures, several uncertainties exist that are related to the imaging modality used for data acquisition. Proper acquisition of CT data is challenging in that numerous factors, including slice thickness, slice spacing, CT number scale, and organ motion, can affect this information resulting in dosimetric and anatomic inaccuracies. A CT image artifact known as partial volume averaging occurs when two structures of different tissue density occupy the same voxel resulting in an averaging of their CT numbers. Unless the appropriate CT slice thickness is used, accurate target delineation can be compromised, as details of contiguous anatomic structures may not be appreciated. CT imaging of a moving organ can lead to significant distortions, particularly when the organ is small compared with the extent of its displacement. When the scan time is protracted, the artifact can be significant enough to render the reconstructed images unrecognizable in relation to its stationary counterpart (20). TG 53 recommends the use of imaging protocols that standardize scan parameters such as patient position and immobilization, CT slice spacing and thickness, the extent of the patient's anatomy to be scanned, breathing techniques for patients with abdominal or thoracic tumors, and the use of contrast agents (14). Some anatomical structures are better visualized using alternate imaging modalities such as Magnetic Resonance Imaging (MRI) or functional imaging such as Positron Emission Tomography (PET) scans. For example, to improve the accuracy of thoracic GTV recognition, PET scans have been used in conjunction with CT-based simulation. Although in some circumstances, the ability to distinguish between thoracic tumor and atelectasis can result in a smaller GTV (21); at other times subclinical mediastinal adenopathy appreciated on PET will require enlarging the treatment field to encompass all active disease (22). When multiple images sets, acquired with different imaging modalities, are used in the planning process, the images must be accurately correlated in a common frame of reference. Typically, the images sets are "fused" onto the CT frame of reference. In visual fusion, the independent images are studied side by side and are visually fused using data from both to outline the GTV. In software fusion, the independent studies are geometrically registered with each other using an overlay of anatomic reference locations. A recent review found that software fusion reduced intra- and interobserver variability and resulted in a more consistent delineation of tumor volume when compared with visual fusion (23). It is imperative to perform QA on the fusion software. Acquiring datasets of a phantom with known geometrical landmarks on all modalities to be tested and performing the fusion process can accomplish this goal. PET/CT scanners that obtain both images simultaneously allowing for self-registration are becoming more widely available and will further facilitate accuracy in contouring.

In addition to uncertainties associated with various imaging modalities, it is well documented that inter- and intraobserver reproducibility exists in GTV delineation, and significant differences in the size of the GTV are noted depending on the imaging modality used (24,25). When contouring CT images, the correct window level settings must be used to appreciate the extent of the tumor shape and its relation to contiguous organs at risk. The treatment planning CT must be carefully reviewed to assess for positional or anatomic anomalies. For example, data acquired in the thoracic or abdominal region should be carefully examined for any sharp discontinuities in the outer contour that might indicate a change in breathing pattern or physical shift of the patient due to coughing, for example. A retrospective review of prostate cancer patients treated with conformal radiotherapy found an association between rectal distension on the planning CT and decreased probability of biochemical cure. Planning with a distended rectum can result in a systematic error in prostate location and was found to have a greater impact on outcome than disease risk group (26).

Once all the relevant organs have been contoured and the target dose and dose constraints have been unambiguously communicated to the dosimetry team, the appropriate combination of beam number, beam direction, energy, and intensity is determined. These parameters are optimized to deliver maximum dose to the CTV and minimum dose to the OAR. Conventional dosimetric calculations known as forward planning involves an experienced planner choosing multiple beams aimed at the isocenter and altering beam orientation and weighting to achieve an acceptable plan. The dose delivered with the chosen beam arrangement will be affected by the interaction of the radiation beam with the patient's tissue density and is calculated by the planning computer. 3D conformal radiotherapy planning uses CT data to generate tumor and normal organ 3D images and displays them from the perspective of different angles using a BEV technique. Optimization of the treatment plan is performed by iteratively adjusting the beam number and direction, selectively adjusting the field aperture, and applying compensators such as wedges. In contrast, IMRT uses inverse planning to deliver a desired dose to the GTV and PTV with constraints to the OAR. Instead of choosing beam directions and then evaluating the resultant dosimetry, the desired dose distribution is stipulated using dose-volume constraints to the PTV and OAR and then the computer algorithm alters the various beam intensities in an attempt to achieve these planning goals.

After the dosimetry team completes their calculations, the proposed treatment plan must be carefully evaluated to confirm the prescription fractional and total doses and to determine whether it satisfies the prescription goal. For conventional treatments, this determination is performed by inspecting 2D isodose displays through one or more cross sections of the anatomy. For 3DCRT and IMRT, BEV data and dose volume histogram (DVH) analysis is used in addition to the isodose displays to evaluate dose minima, maxima, and means of both target and avoidance structures. The DVH that graphically depicts the percentage of a volume of interest that receives a particular dose does not give spatial information regarding dose distribution. If the DVH indicates underdosing, only by reviewing the plan's isodose display can one locate the area of inadequate coverage. Mathematical models that use DVH statistics to estimate the normal tissue complication probability (NTCP) have been developed. These NTCP models have been found to more accurately predict the likelihood of radiation-induced toxicity than point-dose radiation tolerance data. An important task in a quality assurance program is to calculate the fractional and total doses to the OAR in order to estimate the risk of radiation injury. These doses can be described in terms of minimum, maximum, and mean doses to an entire organ or as the volume of an organ receiving greater than a particular dose. In situations where the PTV anatomically overlaps the OAR, clinical judgment must be used to assign a priority to each goal. The location and volume of dosimetric inhomogeneity (both hot spots and cold spots) must be evaluated, which is particularly true for IMRT where dose homogeneity is often sacrificed for dose conformality.

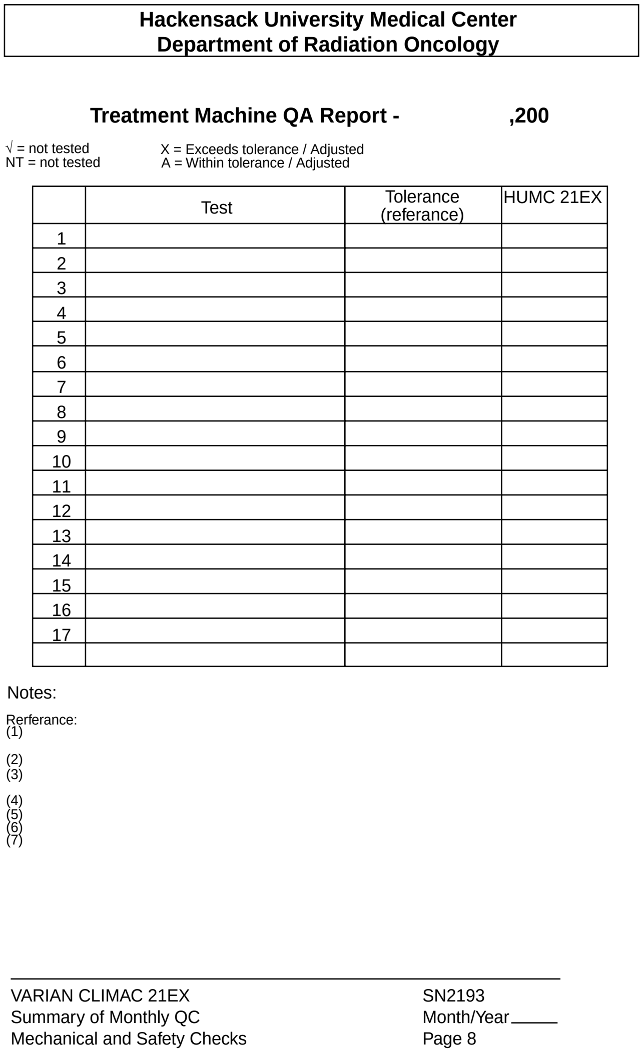
When the plan has been approved by the dosimetrist and the physician, all documented parameters including patient setup, beam configuration, beam intensity, and monitor units are sent to a R&V system either manually or, preferably, electronically. All the data from the plan, printouts, treatment chart, and R&V undergoes an independent review by a qualified medical physicist. This second check entails review of the prescription, the plan's calculation algorithm, wedge placement, dose distribution, DVH, and beam apertures. Hand calculations of a point dose in each field are analyzed to verify the dosimetry. The patient then undergoes a verification simulation to confirm the accuracy and reproducibility of the proposed plan. During this confirmatory simulation, the isocenter position is radiographically confirmed, block geometry is checked, and measurements such as SSD are validated. Finally, a pretreatment port film verification is obtained on the treatment machine to verify reproducibility of set up and to confirm measurements such as the SSD and distance to tabletop.

The verification of patient-specific dose distributions with water phantoms, ionization chambers, diodes, or film dosimetry is an essential component of the QA process. The standard method of evaluation consists of overlaying hardcopy plots of measured and calculated isodose distributions and qualitatively assessing concordance. As a result of the nonuniform intensity inherent in IMRT and the resulting steep dose gradients throughout the treatment field, IMRT plan verification is more challenging. Radiographic film dosimetry can be used to verify the IMRT leaf sequences and monitor units; however, as film sensitivity varies with beam energy, field size, film positioning, and film processing, great care must be taken to normalize the calculated and measured dose. Computer-assisted registration techniques are now available to determine the relative difference between the planned and delivered individual beam fluence or combined dose distributions on a pixel-by-pixel basis in order to score the plan using a predetermined criterion of acceptability (27).

LINEAR ACCELERATOR QA

The quality assurance protocol for linear accelerators is designed to monitor and correct performance of the equipment so that the physical and dosimetric parameters established during commissioning and acceptance testing can be maintained. TG40 (4) described a thorough QA program for linear accelerators with recommended test frequency. The daily tests include checking the safety features such as the door interlock and audiovisual intercom systems. Mechanical performance such as the localizing lasers and the machine's optical distance indicator and dosimetric output such as the X-ray and electron constancy is also checked daily. Monthly checks of the linear accelerator's mechanical accuracy include light-field coincidence; cross hair centering; gantry, collimator, field size, and couch position indicators; latching of the electron cone, wedge, and blocking trays; electron cone interlocks; and the emergency off switch. An example of a monthly mechanical and safety checklist for a linear accelerator is shown in Fig. 2. Monthly checks of the dosimetric accuracy include constancy of the X-ray and electron output, central axis parameters, and X-ray and electron beam flatness. Annual mechanical tests include checks of the safety locks; the tabletop sag; vertical travel of the treatment couch; the collimator's, gantry's, and couch's rotation isocenter; the coincidence of the radiation and mechanical isocenter; and the coincidence of the collimator gantry and couch axes with the isocenter. The annual dosimetry tests check for monitor chamber linearity; wedge transmission factor constancy; off-axis factor constancy; and X-ray and electron output and off-axis constancy dependence on gantry angle. In addition, a subset of the depth-dose and off-axis profile scans acquired at commissioning are performed and compared with the baseline.

The use of multileaf collimators (MLC) in 3D conformal and intensity-modulated radiotherapy requires additional QA measures. When using MLC for 3DCRT, leaf position inaccuracies will have an effect on the resultant dosimetry; however, because of the PTV margins, the effect is minimal. In contrast, when using MLCs for IMRT, a miscalibration of 0.5 mm causing a 1 mm error in radiation portal size can cause a 10% dose error when delivering IMRT with an average field size of 1 cm (15). As MLC function is critical to dosimetric accuracy, rigorous QA protocols are required. The accuracy of the multileaf collimator (MLC) is verified by using radiographic film to measure radiation dose patterns and by checking for a gap between the leaves when they are programmed to be in the closed position.



Hackensack University Medical Center
Department of Radiation Oncology

Treatment Machine QA Report - ,200

√ = not tested
NT = not tested
X = Exceeds tolerance / Adjusted
A = Within tolerance / Adjusted

	Test	Tolerance (referance)	HUMC 21EX
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			

Notes:

Rerferance:
(1)
(2)
(3)
(4)
(5)
(6)
(7)

VARIAN CLIMAC 21EX
Summary of Monthly QC
Mechanical and Safety Checks

SN2193
Month/Year_____
Page 8

Figure 2. Monthly mechanical and safety checklist for a linear accelerator.

TREATMENT DELIVERY QA

After the linear accelerator's function has been verified with a thorough QA program, it is important to confirm that the treatment parameters established during treatment planning are accurately transferred to the treatment machine and that the daily set up and treatment are accurately executed. The R&V system is an important tool that confirms that proper field size, beam arrangement, multileaf collimator settings, collimator angle, gantry angle, beam energy, wedges, and monitor units are used each day. As an R&V system will give a daily validation of the entered parameters, it is essential to verify the set-up data with an independent check on the first day of treatment. The quality assurance mechanism for daily radiation treatments includes laser alignment of the fiducial tattoos and validation of SSD and tabletop measurements as well as weekly port films. These port films are obtained prior to treatment initiation as well as weekly and are compared with the initial simulation film or with the DRR to evaluate isocenter location and block or MLC position. A recent advance in set-up verification and error detection is the development of electronic portal imaging devices (EPID). As a replacement to port films, the EPID provides faster acquisition time, tools that allow digital image enhancement, and tools that can measure the distance of anatomic landmarks to the isocenter or field edge (28). The use of EPID has allowed for the potential of adaptive radiotherapy that collects geometrical uncertain information during the first few treatment fractions and sends it back to the treatment planning system for further optimization. Once the set-up variations have been characterized, the treatment is adapted to either adjust the field size and treatment couch position or adjust the fluence profiles to correct for the set-up error.

SUMMARY

An overview of the major components of quality assurance for external beam radiation therapy has been given. The need to establish a quality assurance program, which provides the organizational structure, responsibilities, procedures, processes, and resources for assuring the quality of patient management, has been demonstrated. The various components that contribute to treatment inaccuracies have been identified as systematic or random variations, and it was established that the impact of systematic errors on target dose and the tumor control probability is much greater than the impact of random variations. The simulation was described as the foundation of treatment planning process and intra- and inter-fraction patient motion was addressed with patient immobilization. QA programs for simulators, linear accelerators, and CT scanners should follow the guidelines provided in AAPM TG40 (4). QA for treatment planning systems is outlined in AAPM TG53 (14). The many factors that contribute to target delineation uncertainty, namely, acquisition parameters, organ motion, imaging modality, image fusion, and intra-observer variability, should all be examined closely for their contributions to treatment uncertainties.

As a result of the ever-changing nature of radiation oncology, this quality assurance program should be reviewed annually. Special attention should be given to new devices and treatment protocols. As adaptive radiation therapy evolves, a whole new component of the quality assurance program will need to be developed.

BIBLIOGRAPHY

1. Yeung TK, et al. Quality assurance in radiotherapy: Evaluation of errors and incidents recorded over a 10 year period. Radiother Oncol 2005;74:283–291.
2. Valli MC, et al. Evaluation of most frequent errors in daily compilation and use of a radiation treatment chart. Radiother Oncol 1994;32:87–89.
3. Huang G, et al. Error in the delivery of radiation therapy: Results of a quality assurance review. Int J Radiation Biol Phys 2005;61:1590–1595.
4. Kutcher GJ, et al. Comprehensive QA for radiation oncology: Report of AAPM radiation therapy committee task group 40. Med Phys 1994;21:581–618.

5. van Herk M, Remeijer P, Lebesque JV. Inclusion of geometric uncertainties in treatment plan evaluation. Int J Radiation Biol Phys 2002;52:1407–1422.
6. van Herk M, et al. The probability of correct target dosage: Dose-population histograms for deriving treatment margins in radiotherapy. Int J Radiation Biol Phys 2000;47:1121–1135.
7. Rosenthal SA, et al. Immobilization improves the reproducibility of patient positioning during six-field conformal radiation therapy for prostate carcinoma. Int J Radiation Biol Phys 1993;27:921–926.
8. Catton C, et al. Improvement in total positioning error for lateral prostatic fields using a soft immobilization device. Radiother Oncol 1997;44:265–270.
9. Kneebone A, et al. A randomized trial evaluating rigid immobilization for pelvic irradiation. Int J Radiation Biol Phys 2003;56:1105–1111.
10. Bayley AJ, et al. A randomized trial of supine vs. prone positioning in patients undergoing escalated dose conformal radiotherapy for prostate cancer. Radiother Oncol 2004;70:37–44.
11. Nagar YS, et al. Conventional 4 field box radiotherapy technique for cancer cervix: Potential for geographic miss without CECT scan based planning. Int J Gyn Ca 2004;14:865–870.
12. Reinstein LE. Patient positioning and immobilization. In: Kahn FM, Potish RA, editors. Treatment Planning in Radiation Therapy. Baltimore, MD: Williams and Wilkins Publishing; 1998. p 55–88.
13. McGee KP, Das IJ. Commissioning acceptance testing and quality assurance of a CT simulator. In: Coia LR, Schultheiss TE, Hanks GE, editors. A Practical Guide to CT Simulation. Madison, WI: Advanced Medical Publishing; 1995.p 5–23.
14. Fraass B, et al. American association of physicists in medicine radiation therapy committee task group 53: Quality assurance for clinical radiotherapy treatment planning. Med Phys 1998; 25:1773–1829.
15. Low DA. Quality assurance of intensity modulated radiotherapy. Semin Radiat Oncol 2002;12:219–228.
16. ICRU Report 50 Prescribing, recording, and reporting photon beam therapy. Bethesda, MD: International Commission on Radiation Units and Measurements; 1993.
17. Rasch C, Steenbakkers R, van Herk M. Target definition in prostate, head and neck. Semin Radiat Oncol 2005;15:136–145.
18. ICRU Report 62. Prescribing, recording, and reporting photon beam therapy (Supplement to ICRU Report 50). Bethesda, MD: International Commission on Radiation Units and Measurements; 1999.
19. Roach M, et al. Penile bulb dose and impotence after three dimensional conformal radiotherapy for prostate cancer on RTOG 9406: Findings from a prospective multi-institutional phase I/II dose escalation study. Int J Radiation Biol Phys 2004;60:1351–1356.
20. Gagne IM, Robinson DM. The impact of tumor motion upon CT image integrity and target delineation. Med Phys 2004;31: 3378–3392.
21. Schmuecking M, et al. Image fusion of F-18 FDG pet and CT- is there a role in 3D radiation treatment planning of non small cell lung cancer? Int J Radiation Biol Phys 2000;48(Suppl): 130.
22. Kiffer JD, et al. The contribution of 18F fluoro-2-deoxy-glucose positron emission tomographic imaging to radiotherapy planning in lung cancer. Lung CA 1998;19:167–177.
23. Fox JL, et al. Does registration of PET and planning CT images decrease interobserver and intraobserver variation in delineating tumor volumes for non-small cell lung cancer? IJROBP 2005;62:70–75.
24. Leunens G, et al. Quality assessment of medical decision making in radiation oncology: Variability in target volume delineation for brain tumors. Radiother Oncol 1993;29:169–175.
25. Roach M, et al. Prostate volumes defined by magnetic resonance imaging and computerized tomographic scans for three-dimensional conformal radiotherapy. Int J Radiation Biol Phys 1996;35:1011–1018.
26. De Crevoisier R, et al. Increased risk of biochemical and local failure in patients with distended rectum on the planning CT for prostate cancer radiotherapy. Int J Radiation Biol Phys 2005;62:965–973.
27. Kapulsky A, Gejerman G, Hanley J. A clinical application of an automated phantom film QA procedure for validation of IMRT treatment planning and delivery. Med Dosim 2004;29: 279–284.
28. Herman M. Clinical use of electronic portal imaging. Semin Radiat Oncol 2005;15:157–167.

See also CODES AND REGULATIONS: RADIATION; RADIOTHERAPY TREATMENT PLANNING, OPTIMIZATION OF; X-RAY QUALITY CONTROL PROGRAM.

RADIATION, ULTRAVIOLET. See ULTRAVIOLET RADIATION IN MEDICINE.

RADIOACTIVE DECAY. See RADIONUCLIDE PRODUCTION AND RADIOACTIVE DECAY.

RADIOACTIVE SEED IMPLANTATION. See PROSTATE SEED IMPLANTS.

RADIOIMMUNODETECTION. See MONOCLONAL ANTIBODIES.

RADIOISOTOPE IMAGING EQUIPMENT. See NUCLEAR MEDICINE INSTRUMENTATION.

RADIOLOGY INFORMATION SYSTEMS

JANICE C. HONEYMAN-BUCK
University of Florida
Gainesville, Florida

INTRODUCTION

If asked to define the functionality of a radiology information systems (RIS) 5–10 years ago, one may have listed order entry, film management, charge capture, billing, patient and examination tracking, and possibly inventory management. These systems were in place long before Picture Archiving and Communication Systems (PACS) and the Electronic Medical Record (EMR) were in widespread use. Although PACS, and especially the EMR, are not globally implemented, it is an accepted premise that they will be globally implemented sometime in the future. Now, people often refer to radiology information systems as a suite of computers serving the myriad of functions required for an electronic radiology practice that might include the classic RIS, PACS, speech recognition, modality workflow, and the radiology portion of the EMR and the Hospital Information System (HIS). Generally, in this article RIS will be defined in the more classic sense, it is the computer that manages all aspects of radiology orders. The RIS must now additionally perform the functions required to automate the workflow in a radiology department including examination ordering and management, modality scheduling, examination tracking, capturing charges, inventory management, electronic signatures, report distribution, and management reporting. Every transaction and interaction with the system must be recorded for auditing purposes and must help maintain the privacy and security of Protected Heath Information (PHI) for a patient. Furthermore, the RIS must interface seamlessly with a PACS, a speech recognition system, an HIS and an EMR. This is a nontrivial task in a multivendor installation.

This article focuses on the RIS as the center of the radiology department workflow management with the interfaces to other systems. Interface standards are introduced with examples and some developing concepts in healthcare as they pertain to radiology are discussed.

INFORMATION SYSTEMS IN RADIOLOGY

Figure 1 shows the typical workflow associated with a radiology study. The referring physician orders the study, typically through the HIS, but often in smaller institutions, through the RIS. The order is then sent to PACS and to the speech recognition system. It is available for the technologists as a virtual worklist and on PACS modalities through DICOM modality worklist. In this case, the modality worklist is supplied by a broker or translation system that creates the correct format from the order. After the examination is completed, it is sent to archives and PACS displays where it is interpreted by the radiologist. The radiologist is working from their own worklist of studies that are available according to their role in the department and which have not been dictated. A role may be defined as a radiologist who interprets chest studies or perhaps a radiologist who interprets CT studies. The radiologist dictates into a speech recognition system that appends the report to the data about the examination that exists in the system. The reports are sent to the RIS, closing the loop, and then are sent to the EMR. The dashed line from the PACS Archive to the EMR indicates that the images may be stored in the PACS archive with links to them in the EMR, so it is unnecessary to store the images in both systems. The PACS displays and databases and the speech recognition system are frequently from different vendors, so an interface between the two must be accomplished to be certain that the radiologist is dictating the report on the study they are viewing. This loose coupling of the report and imaging exam must be carefully developed and tested to be sure the reports are permanently attached to the correct exam. Of course when all the systems are purchased from a single vendor, a tighter coupling of data and images may be easier to accomplish. The magic number in a radiology system that ties all the information, images and reports for a specific study together is the accession number. This number should be unique for a study and should identify the patient, exam, date, time, report, images, contrast used and anything else that goes with that study. An accession number query should only get one accurate result.

THE RIS AND ITS ROLE IN RADIOLOGY WORKFLOW

Although the RIS appears to be a very small actor in radiology workflow, this is far from the truth. Note the number of interfaces, indicated by arrows, with the RIS as opposed to the other systems in Fig. 1. The RIS is the integral part of the total electronic radiology practice and without it there would be no connection with the rest of the healthcare enterprise (1,2).

The other systems in Fig. 1 are also important and must be interfaced carefully. Although at times the lines between the functionality of the various systems blur, each

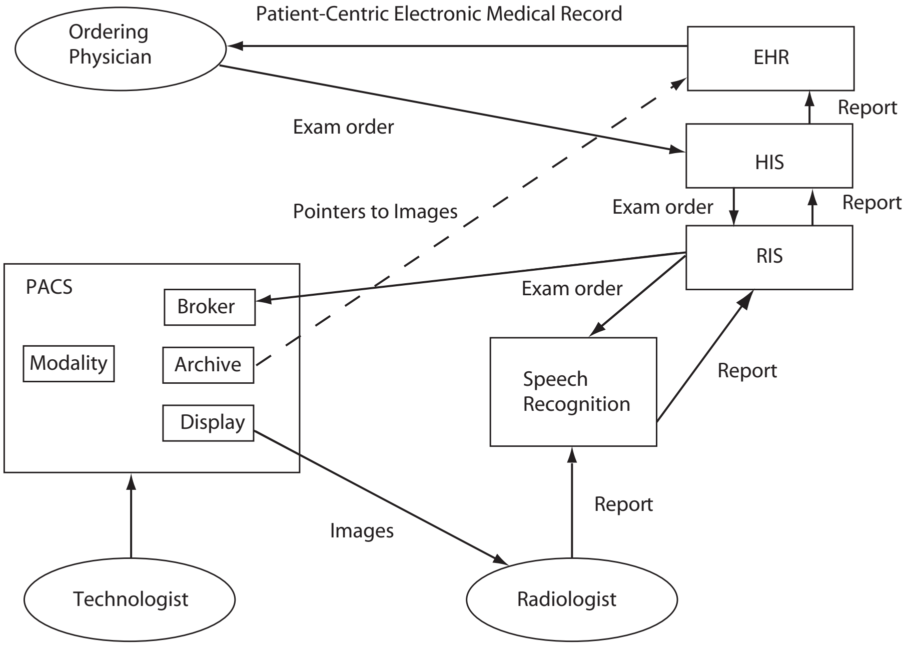


Figure 1. Interactions among information systems during the workflow associated with a radiology study.

system has its purpose in the overall electronic health system. The HIS is generally a system used to capture information about a patient, including but not limited to, their demographics, address, and insurance information (very important). The HIS captures charge information, alerts users when communications with a third-party payer is necessary, and generates bills. Frequently, the HIS is the centralized location for ordering various studies including radiology, lab, speech pathology, physical therapy, and so on. In addition, the HIS can be the center for reporting hospital issues, such as maintenance needs, network failures, and so on. The HIS usually provides extensive administrative reporting tools. The EMR is a patient-centric system containing the electronic record of all a patient's encounters at the institution. The HIS can be used to generate numbers of radiology orders generated in a certain time frame, but when a physician needs to see the total record for his patient with all associated orders, reports, and images to form a diagnosis and treatment plan, the EMR is a better model. The EMR will be discussed with more detail later in the article. The main purpose of PACS is to manage radiology images efficiently. These very large datasets have special needs when it comes to archiving, display, and networks and it makes sense to keep the PACS functionality a little apart from the rest of the usually text-based information systems. Speech recognition is generally developed for specific specialties, such as radiology. The radiology lexicon known by the recognition engine is unique to radiology. Of course, parallel systems exist in other specialties, but the focus here is on radiology. In general the radiologist dictates a report; the speech recognition transforms the voice file into a text file and displays it for confirmation of correctness or for editing.

RIS BASIC FUNCTIONALITY

Most RIS vendors offer a core feature set that captures the information items regarding a patient study in radiology and provides tools for those involved to manage the study. Table 1 is based on the list generated by Aunt Minnie (3) in the section on Radiology Information Systems in their buyer's guide. These functions are probably the minimum required for an RIS purchase. Anyone seeking to purchase an RIS should be aware of the resources available to help them with their decision and should follow a structured Request for Proposal (RFP) or bid format. Buyers should specify how these functions should work in their institution. For example, if the feature "interaction checking between exams" is included, the buyer should write a requirement that matches their workflow. An example of part of the requirement might be, "automatically check patient history for previous exam and warn user at the time of order". If you go back to Fig. 1, it is quite common for the physician or their agent to enter the order on the HIS, then this is transmitted electronically to the RIS, so the RIS would need to alert the HIS, which in turn would alert the physician or their agent. A requirement specification document tells vendors what the buyer expects and forces them to respond to the buyer with respect to their specific workflow. Buyers should be sure to include any special requirements for the institution. Some examples include the length and format of the medical record number or accession number or the requirement that it must be possible to enter a report directly on the RIS in the case of an HIS downtime. The buyer may want to specify their requirements for performance; "a query for a patient record should return results in < 2 s".

Table 1. Core Feature Set for the Typical Current RIS

Function	Comment
Patient registration	This may be performed at an HIS level and transferred to the RIS
Patient tracking	This allows a user to track the patient through the procedure (started, ended)
Order entry	This may be through the HIS, but in the case of HIS downtime, users must be able to enter an order
Merge or reconcile patient information	This is important after a downtime when temporary Medical Record Number (MRN) or accession numbers may have been used or in a trauma situation where a temporary name was used.
Interaction checking between exams	Does a previous exam interfere with the one being ordered?
Single exam code for combined orders	Can you combine Chest/Abd/Pelvis on one exam code with individual accession numbers?
Generate future orders	
Generate recurring orders	Can this also remind the ordering physician that the patient is receiving recurring orders, for example, recurring portable chest exams?
Document imaging	Can you scan paper and include it in the record?
Alerts for pregnancy, diabetes, allergy, and so on	This information will probably be sent from the HIS.
Attaches Prep information to ordered procedures	
Generates an online worklist for technologists	
Supports DICOM modality worklist	This could be a native feature in the RIS or an interface to PACS.
Charge capture	Including examination, supply items
Links CPT and ICD-9 codes to an examination	
Supports distribution of reports	This may be an interface to an HIS or an EHR.
Generates requisitions, labels, and so on	For those still using paper
HIPAA audit capabilities	
Graphic User Interface	

Although an institution may have an HIS and/or an EMR, the RIS will probably also keep an archive of diagnostic reports. From an information sciences storage perspective, a single data repository for specific information is more desirable that multiple, different repositories that have to be synchronized and managed. Since the RIS is rarely the system that is used by the ordering or referring physicians, it serves as an additional archive of diagnostic reports and can be used to track trends, search for diagnoses and impressions, and is used as a backup should other archives fail.

Table 2 contains a (noncomprehensive) list of advanced features for an RIS. Many of these features have become more important since hospitals have added PACS and required integration of all their systems. Bar code support may not be as important in a paperless world, but most of us are not there yet. Bar codes offer a quick and painless way to choose accurate information from a database. One or two bar code entries can locate the right information without the frustration most people experience trying to enter a long string of numbers and letters.

More institutions have merged into one larger entity or have splintered out clinics and imaging centers to locations with easier access. Since it is frequently the case that patients can be seen in various locations in a healthcare system, the RIS needs to support multifacility scheduling, as well as patient tracking. If a group of institutions form an enterprise and each institution can create individual medical record numbers (MRN) identifying patients, it is possible that more than one person will have the same medical record number: from different institutions. The RIS, as well as the EMR and HIS and PACS, must be able to differentiate individuals and track them throughout all the institutions in an enterprise. This can be a difficult and frustrating problem and should be managed carefully. All the interfaces need to be specified in detail. A more comprehensive discussion on the required interfaces is included later in this article. Electronic signatures and addenda need to match the workflow for an institution. In a large teaching hospital, it is common for a resident to dictate a report and for a faculty member to approve and verify it. At this point, two signatures are required. Then, if an addendum is entered on the report, which could happen when comparison studies are received from an outside source, a different resident–faculty member combination could dictate it, resulting in up to four signatures on a report. If multiple addenda are allowed, multiple signatures must also be allowed.

It is important for report distribution, and especially in the case where critical results have been found, that referring physician information is stored somewhere. When an RIS is in a stand-alone installation or the RIS is the distribution entity for reports, then each ordering physician and ordering service must have a unique profile and protocol rules for the communication of reports and especially for critical results. When the RIS is part of a larger enterprise system, the HIS or EMR will distribute the reports, but some triggering mechanism must alert the physician if a critical result or unusual finding is present. This will usually be a function of the RIS. The system must keep an audit of the successful communication with the ordering physician or service as part of the Joint Commission on the Accreditation of Health Care Organizations (JCAHO) recommendations for taking specific steps to improve communications among caregivers (4). This organization requires that each accredited institution have a method for rapid communication of results for both critical results and critical tests in place. In addition, each institution must have a way to monitor and report the efficiency and effectiveness of the communication for a subset of the results considered critical. For radiology, this will most likely be a function of the RIS.

The RIS should be able to generate administrative reports on the numbers of studies performed by date, the modalities used for studies, performance figures for technologists, costs of examinations, and reimbursement trends. In addition, the system should have a query interface so custom reports can be generated. It is very common for this query interface to use the Structured Query Language (SQL) that may have a steep learning curve. Some

Table 2. Advanced Feature Set for the Typical Current RIS

RIS Options	Comments
Bar code support	For quickly entering data in the environment where paper is still used
Supports multifacility scheduling	Can a single person be tracked through multifacility visits?
Interface to PACS	This needs careful specification and configuration
PACS included	
Appointment confirmation / reminders	
Information about referring physicians available	This may be better in the HIS, but will require close coordination
Technologist comments stored	Are they available in the speech recognition system or in PACS? How will the radiologist see them?
Electronic signature	
Proxy signature	
Dual signature	
Different signature on an addendum	
Multiple addenda	
Web based user interface	
Integration with speech recognition	
Interface to HIS	
DICOM modality worklist	This may be provided by the RIS instead of the broker in PACS

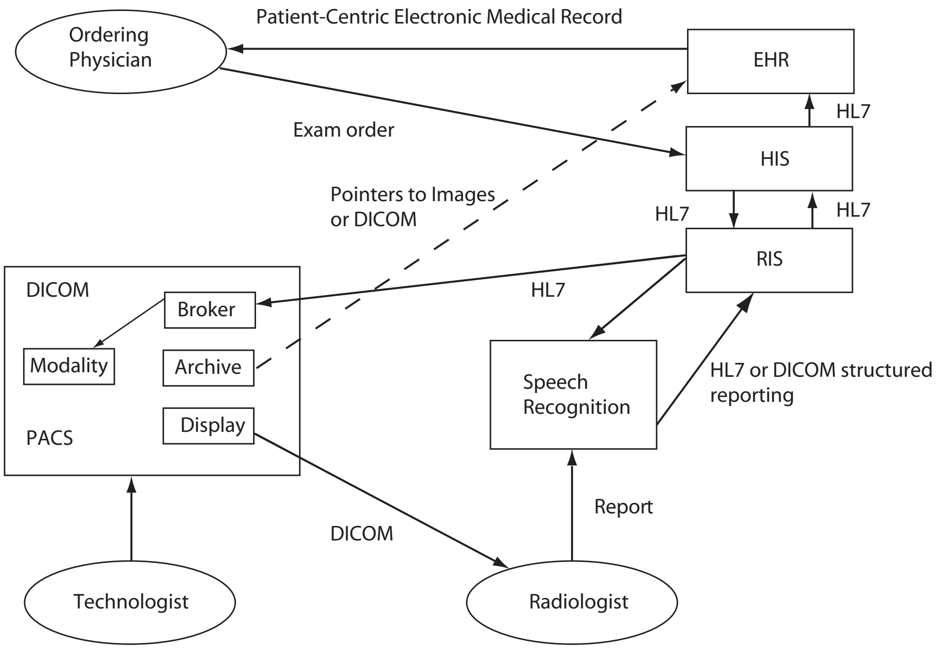


Figure 2. Current standards used as interfaces among information systems associated with a radiology study.

vendors offer a graphical user interface for custom queries that may make queries and reports easier to generate. Many of the RIS systems on the market today have a web-based interface that seems to be intuitive to the current internet-literate generation.

It is increasingly common to find RIS vendors storing images and providing PACS services, such as study display. For some users, this may be a good solution to an electronic radiology practice, for others, the RIS vendors may not have the sophisticated tools for image manipulation that PACS vendors have traditionally supplied. In addition, to complicate matters, PACS vendors are now offering more RIS functionality and once again the borders between the two systems are becoming blurred. It will be up to the institution to decide whether to use one of these hybrid systems or to interface two dedicated systems.

INTERFACING SYSTEMS

In radiology, there are two main interface standards, Health Level 7 (HL7) (5) and Digital Imaging Communications in Medicine (DICOM) (6). The HL7 is a formatted text-based standard that typically specifies the content of messages that are communicated among information systems in a healthcare institution, such as admission information, radiology results, and operating room notes. An RIS typically contains information about radiology schedules, orders, billing and reports, mostly text values. A complete EMR needs reports and images from radiology. Reports can be sent using HL7, but there is no way to encode image data in the current versions of HL7. Because of this limitation, DICOM was developed as the standard to handle transfer of images between systems that acquire, store, and display them.

Figure 2 illustrates the usual standards used for communication in the electronic medical practice first introduced in Fig. 1. Following the information flow, the attending physician places their order for a radiology study on the HIS, which is sent to the RIS via HL7. The order information is sent to PACS and the study is scheduled. The HL7 information is converted to DICOM in PACS using a translation program or broker and the patient demographics are attached to a study produced by PACS acquisition units (CT scanners, MRIs, computed radiology units, etc.). The radiologist using the DICOM viewer views the images and a report is generated using speech recognition. Note that the speech recognition system also receives the ordering information from the RIS, so at this point, there should be a single pointer or index to match the images with the report. The output from speech recognition could be in an HL7 format or a DICOM structured report. The report is usually sent to the RIS and then to the EMR.

The HL7 offers a rich data definition and is widely used to communicate information about the status of the patient in the form of an Admission Discharge Transfer (ADT) packet that informs all relevant systems about the location and identity of a patient, an order packet (ORU), and a report packet (ORP). There are of course many other types of packets that are part of normal transactions of an institution, but these are the ones most often used in the transfer of information between the RIS, PACS, and the speech recognition system. An example of an HL7 message is shown in Table 3. The field entries have been created for this example and this does not reflect an actual patient, referring physician, or radiologist. However, if this were an actual patient, the medical record number would be 0000123456, and the patient's name would Robert Richards, who was born on September 1, 1991 (19910901). The patient's ordering and attending physician is Forest Wood whose phone number is listed in the order. The accession number for this patient is 4445555 and the examination ordered is a bone age radiograph. The reason for the study is included in the text. For this institution, the procedure number for the routine bone age is 3000 and the location where the study will

Table 3. Example HL7 Simulation Showing a Typical Radiology Order

```
PID||100000123456|000000123456|000004567890|RICHARDS∧ROBERT
||19910901|M||B|555 NW 2ND ST
PV1|1|O|XRY∧|C|||09999∧WOOD∧FOREST∧(352)555-
1212|09999∧WOOD∧FOREST∧(352)555-1212||UF||||XRY,
ORC|NW|444444∧0|00003-
001|AUXR3000|SC||1∧∧∧200309151019∧∧R||200309151022|||09999∧WOO
D∧FOREST∧(352)555-1212||||∧||OBR|1|4445555∧0|00003-
001|UXR∧3000∧∧BONE
AGE|ROUTINE|200309151022|200309151022|||MIL|||ORDERED|2003091
51022||09999∧WOOD∧FOREST∧(352)555-1212|(352)555-
1212||ORT|UEXT||||UXR||∧∧∧∧∧∧∧|1∧∧∧200309151022∧∧R||||BONE AGE
RT WRIST PAIN S/P LT∧DISTAL TIBIA HEMI EPIPHYSIODES IS H/O
MULTIPLE HEREDITARY EX∧STOSES AND LIMB LENGTH
DISCRE∧DISTAL TIBIA HEMI EPIPHYSIODES IS H/O MULTIPLE
HEREDITARY EXO∧STOSES AND LIMB LENGTH
DISCRE|∧∧∧||||200309151022
```

be performed is AUXR. When two or more vertical lines appear in the message, they represent fields that contain no data. Each field of an HL7 message is carefully specified and each system using HL7 understands that the PID–3 field will contain a medical record number. A report message in HL7 also contains specified fields that include many of those in this order packet as well as the diagnostic report, the reporting radiologist, and the verifying or signing radiologist. The accession number, 4445555 will be the key for connecting the order to the report to the images, and also to identify that this study was performed on Robert Richards. Much more information on HL7 and the standards committee may be found on the HL7 home page (5) and on numerous other web sites.

Since HL7 had no current capacity for images, the American College of Radiology (ACR) and the National Electrical Manufacturer's Association (NEMA) developed a standard for image communication among PACS devices. The earliest incarnations of this standard were hardware based with point-to-point connections between a modality and a computer. Originally named the ACR–NEMA standard, it has evolved to become DICOM and is still evolving to meet the ever-changing demands of radiology. The standard not only specified the content of the messages passed, it specified the communication pathways for carrying these messages. A full description of the DICOM standard is beyond the scope of this article. More information may be found on the DICOM home page (6), where DICOM specifies the transfer of images among devices, printing to film and paper, and the creation of a modality worklist, among many other things. Radiology modalities typically do not have HL7 interfaces, they are strictly DICOM enabled computers, and in order for them to have knowledge of an order placed by the RIS, the HL7 order needs to be translated into DICOM. For historical reasons, this translation device or software is often called a broker. The broker receives the HL7 and builds a DICOM modality worklist that may be queried by modalities, such as computed tormography (CT) or magnetic resonance imaging (MRI) units. The modality then compiles a list of studies that have been ordered and the technologist may pick a study from the list as they are setting up the console for the examination. This alleviates the need to reenter information about the patient and assures accurate and complete information that can be tracked back to the original order. The all-important accession number is thus associated with the study being performed. Table 4 shows a partial DICOM message attached to the image produced for the study ordered previously. This message is a series of groups, elements, element sizes, and element contents that are again very specific. The accession number is always located in group 8, element 50; the patient's name is always in group 10, element 10; and the patient's medical record number is always in group 10, element 20. The groups contain common elements. Group 8, for example, is information about the examination while group 10 contains information about the patient.

When the diagnostic report is generated on the speech recognition system, the accession number is attached and when the study is completed, the image (Fig. 3) and report (Table 5) are matched correctly. An archive query returns the basic information for the study so a correct selection can be made (Fig. 4).

The basis for PACS and speech recognition working together correctly is the RIS acting as an order entry and management system along with the HL7 and DICOM standards. Without the RIS and the standards, there would be no accurate way to capture all the patient information and to attach the report to the image. The PACS images should never be available without accurate and complete information (7). Every PACS must have a way to link pertinent patient information and the diagnostic report with the image. Although the example presented was fairly simple and straightforward with only one image, consider the case where a patient in an intensive care unit has multiple chest radiographs every day. Without a way to associate a report accurately with an image, an ordering physician could read a report for the wrong time. With accurate and complete information everyone can be assured that the report is attached to the correct study.

OPTIONS FOR THE INTEGRATION OF SYSTEMS

Although the standards are crucial to the success of an integration project, there are usually still issues among

Table 4. Example Portion of a DICOM Message Corresponding to the HL7 Demonstrated in Table 3[a]

0008 0020 DA	8	Study Date	20030915
0008 0021 DA	8	Series Date	20030915
0008 0022 DA	8	Acquisition Date	20030915
0008 0030 TM	6	Study Time	111002
0008 0031 TM	6	Series Time	111002
0008 0032 TM	6	Acquisition Time	111002
0008 0050 SH	8	Accession Number	4445555
0008 0060 CS	2	Modality	CR
0008 0070 LO	4	Manufacturer	AGFA
0008 0080 LO	22	Institution Name	Shands Hospital at UF
0008 0090 PN	12	Referring Physician's Name	WOOD∧FOREST∧
0008 1010 SH	10	Station Name	ADCPLUS03
0008 1030 LO	8	Study Description	BONE AGE
0008 103E LO	8	Series Description	hand PA
0008 1040 LO	26	Institutional Department Name	Shands at UF / Orthopedics
0008 1090 LO	8	Manufacturer's Model Name	ADC_5146
0010 0000 UL	4	Group 0010 Length	70
0010 0010 PN	16	Patient's Name	RICHARDS∧ROBERT
0010 0020 LO	8	Patient ID	00123456
0010 0030 DA	8	Patient's Birth Date	19910901
0010 0040 CS	2	Patient's Sex	M
0018 0000 UL	4	Group 0018 Length	198
0018 0015 CS	4	Body Part Examined	HAND
0018 1000 LO	4	Device Serial Number	1581
0018 1004 LO	4	Plate ID	02
0018 1020 LO	8	Software Versions(s)	VIPS1110
0018 1164 DS	30	Imager Pixel Spacing	1.00000000E-01\1.00000000E-01
0018 1260 SH	6	Plate Type	code 15
0018 1401 LO	12	Acquisition Device Processing	10101Ia713Ra
0018 1402 CS	8	Cassette Orientation	PORTRAIT
0018 1403 CS	8	Cassette Size	8INX10IN

[a] Note the inclusion of the accession number.

vendors. The DICOM can be interpreted in different ways: HL7 fields may be required by one vendor and ignored by another, and problems will arise. Buyers have several options to consider. The easiest way to assure that integration will be successful is to purchase all information systems from the same vendor. Of course, that will not completely assure success because not all vendors can supply all the required systems. Most noticeably, they may not all supply the imaging modalities needed. As soon as another vendor's modality is introduced, an integration project is needed.

Another decision that must be made is the actual storage of the images from PACS and the diagnostic reports associated with these images. The RIS or EMR can supply the storage for both, or the PACS can supply the storage for both, or some combination of storage options can be designed. Neither the RIS nor the PACS would be the preferred method for storing and delivering studies and reports to referring and ordering physicians because the information is all radiography-centric and it is far more desirable to see a total picture of the patient with laboratory results, history and physical notes, nursing documentation, and all other information regarding the health record for a patient.

The Health Insurance Portability and Accountability Act (HIPAA), a landmark law that was passed in 1996, specifies mandates in the transactions between healthcare companies, providers, and carriers (8). This act was originally designed to make the healthcare records for a patient available to healthcare institutions and physicians and to insurance companies using standards. Individuals should be able to give a physician permission to access their healthcare record at any location, quickly and efficiently. The law also protects the privacy of the patient and provides security guidelines to assure the information could not be accessed inappropriately. Only an EMR system build on well-accepted standards can meet these requirements (9–12). The radiology information should be either stored in an EMR or the EMR should have pointers to the information and should be able to communicate that information in a standard format.

The Radiological Society of North America (RSNA), the Healthcare Information and Management Systems Society (HIMSS), and the American College of Cardiology (ACC) are working together to coordinate the use of established standards, such as DICOM and HL7, to improve the way computer systems in healthcare share information. The initiative is called "Integrating the Healthcare Enterprise" or IHE (13). Integrating the Healthcare Enterprises promotes the coordinated use of established standards, such as DICOM and HL7, to address specific clinical needs in support of patient care. Healthcare providers envision a day when vital information can be passed seamlessly from system to system within and across departments (10,11,14). In addition, with IHE, the EMR

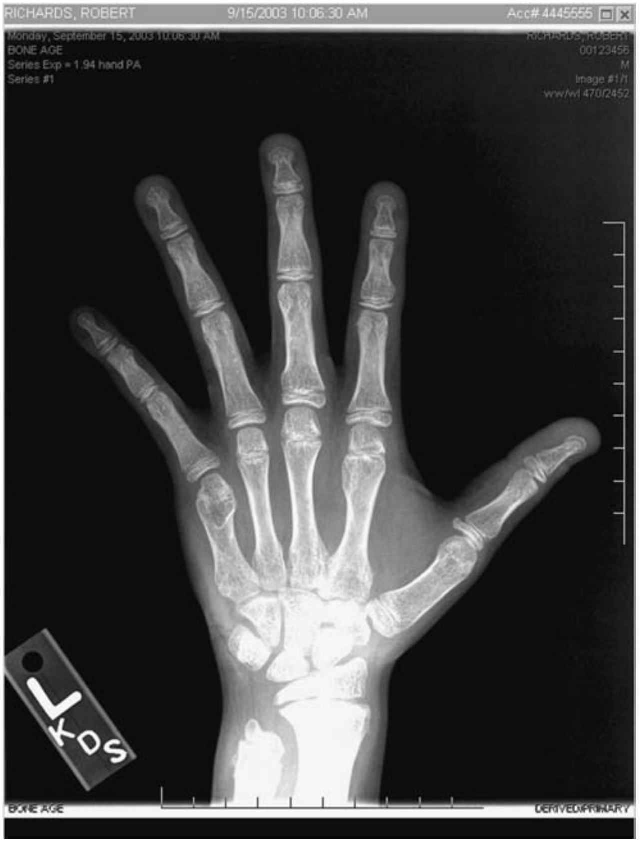


Figure 3. An image associated with the simulated order shown in Table 3. Note the inclusion of the accession number on this image that is required to associate the image with the report.

will be a standard and will facilitate information communication among healthcare venues. Recent research suggests that the United States could realize a savings potential of $78 billion annually if a seamless, fully interoperable healthcare information exchange could be established among key stakeholders in the healthcare delivery system (9).

SELECTING AN RIS

The KLAS company, founded in 1996, is a research and consulting firm specializing in monitoring and reporting the performance of Healthcare's Information Technology's (HIT) vendors. The comprehensive reports they produced are valuable for comparing the vendors they review. Buyers should be aware that not all vendors are included in the reports, but the major ones are represented. The Comprehensive Radiology Information Systems Report, Serving Large, Community, and Ambulatory Facilities was released in January, 2005 by KLAS (15). In the report, eight RIS vendor products were represented and were reviewed by interviewing users of the systems. In an addendum, seven other vendors were presented whose products did not yet meet the KLAS standards for statistical confidence in order to be compared with other vendors in the main body of the report. Vendors were all allowed to prepare overviews and their perceptions of their products. Of course, these overviews stress the strengths of a vendor's product. Performance measurements of the KLAS traditional 40 indicators (Table 6), technology overviews, client win/loss and pricing provide the bases of the provider experience. The KLAS report should be part of an institution's decision-making process when a report is available, and in this case the report is available and very timely.

This report focused on large ($>$ 200 bed) and ambulatory (free standing clinics and imaging centers) facilities. The large facilities were asked additional questions. The larger institutions were asked why a vendor was selected and why a vendor was NOT selected based on the following six criteria: Functionality, Cost, Relationship with Vendor, References/Site Visits/Technology, and Integration/Interfacing. The most cited reason as to why a vendor was selected was their ability to integrate or interface. The most cited reason as to why a vendor was not selected was their perceived limited functionality.

Survey participants from the large institutions were asked questions regarding their RIS and its; (*1*) benefits; (*2*) functional strength of the reporting module; and (*3*) the RIS–PACS integration. The top benefit, reported by $>$ 50% of the survey respondents, was that of More Efficient/Better Workflow. The number two and three benefits identified were Interface–Integration and Manage Department Better. The functional strength of the reporting module on a scale of 1–5 (1 + weak and 5 = strong) was rated with an average of 3.5, with the highest score of 4, which indicates that there are a lot of users who are not totally satisfied with their reporting module. Forty-seven percent of the survey respondents indicated that they have plans to move to an integrated RIS/PACS solution and 81% of these reported that Radiologists and Clinicians were mostly driving this integration. This indicates that there are a large number of institutions without an integrated solution.

THE RIS OF THE FUTURE

In the future, it is likely that speech recognition systems will be incorporated into an RIS solution. Throughout this article, the two systems have been shown as separate entities and indeed, that is the most common incarnation at this time. Speech recognition has become a more important part of the radiology workflow as researchers demonstrate the potential for improving report turnaround time and decreasing costs when compared to systems with manual transcriptionists (16–19). Unfortunately, the increase in report turnaround time does not guarantee efficient personnel utilization. Although the report can be available for the physician immediately after the radiologist verified the dictation, the responsibility for editing the report falls on the radiologist and may make the time required for the process longer than with a transcriptionist performing the typing and editing. Because of this, speech recognition is not eagerly embraced by many radiologists, and therefore

Table 5. Example Portion of a Radiology Report for the Simulated order in Table 3[a]

SHANDS DIAGNOSTIC EXAM: BONE AGE
NAME: RICHARDS, ROBERT
EXAM DATE: 09/15/2003 LOC: XRY- SEX: M
MRN: 00123456 DOB: 09/01/1991
ORDERING MD: WOOD , FOREST
ORD. SERVICE: ORT ORD. LOC: ORT TECH:JJJ

REASON: BONE AGE RT WRIST PAIN S/P LT

EXAM DATE: 09/15/2003 11:18
ACCESSION#: 44455555
UXR 3000 BONE AGE
ICD9 CODES: 719.43380.8
FINDINGS:
Clinical History: 12 year-old male with multiple hereditary exostoses,
limb length discrepancies, and left wrist pain.

Findings: The bone is within 2 standard deviations of the chronological age.
There are multiple exostoses in the right hand.

IMPRESSION:
1. Normal bone age.
2. Findings consistent with multiple hereditary exostoses.

Dictated By: MARK FISCHER, MD
Dictated Date: SEPT 15, 2003 1:54P

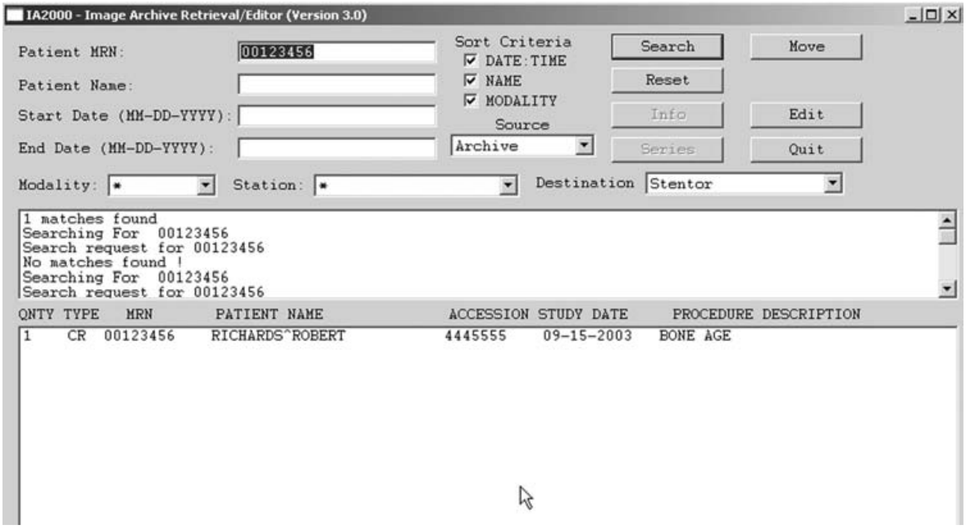
Interpreted By:
WILLIAM JENNINGS, MD MARK FISCHER, MD
This study was personally Resident Radiologist
reviewed by me and I agree
with the report.

---

[a] Note the accession number which associates the report with the study images.

acceptance has been slower than with other information systems used in radiology.

As speech recognition gains more widespread acceptance, researchers are looking into improving the diagnostic reporting methodology by structuring the format of the report. The DICOM Structured Report is a definition for the transmission and storage of clinical documents. A DICOM Structured Report can contain the traditional text report in addition to structured information and links to key images. If you return to Fig. 2, you will notice that the RIS and HIS typically communicate using HL7, not DICOM, so either HL7 will have to change to support these reports or the RIS will have to support DICOM. There is an evolving standard for the HL7 Clinical Document Architecture (CDA) that supports text, images, sounds, and other multimedia content. The DICOM working group 20 (the Integration of Imaging and Information systems Group) and the HL7



Patient MRN: 00123456
Patient Name:
Start Date (MM-DD-YYYY):
End Date (MM-DD-YYYY):
Sort Criteria
☑ DATE:TIME
☑ NAME
☑ MODALITY
Source: Archive
Modality: *
Station: *
Destination: Stentor

```
1 matches found
Searching For  00123456
Search request for 00123456
No matches found !
Searching For  00123456
Search request for 00123456
```

```
QNTY TYPE   MRN       PATIENT NAME              ACCESSION  STUDY DATE    PROCEDURE DESCRIPTION
1    CR   00123456    RICHARDS^ROBERT            4445555    09-15-2003   BONE AGE
```

Figure 4. An archive query associated with the simulated order shown in Table 3.

Table 6. The 40 Success Indicators Used by KLAS in Their Information Systems Reports

10 Product/Technology Indicators[a]
- Enterprise Commitment to Technology
- Product Works as Promoted
- Product Quality Rating
- Quality of Releases and Updates
- Quality of Interface Services
- Interfaces Met Deadlines
- Quality of Custom Work
- Technology Easy to Implement and Support
- Response Times
- Third-Party Product Works with Vendor Product

10 Service Indicators[a]
- Proactive Service
- Real Problem Resolution
- Quality of Training
- Quality of Implementation
- Implementation on Time
- Implementation within Budget/Cost
- Quality of Implementation Staff
- Quality of Documentation
- Quality of Telephone/Web Support
- Product Errors Corrected Quickly

8 Success Indicators[a]
- Worth the Effort
- Lived Up to Expectations
- Vendor is Improving
- Money's Worth□□
- Vendor Executives Interested in You
- Good Job Selling
- Contracting Experience
- Helps Your Job Performance

12 Business Indicators[b]
- Implemented in the Last 3 Years)
- Core Part of IS Plan
- Would You Buy It Again
- Avoids Nickel-and-Diming
- Keeps All Promises
- A Fair Contract
- Contract is Complete (No Omissions
- Timely Enhancement Releases
- Support Costs as Expected
- Would You Recommended to a Friend/Peer
- Ranked Client's Best Vendor
- Ranked Client's Best or Second Best Vendor

[a] Rating 1–9, where 1=poor and 9=excellent.
[b] Rating is a yes or no.

Imaging Integration Special Interest Group (IISIG) are working on harmonizing the existing standards (20,21). Future RIS implementations will most certainly support the DICOM structured report or its HL7 CDA translation.

In addition to supporting the traditional terminal or remote computer, the RIS of the future will be required to support a web interface as well as supporting handheld devices. A radiologist with a handheld Personal Digital Assistant (PDA) can currently access e-mail, the internet, digital media, and documents. With wireless networking available in many hospitals, these devices can support the exchange of information between ordering physician and radiologist including results, they can support exchange of information between radiologist and technologist including protocol selection, they can access history and lab reports for the patient, they can provide a worklist of current studies, and can even display low resolution images. The ability of the PDA to support these functions is only limited by the ability of the RIS to provide PDA support (22).

The RIS of the future may store PACS images, must be able to interface with the EMR, and should provide structured reporting and support for PDAs. As institutions plan for purchasing new systems or updating the old ones, future functionality of the systems must be considered and be part of the request for proposal or bid process. In the past, institutions were able to select an RIS based on individual preferences without consideration of interface issues. As the national health records structure evolves, all institutions must be in a position to interface to an EMR using well developed standards.

BIBLIOGRAPHY

1. Honeyman-Buck JC. PACS Adoption. Seminars in Roentgenology 2003; July, 38(3):256–269.
2. Thrall JH. Reinventing Radiology in the Digital Age. Part I. The All-Digital Department. Radiology 236(2):382–385.
3. Aunt Minnie Buyer's guide for Radiology Information Systems. Aunt Minnie Web Site. Available at http://www.auntminnie.com. Accessed 2005; Aug 2.
4. 2005 National Patient Safety Goals. JCAHO web site. Available at http://www.jcaho.org. Accessed 2005; Aug 11.
5. HL7. HL7 web site. Available at http://www.hl7.org. Accessed 2005; Aug 11.
6. DICOM. DICOM web site. Available at http://medical.nema.org. Accessed 2005; Aug 11.
7. Carrino JA. Digital Imaging Overview. Seminars Roentgenol 2003;38(3):200–215.
8. HIPAA. HIPAA web site. Available at http://www.hipaa.org. Accessed on 2005; Aug 20.
9. Middleton B, Hammond WE, Brennan PF, Cooper GF. Accelerating U.S. HER Adoption: How to Get there From Here, Recommendations Based on the 2004 ACMI Retreat. JAMIA 2005;12(1):13–19.
10. Makoul G, Curry RH, Tang PC. The Use of Electronic Medical Records: Communication Patterns in Outpatient Encounters. JAMIA 2001;8(6):610–615.
11. Stead WW, Kelly BJ, Kolodner RM. Achievable Steps Toward building a National Health Information Infrastructure in the United States. JAMIA 2005;12(2):113–120.
12. Berner ES, Detmer DE, Simborg D. Will the wave Finally break? A brief View of the Adoption of Electronic Medical Recoreds in the United States. JAMIA 2005;12(1):3–7.
13. IHE. The IHE web site. Available at http://www.ihe.net. Accessed on 2005; Aug 20.
14. Channin DS. Driving Market-driven Engineering. Radiology 2003;229:311–313.
15. Comprehensive Radiology Information Systems Report, KLAS Enterprises. Available at www.healthcomputing.com: Accessed 2005.
16. Langer SG. Impact of Tightly coupled PACS/Speech Recognition in Report Turnaround Time in the Radiology Department. J Digital Imaging 2002;15(Supp 1):234–236.
17. Gutierrez AJ, Mullins ME, Novelline RA. Impact of PACS and Voice-Recognition Reporting on the Education or Radiology Residents. J Digital Imaging 2005;18(2):100–108.

18. Zick RG, Olsen J. Voice Recognition software Versus a Traditional Transcription Service for Physician Charting in the ED. Am J Emerg Med 2001;19:295–298.
19. Bramson RT, Bramson RA. Overcoming Obstacles to Work-Changing Technology Such As PACS and Voice Recognition. AJR 2005;184:1727–1730.
20. Hussein R, Schroeter A, Meinzer H-P. DICOM Structured Reporting.RadioGraphics 2004;24(3):891–896.
21. Dolin RH, et al. The HL7 Clinical Document Architecture. JAMIA 2001;8:552–579.
22. Flanders AE, Wiggins RH, Gozum ME. Handheld Computers Radiol. RadioGraphics 2003;23:1035–1047.

See also EQUIPMENT ACQUISITION; PICTURE ARCHIVING AND COMMUNICATION SYSTEMS; RADIOTHERAPY TREATMENT PLANNING, OPTIMIZATION OF; TELERADIOLOGY.

RADIOLOGY, PHANTOM MATERIALS.

See PHANTOM MATERIALS IN RADIOLOGY.

RADIOMETRY.

See THERMOGRAPHY.

RADIONUCLIDE PRODUCTION AND RADIOACTIVE DECAY

SILVIA S. JURISSON
WILLIAM MILLER
J. DAVID ROBERTSON
University of Missouri
Columbia, Missouri

INTRODUCTION

Radiopharmaceuticals, drugs containing radioactive atoms, are used for diagnostic imaging or therapeutic applications in nuclear medicine, depending on their radioactive emissions. Penetrating radiations (gamma rays or annihilation photons from positron emission) are used for diagnostic applications with gamma cameras, single photon emission computed tomography (SPECT), or positron emission tomography (PET) instrumentation. Diagnostic imaging with gamma emitters became a mainstay of nuclear medicine with the advent of the molybdenum-99/technetium-99m (^{99}Mo/^{99m}Tc) generator (the Brookhaven generator), which was developed by Richards in ~ 1961 at Brookhaven National Laboratory (1). This generator made the short half-lived ^{99m}Tc (6.01 h; 140 keV γ-ray) readily available to nuclear medicine departments around the world, and not just at the site of ^{99m}Tc production. Today, ^{99m}Tc accounts for >80% of the diagnostic scans performed in nuclear medicine departments in the United States, with a variety of U.S. Food and Drug Administration (FDA) approved agents for functional imaging of the heart, liver, gallbladder, kidneys, brain, and so on (2). Particle emitters, such as alpha, beta, and Auger electron emitters, are used for targeted radiotherapy since their decay energy is deposited over a very short range in tissue. Radiotherapeutic applications in humans began in the late 1930s with the beta emitters radioiodine (^{128}I, ^{131}I), radiophosphorus (^{32}P) and radiostrontium (^{89}Sr) for cancer treatment (3). Radioiodine (^{131}I iodide) was the first, and continues to be the only, true "magic bullet" through its specific and selective uptake in thyroid. It is used to treat hyperthyroidism and thyroid cancer.

Radionuclides for medical applications are produced either at nuclear reactors or accelerators. The "Availability of Radioactive Isotopes" was first announced from the headquarters of the Manhattan Project (Washington, DC) in Science in 1946 (4), and today medical isotope availability remains an important issue for the nuclear medicine community. The selection of radionuclides for use in radiopharmaceuticals is dependent on their decay properties (half-life, emissions, energies of emissions, dose rates) and their availability (production and cost).

Radionuclides suitable for diagnostic imaging are gamma emitters (with no or minimal accompanying particle emissions) such as ^{99m}Tc and positron emitters (annihilation photons) such as ^{18}F. The half-lives should be as short as possible and still allow preparation (synthesis and purification) of the radiopharmaceutical, administration of the agent to the patient, and the diagnostic imaging procedure. Typically, half-lives on the order of minutes to a week are used, with hours to a day considered optimal for most applications. Gamma energies in the range of 80–300 keV are considered good, although higher energy (e.g., ^{131}I at 364.5 keV) are used, and 100–200 keV is considered optimal. The PET instrumentation is designed for the two 511 keV annihilation photons.

The radionuclides used or under investigation for radiotherapeutic applications are particle emitters, with most of the efforts focusing on the beta emitters (e.g., ^{153}Sm, ^{90}Y, ^{186}Re, ^{188}Re, ^{177}Lu, ^{149}Pm, ^{166}Ho, ^{105}Rh, ^{199}Au) and some on alpha emitters (e.g., ^{212}Bi, ^{213}Bi, ^{225}Ac). In the case of radiotherapy, the half-life of the radionuclide should match the biological half-life for the radiopharmaceutical delivery to its *in vivo* target site (i.e., tumor), typically < 1 day – 1 week. The optimum particle and particle energy remains under investigation, and it is not clear that a higher particle energy translates into a more successful treatment (i.e., radiation dose to nontarget organs will limit the dose allowed for administration). Suitable accompanying gamma emissions will allow the *in vivo* tracking of the radiotherapeutic dose.

Two important factors for the use of radionuclides in nuclear medicine are thus radioactive decay and radionuclide production, both of which are discussed in detail below. Radionuclides used in nuclear medicine applications are used as examples in discussing these topics. Radioactive decay includes the types of decay (e.g., alpha, beta, electron capture, positron, gamma) with focus on those modes with nuclear medicine applications and rates of radioactive decay (e.g., units of radioactivity, exponential decay, activity–mass relationships, parent–daughter equilibria, medical generators).

TYPES OF RADIOACTIVE DECAY

A nuclide is considered to be radioactive if its nuclear configuration (number of neutrons and protons in the nucleus) is unstable. Radioactive decay is then a means for the unstable nucleus to achieve a more stable nuclear configuration although it may or may not form a stable nuclide following a single decay. There are several modes of decay possible for a radioactive nuclide including alpha emission, beta emission, positron emission, electron capture decay and/or gamma emission. An alpha (α) particle is a helium nucleus ($^{4}_{2}\mathrm{He}^{2+}$) and is generally a mode of decay available in heavy nuclei (i.e., $Z \geq 83$). Some lanthanides can show α-decay and heavy nuclides can also undergo decay by β^-, EC, and spontaneous fission. The following equation gives an example of alpha decay:

$$^{213}_{83}\mathrm{Bi} \xrightarrow{45.59\,\mathrm{min}} {}^{209}_{81}\mathrm{Tl} + {}^{4}_{2}\mathrm{He}^{2+} + \gamma + Q$$

where Q is the energy released during the decay process and γ is the 727 keV gamma photon emitted in 11.8% of the decays. Most of the decay energy in this process is in the form of the kinetic energy of the α-particle, which is rapidly deposited in matter (e.g., tissue) because of its relatively high charge and mass.

Beta (β^-) decay occurs when the nucleus has a proton/neutron ratio that is too low relative to the proton/neutron ratio in the stable nuclei of that element, and during this process a neutron is converted into a proton resulting in a nuclide that is one atomic number (Z) higher than its parent. An example of beta decay is shown below,

$$^{188}_{75}\mathrm{Re} \xrightarrow{16.98\,\mathrm{h}} {}^{188}_{76}\mathrm{Os} + {}^{0}_{-1}\beta + \tilde{\nu}_e + \gamma + Q$$

where Q is the energy released, $\tilde{\nu}_e$ is an antineutrino, and γ is the 155 keV gamma photon emitted in 15% of the decays. These neutron-rich radionuclides are produced in nuclear reactors as will be discussed in the section Radionuclear Production.

When a radioactive nucleus has a proton/neutron ratio that is too high relative to that of the stable nuclei of the same element, two modes of decay are possible (positron emission and/or electron capture decay), with both converting a proton into a neutron resulting in a nuclide that is one atomic number lower than its parent. Electron capture decay arises from the overlap of nucleon orbitals with electron orbitals. It is a result of the "weak force", which is very short ranged. During electron capture decay (ε), essentially an inner-shell electron (usually K-shell) is incorporated into the nucleus converting a proton into a neutron. This process creates an orbital vacancy and results in a cascade of outer-shell electrons filling the lower energy inner-shell vacancies, with the excess energy released as X-rays and/or Auger electrons. The following equation shows the electron capture process:

$$^{201}_{81}\mathrm{Tl} + {}^{0}_{-1}\mathrm{e}^- \xrightarrow{73.1\,\mathrm{h}} {}^{201}_{80}\mathrm{Hg} + \nu_\mathrm{e} + Q + \mathrm{Auger\,e^-/X\,rays}$$

where Q is the energy released and ν_e is a neutrino. The 80 keV X rays emitted are used for myocardial imaging in stress–rest tests performed in nuclear medicine departments. Positron emission is only possible when Q is > 1.022 MeV, the energy equivalent of the mass of two electrons. During positron decay, a proton is converted into a neutron with simultaneous emission of a positron (β^+), which is a positive electron. Since a neutron has greater mass than a proton, the energy equivalent of two electrons (one to convert a proton to a neutron in the nucleus, and one emitted as a positron) must be available for this decay mode to be possible. Practically, positron emission does not occur unless Q is ≥ 2 MeV. An example of the positron decay process is shown below.

$$^{18}_{9}\mathrm{F} \xrightarrow{110\,\mathrm{min}} {}^{18}_{8}\mathrm{O} + {}^{0}_{1}\beta^+ + \nu_\mathrm{e} + Q$$
$$^{0}_{1}\mathrm{e}^+ + {}^{0}_{-1}\mathrm{e}^- \longrightarrow 2\text{–}511\,\mathrm{keV\ annihilation\ photons}$$

where Q is the decay energy (including the maximum positron energy plus 1.022 MeV), ν_e is a neutrino, and the two 0.511 MeV photons result from positron annihilation (often called annihilation photons) and are emitted 180° opposite each other. They are the basis of PET imaging. The ^{18}F radiolabeled fluorodeoxyglucose (^{18}F-FDG) is used in nuclear medicine departments for imaging glucose metabolism, such as found in growing tumors. In some cases, both electron capture and positron emission can occur. These proton-rich radionuclides are produced by accelerators which is discussed below.

Gamma (γ) emission can accompany any other decay process (i.e., alpha, beta, electron capture, positron decay) or it can occur without any particle emission. In the latter case, it is called isomeric transition (IT) and occurs when a metastable radioisotope [higher energy excited state of a nucleus with a measurable lifetime (≥ns)] decays to the ground state (lower energy state of the nucleus). Isomeric transition is accompanied by energy release without any other change occurring in the nucleus (i.e., the number of protons and neutrons in the nucleus does not change during IT). The decay of $^{99\mathrm{m}}$Tc is an example of an isomeric transition.

$$^{99\mathrm{m}}_{43}\mathrm{Tc} \xrightarrow{6.01\,\mathrm{h}} {}^{99}_{43}\mathrm{Tc} + \gamma + Q$$

where Q is the decay energy and γ is the 140 keV photon released from the nucleus. The $^{99\mathrm{m}}$Tc in various chemical formulations is used routinely for diagnostic imaging of a variety of diseases and/or organ functions.

Whenever a gamma photon is released from the nucleus, the emission of conversion electrons is also possible. A conversion electron is the emission of an electron that has the energy of the gamma photon minus the electronic atomic binding energy, and it is emitted in place of the gamma photon. The probability of conversion electron emission rather than gamma emission increases as the energy of the gamma photon decreases and it increases with increasing nuclear charge. For example, the 140 keV gamma photon of $^{99\mathrm{m}}$Tc is emitted with 89% abundance; conversion electrons, rather than gamma photons, are emitted 11% of the time when $^{99\mathrm{m}}$Tc decays to ^{99}Tc.

There are radionuclides in which more than one type of radioactive decay occurs. An example is the decay of ^{64}Cu, which undergoes beta decay (39% of the time), positron emission decay (19% of the time), and electron capture decay (42% of the time) with a weighted average half-life of 12.7 h (the specific half-lives are shown below). This radionuclide has both diagnostic imaging (β^+ emission via the annihilation photons) and radiotherapy ($\beta^{-/+}$ and Auger electrons) applications.

$$^{64}_{29}\mathrm{Cu} \xrightarrow{33.4\,\mathrm{h}} {}^{64}_{30}\mathrm{Zn} + {}^{0}_{-1}\beta^{-} + \overline{\nu}_{\mathrm{e}} + \gamma + Q$$

$$^{64}_{29}\mathrm{Cu} \xrightarrow{66.8\,\mathrm{h}} {}^{64}_{28}\mathrm{Ni} + {}^{0}_{+1}\beta^{+} + \nu_{\mathrm{e}} + Q + 2 - 0.511\,\mathrm{MeV\,photons}$$

$$^{64}_{29}\mathrm{Cu} + \mathrm{e}^{-} \xrightarrow{30.2\,\mathrm{h}} {}^{64}_{28}\mathrm{Ni} + \nu_{\mathrm{e}} + Q + \mathrm{Auger\,e}^{-}/\mathrm{X\,rays}$$

All of the above modes of radioactive decay have been utilized or are under investigation for utilization in the development of radiopharmaceuticals for nuclear medicine applications (either diagnostic imaging or radiotherapy).

RATES OF RADIOACTIVE DECAY

Radioactive decay is a random process where the number of nuclei in an isotopically pure sample that decay during a given time period is proportional to the number of nuclei (N) present in the sample. If radioactivity is defined as a measure of the rate at which the radioactive nuclei disintegrate, then the number of disintegrations per unit time (dN/dt) from a sample is given by the following expression where λ is the decay constant (the probability of decay per unit time) and negative sign indicates the loss of the radionuclide through the disintegration. This results in a radioactive decay process that follows a first-order rate law:

$$\frac{dN}{dt} = -\lambda N$$

From this definition, the number of disintegrations per unit time (dN/dt) is also known as the activity. For an isotopically pure sample, the number of nuclei N can be calculated knowing Avogadro's number (N_{A}), the mass number of the isotope (A) in daltons or grams per mole and the sample's weight (wt) in grams:

$$\mathrm{Activity} = \lambda N = \lambda \frac{N_{\mathrm{A}}}{A}\mathrm{wt}$$

As with all measurements, several systems of units can be used to define the amount of radioactivity in a particular sample. And, as with many measuring systems, there are both "traditional" units and new SI units (International System of Units). The original units are referenced to the curie (Ci), named in honor of Madam Curie. A curie is 3.7×10^{10} (37 billion) radioactive disintegrations per second, or approximately the amount of radioactivity in a gram of radium, one of the natural radioactive elements with which Madam Curie did her research. For many applications of radiation in medicine or industry, the curie is a relatively large quantity, and so the units of mCi (1/1000th) and microcurie (μCi) (1/1,000,000th) are utilized. On the other hand, a nuclear reactor contains millions of curies of radioactivity and units of kilocurie (kCi) (1000) and megacurie (MCi) (1,000,000) are sometimes used. Although the unit of the curie is being supplemented with the newer SI unit, it is still very much in common use.

The SI unit of radioactivity, the becquerel (Bq), is named after Henri Becquerel, the discoverer of radioactivity. It is defined as one disintegration per second. Thus it is much smaller than the Ci. Multiples of becquerel are the kilobecquerel (kBq) (1000), megabecquerels (MBq) (1,000,000), and so on. These units are summarized in Table 1.

Table 1. Units of Activity

Curies	Becquerels or disintegrations s^{-1}	Becquerels or disintegrations s^{-1}	Curies
1 MCi	3.7×10^{16}	1 Bq	2.7×10^{-11}
1 kCi	3.7×10^{13}	1 kBq	2.7×10^{-8}
1 Ci	3.7×10^{10}	1 MBq	2.7×10^{-5}
1 mCi	3.7×10^{7}	1 GBq	0.027
1 μCi	3.7×10^{4}		

The rate at which radioactive decay occurs can be defined as the half-life, or the amount of time that it takes for one-half of the radiation to decay. Unfortunately, two half-lives do not eliminate it (i.e., one-half decaying in one half-life and then the second one-half decaying in a second half-life), but rather the half-life is always referring to how much is left at any given point in time. Thus, one half-life reduces the radioactivity to one-half or 50%; two half-lives to one-half of one-half, one-fourth or 25%; three half-lives to one-eighth or 12.5%; and so on This leads to a kinetic model that is described by an exponential function (see Fig. 1), which is the solution to the previous equation:

$$N(t) = N_0 e^{-\lambda t}$$

The half-life is related to the physical decay constant (λ) by the simple expression:

$$t_{1/2} = \frac{0.693}{\lambda}$$

As a radionuclide decays, the number of nuclei (N) changes, and the amount of radioactivity present at any given time "t" remaining from an initial amount A_{o} (expressed in either Bq or Ci) can be given by any of the following:

$$A(t) = A_0 e^{-\lambda t}$$

$$A(t) = A_0 e^{-(0.693/t_{1/2})t}$$

$$A(t) = A_0 \left(\frac{1}{2}\right)^{t/t_{1/2}}$$

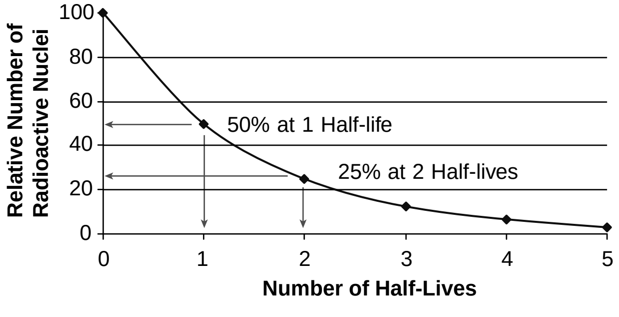


Figure 1. Relative number of radioactive nuclei remaining after various half-lives of decay.

Radioactive nuclei have half-lives that range from fractions of a second to billions of years. The radioactive material introduced into the body for medical purposes would typically have half-lives on the order of a few hours to a few days, so that it decays away to negligible levels in a relatively short amount of time. At the other extreme, some of the naturally radioactive nuclei in our environment have half-lives of billions of years and the reason they are still present is that they have not had enough time to decay to negligible levels since the earth was formed. An example of such a radionuclide is ^{40}K, which makes up 0.0117% of naturally occurring potassium and has a half-life of 1.27×10^9 years.

Another interesting case of radioactive decay involves parent–daughter relationships in which a radioactive parent decays to a radioactive daughter. This process is utilized extensively in the medical profession to provide a long-lived source for a short-lived radioisotope in what is known as a generator. The daughter can be obtained from the generator by “milking” it, taking advantage of the differences in chemistry between the parent and daughter elements to extract the daughter off of an ion column, which holds the parent.

The use of $^{99\mathrm{m}}$Tc, which has a 6 h half-life, is a case in point. Within a 24 h period it has decayed through four half-lives and is only one-sixteenth of its original value. Thus, $^{99\mathrm{m}}$Tc would have to be made in a nuclear facility and shipped daily to meet hospital needs. Fortunately, the daughter $^{99\mathrm{m}}$Tc is produced by the decay of a parent isotope ^{99}Mo, which has a 2.75 day half-life. A supply of $^{99\mathrm{m}}$Tc can thus be obtained over a period of ~1 week from the more slowly decaying ^{99}Mo.

$$ {}^{99}_{42}\mathrm{Mo} \xrightarrow{2.75\,\mathrm{days}} {}^{99\mathrm{m}}_{43}\mathrm{Tc} + \beta^- + \bar{\nu}_\mathrm{e} \xrightarrow{6\,\mathrm{h}} {}^{99}_{43}\mathrm{Tc} + \gamma $$

The kinetics of the quantity of a daughter isotope (d) available from a parent isotope (p) can be readily solved using a first-order differential equation resulting in a straightforward algebraic expression:

$$ \frac{A_\mathrm{daughter}}{A_\mathrm{parent}} = \frac{\lambda_\mathrm{d}}{\lambda_\mathrm{d} - \lambda_\mathrm{p}} \left(1 - e^{-(\lambda_\mathrm{d} - \lambda_\mathrm{p})t}\right) $$

For most practical cases where the parent is longer lived than the daughter, the daughter activity reaches a value close to the activity of the parent after approximately four half-lives. Thus, in an undisturbed generator that has been allowed to reach equilibrium, the quantity of daughter present (in Ci or Bq) is approximately equal to the activity of the parent at that time. Once the daughter has been extracted (or milked), it immediately begins building up again to a new equilibrium value as shown in Fig. 2.

Also of importance is the time it takes for the daughter to reach its maximum value, which is determined by the decay constants of the two isotopes involved:

$$ t_\mathrm{max} = \frac{\ln\left(\frac{\lambda_\mathrm{d}}{\lambda_\mathrm{p}}\right)}{\lambda_\mathrm{d} - \lambda_\mathrm{p}} $$

Again assuming a typical case where the parent is longer lived than the daughter, this is largely determined by the half-life of the daughter. Thus, the daughter will reach ~ 50% of the equilibrium value in one half-live, 75% of the equilibrium value in two halve-lives, and so on. The time between subsequent extractions of a daughter from a generator is set by this regeneration time. For the ^{99}Mo → $^{99\mathrm{m}}$Tc system, the regeneration time needed to reach the maximum amount of daughter $^{99\mathrm{m}}$Tc is very close to 24 h, a very convenient amount of time for routine hospital procedures and is just a fortuitous consequence of the half-lives of the two isotopes involved.

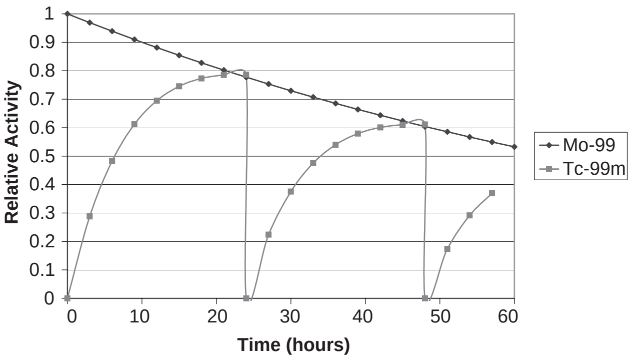


Figure 2. Shows build-up of $^{99\mathrm{m}}$Tc activity from ^{99}Mo initially and following elution.

Other examples of parent–daughter generators that are currently used or are under development for future use are ^{188}W/^{188}Re generators, ^{90}Sr/^{90}Y generators, ^{82}Sr/^{82}Rb direct infusion generators, ^{62}Zn/^{62}Cu generators, ^{224}Ra/^{212}Bi or ^{224}Ra/^{212}Pb generators, which also are a source of ^{212}Bi, the daughter product of ^{212}Pb, ^{225}Ac/^{213}Bi generators and ^{68}Ge/^{68}Ga generators. (For more information on generators see Ref. 5 and references cited therein.)

Specifically,

$$ {}^{82}_{38}\mathrm{Sr} \xrightarrow{25.4\,\mathrm{days}} {}^{82}_{37}\mathrm{Rb} \xrightarrow{1.6\,\mathrm{min}} {}^{82}_{36}\mathrm{Kr} + \beta^+ $$

$$ {}^{68}_{32}\mathrm{Ge} \xrightarrow{271\,\mathrm{days}} {}^{68}_{31}\mathrm{Ga} \xrightarrow{68\,\mathrm{min}} {}^{68}_{30}\mathrm{Zn} + \beta^+ $$

are two generators that can produce short lived PET isotopes, with ^{82}Rb being used for PET studies of myocardial function (rubidium acts as a potassium ion mimic).

The following three are examples of generators for beta or alpha emitting radioisotopes as possible therapeutic agents:

$$ {}^{188}_{74}\mathrm{W} \xrightarrow{69.4\,\mathrm{days}} {}^{188}_{75}\mathrm{Re} \xrightarrow{17\,\mathrm{h}} {}^{188}_{76}\mathrm{Zn} + \beta^- $$

$$ {}^{166}_{66}\mathrm{Dy} \xrightarrow{3.4\,\mathrm{days}} {}^{166}_{67}\mathrm{Ho} \xrightarrow{26.8\,\mathrm{h}} {}^{166}_{68}\mathrm{Er} + \beta^- $$

$$ {}^{212}_{82}\mathrm{Pb} \xrightarrow{10.64\,\mathrm{h}} {}^{212}_{83}\mathrm{Bi} \xrightarrow{60.6\,\mathrm{min}} {}^{208}_{81}\mathrm{Tl} + \alpha $$

RADIONUCLIDE PRODUCTION

The radionuclides used in medicine and the life sciences are produced through neutron and charged particle induced nuclear reactions. As with chemical reactions, the yield of the product (radioisotope) of interest will depend on the nuclear reaction employed, the energetics of the reaction, the probability of competing reaction pathways, and the ability to separate the desired product from the reactants (target) and any additional nuclear reaction products. For a complete overview of the practice and theory of nuclear reactions, the interested reader is referred to an introductory text on nuclear and radiochemistry (6–9).

Proton-rich radionuclides that decay by positron emission and/or electron capture are produced at cyclotrons and accelerators through charged particle induced nuclear reactions. For example, the production of the commonly used PET radionuclide ^{18}F through the bombardment of a target of ^{18}O-enriched water with high energy protons is produced by the following nuclear reaction:

$$^{1}_{1}\mathrm{p} + {}^{18}_{8}\mathrm{O} \rightarrow {}^{18}_{9}\mathrm{F} + {}^{1}_{0}\mathrm{n} + Q$$

that is written in a short-hand notation as ^{18}O(p,n)^{18}F. Note that, as in radioactive decay, charge (number of protons), mass number, and total energy are conserved in the nuclear reaction. The Q value for the reaction is the difference in energy (mass) between the reactants and products and is readily calculated from the measured mass excess (Δ) values for the reactants and products

$$Q = \sum \Delta_{\text{reactants}} - \sum \Delta_{\text{products}}$$

If Q is < 0, then the reaction is endoergic and energy must be supplied to the reaction (through the kinetic energy of the projectile) and if Q is >0, the reaction is exoergic and energy is released in the nuclear reaction. For the ^{18}O(p,n)^{18}F reaction,

$$\begin{aligned} Q &= \Delta^{1}_{1}\mathrm{p} + \Delta^{18}_{8}\mathrm{O} - (\Delta^{1}_{0}\mathrm{n} + \Delta^{18}_{9}\mathrm{F}) \\ &= -0.782 + 7.289 - (8.071 + 0.873) \\ &= -2.44\ \mathrm{MeV} \end{aligned}$$

and the proton must supply 2.44 MeV of energy to the reaction. In practice, the actual proton bombarding energy is higher than the Q value because (1) not all of the kinetic energy of the proton is available for the nuclear reaction because of momentum conservation in the collision and (2) the probability of the reaction is typically very low at the threshold energy (Q). Typical conditions for the ^{18}O(p,n)^{18}F reaction on a water target with an in-house cyclotron are 16.5 MeV at 100 μA yielding 3–4 Ci/h. Even in those cases when the Q value for the reaction is >0, the incoming charged particle must have sufficient kinetic energy to overcome the coulomb or charge barrier between the projectile and the target nucleus and any angular momentum barrier that might exist for the reaction.

The probability that the projectile will strike the target nucleus and produce the radioisotope of interest is quantified with the reaction cross-section (σ) that has the dimensions of area. While quite sophisticated models exist for predicting reaction cross-sections based upon the underlying nuclear physics, the simple physical analogy for the cross-section is the area that the target nucleus presents to the incoming beam of projectiles. The SI unit for cross-section is m^2, but the more common unit is a barn (b); one barn is equal to 10^{-24} cm^2 (or 10^{-28} m^2). The magnitude of a barn can be understood from the fact that a target nucleus with mass number 100 has a radius on the order of 6.5×10^{-15} m and a "cross-sectional area" of 1.3×10^{-28} m^2 or 1.3 b. In the simplest case when a charged particle beam is bombarding a target that is "thin" enough so that the beam does not lose any appreciable energy in passing through the target, then the production rate (R) for the radioisotope of interest is equal to

$$R = n \times I \times t \times \sigma$$

where n is the target nuclei density (nuclei cm^{-3}), I is the number of incident particles per unit time (particles s^{-1}), t is the target thickness (cm), and σ is the reaction cross-section (cm^2). In most cases, radioisotope production is performed using a thick target and an estimate of the reaction production rate takes into account the variation in the reaction cross-section with projectile energy, since the charged-particle beam loses energy as it passes through or stops in the thick target. Charged particle induced reactions used to produce medical radioisotopes have maximum cross-sections on the order of millibarns.

Neutron-rich radioisotopes that decay by negatron or β^- emission are produced primarily through neutron-induced nuclear reactions at nuclear reactors. The most commonly used reactions are direct production through single or double neutron capture, direct production through neutron induced fission, and indirect production through neutron capture followed by radioactive decay. An example of direct neutron capture is the production of the ^{166}Ho through the irradiation of ^{165}Ho in a nuclear reactor by the following reaction:

$$^{1}_{0}\mathrm{n} + {}^{165}_{67}\mathrm{Ho} \rightarrow {}^{166}_{67}\mathrm{Ho} + Q$$

Like all neutron capture reactions used to produce medical radioisotopes, this reaction is exoergic with a reaction Q value 6.24 MeV. In contrast to charged-particle induced reactions, there is no coulomb or charge barrier for neutron induced reactions and the cross-section for the neutron capture reaction increases as the energy of the neutron decreases. This increase in cross-section can be understood in that the wavelength of the neutron, and hence its probability of interacting with the target nucleus, increases as the kinetic energy or velocity of the neutron decreases. The maximum yield for most neutron capture reactions is obtained by irradiating the target material in a region of the nuclear reactor where the high energy neutrons produced by fission have been slowed down or moderated so that they have an average kinetic energy of 0.025 eV. One advantage of producing radioisotopes through neutron capture reactions is that the cross-sections are, at 0.025 eV, on the order of barns, whereas charged particle induced cross-sections have peak values on the order of millibarns. A second advantage is that many different targets can be irradiated at the same time in a nuclear reactor, while most accelerator facilities can only irradiate one or two targets at a time. The obvious disadvantage of neutron induced reactions is that the isotope production must be performed off-site at a nuclear reactor, whereas compact cyclotrons can be sited at or near the medical facility making it possible to work with quite short-lived proton-rich isotopes.

The production rate (R) of a radionuclide during irradiation with the moderated or "thermal" neutrons in a nuclear reactor is given by

$$R = N \times \Phi \times \sigma$$

where N is the number of nuclei of the target isotopes in the sample, Φ is the flux of thermal neutrons ($n{\cdot}cm^{-2}{\cdot}s^{-1}$), and σ is the thermal neutron reaction cross-section (cm^2). Because those radionuclides created during irradiation can decay during the production process, the activity (A) in Bq of a radionuclide with decay constant λ produced by irradiating a sample either in a reactor or with a charged particle beam is given by

$$A = R(1 - e^{-\lambda t_i})e^{-\lambda t_d}$$

where R is the production rate for the reaction, t_i is the irradiation time, and t_d is the amount of time the sample has been allowed to decay following the irradiation. Using again the example of ^{165}Ho(n,γ) ^{166}Ho reaction, ^{166}Ho has a half-life of 1.12 days, ^{165}Ho has a thermal cross-section of 58 b and an isotopic abundance of 100%. Irradiation of 87 mg of ^{165}Ho [100 mg target of holmium oxide (Ho_2O_3)] for 5 days in a thermal neutron flux of 1×10^{14} $n{\cdot}cm^{-2}{\cdot}s^{-1}$ will produce $\sim 1.8 \times 10^{12}$ Bq or 48 Ci of ^{166}Ho with a specific activity of 0.55 Ci of ^{166}Ho mg^{-1} of Ho in the target. The specific activity that can be achieved in the overall (nuclear and chemical) production process is a critical consideration when the radionuclide is used for therapy because it represents the fraction of the atoms in the sample that will have radio-therapeutic activity. In the example of ^{166}Ho, a specific activity of 0.55 $Ci{\cdot}mg^{-1}$ means that only 1 out of every 1300 Ho atoms in the sample that will be incorporated into the therapeutic agent are radioactive. A disadvantage of using direct neutron capture reactions is that the specific activity of the product radioisotope cannot be improved through chemical means (the product is an isotope of the target).

The parent material of the most commonly used medical radionuclide (^{99m}Tc) is produced in high specific activity through the neutron induced fission of ^{235}U. On average, every 100 thermal neutron induced fissions of ^{235}U produce six ^{99}Mo atoms. Because the molybdenum produced in the uranium target can be chemically separated from the uranium and other fission products, and because all of the other molybdenum isotopes have much shorter half-lives than ^{99}Mo, the process results in a sample in which nearly every molybdenum atom is ^{99}Mo. While there are a number of advantages from using this "carrier free" (i.e., 100% of the Mo is radioactive ^{99}Mo) ^{99}Mo to create the commercial ^{99}Mo/^{99m}Tc generators, the process does create significant amounts of waste that must be appropriately disposed. Neutron-rich radionuclides can also be produced in high specific activity through an indirect method that utilizes neutron capture followed by radioactive decay. Consider for example the production of ^{105}Rh from the irradiation of ruthenium target. Neutron capture on ^{104}Ru produces ^{105}Ru through the following reaction:

$${}^{1}_{0}n + {}^{104}_{44}Ru \rightarrow {}^{105}_{44}Ru + Q$$

^{105}Ru is radioactive ($t_{1/2} = 4.44$ h) and beta decays into ^{105}Rh, which has a half-life of 35.4 h. As in the fission case, the rhodium in the target can be chemically separated from the ruthenium target to produce a sample that is essentially carrier free.

SUMMARY

The use of radionuclides in medicine has been, and continues to be, important to diagnosis of disease and radiotherapy. Many diagnostic agents based on ^{99m}Tc are part of the arsenal of radiopharmaceuticals available to the physician. The last 10 years has seen the FDA approval of three new radiotherapeutic agents, namely, Quadramet (a bone pain palliation agent containing ^{153}Sm) and the first two FDA approved radioimmunoconjugates, Zevalin [a ^{90}Y labeled monoclonal antibody that specifically targets the CD20 antigen expressed on >90% of non-Hodgkin's lymphomas (NHL), and Bexxar (an ^{131}I labeled monoclonal antibody that specifically targets the CD20 antigen expressed on >90% of non-Hodgkin's lymphomas]. The development of new radiodiagnostic and radiotherapeutic agents will continue and will undoubtedly take advantage of the advances occurring in molecular biology and genomics. Radioactive decay and radionuclide production are two important aspects in the design of new radiopharmaceuticals. For more detailed discussions on these topics, the reader is referred to general textbooks on nuclear and radiochemistry (6–8).

BIBLIOGRAPHY

1. Steigman J, Eckelman WC. Nuclear Science Series (NAS-NS-3204). Nuclear Medicine, The Chemistry of Technetium in Medicine. National Academy Press; Washington (D.C.): 1992.
2. Jurisson SS, Lydon JD. Potential Technetium Small Molecule Radiopharmaceuticals. Chem Rev 1999;99:2205–2218.
3. Brucer M. In: Sorenson JA. et al., editors. The Heritage of Nuclear Medicine. The Society of Nuclear Medicine; New York: 1979.
4. Science 1946;103(2685): 698–705.
5. Lever SZ, Lydon JD, Cutler CS, Jurisson SS. Radioactive Metals in Imaging and Therapy. In: Meyer T, McCleverty J, editors. Comprehensive Coordination Chemistry II Volume 9. London: Elsevier Ltd.; 2004. pp 883–911.
6. Friedlander G, Kennedy JW, Macias ES, Miller JM. Nuclear and Radiochemistry. 3rd ed. New York: Wiley; 1981.
7. Choppin G, Rydberg J, Liljenzin JO. Radiochemistry and Nuclear Chemistry. 2nd ed. Oxford: Butterworth-Heinemann Ltd; 1995.
8. Ehmann WD, Vance DE. Radiochemistry and Nuclear Methods of Analysis. New York: Wiley; 1991.
9. Loveland WD, Morrissey D, Seaborg GT. Modern Nuclear Chemistry. New York: Wiley; 2005.

See also NEUTRON ACTIVATION ANALYSIS; TRACER KINETICS.

RADIOPHARMACEUTICAL DOSIMETRY

Hubert M.A. Thierens
University of Ghent
Ghent, Belgium

INTRODUCTION

In nuclear medicine, radiopharmaceuticals are administered to patients for diagnosis or treatment purposes. Each pharmaceutical compound has its specific biodistribution over organs and tissues in the body with related retention times. One method of calculating absorbed dose values delivered internally was developed in the 1960s by the medical internal radiation dose (MIRD) committee of the American Society of Nuclear Medicine (1,2). The original aim was to develop a dosimetry methodology for diagnostic nuclear medicine, but the method can be applied to dosimetry for radionuclide therapy, where the need for an accurate dosimetry is more imperative in view of the high activity levels administered to the patient. The MIRD dosimetry protocol is applied by different international organizations. The International Commission on Radiological Protection (ICRP) has published catalogs of absorbed doses to organs and tissues per unit activity administered, calculated using this dosimetric approach, for most diagnostic radiopharmaceuticals commonly applied (3,4). These tables are established for patients with standard biokinetics of the radiopharmaceutical. For application of the MIRD protocol in the nuclear medicine department when taking into account patient-specific biokinetics a user-friendly computer program called MIRDOSE was developed by Stabin (5). This software has been replaced recently by the authors by an U. S. Food and Drug Administration (FDA) approved program OLINDA (Organ Level Internal Dose Assessment) (6).

By combining well-selected β-emitting radionuclides with disease-specific pharmaceuticals, administration of radiolabeled drugs can provide efficient internal radiotherapy for localized disease as well as for metastatic cancer. As a result, an increasing number of radioactive therapeutic agents are being used in nuclear medicine for the treatment of a large variety of diseases (7–12). For these medical applications of radioactive compounds accurate patient-specific internal dosimetry is a prerequisite. Indeed, the basic goal of the majority of these types of metabolic radiotherapy is to ensure a high absorbed dose to the tumoral tissue without causing adverse effects in healthy tissues. In a curative setting an optimized activity has to be calculated and administered to the patient to ensure the delivery of a predetermined absorbed dose to the tumor resulting in complete tumor control, while minimizing the risk of normal tissue complications. The determination of latter activity necessitates a patient specific dosimetry with respect to drug pharmacokinetics and if possible patient-specific anatomical data.

Nowadays, for most applications patient-specific biokinetics are derived from sequential images after administration of a tracer activity and combined with the MIRD methodology to calculate absorbed doses to target and critical tissues (13). For a more complete dosimetric analysis as in the case of clinical trials, the information from imaging is completed by data obtained from blood sampling and urinalysis. In general, patient anatomy is represented by a standard anthropomorphic phantom (14). However, in a complete patient-specific dosimetry approach the individual patient anatomy is also taken into account and derived from computed tomography (CT) or magnetic resonance imaging (MRI). Three dimensional (3D) absorbed dose estimates are then determined from single-photon emission computed tomography (SPECT) or positron emission tomography (PET) activity data using dose-point kernel convolution methods (15,16), or by direct Monte Carlo calculation (17–20). Dose point kernels describe the pattern of energy deposited by the radiation at various radial distances from a point source. Convolution of the dose point kernel of the considered radionuclide with the activity distribution in the patient results in the absorbed dose distribution in the patient. The general idea of Monte Carlo analysis is to create a model as similar as possible to the real physical system and to create interactions within that system based on known probabilities of occurrence, with random sampling of the probability density functions. For dosimetric applications, differential cross-section data of the interactions of the ionizing particles with matter are used and the path of each particle emitted by the radioactive material is simulated until it is completely stopped. The energy deposited in the medium along the path of the ionizing particles results in the absorbed dose distribution. The advantage of direct calculation by Monte Carlo techniques is that this method allows media of inhomogeneous density to be considered. More information on the application of Monte Carlo techniques in dosimetry can be found in Ref. 21. Several software packages have been devised and validated by different groups for patient specific dose calculations. Typical examples are the 3D-ID code from the Memorial Sloan-Kettering Cancer Center (22), the RMDP-MC code from the Royal Marsden hospital (UK) (23) and the VOXEL-DOSE code from Rouen (France) (24). These programs are based on general Monte Carlo codes also used in other medical applications of ionizing radiation as external beam radiotherapy: EGSnrc (25) and GEANT (26).

Radiolabeled pharmaceuticals are also used in the development of new drugs. Before a drug can be applied to patients in phase I and II clinical trials, different steps have to be taken in the investigation of the toxic effects of the new (radio)pharmaceutical compound. This involves firstly a number of animal studies followed by the administration of the pharmaceutical to a restricted number of volunteers. In general, for these animal and volunteer studies, a radiolabeled formulation of the newly elaborated drug is used with ^{3}H or ^{14}C as radionuclide. Sacrifice of the animals at different time points postadministration and quantitative whole-body autoradiography allow the determination of the biodistribution with metabolite profiling, the retention in the different organs and tissues, and the study of excretion pathways of the pharmaceutical in the animals. In the development of radiopharmaceuticals specifically for nuclear medicine imaging purposes, the new compound can be labeled with gamma-emitting radionuclides and the biokinetics are derived from serial imaging of animals. To this end, dedicated (micro)SPECT and (micro)PET systems were constructed (27–29). The animal activity data are extrapolated to humans to determine the maximal activity of the radiolabeled compound allowed to be administered to healthy volunteers. Criterion is here that the effective dose may not exceed the limits for the considered risk category of the volunteers following the ICRP Publication 62 categories (30). In general, the risk category IIa (risk $\sim 10^{-5}$) with a maximal effective dose of 1 mSv is appropriate for volunteers in testing new drugs. This corresponds to an intermediate-to-moderate level of social benefit for a minor to intermediate risk level for the volunteers. This evaluation procedure necessitates a reliable dosimetry estimate based on the extrapolation of animal activity data to humans. To obtain this dose estimate generally the MIRD formalism is applied.

THE MIRD SCHEMA

Basic Principles and Equations

For patient dosimetry in nuclear medicine diagnostic procedures, the MIRD schema is applied. This formalism is also used systematically for dose calculations in administration of radiolabeled drugs of volunteers in the framework of drugs development.

In the MIRD protocol, organs and tissues in the body with a significant uptake of the radiopharmaceutical are considered as source organs. On the other hand all organs and tissues receiving an absorbed dose are considered as target organs. This is illustrated in Fig. 1 for iodine isotopes as ^{131}I with the thyroid as source organ and the lungs as the target organ on the anthropomorphic phantom developed by Snyder et al. (31). The absorbed dose to a particular target organ t from a source organ s, $D_{t \leftarrow s}$, can be obtained using the following equation (1,2) :

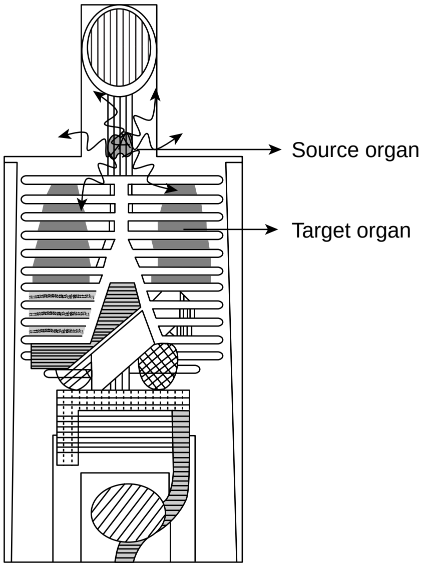


Figure 1. Illustration of the MIRD method for the γ rays emitted in the decay of iodine isotopes as ^{131}I with the thyroid as source organ and the lungs as target organ in the anthropomorphic phantom developed by Snyder et al. (31).

$$D_{t \leftarrow s} = \tilde{A}_s S_{t \leftarrow s}$$

with $\tilde{A}_s$ is the cumulated activity in the source organ and $S_{t \leftarrow s}$ is the mean dose to the target organ per unit cumulated activity in the source organ.

The cumulated activity $\tilde{A}_s$ is the total number of disintegrations of the radioactivity present in the source organ integrated over time and expressed in units Bq.s. It depends on the activity administered, the uptake, retention and excretion from the source organ, and the physical decay of the radionuclide. The $S_{t \leftarrow s}$ values depend on the decay modes of the considered radionuclide and the source-target geometry. The $S_{t \leftarrow s}$ values are tabulated for standard men and children anthropomorphic phantoms (14).

The cumulated activity $\tilde{A}_s$ is the time integral of the activity in the source organ $A_s(t)$:

$$\tilde{A}_s = \int_0^\infty A_s(t)dt$$

The biological retention in the source organ is generally derived from sequential scintigraphies with a gamma camera. For this, opposing planar views or SPECT are used with a calibrated source in the field of view. Corrections are needed for patient attenuation and scatter of the γ radiation. The cumulated activity in the different source organs $\tilde{A}_s$ allows to calculate the residence time τ being the average time the administered activity A_0 spends in the considered source organ:

$$\tau = \frac{\tilde{A}_s}{A_0}$$

In the MIRDOSE software, the cumulated activity in the different source organs is introduced by the values of the residence time (5).

The mean dose to the target organ per unit cumulated activity in the source organ, $S_{t \leftarrow s}$, is given by the expression:

$$S_{t \leftarrow s} = \frac{1}{m_t} \sum_i \Delta_i \phi_i (t \leftarrow s)$$

with m_t mass of the target organ, Δ_i the mean energy emitted per disintegration for radiation of type and energy i, and $\varphi_i(t \leftarrow s)$ the absorbed fraction for radiation of type and energy i. The absorbed fraction $\varphi_i(t \leftarrow s)$ is defined as the fraction of the radiation of type i emitted by the source organ s absorbed in the considered target organ t. The Δi values are obtained from the decay scheme of the considered radionuclide. The values of the specific absorbed fractions $\varphi_i(t \leftarrow s)/m_t$ were calculated by Monte Carlo methods (32). The $S_{t \leftarrow s}$ values for commonly used isotopes in nuclear medicine calculated in this way for a number of standard anthropomorphic phantoms including children of different ages (14) are tabulated and included in the data base of the MIRDOSE package (5). This procedure assumes a uniform distribution of the activity over the source organs and a standard anatomy of the patient.

Table 1. Effective Dose Values Per Unit Activity Administered For A Number Of Radiopharmaceuticals Commonly Applied In Nuclear Diagnostics [a]

Radiopharmaceutical	Effective Dose Per Unit Activity Administered (mSv/MBq)				
	1 year	5 years	10 years	15 years	Adult
^{18}F FDG	0.095	0.050	0.036	0.025	0.019
^{67}Ga citrate	0.64	0.33	0.20	0.13	0.10
^{99m}Tc-DTPA	0.016	0.0090	0.0082	0.0062	0.0049
^{99m}Tc-HMPAO	0.049	0.027	0.017	0.011	0.0093
^{99m}Tc-MIBI	0.053	0.028	0.018	0.012	0.0090
^{99m}Tc-MDP	0.027	0.014	0.011	0.0070	0.0057
^{99m}Tc-pertechnetate	0.079	0.042	0.026	0.017	0.013
^{99m}Tc-leucocytes	0.062	0.034	0.022	0.014	0.011
^{111}In-octreotide	0.28	0.16	0.10	0.071	0.054
^{123}I uptake 35%	2.05	1.08	0.51	0.34	0.22
^{123}I-MIBG	0.068	0.037	0.026	0.017	0.013
^{201}Tl-chloride	2.80	1.70	1.20	0.30	0.22

[a] See Ref. 4.

By using the MIRD working procedure, the absorbed doses of the different target organs for frequently used radiopharmaceuticals per unit activity administered were calculated by the ICRP for an adult and children of 1, 5, 10, and 15 years assuming standard biokinetics and were tabulated (3,4). As measure of the radiation burden patient the effective dose E is calculated by summing up the tissue equivalent doses H_T using the tissue weighting factors w_T as defined in the ICRP 60 publication (33):

$$E = \sum_{T} w_T H_T$$

For the types of radiation emitted by the radionuclides used in nuclear medicine the radiation weighting factor w_R is one except for alpha particles, where w_R equals 20. In Table 1 the effective dose values per unit activity administered for a number of radiopharmaceuticals commonly applied in nuclear diagnostics under the assumption of standard biokinetics is summarized. Table 1 shows that in diagnostic pediatric nuclear medicine the patient dose is strongly dependent on patient age for the same administered activity. This is mostly due to the change in patient weight. Weight dependent correction factors for the activity to be administered have been calculated to obtain weight independent effective doses (34,35). The concept of effective dose is intended to estimate the risk for late stochastic radiation effects as radioinduced cancer and leukemia in the low dose range, and by this applicable to nuclear medicine investigations for diagnosis. Its value is not representative for the risk for direct deterministic effects as bone marrow depletion in case of therapeutic applications of radiopharmaceuticals.

Microdosimetric Considerations

The S-values commonly applied at the macroscopic level are calculated assuming a uniform distribution of the activity over the source organ and the target being the whole volume of the target tissue. The use of S-values based on these assumptions can lead to erroneous results at the microscopic level in case of self-dose calculation in an organ (target = source) when the isotope distribution is nonuniform at the cellular level and particles with range of the order of cellular dimensions are emitted in the decay. This is particularly the case when the radionuclide used is an Auger electron or an alpha emitter. A typical example is the dosimetry of lymphocytes labeled with ^{99m}Tc. In Fig. 2, the therapeutic range in soft tissue of the low energy electron groups in the decay of ^{99m}Tc is represented and compared to the dimensions of the DNA helix (2 nm) and a lymphocyte (10 μm). Taking into account the range of the large intensity Auger electron groups (2–100 nm) the dose to the DNA in the cell nucleus, which is the biological radiation target in the cell, is strongly dependent if we consider intra- or extracellular distribution of a ^{99m}Tc radiopharmaceutical. In most radiopharmaceuticals, ^{99m}Tc is located extracellularly and the radiation burden from the Auger electrons to the nucleus is very low. In those cases, the cellular dose is due to the 140 keV γ-emission and the macroscopic S-values assuming a uniform distribution of the ^{99m}Tc activity can be used for the dose calculation. However, in case of intracellular labeling as in the case of labeling of lymphocytes with ^{99m}Tc-HMPAO the dose to the lymphocytes due to the Auger electrons is very high, which leads to radiotoxic effects in these cells (36).

For dose assessment in case of intracellular labeling or labeling of the membrane with an Auger electron emitter a microdosimetric approach based on Monte Carlo calculation methods is indicated. In those cases, the MIRD method can be applied at the cellular level for the calculation of the cell nucleus dose from activity uniformly present in the nucleus, the cell cytoplasm or the cell surface using appropriate microscopic $S_{t \leftarrow s}$ values tabulated for all radionuclides (37). To cope with nonuniform activity distributions in source organs, determined by PET and SPECT imaging, radionuclide voxel S-values are tabulated for five radionuclides for cubical voxels of 3 and 6 mm (38).

DOSIMETRY FOR RADIONUCLIDE THERAPY

Methodology

In external beam radiotherapy, there is a long tradition in performing treatment planning calculations for each

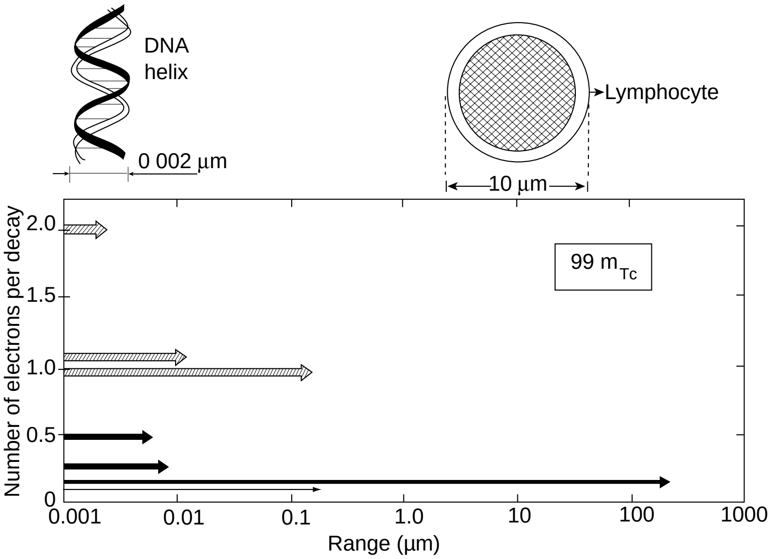


Figure 2. The range in soft tissue of the low energy electron groups in the decay of ^{99m}Tc, compared to the dimensions of the DNA helix and a lymphocyte. The latter comparison is relevant for dosimetry in the intracellular labeling of lymphocytes with ^{99m}Tc-HMPAO.

individual patient. The dosimetry protocols necessary for metabolic radiotherapy, however, are far more complex than those used in external beam therapy. In fact, the *in vivo* activity distribution initially is patient-specific and unknown in both space and time. For the determination of the patient specific drug pharmacokinetics, a tracer activity of the radiopharmaceutical is administered to the patient and quantitative imaging at multiple time points is employed to establish patient-specific biokinetics (13). Here, nuclear medicine imaging with proper correction for photon attenuation, scatter, and collimator resolution is needed to obtain the most accurate activity maps possible. The patient-specific biokinetics can then be combined with the MIRD methodology, described above, to calculate absorbed doses to organs and tissues. The MIRDOSE software allows to calculate the self-absorbed dose to a sphere representing the tumor in case of oncological applications. From these dosimetric calculations, the activity of the radiopharmaceutical to be administered to deliver the prescribed absorbed dose level to the considered tissues is then calculated by extrapolation. This approach does not take into account the patient anatomy. Instead, anatomical data of the average male, female, and children of different ages are introduced by anthropomorphic phantoms (14).

More accurate dosimetric calculations require that the individual patient anatomy derived from CT or MRI images is converted into a 3D voxel representation as in external beam radiotherapy. The 3D absorbed dose estimates from the tracer activity administration are then determined from SPECT or PET activity imaging using dose-point kernel convolution methods, or by direct Monte Carlo calculation (15–24). This approach necessitates image fusion between the different imaging modalities lused. The advent of combined SPECT–CT and PET–CT equipment allows a more general application of this complete patient-specific dosimetry (39–43). In this setting, the CT data may be used as an attenuation map, which is an important improvement for accurate quantification (39).

A dosimetry calculation can be useful not only for assessment of the amount of activity to be administered before radionuclide therapy, but also after the performed radionuclide therapy. First, it is important to verify the predicted absorbed dose distribution. Second, the dosimetry results of a patient population can be combined with the outcome of the therapy to analyse the dose-response of the radionuclide therapy and to make changes in the therapy protocol when necessary (e.g., the predetermined target dose level). As was the case for the pretherapy calculation of the administered activity, posttherapy dosimetry can be performed at different levels of sophistication.

Dosimetry is not only important in the framework of therapy prediction, but also in the dose assessment of organs at risk. In radiopeptide therapy, the kidneys are the dose-limiting organ (44,45). Radiopeptides are cleared physiologically via the kidneys. Most peptides are cleaved to amino acids as metabolites in the kidneys with a high and residualizing uptake in the tubular cells. Damage to the kidneys induced by the radiolabeled metabolites can cause nephropathy after therapeutic application of radiopeptides (46). Application of basic amino acids can reduce the renal accretion of radiolabeled metabolites and the kidney dose (47).

Dosimetry of Radioiodine Therapy for Thyrotoxicosis

The most common application of radionuclide therapy is treatment of hyperthyroidism as observed in Graves' disease or Plummer's disease (toxic nodular goiter) by oral administration of ^{131}I. The rationale behind dosimetry for this kind of treatment is that at long-term hypothyroidism may be the outcome for patients treated with radioiodine and that the incidence of this inverse effect is higher with an earlier onset for patients treated with higher activities (48). A large variation exists in the literature on the value of target dose to be delivered to the hyperthyroid tissue to become euthyroid. Howarth et al. (49) reported that doses of 60 and 90 Gy cured 41 and 59 % of patients after 6 months. Guhlmann et al. (50) cured hyperthyroidism in 83 % of patients at 1 year post-treatment with a dose of 150 Gy. According to Willemsen et al. (51) hyperthyroidism is eliminated in all patients 1 year post-treatment with a dose of 300 Gy, but at this high dose level 93% of patients became hypothyroid.

For dose calculation in general an adapted version of the Quimby-Marinelli formula (52) has to be used

$$A(\mu\text{Ci}) = \frac{6.67 \times \text{Dose}\,(\text{cGy})\,\text{mass}\,(\text{g})}{T_{1/2\text{eff}}(\text{days}) \times \%\,\text{uptake}(24\,\text{h})}$$

Application of this protocol for individual patient dosimetry necessitates the determination of the following important variables: percentage uptake 24 h after administration, effective half-life of the radioiodine, and mass of the thyroid gland. For uptake and kinetics assessment, serial scintigraphies or probe measurements of the patient's thyroid after administration of a tracer dose have to be performed. This approach assumes that the kinetics of a tracer and a therapeutical amount of administered activity are the same. According to some authors, a pretherapeutic tracer dose may induce a stunning effect limiting the uptake of the therapeutic activity in the thyroid afterward (53). The thyroid mass is generally determined by the pretherapeutic scintigraphy, by ultrasonography or by MRI (54). A ^{124}I PET image also allows measurement of the functioning mass of the thyroid (55). Dosimetry protocols exist based on only a late uptake measurement at 96 or 192 h after tracer activity administration (56). A thorough discussion of the activity to be administered and the dosimetry protocol to follow can be found in Refs. 57,58.

Dosimetry of Radioiodine Therapy for Differentiated Thyroid Cancer

Radioiodine is also administered frequently to patients for differentiated thyroid cancer to ablate remnant thyroid tissue in the early postoperative period, for locoregional recurrences, and for distant metastases. Although most centers administer standard activities, typically 2.8–7.4 GBq (75–200 mCi), because of the practical difficulties to determine the target absorbed dose, absorbed dose-based protocols are also applied (59). For the calculation of the activity to be administered to give a predetermined tumor absorbed dose protocols as for thyrotoxicosis treatment described earlier are used. As predetermined absorbed dose-to-remnant thyroid tissue a value of 300 Gy is considered to be sufficient (60). For treatment of metastases lower doses giving a complete response have been reported: 85 Gy (61) and 100-150 Gy (62). This approach necessitates determination of the remnant mass of thyroid tissue or metastases by the methods described earlier, which is now more difficult in practice. This introduces in general a large uncertainty on the activity to be administered to ensure the desired dose to the target tissue. Also, the radioiodine kinetics with the 24 h uptake and the effective half-life has to be determined for the patient by administration of a tracer dose. Because of the relatively high activities necessary for quantitative imaging of the target thyroid tissue for this application (at least 37 MBq-1 mCi ^{131}I) complication of the therapy by stunning introduced by the tracer activity mentioned earlier, is more critical here. Because of this and the inaccuracy in the target mass determination, dosimetry protocols based on target dose levels remain difficult for treatment of differentiated thyroid cancer. The ^{124}I PET imaging allows a more exact *in vivo* determination of iodine concentration and volume determination. By using this method, radiation doses to metastases ranging between 70 and 170 Gy were delivered to the lesions (63).

Instead of target absorbed dose-based protocols, dosimetry protocols based on the largest safe approach are also applied. This approach based on the dose to the critical tissues allows the administration of the maximum possible activity to achieve the maximum therapeutic efficacy. Application of this method necessitates serial total body scintigraphy after the administration of a tracer dose. The dose to the bone marrow, the lungs, and the thyroid tissue or metastases is then calculated by the MIRD formalism. From the absorbed doses obtained by the tracer activity imaging the amount of activity giving the maximal tolerable absorbed dose to the critical tissues is calculated. It has been generally accepted that the activity that delivers 2 Gy whole body dose as a surrogate for the bone marrow dose with a whole body retention < 4.44 GBq (120 mCi) at 48 h postadministration is safe with respect to bone marrow suppression (64). In some departments, the tolerable dose level of the bone marrow in radioiodine treatment of patients with metastatic differentiated thyroid cancer is even increased to 3 Gy based on the $LD_{5/5}$ data of external beam radiotherapy with $LD_{5/5}$ being the dose for the red marrow giving a 5% risk of severe damage to the blood-forming system within 5 years after administration (62). Very high activities of ^{131}I in the range 7.4–37.9 GBq (200–1040 mCi) are then administered for treatment of metastases. In a retrospective study of patients treated with this protocol over a period of 15 years, transient bone marrow depression with thrombopenia and leukopenia was observed recovering after a few weeks (62). No permanent damage was observed. In $\sim$ 10% of the patients the dose-limiting organ were the lungs for which a limit of 30 Gy was adopted from $LD_{5/5}$ data.

A recent review of the evolving role of ^{131}I for the treatment of differentiated thyroid carcinoma can be found in Ref. 9. Preparation of patients by administration of recombinant human thyroid-stimulating hormone (rhTSH) may allow an increase in the therapeutic radioiodine activity while preserving safety and tolerability (65). As side effects of the ^{131}I therapy, impairment of the spermatogenesis in males (66) and earlier onset of menopause in older premenopausal women (67) are reported. With respect to pregnancy, it is recommended that conception be delayed for 1 year after therapeutic administrations of ^{131}I and until control of thyroid hormonal status has been achieved. After this period there is no reason for patients exposed to radioiodine to avoid pregnancy (68).

Dosimetry of ^{131}I-MIBG Therapy

Another radionuclide therapy application for which the importance of patient-specific dosimetry is generally accepted is treatment of pediatric neuroblastoma patients with ^{131}I-MIBG. Neuroblastoma is the most common extracranial solid tumor of childhood with an incidence of 1/70000 children under the age of 15 (69). Neuroblastoma cells actively take up nor-adrenalin via an uptake-1 system. The molecule *meta*-iodo benzyl guanidine (MIBG), radiolabeled with ^{131}I, has a similar molecular structure, uptake, and storage in the cell as nor-adrenalin. Since 1984, ^{131}I-MIBG has been used therapeutically in neuroblastoma patients (70,71). Aside from the tumor, ^{131}I-MIBG is also taken up in the liver, heart, lungs, and adrenal glands. The bladder is irradiated by the metabolites of ^{131}I-MIBG. For patient dosimetry the largest safe dose approach is applied in ^{131}I-MIBG therapy with bone marrow as dose limiting organ. In practice, the whole body absorbed dose is also used in this setting as an adequate representation or index of bone marrow toxicity. Most treatment regimens consider the maximal activity to be administered limited by rendering a bone marrow dose of 2 Gy. Prediction of whole body doses is based on a pretherapeutic administration of ^{123}I-MIBG. In Fig. 3, predicted whole body doses based on pretherapeutic ^{123}I-MIBG scintigraphies are compared to doses received by patients after ^{131}I-MIBG therapy (72). The received dose values were derived from post-therapy scans. This figure shows also that in the case of repeated therapies pre-therapy scans do not need to be repeated before each therapy except when the biodistribution of ^{131}I-MIBG is expected to change rapidly (e.g., for patients where bone marrow invasion is present). It has also been shown that the accuracy of whole body dosimetry improves when half-life values of tracer and therapy radionuclides are matched (73).

In ^{131}I-MIBG therapy, protocols with administration based on fixed activity per unit mass protocols are also applied. Matthay et al. (74) reported on dosimetry performed in a dose escalation study of patients treated with ^{131}I-MIBG for refractory neuroblastoma with a fixed activity per unit mass ranging from 111 to 666 MBq·kg^{-1} (3–18 mCi·kg^{-1}). Patients treated with a specific activity < 555 MBq·kg^{-1} did not require hematopoietic stem cell support, while this was necessary for one-half of the patients treated with a higher specific activity. The median whole body dose of the group of patients requiring hematopoietic stem cell support was 3.23 Gy (range 1.81–6.50 Gy) while for the other patients the median dose was 2.17 Gy (range 0.57–5.40 Gy).

In order to improve the results of ^{131}I-MIBG therapy for patients refractive of extensive chemotherapy treatment, a high activity ^{131}I-MIBG schedule is now being used in combination with topotecan as radiosensitizer for the therapy of neuroblastoma in a controlled ESIOP (European International Society of Pediatric Oncology for Neuroblastoma) study protocol (75). The aim here is to administer in two fractions the amount of activity needed to reach a combined total body dose of 4 Gy. These kinds of high doses will inevitably invoke severe side effects, thus frequently necessitating hematopoietic stem cell support and even bone marrow transplantation. However, a contemporary oncological department is well equipped to deal with this kind of treatments. The first amount of ^{131}I-MIBG activity is administered based on a fixed activity per unit body mass (444 MBq·kg^{-1}–12 mCi·kg^{-1}) protocol. Total body dosimetry is carried out using serial whole body scintigraphies after the first administration. Radionuclide kinetics are followed by a whole body counter system mounted on the ceiling of the patient's isolation room. These dosimetry results are then used to calculate the activity of the second administration of ^{131}I-MIBG giving a total body dose of 4 Gy over the two administrations. The first results of this study indicate that *in vivo* dosimetry allows for an accurate delivery of the specified total whole body dose and that the treatment schedule is safe and practicable (75). The approach has now to be tested for efficacy in a phase II clinical trial.

Dosimetry in Radioimmunotherapy

In general, red marrow is the dose-limiting tissue in non-myeloablative and lung for myeloablative radioimmunotherapy (RIT). Administration protocols are applied based on absorbed-dose values of the dose-limiting tissue and on an activity per body weight basis. Typical examples

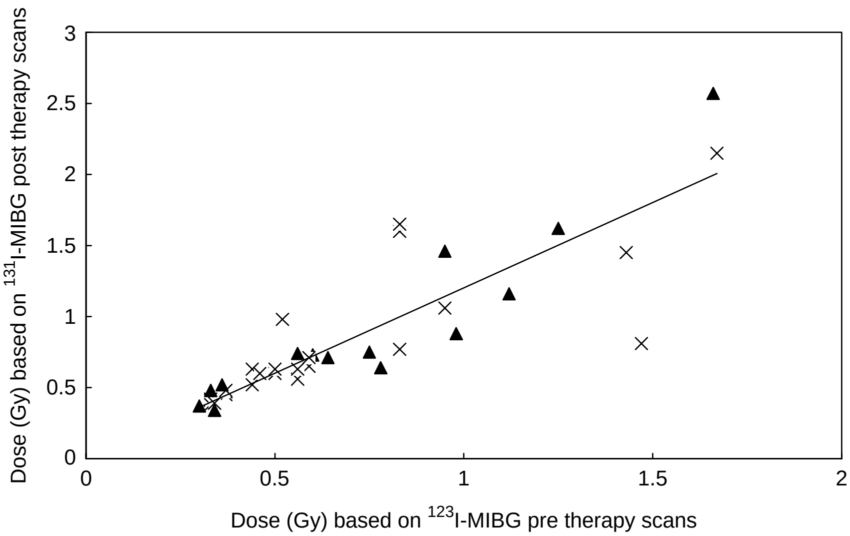


Figure 3. Correlation between the whole body dose estimate based on ^{123}I-MIBG pretherapy scans and the dose derived from ^{131}I-MIBG posttherapy scans in patients treated for neuroendocrine tumors. The triangles represent the data of the first therapies, the crosses the data of retreatments. The straight line is the result of a linear regression to all data ($R^2 = 0.73$).

of these protocols are the ^{131}I-labeled anti-CD20 antibody, tositumomab (Bexxar; Glaxo-SmithKline) (76) and the ^{90}Y-labeled anti-CD20 ibritumomab tiuxetan (Zevalin; Biogen Idec) (77), respectively. These radiolabeled antibodies are used for treatment of non-Hodgkin's lymphoma. The choice for an activity-based protocol for the ^{90}Y-labeled antibody is based on the lack of correlation between absorbed dose and toxicity in the early studies. The explanation for the absence of a dose-response relationship can be found in different sources. In contrast to ^{131}I, ^{90}Y is a pure β-emitter, and ^{90}Y kinetics have to be derived from surrogate ^{111}In imaging. Another point is that prior treatment of these patients and the bone marrow reserve have a strong effect on the bone marrow toxicity in this case. As patients undergoing RIT have been treated previously by chemotherapy, the impact of such prior therapy on the hematopoietic response to the RIT is important.

Although the necessity of patient-specific dosimetry is questionable in some applications of RIT where the dose-response observations for toxicity are poor, there is a general agreement that complete radiation dosimetry is necessary for each new application of a radiolabeled antibody in phase I and most probably also in phase II studies especially for safety reasons (78). An important argument for absorbed dose driven protocols in clinical phase I trials is that many patients are treated below the biologically active level due to the interpatient variability in activity based administration protocols. This implies data difficult to interpret in antitumor response and toxicity.

In view of the central role of red marrow toxicity in RIT methodologies, bone marrow dosimetry got already a lot of attention in the literature (79–85). In general, methods based on imaging as described in the section on ^{131}I-MIBG therapy are used. Also, approaches to calculate the bone marrow dose based on blood activity measurements have been described, but these methods yield only reliable results when the activity does not bind specifically to blood or marrow components including tumor metastases in the marrow (79). By assuming rapid equilibrium of radiolabeled antibodies in the plasma and the extracellular fluid of the red marrow, a red marrow/blood concentration ratio of 0.3–0.4 can be derived. All red marrow dosimetry performed up to now uses a highly stylized representation of the red marrow over the body. More detailed representations are being generated especially for Monte Carlo calculations enhancing accuracy and reliability of the bone marrow doses (86).

Several studies have investigated the relation between the tumor dose and response especially in RIT of non-Hodgkin's lymphoma (87–89) but the results are negative. Possible explanations are the therapeutic effect of the antibody, different confounding biological factors and the accuracy of tumor dosimetry. Here, standardization of data acquisition as presented in MIRD pamphlet No. 16 (13) may help in dose-response investigations. As discussed earlier in the section on ^{131}I-MIBG therapy, a full patient-specific 3D dosimetric approach with imaging data from the combined SPECT–CT systems will improve substantially the accuracy of the tumor dosimetry results.

DOSIMETRY IN THE DEVELOPMENT OF NEW DRUGS

For the study of the absorption, metabolism, and excretion pathways of new drugs a ^{3}H- or ^{14}C-radiolabeled formulation of the drug is administered to healthy volunteers. A dosimetric evaluation of the radiation burden of the volunteers based on animal biodistribution, retention, and excretion data is necessary and presented to an ethical committee before the radiolabeled drug can be administered. This procedure has to ensure that the effective dose will not exceed the limits for the considered risk category of the volunteers according to the ICRP publication 62 categories (30). For testing new drugs mostly a risk category IIa (risk $\sim 10^{-5}$) is adopted corresponding to a maximal effective dose of 1 mSv. Based on this criterion, the activity to be administered is calculated from the dosimetric evaluation.

For the calculation of the dose estimate of the volunteers the MIRD formalism for an administration of a standard activity (37 kBq/1 μCi) of the radiolabeled pharmacon is applied. Animal biodistribution data are used to calculate the residence time in the source organs and tissues based on the maximum uptake *f* and biological half-life. In general a rat strain is used as animal model. As the organ weights in the rat and man are different an important correction of the animal data is necessary to estimate the *f*-values in humans. For each organ, dosimetric calculations are performed assuming (*1*) the same fraction of activity is absorbed by the organs in rat and humans irrespective of the difference in relative weight or (*2*) the fraction of activity absorbed by each organ is proportional to the relative organ weight in rat and humans. The latter assumption means that the uptake per kilogram of organ weight normalized to the whole body weight is the same for both species. Table 2 gives an overview of the organ and tissue weights in a male Wistar rat of 250 g reported in the literature (90) and in the standard human of 70 kg (32). For each organ or tissue two dose values are obtained by assuming a species independent organ uptake and an uptake proportional to the relative organ weight in different species. The highest dose estimate of both is restrained for each organ. If the retention for the individual organs is not known the whole body retention is adopted.

As model for the liver and biliary excretion it is generally assumed that a fraction of the radiopharmaceutical is taken up by the liver. Part of this activity goes directly to the small intestine while the resting part goes to the gallbladder, from where it is cleared to the small intestine. For the total fraction of activity excreted in this way by the gastrointestinal tract, the fraction of the activity retrieved in the feces is adopted from animal data. In general, data are available for different species and the maximal value is retained. For the dose calculation of the sections of the gastrointestinal tract, the kinetic model of the ICRP publication 53 is adopted (3). The kidney–bladder model described in this publication is also used to calculate the dose to the urinary bladder. Urine activity measurements in animals are used to estimate the fraction of the activity eliminated through the kidneys and again the maximal value is adopted if data are available for different species.

The dose estimates to organs and tissues of humans extrapolated in this way from animal data are combined

Table 2. Organ Weights of a Male Wistar Rat of 250 g[a] and Human of 70 kg[b]

Organ	RAT Weight, g	RAT Rel. Weight, %	Humans Weight, kg	Humans Rel. Weight, %
Adrenal glands	0.085	0.034	0.014	0.020
Blood	15	6	5.5	7.86
Bone	12.99	5.19	5	7.14
Bone marrow	5.59	2.24	3	4.29
Brain	1.43	0.574	1.4	2.0
Heart	0.835	0.334	0.33	0.47
Kidneys	1.873	0.749	0.31	0.44
Large intestine	2.635	1.054	0.37	0.53
Liver	10.675	4.27	1.8	2.57
Lungs	1.618	0.647	1	1.43
Oesophagus	0.11	0.044	0.04	0.057
Pancreas	0.913	0.365	0.1	0.14
Plasma	10	4	3.1	4.43
Prostate	0.3	0.12	0.016	0.023
Small intestine	7.30	2.92	0.64	0.91
Spleen	0.738	0.295	0.18	0.26
Stomach	1.23	0.492	0.15	0.21
Testes	1.815	0.726	0.035	0.050
Thymus	0.593	0.237	0.02	0.029
Thyroid	0.02	0.008	0.02	0.029

[a] See Ref. 90.
[b] See Ref. 32.

with the tissue weighting factors to obtain the effective dose after the administration of the standard activity (37 kBq/1 μCi) as described earlier (33). Based on the effective dose estimate obtained in this way and the dose limits proposed in the ICRP 62 publication (30) the activity of the radiolabeled drug to be administered to the volunteers is obtained.

CONCLUSIONS

To estimate the risk for late radiation effects as cancer and leukemia in patients after administration of radiopharmaceuticals for diagnosis the MIRD formalism with standard human anatomy and biokinetics is generally applied. This holds also for the estimation of the same risk of volunteers after the administration of a radiolabeled formulation of newly developed drugs. However, therapeutic applications of radiopharmaceuticals necessitate a reliable patient specific approach at least with respect to the biokinetics and if possible also for the patient-specific anatomical data. For curative treatment of malignant diseases there is now a tendency to use the largest safe dose approach with administration of the maximum possible activity based on the dose to the critical tissues. On the other hand, the advent of combined SPECT-CT or PET-CT imaging means an essential step forward toward an accurate 3D tumor dosimetry, the basic need for the administration protocols with the calculated activity based on a tumor dose prescription as used in external beam radiotherapy.

BIBLIOGRAPHY

1. Loevinger R, Berman M. A schema for absorbed-dose calculations for biologically distributed radionuclides. MIRD pamphlet No. 1. J Nucl Med 1968;9(Suppl.1):7–14.
2. Loevinger R, Budinger TF, Watson EE. MIRD primer for absorbed dose calculations, New York: Society of Nuclear Medicine; revised 1991.
3. ICRP publication 53. Radiation dose to patients from radiopharmaceuticals. Annals of the ICRP Vol 18. Oxford: Pergamon; 1987.
4. ICRP publication 80. Radiation dose to patients from radiopharmaceuticals. Annals of the ICRP Vol 28. Oxford: Pergamon press; 1998.
5. Stabin MG. MIRDOSE: personal computer software for internal dose assessment in nuclear medicine. J Nucl Med 1996;37:538–546.
6. Stabin MG, Sparks RB, Crowe E. OLINDA/EXM: The second generation personal computer software for internal dose assessment in nuclear medicine. J Nucl Med 2005;46:1023–1027.
7. McDougall IR. Systemic radiation therapy with unsealed radionuclides. Sem Rad Oncol 2000;10:94–102.
8. Knox SJ, Meredith RF. Clinical radioimmunotherapy. Sem Rad Oncol 2000;10:73–93.
9. Robbins RJ, Schlumberger MJ. The evolving role of 131I for the treatment of differentiated thyroid carcinoma. J Nucl Med 2005;46:28S–37S.
10. Valdés-Olmos RA, Hoefnagel CA. Radionuclide therapy in oncology: the dawning of its concomitant use with other modalities. Eur J Nucl Med Mol Imaging 2004;32:929–931.
11. Larson SM, Krenning EP. A pragmatic perspective on molecular targeted radionuclide therapy. J Nucl Med 2005; 46:1S–3S.
12. Kwekkeboom DJ, et al. Overview of results of peptide receptor radionuclide therapy with 3 radiolabeled somatostatin analogs. J Nucl Med 2005;46:62S–66S
13. Siegel JA, et al. MIRD Pamphlet No. 16: Techniques for quantitative radiopharmaceutical biodistribution data acquisition and analysis for use in human radiation dose estimates. J Nucl Med 1999;40:37S–61S.
14. Cristy M, Eckerman KF. Specific absorbed fractions of energy at various ages from internal photon sources. ORNL Report ORNL/TM-8381. Oak Ridge: Oak Ridge National Loboratory; 1987.

15. Giap HB, Macey DJ, Bayouth JE, Boyer AL. Validation of a dose-point kernel convolution technique for internal dosimetry. Phys Med Biol 1995;40:365–381.
16. Furhang EE, Sgouros G, Chui CS. Radionuclide photon dose kernels for internal emitter dosimetry. Med Phys 1996;23: 759–764.
17. Furhang EE, Chui CS, Sgouros G. A Monte Carlo approach to patient-specific dosimetry. Med Phys 1996;23:1523.
18. Liu A, Wiliams LE, Wong JYC, Raubitscek AA. Monte Carlo assisted voxel source kernal method (MAVSK) for internal dosimetry. J Nucl Med Biol 1998;25:423–433.
19. Yoriyaz H, Stabin MG, dos Santos A. Monte Carlo MCNP-4B-based absorbed dose distribution estimates for patient-specific dosimetry. J Nucl Med 2001;42:662.
20. Zaidi H, Sgouros G, editors. Therapeutic applications of Monte Carlo calculations in Nuclear Medicine. Bristol (UK): Institute of Physics Publishing; 2002.
21. Andreo A. Monte Carlo techniques in medical radiation physics. Phys Med Biol 1991;36:861–920.
22. Sgouros G, et al. Three-dimensional dosimetry for radioimmunotherapy treatment planning. J Nucl Med 1993;34:1595–1601.
23. Guy MJ, Flux GG, Papavasileiou P, Flower MA, Ott RJ. RMDP-MC: a dedicated package for I-131 SPECT quantification, registration, patient-specific dosimetry and Monte Carlo. Seventh International Radiopharmaceutical Dosimetry Symposium. Proceedings of the International Symposium Nashville; (TN): Oak Ridge Associated Universities; 2002.
24. Gardin I. Voxeldose: a computer program for 3D dose calculation in therapeutic nuclear medicine. Seventh International Radiopharmaceutical Dosimetry Symposium. Proceedings of the International Symposium Nashville; (TN): Oak Ridge Associated Universities; 2002.
25. Kawrakow I, Rogers DWO. The EGSnrc code system: Monte Carlo simulation of electron and photon transport. NRC Report PIRS-701. Ottawa: National Research Council of Canada; 2000.
26. Carriers JF, Archembault L, Beaulieu L. Validation of GEANT4, an object-oriented Monte Carlo toolkit for simulations in medical physics.Med Phys 2004;31:484–492.
27. Weber S, Bauer A. Small animal PET: aspects of performance assessment. Eur J Nucl Med Mol Imaging 2004;31:1545–1555.
28. Sossi V, Ruth TJ. Micropet imaging: *in vivo* biochemistry in small animals. J Neural Transmission 2005;112:319–330.
29. Tai YC, et al. Performance evaluation of the microPET focus: a third generation microPET scanner dedicated to animal imaging. J Nucl Med 2005;46:455–463.
30. ICRP Publication 62. Radiological protection in biomedical research. Annals of the ICRP Vol. 22. Oxford: Pergamon; 1991.
31. Snyder WS, Ford MR, Warner GG, Watson SB. MIRD Pamphlet No. 11. “S”, Absorbed dose per unit cumulated activity for selected radionuclides and organs. New York: Society of Nuclear Medicine; 1975.
32. ICRP Publication 23. Report of the task group on reference man. Oxford: Pergamon; 1975.
33. ICRP Publication 60. 1990 Recommendations of the International Commission on Radiological Protection. Oxford: Pergamon; 1991.
34. Piepsz A, et al. A radiopharmaceutical schedule for imaging in paediatrics. Eur J Nucl Med 1990;17:127–129.
35. Jacobs F, et al. Optimized tracer-dependent dosage cards to obtain weight-independent effective doses. Eur J Nucl Med Mol Imaging 2005;24:
36. Thierens H, Vral A, Van Haelst JP, Van de Wiele C, Schelstraete K, De Ridder L. Lymphocyte labeling with Technetium-99m-HMPAO: A Radiotoxicity Study using the Micronucleus Assay. J Nucl Med 1992;33:1167–1174.
37. Goddu SM, et al. MIRD Cellular S values. New York: Society of Nuclear Medicine; 1997.
38. Bolch WE, et al. MIRD pamphlet No. 17: the dosimetry of nonuniform activity distributions-radionuclide S values at the voxel level. Medical Internal Radiation Dose Committee. J Nucl Med 1999;40:11S–36S.
39. Seo Y, et al. Correction of photon attenuation and collimator response for a body-contouring SPECT/CT imaging system. J NucL Med 2005;46:868–877.
40. Boucek JA, Turner JH. Validation of prospective whole-body bone marrow dosimetry by SPECT/CT multimodality imaging in I-131-anti-CD20 rituximab radioimmunotherapy of non-Hodgkin’s lymphoma. Eur J Nucl Med Mol Imaging 2005;32: 458–469.
41. Coleman RE, et al. Concurrent PET/CT with an intergrated imaging system: Intersociety dialogue from the joint working group of the American College of Radiology, the Society of Nuclear Medicine, and the Society of Computed Body Tomography and Magnetic Resonance. J Nucl Med 2005;46:1225–1239.
42. Mawlawi O, et al. Performance characteristics of a newly developed PET/CT scanner using NEMA standards in 2D and 3D modes. J Nucl Med 2004;45:1734–1742.
43. Keidar Z, Israel O, Krausz Y. SPECT/CT in tumor imaging: technical aspects and clinical applications. Sem Nucl Med 2003;33:205–218.
44. Otte A, et al. Yttrium-90 DOTATOC: first clinical results. Eur J Nucl Med 1999;26:1439–1447.
45. Bodei L, et al. Receptor-mediated radionuclide therapy with 90Y-DOTATOC in association with amino acid infusion: a phase I study. Eur J Nucl Med Mol Imaging 2003;30:207–216.
46. Otte A, Cybulla M, Weiner SM. ^{90}Y-DOTATOC and nephrotoxicity. Eur J Nucl Med Mol Imaging 2002;29:1543.
47. Jamar F, et al. (86Y-DOTAA0)-D-Phe1-Tyr3-octreotide (SMT487): a phase I clinical study-pharmacokinetics, biodistribution and renal protective effect of different regimes of amino acid co-infusion. Eur J Nucl Med Mol Imaging 2003;30: 510–518.
48. Clarke SEM. Radionuclide therapy of the thyroid. Eur J Nucl Med 1991;18:984–991.
49. Howarth D, et al. Determination of the optimal minimum radioiodine dose in patients with Graves’disease: a clinical outcome study. Eur J Nucl Med 2001;28:1489–1495.
50. Guhlmann CA, Rendl J, Borner W. Radioiodine therapy of autonomously functioning thyroid nodules and Graves’disease. Nuklearmedizin 1995;34:20–23.
51. Willemsen UF, et al. Functional results of radioiodine therapy with a 300 Gy absorbed dose in Graves’disease. Eur J Nucl Med 1993;20:1051–1055.
52. Silver S. Radioactive nuclides in medicine and biology. Philadelphia: Lea & Febiger; 1968.
53. Coakley AJ, Thyroid stunning. Eur J Nucl Med 1998;25:203–204.
54. van Isselt JW, et al. Comparison of methods for thyroid volume estimation in patients with Graves’disease. Eur J Nucl Med 2003;30:525–531.
55. Crawford DC, et al. Thyroid volume measurement in thyrotoxic patients: comparison between ultrasonography and iodine-124 positron emission tomography. Eur J Nucl Med 1997;24:1470–1478.
56. Bockisch A, Jamitzky T, Derwanz R, Biersack HJ. Optimized dose planning of radioiodine therapy of benign thyroidal diseases. J Nucl Med 1993;34:1632–1638.
57. Kalinyak JE, McDougall IR. Editorial: How should the dose of iodine-131 be determined in the treatment of Graves’ hyperthyroidism ? J Clin Endocrinol Metab 2003;88:975–977.
58. Leslie WD, et al. A randomized comparison of radioiodine doses in Graves’hyperthyroidism. J Clin Endocrinol Metab 2003;88:978–983.

59. Thomas SR. Options for radionuclide therapy: from fixed activity to patient-specific treatment planning. Cancer Bioth Radiopharm 2002;17:71–81.
60. Maxon HR, Thomas SR, Hertzberg VS. Relation between effective radiation dose and outcome of radioiodine therapy for thyroid cancer. N Engl J Med 1983;309:937–941.
61. Maxon HR, Englaro EE, Thomas SR. Radioiodine-131 therapy for well differentiated thyroid cancer- a quantitative radiation dosimetric approach: outcome and validation in 85 patients. J Nucl Med 1992;33:1132–1136
62. Dorn R, et al. Dosimetry guided radioactive iodine treatment in patients with metastatic thyroid cancer: largest safe dose using a risk-adapted approach. J Nucl Med 2003;44:451–456
63. Eschmann SM, et al. Evaluation of dosimetry of radioiodine therapy in benign and malignant thyroid disorders by means of iodine-124 and PET. Eur J Nucl Med 2002;29:760–767.
64. Benua R, Cical N, Sonenberg M, Rawson R. The relation of radioiodine dosimetry to results and complications in the treatment of metastatic thyroid cancer. Am J Roentgenol 1962;87:171–179.
65. de Keizer B, et al. Bone marrow dosimetry and safety of high I-131 activities given after recombinant human thyroid-stimulating hormone to treat metastatic differentiated thyroid cancer. J Nucl Med 2004;45:1549–1554.
66. Pacini F, et al. Testicular function in patients with differentiated thyroid carcinoma treated with radioiodine. J Nucl Med 1994;35:1418–1422.
67. Ceccarelli C, et al. ^{131}I therapy for differentiated thyroid cancer leads to an earlier onset of menopause: results of a retrospective study. J Clin Endocrinol Metab 2001;86:3512–3515.
68. Schlumberger M, et al. Exposure to radioactive iodine-131 for scintigraphy or therapy does not preclude pregnancy in thyroid cancer patients. J Nucl Med 1996;37:606–612.
69. Young JL Jr, Miller RW. Incidence of malignant tumors in US children. J Pediatr. 1975;86:254.
70. Hoefnagel CA, Voute PA, De Kraker J, Marcuse HR. Radionuclide diagnosis and therapy of neural crest tumors using iodine-131-Metaiodobenzylguanidine. J Nucl Med 1987;28: 308–314.
71. Hoefnagel CA. Radionuclide therapy revisited. Eur J Nucl Med 1991;18:408–431.
72. Monsieurs M, et al. Patient dosimetry for neuroendocrine tumors based on ^{123}I-MIBG pre-therapy scans and ^{131}I-MIBG post therapy scans. Eur J Nucl Med 2002;29(12):1581–1587.
73. Flux GD, et al. Estimation and implications of random errors in whole-body dosimetry for targeted radionuclide therapy. Phys Med Biol 2002;47:3211–3223.
74. Matthay KK, et al. Correlation of tumor and whole-body dosimetry with tumor response and toxicity in refractory neuroblastoma treated with ^{131}I-MIBG. J Nucl Med 2001; 42:1713–1721.
75. Gaze MN, et al. Feasibility of dosimetry-based high dose ^{131}I-meta-iodobenzylguanidine with topotecan as a radiosensitizer in children with metastatic neuroblastoma. Cancer Biother Radiopharm 2005;20:195–199.
76. Wahl RL. The clinical importance of dosimetry in radioimmunotherapy with tositumomab and iodine I-131 tositumomab. Semin Oncol 2003;30:31–38
77. Wiseman GA, et al. Radiation dosimetry results and safety correlations from ^{90}Y-Ibritumomab Tiuxetan radioimmunotherapy for relapsed or refractory non-Hodgkin's lymphoma: combined data from 4 clinical trials. J Nucl Med 2003;44:465–474.
78. DeNardo GL, Hartmann Siantar CL, DeNardo SJ. Radiation Dosimetry for Radionuclide Therapy in a Nonmyeloablative Strategy. Cancer Biother Radiopharm 2002;17(1):107–118.
79. Sgouros G. Bone marrow dosimetry for radioimmunotherapy: theoretical considerations. J Nucl Med 1993;34: 689–694.
80. Shen S, Denardo GL, Sgouros G, O'Donnell RT, DeNardo SJ. Practical determination of patient-specific marrow dose using radioactivity concentration in blood and body. J Nucl Med 1999;40:2102–2106.
81. Sgouros G, Stabin M, Erdi Y. Red marrow dosimetry for radiolabeled antibodies that bind to marrow, bone or blood components. Med Phys 2000;27:2150–2164
82. Stabin MG, Siegel JA, Sparks RB. Sensitivity of model-based calculations of red marrow dosimetry to changes in patient-specific parameters. Cancer Biother Radiopharm 2002;17: 535–543.
83. Behr TM, Behe M, Sgouros G. Correlation of red marrow radiation dosimetry with myelotoxicity: empirical factors influencing the radiation-induced myelotoxicity of radiolabeled antibodies, fragments and peptides in pre-clinical and clinical settings. Cancer Biother Radiopharm 2002;17:445–464.
84. Shen S, Meredith RF, Duan J, Brezovich I, Khazaeli MB, Lobuglio AF. Comparison of methods for predicting myelotoxicity for non-marrow targeting I-131-antibody therapy. Cancer Biother Radiopharm 2003;18:209–215.
85. Shen S, Meredith RF. Clinically useful marrow dosimetry for targeted radionuclide therapy. Cancer Biother Radiopharm 2005;20:119–122.
86. Stabin MG, et al. Evolution and status of bone and marrow dose models. Cancer Biother Radiopharm 2002;17:427–433.
87. Sgouros G, et al. Patient-specific, 3-dimensional dosimetry in non-Hodgkin's lymphoma patients treated with ^{131}I-anti-B1 antibody: assessment of tumor dose-response. J Nucl Med 2003;44:260-268.
88. Koral KF, et al. Volume reduction versus radiation dose for tumors in previously untreated lymphoma patients who received iodine-131 tositumomab therapy. Conjugate views compared with the hybrid method. Cancer 2002;94:1258–1263.
89. Sharkey RM, et al. Radioimmunotherapy of non-Hodgkin's lymphoma with ^{90}Y-DOTA humanized anti-CD22 IgG (^{90}Y-Epratuzumab): do tumor targeting and dosimetry predict therapeutic response? J Nucl Med 2003;44:2000–2018.
90. Lewi PJ, Marsboom RP. Toxicology reference data Wistar rat. Amsterdam-Holland: Elsevier/North Holland Biomedical Press; 1981.

See also Nuclear medicine, computers in; pharmacokinetics and pharmacodynamics.

RADIOSURGERY, STEREOTACTIC

Thomas H. Wagner
Sanford L. Meeks
M. D. Anderson Cancer Center
Orlando
Orlando, Florida

Frank J. Bova
University of Florida
Gainesville, Florida

INTRODUCTION

Conventional external beam radiotherapy, or teletherapy, involves the administration of radiation absorbed dose to cure disease. The general teletherapy paradigm is to irradiate the gross lesion plus an additional volume suspected of containing microscopic disease not visible through physical examination or imaging, to a uniform dose level. External photon beams with peak photon energy in excess of 1 MeV are targeted upon the lesion site by registering external anatomy and internal radiographic anatomy to the radiation (beam) source.

Due to uncertainty and errors in positioning the patient, the radiation beam, which is directed at the lesion, may need to be enlarged to ensure that errors and uncertainty in patient positioning do not cause the radiation beam to miss some or all of the target. Unfortunately, enlarging the radiation beam results in a relatively large volume of nondiseased tissue receiving a significant radiation dose in addition to the target. For example, expansion of a 24 mm diameter spherical target volume to 26 mm to ensure that the target is fully irradiated in the presence of a 2 mm positional error will increase the irradiated volume by 60% (1). As a consequence, non-cancerous (normal) tissue in the expansion region will receive the same high dose that the target will receive. To minimize the normal tissue toxicity, the total radiation dose is delivered in many small increments (fractions), a principle first discovered by Bergonie and Tribondeau in the early twentieth century (2) and used routinely for the majority of external radiotherapy treatments.

In contrast to conventional, fractionated radiotherapy, stereotactic radiosurgery (SRS) involves the spatially precise and conformal administration of a relatively large, single dose of radiation (10–20 Gy) to a small volume of disease, thereby abandoning the advantages provided by fractionation. Hence, it is imperative to minimize the amount of normal tissue irradiated to a high dose using such an approach. Radiosurgery is commonly used to treat intracranial lesions including brain metastases, arteriovenous malformations, benign brain tumors (acoustic schwannoma, meningioma), and primary malignant brain tumors (astrocytoma, glioma, glioblastoma).

Leksel, first conceived radiosurgery for intracranial targeting in 1950 (3). His initial goal was to produce a lesion similar to one created by a radiofrequency probe but without the need to physically introduce a probe into the brain. The lesion was to be created by a very concentrated single high dose of radiation. Stereotactic targeting and arc-centered stereotaxis methods were already known to Leksell. In his initial design, Leksell mounted a therapeutic X-ray tube onto an arc-centered frame with the axis of rotation positioned in the target tissues. This approach allowed many different paths of radiation to converge on the target tissues producing a highly concentrated dose at the intersection point. While this concept did provide a concentrated dose, Leksell continued to investigate alternate radiation delivery systems in hopes of finding a better system that could produce a high concentration of radiation at the target tissues while providing more normal tissue sparing. In later designs Leksell attempted to use particle beams to take advantage of the known Bragg peak effect. The physical limitations of the particle beam delivery systems as well as the expense of the device encouraged Leksell to continue his development finally arriving at a design based on 201 pencil thin cobalt-60 gamma radiation sources arranged on a hemisphere and focused at a single point. This device, known as the Gamma Knife (Elekta Oncology Systems), was used to treat both benign and malignant intracranial targets.

In the 1980s, several groups began to develop technology that would adapt more generally available medical linear accelerators (linacs), to deliver radiosurgical style dose distributions, thereby placing radiosurgical capabilities within the reach of many radiation therapy clinics. Betti (4) and Columbo (5) both developed linear accelerator based radiosurgery systems. Although these early linac-based systems did allow the concentration of radiation dose there existed a question as to how accurately the radiation dose could be delivered to the targeted stereotactic coordinates. Winston and Lutz addressed this issue through the use of a stereotactically positioned phantom target system (6,7). It was found that linac-based systems could maintain the accuracy of radiation beam to target coordinates to within a millimeter or two (Fig. 1 and 2). While some felt that this accuracy was adequate, it fell short of the GammaKnife claim of 0.3 mm isocentric beam accuracy. In the late 1980s Friedman and Bova (8) developed an isocentric subsystem that enabled a routine linac to achieve an

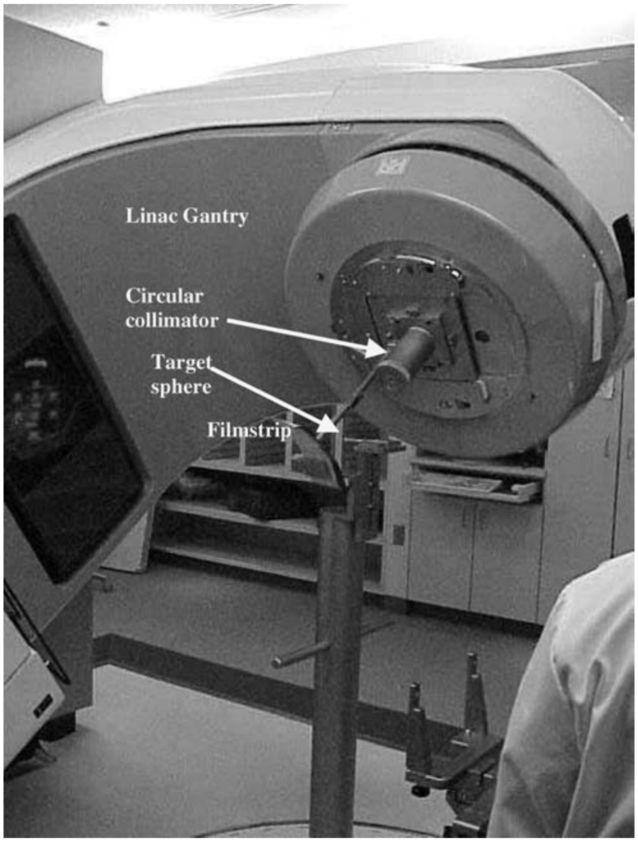


Figure 1. Modified Winston–Lutz test setup, for testing the spatial alignment of the circular radiosurgery X-ray beam with a spherical target ball. The test target sphere should be precisely aligned to the center of the X-ray field defined by the circular collimator. Several film exposures at different linac gantry rotation angles are taken.

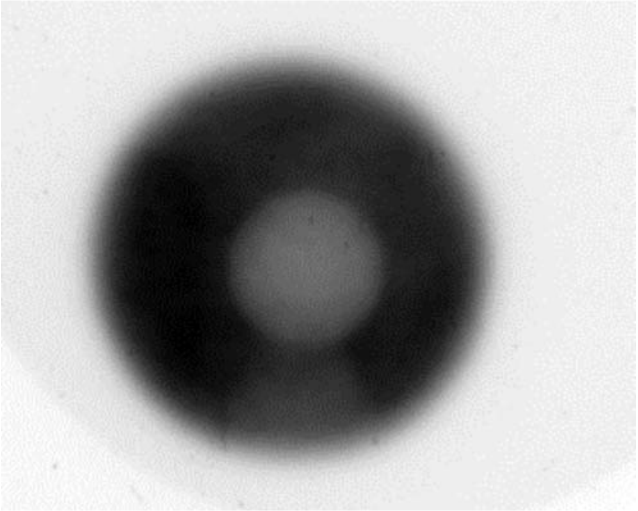


Figure 2. Image of a developed film from the modified Winston–Lutz test, showing the alignment between the circular radiosurgery X-ray beam with a spherical target ball. Analysis of this film image shows that in this case, the target sphere is misaligned from the center of the 20 mm diameter X-ray field by 1.06 mm in one direction, and by 0.97 mm in the other direction. Repeating this test with at least one pair of linac gantry angles gives an estimate of the spatial error inherent in the treatment delivery system.

isocentric beam accuracy of 0.2 mm, thereby matching the accuracy of the GammaKnife.

All early radiosurgery systems used a similar delivery method, namely multiple circular cross-section radiation beams converging to a common point, called the isocenter, located in the center of the target, volume with the directions chosen to minimize the overlap of beams outside the target and hence the normal tissue dose. This scheme worked well for spherical targets using a single isocenter. Non-spherical targets required the use multiple circular collimator diameters focused at multiple isocenters distributed throughout the target volume in an effort to "fill" the volume with dose. The ability to properly select the optimal set of sphere as well as their spacing and weighting were provided by treatment planning systems specially designed to optimize radiosurgical planning.

The next level of advance occurred in the 1990s when new computer controlled collimation devices known as multileaf collimators (MLC) were introduced that were coupled to stereotactic treatment planning software and delivery hardware designed for these new devices. While the early attempts at using MLCs had problems matching the dose conformality and steep dose gradients achieved by multiple isocentric circular collimation techniques, they nevertheless allowed complex targets to be treated more rapidly. Recently, new techniques have been developed that allow both the conformality of multiple isocenters as well as the speed of MLC delivery (9,10).

The majority of medical linear accelerators use microwave radiation in the S band to accelerate electrons and produce X rays. During the 1990s a compact medical X band linear accelerator was developed by Accuray, Inc. Adler placed this X ray source on an industrial robot gantry to create a novel stereotactic radiosurgery system called the CyberKnife (11,12). This system based its stereotactic targeting on an integrated orthogonal X-ray system that performed real-time imaging and correction of beam orientation to compensate for patient motion during treatment. Although this method of targeting was novel in the early 1990s, targeting systems have since been introduced for use on S band medical linacs. Unlike the GammaKnife, which by design is limited to intracranial targets, the CyberKnife can be applied to targets anywhere in the body.

Although radiosurgery based on gamma and X ray sources predominate, the theoretical advantages of proton therapy beams have stimulated great interest in advancing the use of protons to treat intracranial tumors. The physics of proton beam interactions are quite different than those of a photon beam. Unlike X rays, protons have mass and charge that result in a finite range of penetration. Additionally, the density of ionization (linear energy transfer) along the track of a proton beam is greater than that of an X-ray beam, with a region of high ionization density at the end of the track known as a Bragg peak (13,14). The finite range of penetration results in zero radiation dose beyond the Bragg peak, which in theory further allows the concentration of radiation dose to a deep-seated target while sparing underlying radiosensitive structures (15–17). The theoretical advantage of the proton beam Bragg peak is somewhat tempered by practical issue that its width is usually not large enough to encompass an entire radiosurgery target, so that it becomes necessary to superimpose a multitude of proton beams of varying energies to produce a composite depth dose distribution that covers the entire target (16). Historically, proton facilities produced only stationary beams, making it very difficult to bring multiple converging beams upon the patient's lesion. More recently rotating gantry delivery systems for protons have been introduced, which offer more flexibility in selecting beam orientations (18). Nevertheless the high cost (>$50M) of these facilities currently limits their availability to a few large metropolitan centers.

Early stereotactic targeting systems relied upon orthogonal radiographs for target localization, however, stereotactic procedures did not gain wide acceptance until the late 1980s with the development of three-dimensional (3D) treatment planning based on the use of computerized tomography (CT) imaging to obtain a 3D model of the patient. By the 1990s CT based target definition was augmented with magnetic resonance (MR) imaging that provided superior anatomical definition of the central nervous system. Because of difficulties related to MR incompatibility of stereotactic head fixation systems, MR imaging is performed without such hardware and the image set aligned or fused with CT images obtained with head fixation.

Initially, all intracranial stereotactic procedures used a rigid stereotactic head ring, or frame, screwed into the patient's skull to achieve a rigid, reproducible geometry for CT imaging and treatment. While frame-based procedures are still the most precise radiosurgery method they obviously are invasive to the patient and place a time limit on the completion of the procedure within hours of the frame placement. Noninvasive frameless head fixation systems were subsequently developed to address these issues. While less precise then ring-based approaches, they can be used in certain radiosurgery procedures where extreme accuracy is not required, such as treatment of brain metastases that are not located near critical structures like the brain stem or optical apparatus. Some of these systems are based upon the fitting of patients with thermoplastic face masks (19), while other systems separate the fixation and localization processes through the used of biteplates and thermoplastic masks (20,21).

Extracranial stereotactic targeting of lesions outside of the skull has been made possible through the development of a number of new technologies. One of the first was the use of ultrasound to allow the clinician to obtain a two-dimensional (2D) or 3D image set of the patient in position for radiosurgery (22). The ultrasound probe is tracked during image acquisition and the image voxels are mapped to a reference that allows precise targeting of the radiation beam. To enhance the ability to target tissues these scans are often registered to pretreatment CT and MR scans. Other methods involving fixed stereotactic X-ray tubes with image intensifiers have been developed by Accuray (23) and BrainLab (24). These systems function by obtaining either orthogonal or stereoscopic radiographs that are registered to the projection of a previously obtained 3D CT dataset. While these planar X-ray localization methods work well for bony anatomy the poorer contrast of soft tissues makes them less useful for localizing targets that are not rigidly fixed to bone. More recently the development of large format amorphous silicon detectors has facilitated the development and integration of cone beam CT scanning systems onto medical linear accelerators allowing stereotactic localization and registration of soft tissue anatomy to the linear accelerator's reference coordinate system. These new units have promise of providing unprecedented targeting accuracy to extra cranial targets.

STEREOTACTIC IMAGING AND LOCALIZATION

The uncertainty inherent to the imaging modality used can be the largest source of uncertainty the radiosurgery process. Poor imaging techniques increase this uncertainty and nullify the efforts to improve accuracy in treatment planning and delivery. Therefore, it is important to understand stereotactic imaging techniques, the increased quality assurance demands that are placed on the diagnostic imaging apparatus used, and the inherent limitations associated with each modality. Following are brief explanations of the three stereotactic imaging techniques used in radiosurgery: computed tomography, magnetic resonance imaging, and angiography.

Computed tomography is the primary modality used for radiosurgery treatment planning due to its spatial accuracy and electron density information that are both useful for accurate dose calculation and targeting. Stereotactic CT images can be obtained with a CT-compatible localizer attached to the stereotactic head ring such as the Brown–Roberts–Wells (BRW) (Fig. 3) or other commercially available designs (25,26). Since the geometry of the localizer is known relative to the head ring, stereotactic coordinates of any point in a volumetric CT image set may be accurately calculated, using the localizer fiducial markers in each axial image (Fig. 4). For example, the characteristic N-shaped rods of the BRW localizer allow the *x–y–z* coordinates of any point in space to be mathematically determined relative to the head ring rather than relying on the CT coordinates. This method provides more accurate spatial localization, and minimizes the CT scanner quality assurance requirements. In order to minimize the inaccuracies associated with the stereotactic imaging, it is important to obtain all imaging studies with the best available spatial resolution. Typically this means reducing the uncertainty to < 1 mm by acquiring CT images at an

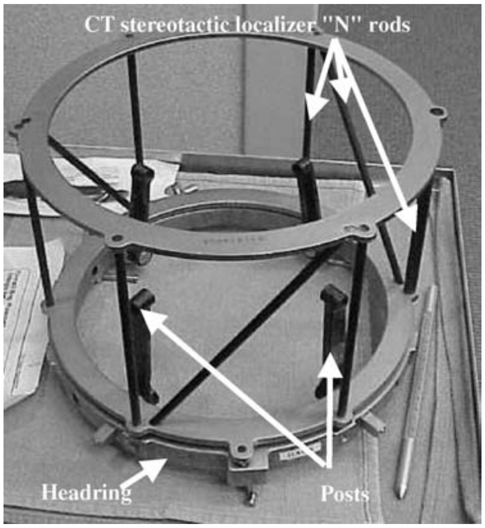


Figure 3. Computed tomography localizer attached to frame.

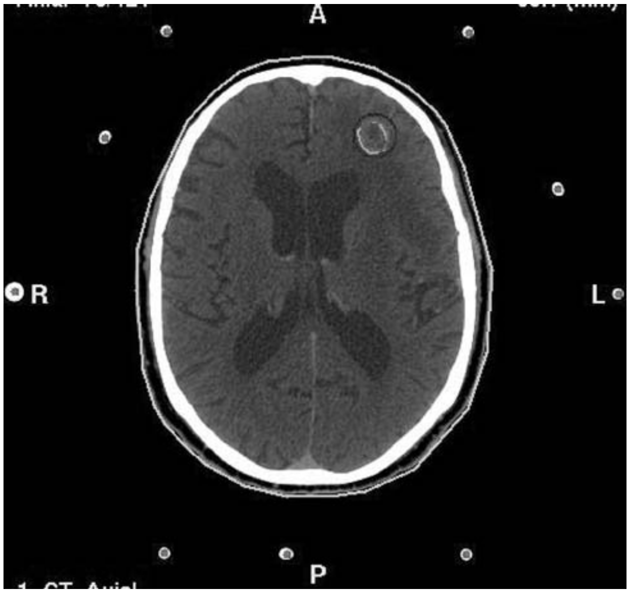


Figure 4. Axial CT image of patient with BRW stereotactic headframe and CT localizer attached. There are three sets of N-shaped rods; the location of any point in a CT image that contains all nine CT localizer rods, can be computed using the interrod distances to precisely define the plane of the image with respect to the BRW headframe and its coordinate system.

image resolution and slice spacing less than this amount. For example, a minimum 34.5 cm field of view is just large enough to image all of the stereotactic fiducials of a BRW localizer. This FOV corresponds to a pixel size of 0.67 mm for a 512 × 512 image matrix. In addition, current multislice diagnostic helical CT scanners can obtain CT images at 0.5–1 mm splice spacing.

Magnetic resonance (MR) imaging often provides superior tumor visualization, but spatial distortion inherent in the MR images due to magnetic field non-uniformities and patient-specific artifacts, and secondarily the lack of electron density information makes the use of MR images less desirable than CT images for radiotherapy dose calculations. Introducing a stereotactic frame and localizer into an MR imager will perturb the magnetic field producing image distortions on the order of 0.7–4 mm in each orthogonal plane (axial, sagittal, coronal) of a stereotactic MRI (27,28). Furthermore the size of the stereotactic head frame may be incompatible with the geometry of standard MRI head coil necessitating the use of a larger MRI coil, such as the standard body coil, with a consequent degradation in image quality due to a reduced signal/noise ratio. These problems are overcome by eliminating the head frame during the MR imaging procedure and using image fusion techniques to register the MR image volume to the CT image volume of the patient in the head frame. For frame-based radiosurgery, the 3D volumetric MR scan is acquired prior to head ring placement using a pulse sequence that allows a fast image acquisition to minimize image distortion due to patient movement. All currently available image correlation routines consider the MR images as rigid bodies, and do not remove local image distortions that can exist in the MR data, hence careful review of the coregistered MRI and CT image sets is essential. This comparison should focus on internal anatomy, such as the ventricles, tentorium, sulci. and avoid the external contour since it can be shifted 3–4 mm due to the fat shift (a distortion resulting in the difference in the resonant frequency of protons in fat relative to their resonant frequency in water).

The third imaging modality important to radiosurgery, angiography, is used for diagnosis and anatomic characterization of cerebral arteriovenous malformations (AVMs). Unlike volumetric CT and MR tomography, stereotactic planar angiography utilizes a set of orthogonal radiographs of a special localizer attached to the stereotactic head ring bearing radio-opaque fiducials. The stereotactic coordinates of any point within the localizer may be calculated very accurately since the geometry of the fiducials is known relative to the head ring. The orthogonal film pair is obtained with contrast injected rapidly at the location of the AVM nidus allowing excellent visualization of fine vasculature and fiducials.

The use of orthogonal images as the sole localization method for treatment planning is inadequate for accurately determining the shape, size and location of an arbitrarily shaped AVM nidus (29–31). Furthermore, overlapping structures, such as feeding or draining blood vessels, may obscure the view of the AVM nidus and will result in unnecessary irradiation of normal tissue if these blood vessels are included in the targeted volume. Because of these issues a volumetric CT angiography image dataset (1 mm slice thickness; intravenous contrast infused at a rate of 1 $cm^3 \cdot s^{-1}$) is always acquired in addition to, or in replacement of, stereotactic angiography. The resultant CT images provide an accurate 3D description of the AVM nidus, along with the feeding and draining vessels.

RADIOSURGERY DELIVERY TECHNIQUES

Numerous radiosurgery techniques have been devised based noncoplanar configurations static beams or arcs. The majority of radiosurgery treatments use circular collimators to create spherical regions of high dose. The classic example of a static beam delivery system is the GammaKnife unit, which consists of 201 narrow-beam cobalt-60 sources arranged on a hemisphere. A collimation helmet containing 201 circular collimators, each of the same diameter, is placed between the hemisphere of sources and the patient's head with the collimator's focal point centered on the intracranial target. This produces a spherical dose distribution or shot in GammaKnife parlance. An irregular volume is treated with multiple shots whose diameters are selected based on the available helmet collimators (32). The CyberKnife robotic radiosurgery unit is also used in a similar manner to deliver treatments from fixed beam orientations using a circular collimator.

Alternatively, a conventional medical linear accelerator can be outfitted with a circular collimator and multiple (5–9) noncoplanar arc delivery used to achieve a spherical dose distribution. When used with linear accelerators, the circularly collimated beam is rotated around the target at isocenter by moving the gantry in arc mode while the patient and treatment couch are stationary, producing a parasagittal beam path around the target. Betti and Derichinsky developed their linac radiosurgery system with a special chair, the Betti chair, which moved the patient in a side to side arc motion under a stationary linac beam, and which produced a set of para-coronal arcs (4). With modern, computer controlled linear accelerators, more complex motions other than these simple arcs are possible. The Montreal technique, which involves synchronized motion of the patient couch and the gantry while the radiation beam is on, is an example of this, producing a baseball seam type of beam path (33). The rationale of using arcs with circular collimators is to concentrate radiation dose upon the target, while spreading the beam entrance and exit doses over a larger volume of nontarget tissue, theoretically reducing the overall dose and toxicity to nontarget tissue.

Radiosurgery based on circular collimators produces a spherical region of high dose with steep falloff, or gradient, that is adequate for spherical targets. Irregular target volumes require the use of multiple spheres, or isocenters, abutted together to conform the dose more closely to the shape of the target so as to minimize nontarget tissue dose (34). A consequence of the multiisocenter approach is that the shape of the total dose distribution is very sensitive to the abutment of the spherical dose distributions due to their steep falloff. For this reason, it is common practice to accept 30% or greater dose variation over the irregular target volumes using circular collimation.

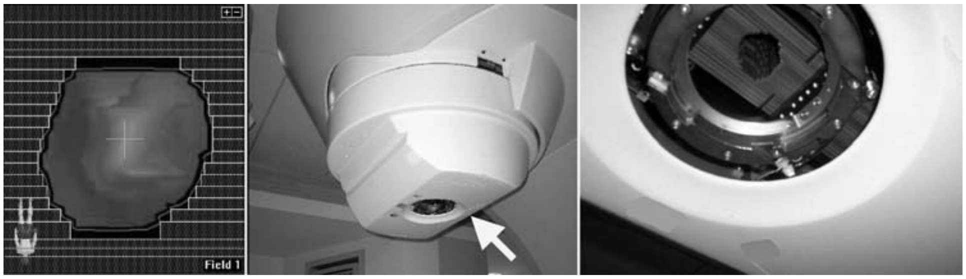


Figure 5. Beam's eye view showing target shape. Instead of constructing a custom block for the continuous shape of the target, at left a MLC (narrow rectangles) approximates the shape of the conformal beam. Each rectangle represents a tungsten leaf which moves left and right across the field of view shown under computer control. In this example, the MLC leaves would remain stationary while the treatment beam is on, providing a dose distribution very similar to a custom block. At center, the position of the MLC (arrow) on the linac gantry is shown; X rays emerge from the MLC-shaped aperture. At right, close-up photograph of the actual MLC, whose leaves are shaped to the field shown in the left image.

The linear accelerator offers additional flexibility in that tertiary computer-controlled multileaf collimators (MLCs) may be used to produce noncircular beams and beams with nonuniform intensity profiles that conform dose distributions more closely to irregular target volumes with less dose non uniformity. The most common type of MLC consists of two banks of opposed high density metal plates, or leaves that can be moved in a plane perpendicular to the beam's direction. The MLC can be rotated with the treatment machine's collimator in order to align the leaves for the best fit to the target's projected shape. The simplest use of an MLC is simply as a functional replacement for custom made beam shaping blocks, in which the rectangular MLC edges are used to approximate a continuous target outline shape (Fig. 5) (35). This field shaping can be used for either static field treatments or for dynamic arcs in which the MLC shape is continually changed to match the beam's-eye-view projection of the target volume. Moss investigated the efficacy of performing radiosurgery treatments with a dynamically conforming MLC in arc mode, and concluded that dynamic arc MLC treatments offered target coverage and normal tissue sparing comparable to that offered by single and multiple isocenter radiosurgery (36). Nedzi (37) showed that even crude beam shaping devices offered some conformal benefit over single isocenter treatments with circular collimators. Since the mid- to late-1990s, the use of miniature multileaf collimators (MLCs with a leaf width projected to isocenter of 5 mm or less) has become increasingly common.

However, the MLC may be used in a more sophisticated fashion to form many different beam shapes of arbitrary size and intensity (by varying the amount of radiation applied through each beam aperture). In this manner, radiation fields with a similar dose profile as a shaped, wedged field may be delivered using only the computer-controlled MLC, shall can also deliver intensity modulated dose profiles similar to those achievable using custom beam compensators, but without the disadvantages of fabrication time or of needing to manually change a physically mounted beam filter between each treatment field (38). Thus, a computer-controlled MLC and treatment machine offer the potential to deliver more sophisticated radiation treatments to each patient with the same time and cost resources available.

The MLC-based solutions are available for both static multiple beams and arc-based delivery. Radionics initially introduced the use of a mini-MLC for defining static beam shapes that conformed to the projected shape of the target volume in the beam's eye view (35,39,40). The device consisted of multiple thin plates, or MLC leaves, that were mechanically clamped together to form an irregular beam shape defined by a plastic template corresponding to the projected shape of the target. Subsequent developments by other vendors added computer-controlled motorization to the leaves so that treatments could be carried out more efficiently. While most mini-micro-MLC implementations were based on static delivery of few fixed beams, NOMOS, Inc. and 3D Line, Inc. developed specialized arc-based intensity-modulated radiosurgery (IMRS) systems. Most, if not all, tertiary MLC vendors have now developed integrated treatment planning systems designed specifically for their MLCs and treatment applications, including IMRS.

The potential for improvement presented by some of these newer and more sophisticated treatment delivery methods has spurred interest in their evaluation relative to the more traditional linac SRS methods of multiple intersecting arcs and circular collimators. These studies are usually conducted by those who have had difficulty achieving the conformality routinely published by those experienced in multiisocenter planning. These comparisons generally demonstrate that for small to medium (up to $\sim 20\ \mathrm{cm}^3$) intracranial targets multiple static beams offer conformity with ratios of normal tissue to target tissue treatments in the range of 1.5–2.0, with target dose homogeneity on the order of 10–20%, while offering a more standard radiation therapy treatment planning interface and process (39,41–43). These studies go on to show that static beam IMRT techniques generally performed comparable to or better than static beam plans, usually increasing the dose homogeneity and possibly conformality (44,45).

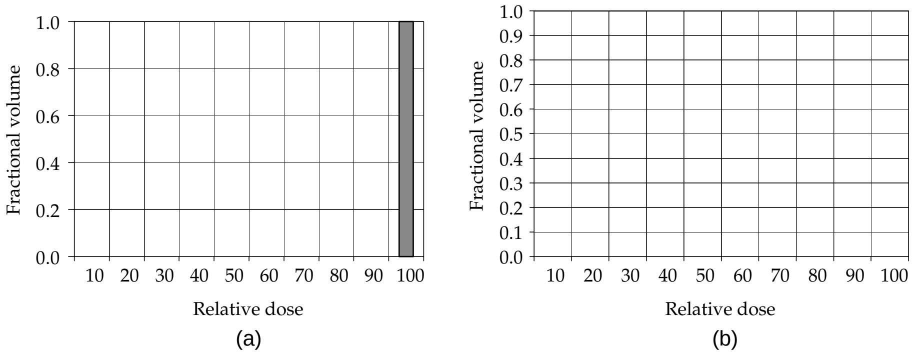


Figure 6. Ideal target (a) and nontarget volume (b) direct DVHs. Note that in the ideal direct DVH of the nontarget volume (right side), the plot is empty, since there is no nontarget volume receiving any dose in the ideal case.

A potential problem with these comparison studies is that they may not equitably compare the full potential of multiple isocenter radiosurgery with circular collimators. A qualitative inspection of the multiple isocenter dosimetric results shown in these comparisons leads one to suspect that in many cases, suboptimal multiple isocenter plans are being compared with reasonably optimized static beam and dynamic MLC arcs–IMRT plans. Although the multiple isocenter treatment plans in these comparisons in the literature may represent a level of plan quality achievable by an average or unfamiliar user, they do not always represent the experience of expert users. Some expert users have reported on the use of multiple isocenter–circular collimator radiosurgery systems to plan and deliver tightly conformal dose distributions to irregularly shaped targets near radiosensitive structures, while maintaining a sharp dose gradient away from the target toward radiosensitive structures (34,46–48).

TOOLS FOR EVALUATING RADIOSURGERY TREATMENT PLANS

The clinical objective of radiosurgery is to deliver a tumorcidal radiation dose to a target volume while minimizing the dose to surrounding tissues. The following tools are available to the treatment planner to evaluate a 3D dose distribution in order to quantify the degree to which this objective is achieved: (*1*) 2D isodose curves and 3D isodose surfaces, (*2*) dose–volume histograms, and (*3*) physical dose–volume figures of merit. The following sections explain the use of each of these tools in radiation therapy and radiosurgery treatment planning.

It is possible to display 3D semitransparent surface renderings of constant dose levels overlaid on 3D renderings of the target volume to determine if the target adequately covered, but these can be difficult to analyze quantitatively. For this reason, 2D cross-sections of the 3D dose distribution are evaluated making it easier to quantitative assess target coverage. The 2D dose cross-sections are displayed as isocontour plots (isodose plots) overlaid on the patient's CT and MR images to allow visual assessment of dose coverage to an accuracy of within one image pixel. Although this implies submillimeter precision, the 1 pixel uncertainty in isodose position can result in a large uncertainty in dose coverage for small intracranial targets. In the case of a 0.67 mm pixel, a 20 mm sphere, equal to 4.2 cm^3, would apparently be equally well covered by an isodose surface ranging in volume from 3.8 to 4.6 cm^3 corresponding to a 10% uncertainty in volume. Hence, although visual inspection of isodose plots on multiple images is commonly performed, it is cumbersome and there is a large uncertainty in assessing the dose coverage that is associated representing small targets using finite size pixels.

One commonly used solution to this problem is to use dose–volume histograms (DVHs). DVHs are a method of condensing 3D dose information into a more manageable form for analysis. The simplest type of DVH is a differential histogram of volume versus dose (49). This is simply a histogram showing the number of occurrences of each dose value within a 3D volume. A second more common representation is the cumulative DVH, which is the integral of the differential DVH as a function of dose. Unfortunately, in either type of DVH, the spatial information of which specific volumes are exposed to each dose level is lost in the process of constructing a DVH. For this reason, DVHs are generally used clinically in conjunction with the evaluation of multiple isodose plots as mentioned earlier.

The ideal treatment planning situation is one in which the target volume receives a uniform dose equal to the maximum dose, and the nontarget volume receives zero dose. This would correspond to ideal differential DVHs for target and nontarget volumes as shown in Fig. 6. Clinically realistic differential DVHs for target and nontarget volumes for a more typical (non-ideal) radiosurgery dose distribution are shown in Fig. 7. Figure 8 shows two differential DVHs from competing radiosurgery plans plotted on a common axes to allow a direct comparison of the plans. Note that it can difficult to evaluate competing plans using such differential histograms (50), as demonstrated in Fig. 8. Above ~ 40 units of dose, both plans appear to be identical, but the two plans expose differing volumes of brainstem at doses less than ~ 40 units. For this reason cumulative DVH analysis is more commonplace. Transforming the differential DVHs into cumulative DVHs, by plotting the volume receiving at least a certain dose versus dose, makes it simpler to evaluate the differences in the dose distributions, as shown in Fig. 9.

Optimal cumulative DVH curves for target structures will be as far toward the upper right hand corner of the plot as possible, while the those for nontarget structures will be as close as possible to the lower left hand corner of the

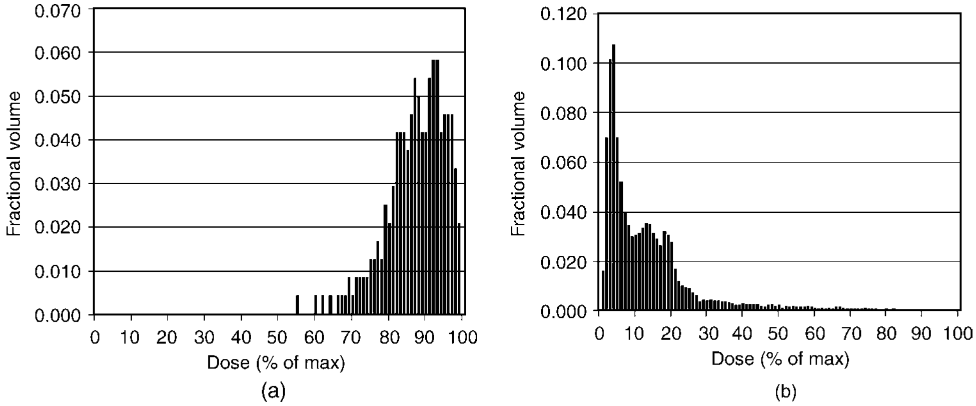


Figure 7. Typical (nonideal) radiosurgery direct DVHs for target volume (a) and nontarget volume (b).

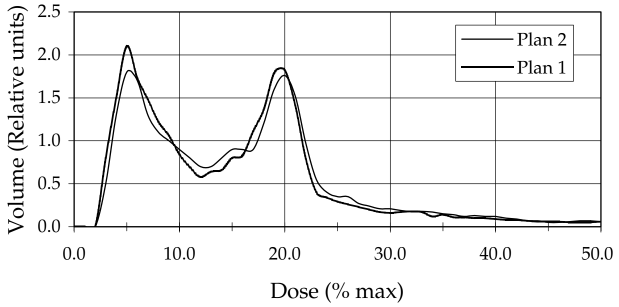


Figure 8. Direct DVHs for a radiosensitive nontarget structure in two hypothetical treatment plans.

plot as shown in Fig. 10. Considering the two completing plans shown in Fig. 9, the better plan will have its target cumulative DVH further to the upper right corner and its nontarget cumulative DVH further to the lower left corner than the poorer plan. Plan 1 is the preferred plan, since its nontarget DVH for the brainstem lies below and to the left of that for Plan 2. The relative ease of this comparison underscores the general utility of cumulative DVHs over differential DVHs (49,51). Unfortunately, it is rare for the cumulative DVHs of rival treatment plans to separate themselves from one another so cleanly. Typically, the DVHs cross one another, perhaps more than once as shown in Fig. 11. The simple rules for evaluating DVHs cannot resolve this situation, in which case other means must be used to evaluate the treatment plans by applying a score to each plan derived from a clinically relevant figure of merit.

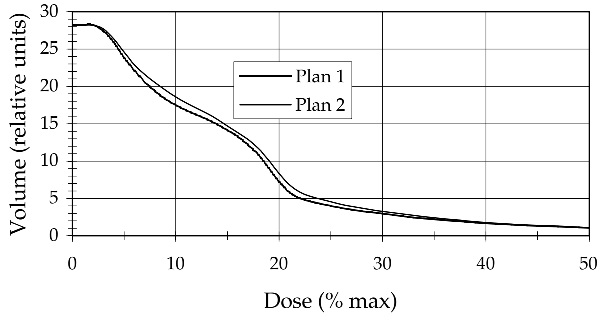


Figure 9. Cumulative DVH plot of the direct DVH data shown in Fig. 8.

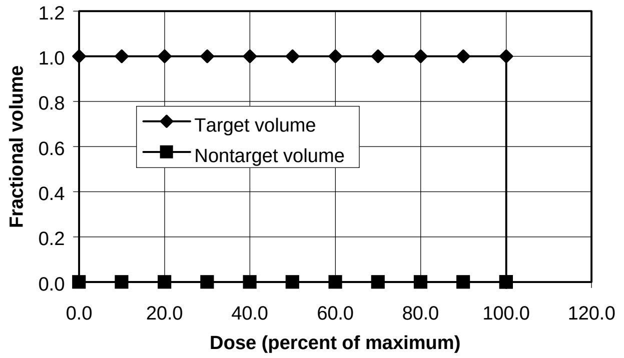


Figure 10. Ideal cumulative DVH curve for target and nontarget volumes.

The three properties of radiosurgery dose distributions that have been correlated with clinical outcome and that lend themselves to clinical figures of merit are (*1*) dose conformity, (*2*) dose gradient, and (*3*) dose homogeneity (34). The conformity of the dose distribution to the target volume may be simply expressed as the ratio of the prescription isodose volume to the target volume, frequently referred to as the PITV ratio (52).

$$\text{PITV} = \text{Prescription isodose volume/target volume} \quad (1)$$

Perfect conformity of a dose distribution to the target, that is, PITV = 1.00, is typically not achievable, and some

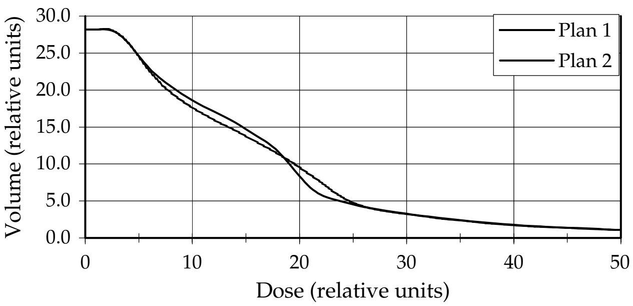


Figure 11. Crossing cumulative DVH curves.

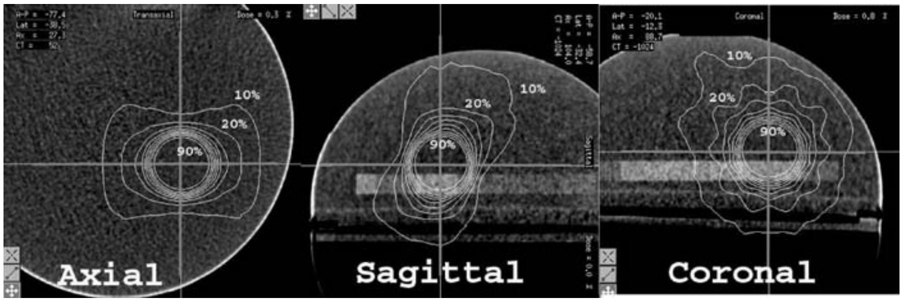


Figure 12. Transaxial, sagittal, and coronal isodose distributions for five arcs of 100° each delivered with a 30 mm collimator. Isodose lines in each plane increase from 10 to 90% in 10% increments, as indicated. The isocenter is marked with cross-hairs.

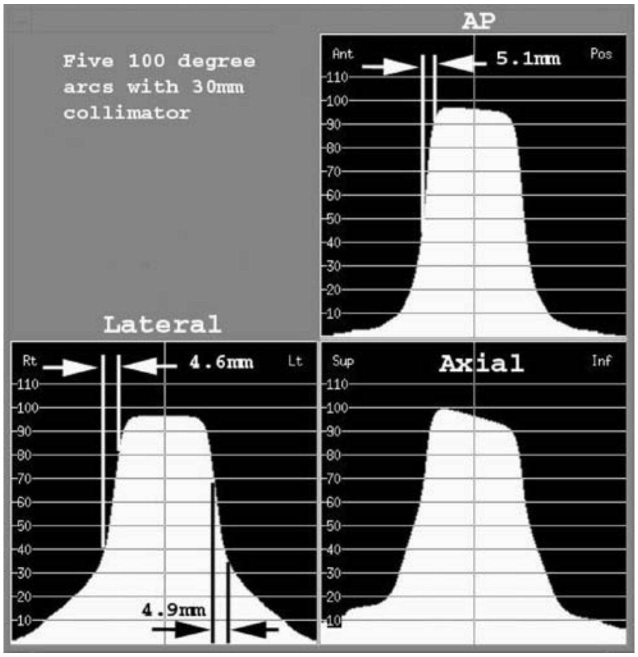


Figure 13. Dose cross-plots through the isocenter, corresponding to the isodose distributions shown in Fig. 12. The sharpest dose fall-off, from dose D to half-dose 0.5D, occurs between dose D of 80% to 0.5D = 40%, which occurs in a distance of 4.6 mm. The D to 0.5D fall-off distance is larger for 90–45% (5.1 mm) and for 70–35% 4.9 mm) doses.

volume of nontarget tissue must be irradiated to the same dose level as the target, resulting in PITV ratios greater than unity. The most conformal treatment plans are those with the lowest PITVs, if all of the plans under comparison provide equivalent target coverage. This stipulation is necessary because the definition of PITV does not specify how the prescription isodose is determined. It is possible (but undesirable) to lower, and thus improve, the PITV by selecting an isodose level that incompletely covers the target as the prescription isodose, and therefore reduces the numerator of Eq. 1. Many investigators report isodose shells that cover in the neighborhood of at least 95% of the target volume or 99% of the target volume (17,34,46,48,53–57). This ensures a more consistent basis of comparisons for all treatment plans.

A sharp dose gradient (fall off in dose with respect to distance away from the target volume) is an important characteristic of radiosurgery and stereotactic radiotherapy dose distributions. Dose gradient may be characterized by the distance required for the dose to decrease from a therapeutic (prescription) dose level to one at which no ill effects are expected (half prescription dose). For illustrative purposes, a typical dose distribution in a hemispherical water phantom for a single isocenter delivered with five converging arcs and a 30 mm collimator is depicted in Fig. 12. A quantitative measure of gradient is obtained from examining the dose profiles along orthogonal directions in the principal anatomical planes (transaxial, sagittal, and coronal), as shown by cross-plots in Fig. 13. As in this example the steepest dose gradient (4.6 mm) occurs between the 80% and 40% isodose shells, and for this reason single isocenter dose distributions are prescribed to the 80% isodose shell (34). Table 1 lists dose gradient information between the 80 and 40% isodose shells for single isocenter spherically symmetric dose distributions with 10–50 mm diameter collimators.

A method has been proposed that uses easily obtainable DVH information to generate a numerical measure, or score, of the overall dose gradient for evaluating the dose conformality of radiosurgery dose distributions. The Conformity–Gradient Index score, or CGIg, has been proposed as a metric for quantifying dose gradient of a stereotactic treatment plan (58,59). From treatment planning experience at the University of Florida, it has been observed that it is possible to achieve a dose distribution that decreases from the prescription dose level to half of prescription dose in a distance of 3–4 mm away from the target. Taking this as a guide, a gradient score CGIg may be computed as

$$\mathrm{CGIg} = 100 - \{100 \times [(R_{\mathrm{eff,50\%Rx}} - R_{\mathrm{eff,Rx}}) - 0.3\,\mathrm{cm}]\} \quad (2)$$

where $R_{\mathrm{eff50\%Rx}}$ is the effective radius of the half-prescription isodose volume, and R_{effRx} is the effective radius of the prescription isodose volume. The effective radius of a

Table 1. Single Isocenter (Five Converging Arcs) Dose–Volume and Gradient Information for 10–50 mm Diameter Circular Collimators

Coll.	$V_{80\%}$, cm^3	$R_{\mathrm{eff80\%}}$, mm	$V_{40\%}$, cm^3	$R_{\mathrm{eff40\%}}$, mm	Eff. Gradient, mm	CGIg
10	0.3	4.2	1.2	6.7	2.4	106
20	3.9	9.8	9.7	13.2	3.5	95
30	13.9	14.9	30.8	19.4	4.5	85
50	67.4	25.2	111.6	29.9	4.6	84

volume is the radius of a sphere of the same volume, so that R_{eff} for a volume V is given by

$$R_{\mathrm{eff}} = (3V/4\pi)^{-1/3} \tag{3}$$

The volumes of the prescription isodose shell and the half prescription isodose shell are obtained from a DVH of the total volume (or a sizeable volume that completely encompasses the target volume and a volume that includes all of the half prescription isodose shell) within the patient image dataset. The CGIg score is a dimensionless number that exceeds 100 for dose gradients < 3 mm (steeper falloff from prescription to half-prescription dose level), and which decreases < 100 as a linear function of the effective distance between the prescription and half-prescription isodose shells.

Dose conformity is another important characteristic of a radiosurgery treatment plan that should be considered in plan evaluation. The Conformity-Gradient Index (conformal), or CGIc, is defined as (58):

$$\mathrm{CGIc} = 100 \times (\mathrm{PITV})^{-1} \tag{4}$$

The CGIc converts PITV into a numerical score expressing the degree of conformity of a dose distribution to the target volume. The CGIg score increases as the dose gradient improves, and the CGIc score increases as dose conformity improves. Perfect conformity (assuming the target is adequately covered) of the prescription isodose volume to the target is indicated by a PITV = 1.00 and a CGIc = 100.

As dose gradient and dose conformity are both important parameters in judging a stereotactic radiosurgery or radiotherapy plan, an overall figure of merit for judging radiosurgery plans should incorporate both of these characteristics. Since clinical data to indicate the relative importance of conformity versus gradient is currently lacking, an index, the Conformity-Gradient Index (CGI) is proposed that assigns equal importance to both of these factors. The overall Conformity-Gradient Index score, or CGI, for a radiosurgery or radiotherapy plan is the average of the CGIc and CGIg scores:

$$\mathrm{CGI} = 0.5 \times (\mathrm{CGIc} + \mathrm{CGIg}) \tag{5}$$

A final measure of plan quality considered by some to be an important factor in evaluating treatment plans is dose homogeneity. While is a homogenous dose is desirable for conventional, fractionated radiotherapy (60), its role is less clear in radiosurgery. Several studies have associated large radiosurgical dose heterogeneity (maximum dose to peripheral dose ratio, or MDPD, > 2.0) with an increased risk of complications (61,62). However, some radiosurgeons have hypothesized that the statistically significant correlation between large dose inhomogeneities and complication risk may be associated with the relatively nonconformal multiple isocenter dose distributions with which some patients in these studies were treated, and not with dose inhomogeneity alone. One theory is that the extreme hot spots associated with large dose heterogenities may be acceptable, if the dose distribution is very conformal to the target volume and the hot spot is contained within the target volume. Nonconformal dose distributions could easily cause the hot spots to occur outside of the target, greatly increasing the risk of a treatment complication. The extensive successful experience of gamma unit treatments administered worldwide (almost all treatments with MDPD ≥ 2.0) lends support to this hypothesis (63). Therefore, as a general principle, one strives for a homogeoneous radiosurgery dose distribution, but this is likely not as important a factor as conformity of the high dose region to the target volume, or the dose gradient outside of the target.

SUMMARY

While the use of radiosurgery is now in its fifth decade the basic principles of dose prescription and delivery have changed very little from those first conceived by Leksell. The primary, and still most effective method to treat relatively small target tissues with a high dose and to maintain a very steep dose gradient is through the use of many beams that all converge on the target tissue and diverge along independent paths while approaching and leaving the target region. Other dose targeting and restriction techniques, such as intensity modulation, provide the ability to position beams that geometrically avoid tissue and potentially provide a more powerful tool for dose optimization. These techniques are often combined to allow for the best of both optimized dose planning and efficient dose delivery.

BIBLIOGRAPHY

Cited References

1. Bova FJ, Meeks SL, Friedman WA. Linac Radiosurgery: System Requirements, Procedures and Testing. In: Treatment Planning in Radiation Oncology. Khan FM, Potish RA, editors. Baltimore: Williams and Wilkins; 1998. p 215–241.
2. Hall EJ. Time, dose, and fractionation in radiotherapy. Radiobiology for the Radiologist. Philadelphia: J.B. Lipponcott; 1994. p 211–229.
3. Lindquist C. Gamma Knife Radiosurgery. Semin Radiat Oncol 1995;5(3):197–202.
4. Betti O, Derechinsky V. [Multiple-beam stereotaxic irradiation]. Neurochirurgie 1983;29(4):295–298.
5. Colombo F, et al. External stereotactic irradiation by linear accelerator. Neurosurgery 1985;16(2):154–160.
6. Winston K, Lutz W. Linear Accelerator as a Neurosurgical Tool for Stereotactic Radiosurgery. Neurosurgery 1988;22(3): 454–464.
7. Lutz W, Winston KR, Maleki N. A System for Stereotactic Radiosurgery with a Linear Accelerator. Int J Radiat Oncol Biol Phys 1988;14(2):373–381.
8. Friedman WA, Bova FJ. The University of Florida radiosurgery system. Surg Neurol 1989;32(5):334–342.
9. St John T, et al. Intensity-Modulated Radiosurgery Treatment Planning By Fluence Mapping Multi-isocenter Plans. Med Phys 2001;28(6):1256.
10. St John T, Wagner TH, BFJ, FWA, MSL. A geometrically based method of step and shoot stereotactic radiosurgery with miniature multileaf collimator. Phys Med Biol 2005;50: 3263–3276.
11. Adler JR Jr, et al. The Cyberknife: a frameless robotic system for radiosurgery. Stereotact Funct Neurosurg 1997;69(1–4 Pt. 2): 124–128.

12. Murphy MJ, Cox RS. The accuracy of dose localization for an image-guided frameless radiosurgery system. Med Phys 1996; 23(12):2043–2049.
13. Stanton R, Stinson D. Applied Physics for Radiation Oncology. Madison, (WI): Medical Physics Publishing; 1996. p 366.
14. Moyers MF. Proton Therapy. In The Modern Technology of Radiation Oncology. Van Dyk J, editor. Madison, (WI): Medical Physics Publishing; 1999. p 823–869.
15. Baumert BG, Lomax AJ, Miltchev V, Davis JB. A comparison of dose distributions of proton and photon beams in stereotactic conformal radiotherapy of brain lesions. Int J Radiat Oncol Biol Phys 2001;49(5):1439–1449.
16. Bussiere MR, Adams JA. Treatment planning for conformal proton radiation therapy. Technol Cancer Res Treat 2003;2(5): 389–399.
17. Verhey LJ, Smith V, Serago CF. Comparison of radiosurgery treatment modalities based on physical dose distributions. Int J Radiat Oncol Biol Phys 1998;40(2):497–505.
18. Chapman PH, Loeffler JS. Proton Radiosurgery. In Youman's Neurological Surgery. Winn HR, editor. Philadelphia: Saunders; 2004. p 4123–4130.
19. Willner J, Flentje M, Bratengeier K. CT simulation in stereotactic brain radiotherapy—analysis of isocenter reproducibility with mask fixation. Radiother Oncol 1997;45(1):83–88.
20. Bova FJ, et al. The University of Florida frameless high-precision stereotactic radiotherapy system. Int J Radiat Oncol Biol Phys 1997;38(4):875–882.
21. Meeks SL, et al. IRLED-based patient localization for linac radiosurgery. Int J Radiat Oncol Biol Phys 1998;41(2): 433–439.
22. Meeks SL, et al. Ultrasound-guided extracranial radiosurgery: technique and application. Int J Radiat Oncol Biol Phys 2003; 55(4):1092–1101.
23. Chang SD, et al. An analysis of the accuracy of the CyberKnife: a robotic frameless stereotactic radiosurgical system. Neurosurgery 2003;52(1):140–146; discussion 146–147.
24. Yan H, Yin FF, Kim JH. A phantom study on the positioning accuracy of the Novalis Body system. Med Phys 2003;30(12): 3052–3060.
25. Brown RA. A stereotactic head frame for use with CT body scanners. Invest Radiol 1979;14(4):300–304.
26. Saw CB, Ayyangar K, Suntharalingam N. Coordinate transformations and calculation of the angular and depth parameters for a stereotactic system. Med Phys 1987;14(6):1042–1044.
27. Burchiel KJ, Nguyen TT, Coombs BD, Szumoski J. MRI distortion and stereotactic neurosurgery using the Cosman-Roberts-Wells and Leksell frames. Stereotact Funct Neurosurg 1996;66(1–3):123–136.
28. Kitchen ND, Lemieux L, Thomas DG. Accuracy in frame-based and frameless stereotaxy. Stereotact Funct Neurosurg 1993;61(4):195–206.
29. Spiegelmann R, Friedman WA, Bova FJ. Limitations of angiographic target localization in planning radiosurgical treatment. Neurosurgery 1992;30(4):619–623; discussion 623–624.
30. Bova FJ, Friedman WA. Stereotactic angiography: an inadequate database for radiosurgery? Int J Radiat Oncol Biol Phys 1991;20(4):891–895.
31. Blatt DR, Friedman WA, Bova FJ. Modifications based on computed tomographic imaging in planning the radiosurgical treatment of arteriovenous malformations. Neurosurgery 1993;33(4):588–595; discussion 595–596.
32. Maitz AH, Wu A. Treatment planning of stereotactic convergent gamma-ray irradiation using Co-60 sources. Med Dosim 1998;23(3):169–175.
33. Wasserman TH, Rich KM, Drzymala RE, Simpson JR. Stereotactic irradiation. In: Principles and Practice of Radiation Oncology. Perez CA, Brady LW, editors. Philadelphia: Lippencott-Raven; 1996. p 387–404.
34. Meeks SL, et al. Treatment planning optimization for linear accelerator radiosurgery. Int J Radiat Oncol Biol Phys 1998;41(1): 183–197.
35. Brewster L, et al. Three dimensional conformal treatment planning with multileaf collimators. Int J Radiat Oncol Biol Phys 1995;33(5):1081–1089.
36. Moss DC. Conformal stereotactic radiosurgery with multileaf collimation. In Nuclear Engineering Sciences. Gainesville, (FL): University of Florida; 1992.
37. Nedzi LA, et al. Dynamic field shaping for stereotactic radiosurgery: a modeling study. Int J Radiat Oncol Biol Phys 1993; 25(5):859–869.
38. Sternick ES, Carol MP, Grant W. Intensity-modulated radiotherapy. In: Treatment Planning in Radiation Oncology. Khan FM, Potish RA, editors. Baltimore: Williams and Wilkins; 1998. p 187–213.
39. Shiu AS, et al. Comparison of miniature multileaf collimation (MMLC) with circular collimation for stereotactic treatment. Int J Radiat Oncol Biol Phys 1997;37(3):679–688.
40. Leavitt DD. Beam shaping for SRT/SRS. Med Dosim 1998; 23(3):229–236.
41. Laing RW, et al. Stereotactic radiotherapy of irregular targets: a comparison between static conformal beams and non-coplanar arcs. Radiother Oncol 1993;28(3):241–246.
42. Cardinale RM, et al. A comparison of three stereotactic radiotherapy techniques; ARCS vs. noncoplanar fixed fields vs. intensity modulation. Int J Radiat Oncol Biol Phys 1998; 42(2): 431–436.
43. Hamilton RJ, et al. Comparison of static conformal field with multiple noncoplanar arc techniques for stereotactic radiosurgery or stereotactic radiotherapy. Int J Radiat Oncol Biol Phys 1995;33(5):1221–1228.
44. Woo SY, et al. A comparison of intensity modulated conformal therapy with a conventional external beam stereotactic radiosurgery system for the treatment of single and multiple intracranial lesions. Int J Radiat Oncol Biol Phys 1996;35(3): 593–597.
45. Kramer BA, et al. Dosimetric comparison of stereotactic radiosurgery to intensity modulated radiotherapy. Radiat Oncol Investig 1998;6(1):18–25.
46. Meeks SL, et al. Potential clinical efficacy of intensity-modulated conformal therapy. Int J Radiat Oncol Biol Phys 1998; 40(2):483–495.
47. Meeks SL, et al. Linac scalpel radiosurgery at the University of Florida. Med Dosim 1998;23(3):177–185.
48. Wagner T, et al. A Geometrically Based Method for Automated Radiosurgery Planning. Int J Radiat Oncol Biol Phys 2000; 48(5):1599–1611.
49. Lawrence TS, Kessler ML, Ten Haken RK. Clinical interpretation of dose-volume histograms: the basis for normal tissue preservation and tumor dose escalation. Front Radiat Ther Oncol 1996;29:57–66.
50. Drzymala RE, et al. Dose-volume histograms. Int J Radiat Oncol Biol Phys 1991;21(1):71–78.
51. Kutcher GJ, Jackson A. Treatment plan evaluation. In: Treatment Planning in Radiation Oncology. Khan FM, Potish RA, editors. Baltimore: Williams and Wilkins; 1998 p 281–294.
52. Shaw E, et al. Radiation Therapy Oncology Group: radiosurgery quality assurance guidelines. Int J Radiat Oncol Biol Phys 1993;27(5):1231–1239.
53. Wagner T, et al. Isotropic beam bouquets for shaped beam linear accelerator radiosurgery. Phys Med Biol 2001;46(10): 2571–2586.
54. Wagner TH. Optimal delivery techniques for intracranial stereotactic radiosurgery using circular and multileaf collimators. Nuclear and Radiological Engineering. Gainesville, (FL): University of Florida; 2000. p 306.

55. Tome WA, et al. A high-precision system for conformal intracranial radiotherapy. Int J Radiat Oncol Biol Phys 2000;47(4): 1137–1143.
56. Tome WA, et al. Optically guided intensity modulated radiotherapy. Radiother Oncol 2001;61(1):33–44.
57. Smith V, Verhey L, Serago CF. Comparison of radiosurgery treatment modalities based on complication and control probabilities. Int J Radiat Oncol Biol Phys 1998;40(2):507–513.
58. Wagner TH, et al. A simple and reliable index for scoring rival stereotactic radiosurgery plans. Int J Radiat Oncol Biol Phys 2003;57(4):1141–1149.
59. Bova FJ, Meeks SL, Friedman WA, Buatti JM. Stereotactic plan evaluation tool, the "UF Index". Int J Radiat Oncol Biol Phys 1999;45(3(S)):188.
60. Landberg T, et al.. Prescribing, recording, and reporting photon beam therapy, ICRU Report 50. Bethesda, (MD): International Commission on Radiation Units and Measurements; 1993.
61. Nedzi LA, et al. Variables associated with the development of complications from radiosurgery of intracranial tumors. Int J Radiat Oncol Biol Phys 1991;21(3):591–599.
62. Shaw E, et al. Radiosurgery for the treatment of previously irradiated recurrent primary brain tumors and brain metastases: initial report of radiation therapy oncology group protocol (90-05). Int J Radiat Oncol Biol Phys 1996;34(3):647–654.
63. Flickenger JC, Kondziolka D, Lunsford LD. What is the effect of dose inhomogeneity in radiosurgery?, In: International Stereotactic Radiosurgery Society 3rd Meeting. Kondziolka D, editor. Madrid: Karger; 1997. p 206–213.

See also Gamma knife; stereotactic surgery.

RADIOTHERAPY ACCESSORIES

James A. Purdy
UC Davis Medical Center
Sacramento, California

INTRODUCTION

The use of radiation to treat cancer is a complex process that involves many trained professionals and a variety of interrelated functions. Radiation oncologists, medical physicists, dosimetrists, and radiation therapists have for many years sought apparatus and devices that aid in this process, particularly in regard to tumor localization, treatment planning, treatment delivery, and verification (1). Precision radiation therapy is necessary as clinical and experimental results show that tumor control and normal tissue response can be a very steep function of radiation dose, and hence, small changes in the dose delivered can result in a dramatic change in the local response of the tumor and/or normal tissues. Moreover, the prescribed curative tumor doses are often, by necessity, close to the doses tolerated by the normal tissues. Thus, for optimum treatment, the radiation dose must be delivered with a high degree of accuracy; a value of ±5% has been recommended by the International Commission on Radiation Units and Measurements (ICRU) (2).

Since the first edition of this encyclopedia, the field of radiation oncology has undergone dramatic changes. At that time, radiation oncologists were trained to plan and treat patients using what has been labeled a two-dimensional (2D) approach. This approach emphasizes the use of a conventional X-ray simulator for designing beam portals that are based on standardized beam arrangement techniques and the use of bony landmarks visualized on planar radiographs. This approach, while still used by some clinics, has been largely replaced by a three-dimensional (3D) approach in modern day radiotherapy clinics. This was made possible by the introduction of commercial 3D treatment planning systems in the early 1990s (3). In contrast to the 2D method, 3D treatment planning emphasizes an image-based virtual simulation approach for defining tumor volumes and critical organs at risk for the individual patient (4). The new 3D process puts new demands on the radiation oncologist to specify target volumes and organs at risk with far greater accuracy than before, and also on the medical physicist to provide effective quality assurance (QA) processes to ensure safe use of the new image-based planning and computer controlled treatment delivery approach (5).

This article presents a review of the devices that have been designed to help achieve the high degree of precision and accuracy needed in the radiation treatment (for both the 2D and 3D approaches) of the cancer patient. These devices have been arranged in the following general categories: tumor localization and treatment simulation devices, patient setup and restraint–repositioning devices, field-shaping devices–dose-modifying devices, and treatment verification and quality assurance devices.

TUMOR LOCALIZATION AND TREATMENT SIMULATION DEVICES

Devices and apparatus in this category are designed to aid in visualizing and determining the extent of the tumor in relation to the treatment geometry (target volume localization) and to obtain measurements of the patient's body contours and thicknesses. In the past, target volume localization was usually accomplished by physical examination and the use of a device called an X-ray simulator, which combines radiographic and fluoroscopic capability in a single machine that mimics the actual treatment unit geometries (Fig. 1). The simulation process itself may be supplemented with other diagnostic imaging studies including computed tomography (CT), magnetic resonance images (MRI), and, more recently, positron emission tomography (PET).

Devices used to aid the conventional simulation process include a radiopaque fiducial grid (Fig. 2) projected on the patient's anatomy, which allows one to determine dimensions of the treatment volume from the simulator plane films. Examples of other devices used in the 2D target volume localization process are magnification rings placed in the irradiated field and lead-tipped rods that can be inserted into body openings, such as the vagina for carcinoma of the cervix or into the rectum for carcinoma of the prostate. The lead tip can be visualized clearly on simulator films or treatment portal films and allows evaluation of treatment field margins.

Note that a new generation of conventional simulators (Fig. 3) has recently been developed in which the image intensifier system has been replaced with an amorphous silicon flat-panel detector. This device produces high resolution, distortion-free digital images, including cone-beam CT.

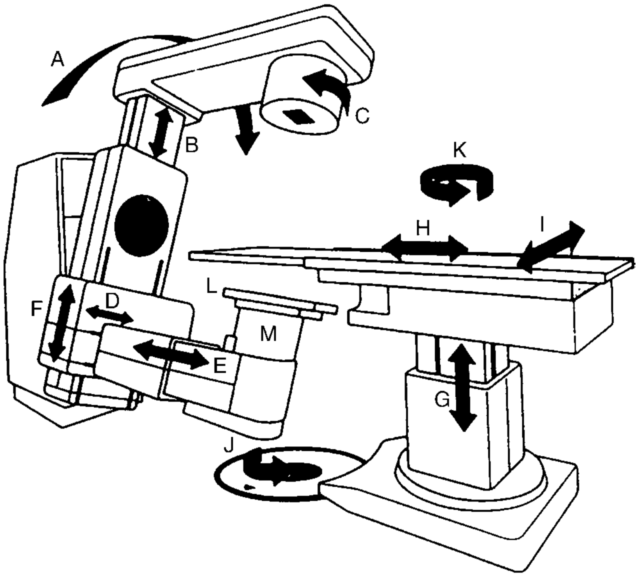


Figure 1. The basic components and motions of a radiation therapy simulator: A, gantry rotation; B, source-axis distance; C, collimator rotation; D, image intensifier (lateral); E, image intensifier (longitudinal); F, image intensifier (radial); G, patient table (vertical); H, patient table (longitudinal); I, patient table (lateral); J, patient table rotation about isocenter; K, patient table rotation about pedestal; L, film cassette; M, image intensifier. Motions not shown include field size delineation, radiation beam diaphragms, and source-tray distance. (See Ref. 6.)

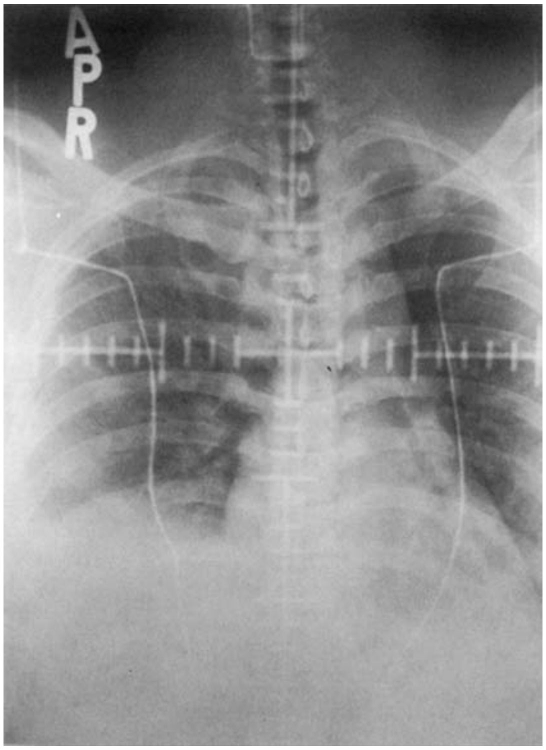


Figure 2. X-ray simulator radiograph showing fiducial grid projected on patient's anatomy. The grid is used for localizing target volume and determining treatment field size.

Figure 3. New generation radiation therapy simulator, in which image intensifier system has been replaced with amorphous silicon flat-panel that produces high resolution, distortion-free images and facilitates a filmless department. (Courtesy of Varian Medical Systems.)

Once the treatment geometry has been determined and the patient is in treatment position, the patient's body thicknesses and contours are measured and recorded for purposes of computing a dose distribution and determining treatment machine settings. Manual methods using calipers, lead solder wire, plaster cast strips, flexible curves, or other devices, such as the contour plotter (see Fig. 4) are the most common methods of obtaining this type of data when using the 2D planning approach.

Fields to be treated are typically delineated in the 2D simulation process using either visible skin markings or marks on the skin visible only under an ultraviolet (UV) light. Some institutions prefer to mark only reference setup points using external tattoos. These skin markings are used to reposition the patient on the treatment machine using the treatment machine's field localization light and optical distance indicator and laser alignment lights mounted in the treatment room that project transverse, coronal, and sagittal light lines on the patient's skin surface (Fig. 5).

In the new 3D era, CT simulators have become the standard of practice; a typical CT simulator facility design is shown in Fig. 6. A volumetric set of CT images is used to define the target volume, critical organs at risk, and skin surface. The CT numbers can be correlated with the electron densities of the tissues imaged to account for heterogeneities when calculating the dose distribution. Numerous studies have documented the improvements in target volume localization and dose distributions achieved with anatomic data obtained from CT scans as compared with the conventional simulation process (7,8). The CT simulators have advanced software features for image manipulation and viewing such as beam's eye view (BEV) display, and virtual simulation tools for setting isocenter, and digital reconstructed radiographs (DRRs)

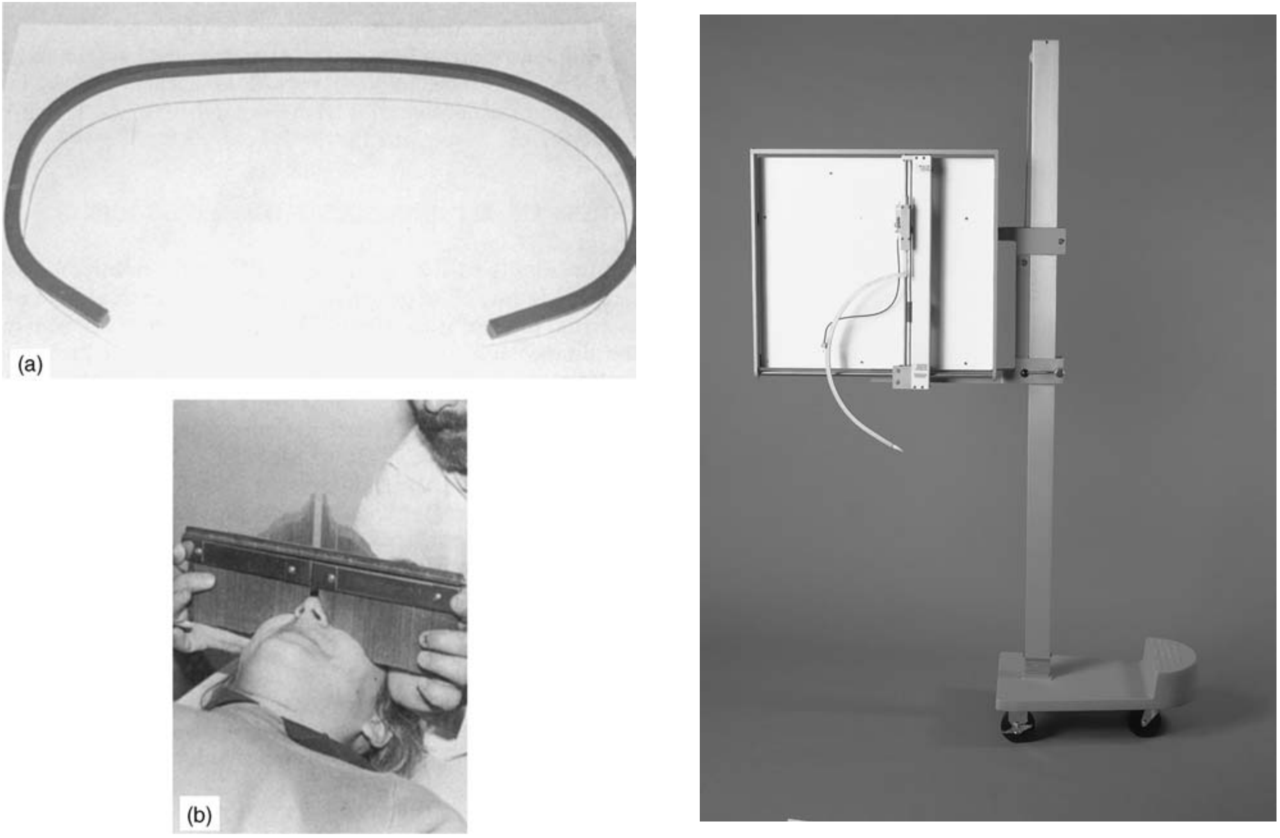


Figure 4. Contour plotter. The device is a simple, easy-to-use precision pantograph that links a drawing pen to a stylus arm and, upon contact with the body, communicates body contours to an overhead drawing board. The contour plotter is suspended on a vertical column and can easily be adjusted and locked securely. A continuous plot is drawn as the operator follows the physical contours of the patient. Marks can be made along the contour to indicate beam entry and laser light locations. (Courtesy of MEDTEC.)

(9,10). Such systems provide all the functionality of a conventional simulator, with the added benefit of increased treatment design options and the availability of software tools to facilitate the understanding and evaluation of treatment options. In addition, the simulation process is more efficient and less traumatic to the patient. Laser alignment lights and repositioning devices registered to the treatment couch are used to facilitate repositioning the patient in the treatment machine coordinate system once the virtual simulation process is complete.

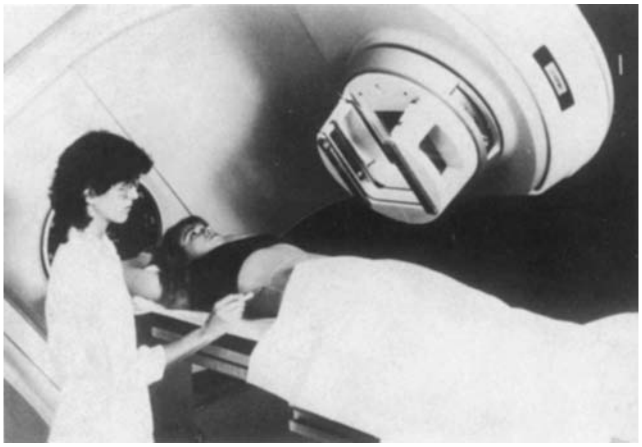


Figure 5. Laser alignment system. Patient in treatment position on treatment couch. Close-up of laser lines imaged on patient skin under typical treatment room lighting conditions. (Courtesy Gammex RMI, Inc.)

PATIENT RESTRAINT AND REPOSITIONING DEVICES

Accurate daily repositioning of the patient in the treatment position and reduction of patient movement during treatment is essential to accurately deliver the prescribed dose and achieve the planned dose distribution. As we will see in this section, modern day immobilization and repositioning systems are designed to be able to be attached to the simulation and treatment couches, so that the immobilization device and the patient are registered to the treatment machine coordinate system. Once the immobilization device has been locked into a specified position, the patient is then aligned to the immobilization system. The end result is that a set of coordinates is obtained from the CT simulator that is used in the virtual simulation process and that can be correlated to the treatment room isocenter.

The anatomic sites most often needing immobilization in radiation therapy are the head and neck, breast,

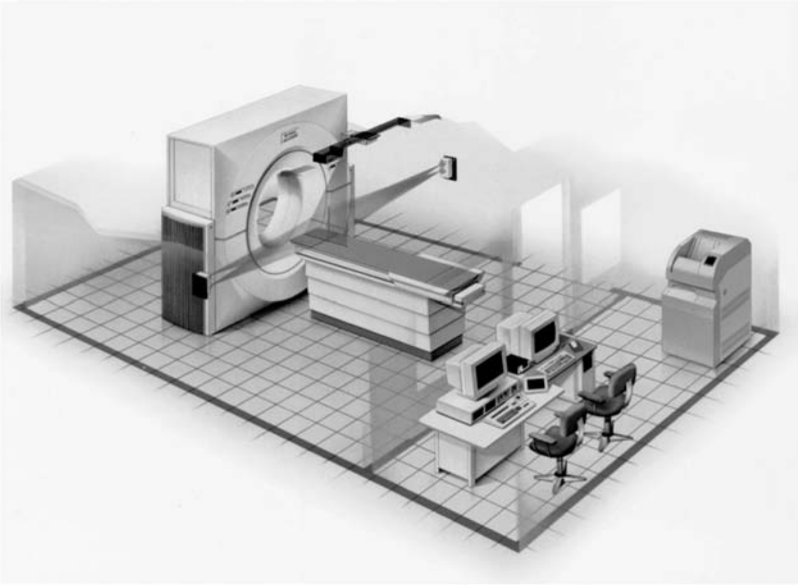


Figure 6. Typical CT simulation suite showing the scanner, flat tabletop, orthogonal laser system, virtual simulation workstation and hardcopy output device. (Courtesy of Philips Medical Systems.)

thorax–esophagus, shoulders and arms, pelvic areas (especially if obese), and limbs rotated to unusual positions. The precision achievable in the daily treatment positions of a patient depends on several factors other than the anatomic site under treatment, such as the patient's age, general health, and weight. In general, obese patients and small children are the most difficult to reposition.

Simple patient restraint and repositioning devices can be used in treating some anatomic sites. For example, the disposable foam plastic head holder shown in Fig. 7 provides stability for the head when the patient is in the supine position. If the patient is to be treated in the prone position, a face-down stabilizer can be used as shown in Fig. 8. This device has a foam rubber lining and a disposable

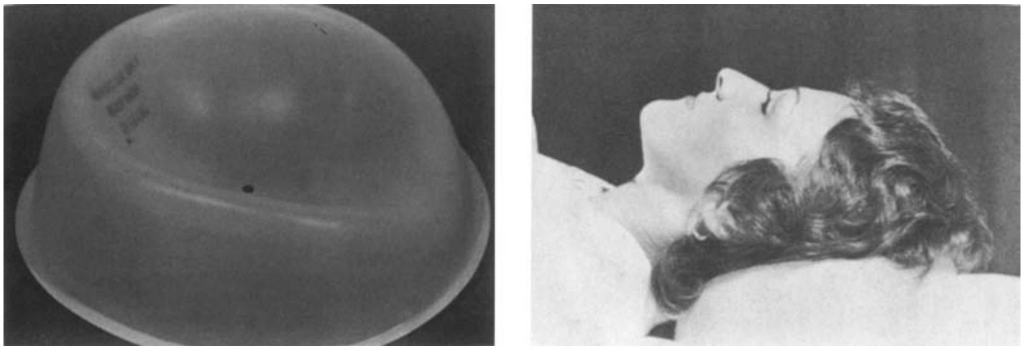


Figure 7. Disposable foam plastic head holder provides stability to the head when the patient is in the supine position.

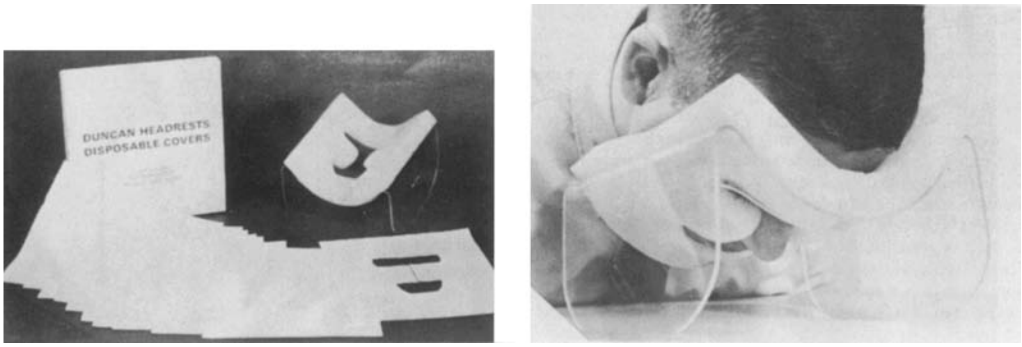


Figure 8. Face-down stabilizer. This formed plastic head holder has a foam rubber lining and disposable paper liner with an opening provided for the eyes, nose, and mouth. It allows comfort and stability as well as air access to the patient in the prone position.

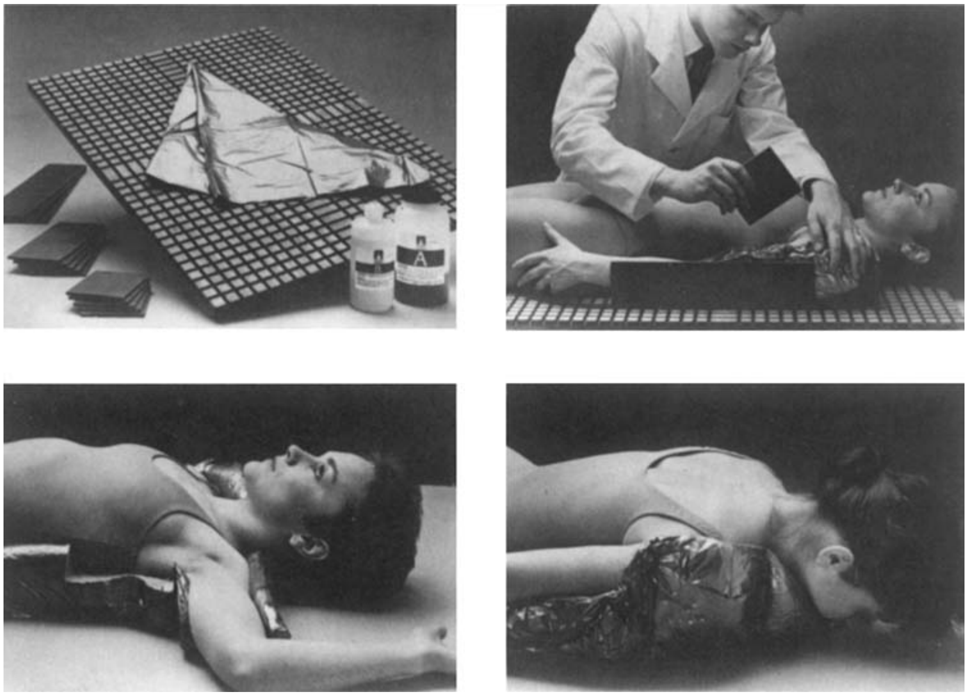


Figure 9. Polyurethane body mold. The chemical mixture is poured into the foam mold under a latex sheet. The patient is positioned in the foam mold as the polyurethane mixture expands to body shape. These body molds are easy to make, save time in patient alignment, and increase patient comfort. (Courtesy Smithers Medical Products, Inc.)

paper lining with an opening provided for the eyes, nose, and mouth of the patient. It allows comfort and stability as well as air access for the patient during treatment in the prone position.

There are now several commercially available body mold systems that are in widespread use as immobilization and repositioning aids. Fig. 9 illustrates one such system that utilizes a foam block cutout of the general anatomic area and polyurethane chemicals, which when mixed expand and solidify to conform to the patient's shape in a matter of minutes. Another widely used system (Fig. 10) consists of a vinyl bag filled with plastic minispheres. The bag is positioned around the patient to support the treatment position and then a vacuum is applied causing the minispheres to come together to form a firm solid support molded to the patient's shape.

Plaster casts are still used in some clinics, but have not gained widespread use in the United States, probably because they are too labor intensive and time consuming. Also, transparent form-fitting plastic shells (Fig. 11) that

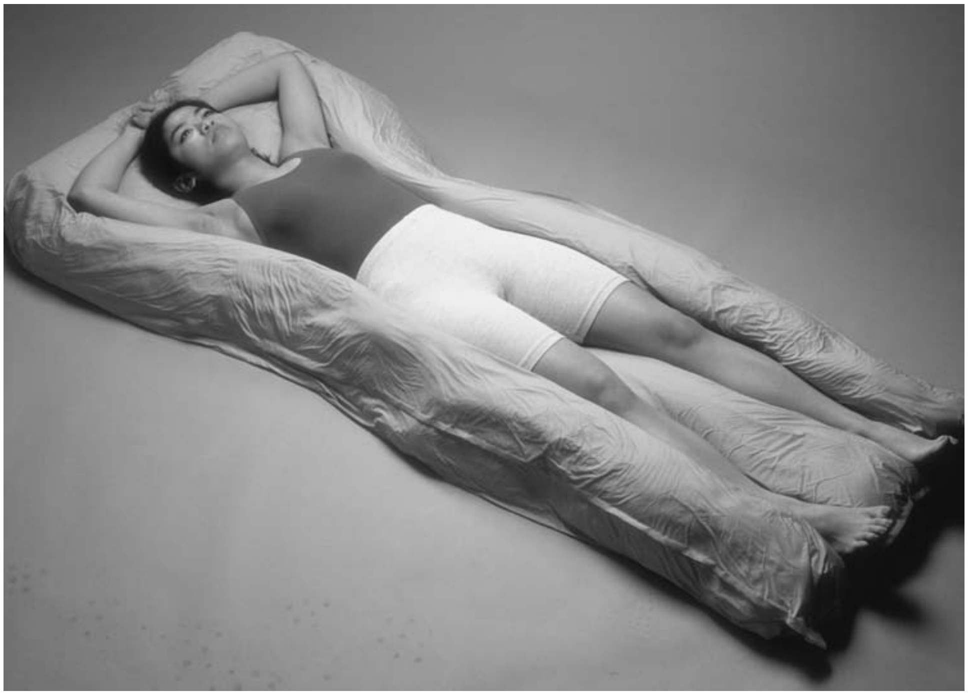


Figure 10. Vacuum-form body immobilizer. The system consists of a plastic mattress filled with microspheres connected to a vacuum pump. Under vacuum, the mattress shapes itself to the body contours. (Courtesy of MEDTEC.)

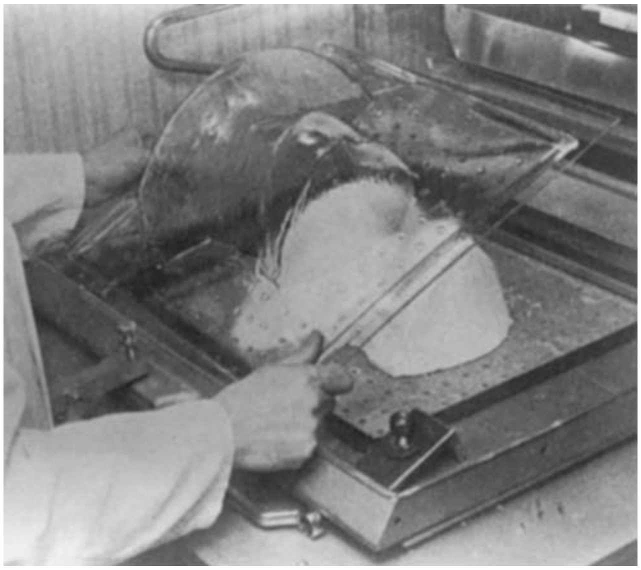


Figure 11. Plastic shell. Transparent form-fitting plastic shells fabricated using a special vacuum device. (See Ref. 11.)

are fabricated using a special vacuum device are also used in some countries (e.g., Great Britain and Canada), but again are very labor intensive and have not gained acceptance in this country. Both methods are described in detail in the book by Watkins (11). In the United States, thermal plastic masks are much more commonly used (Fig. 12). The plastic sheet or mesh is placed in warm water to make it very pliable, and when draped over the patient conforms to the patient's shape and hardens upon cooling.

Fig. 13 illustrates a device called a bite block, which is used as an aid in patient repositioning in the treatment of head and neck cancer. With this device, the patient is placed in the treatment position and instructed to bite into a specially prepared dental impression material layered on a fork which is attached to a supporting device. When the material hardens, the impression of the teeth is recorded. The bite-block fork is connected to a support arm that is attached to the treatment couch and may be used either with or without scales for registration.

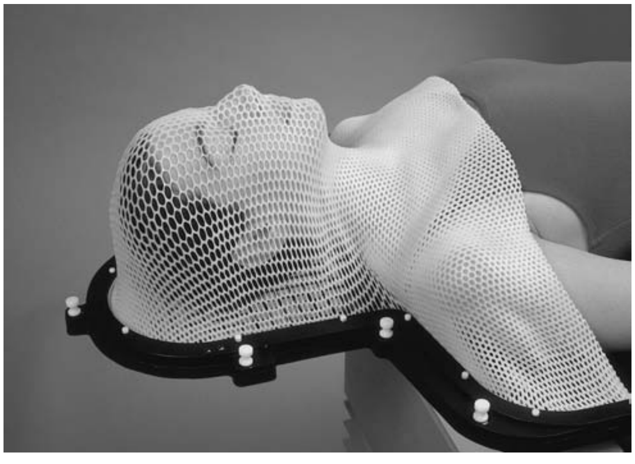


Figure 12. Thermoplastic cast. When placed in a warm (170°F) water bath, thermoplastic material becomes very flexible and can easily be molded to the patient's surface curvature. Immobilization using this material is less labor intensive than the conventional plaster cast or plastic shells and is therefore more readily adaptable on a routine basis for the immobilization of patients during radiation therapy. (Courtesy of MEDTEC.)

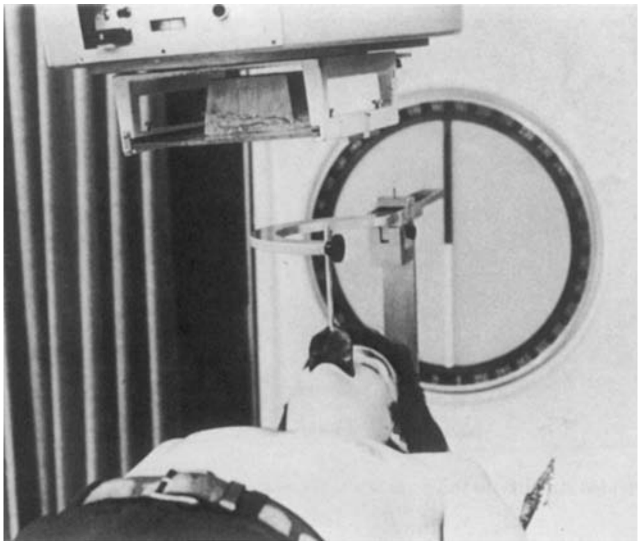


Figure 13. Bite-block system. Placement of the patient in a comfortable supine position and use of bite-block immobilization minimize patient movement for head and neck treatments. Note that the C-arm design allows both lateral and anterior beam arrangements to be used. (Courtesy of Radiation Products Design, Inc.)

There are many other devices used to help in the treatment setup of patients that are site specific. For example, breast patients are usually positioned supine, with the arm on the involved side raised and out of the treatment area. Fig. 14 shows a device called a breast or tilt board that is used to optimize the position of the patient's chest wall (or thorax). The device is constructed with a hinge section that can be positioned and locked into place at various angles to the horizontal treatment table top. Modern breast boards now provide options for head support, arm positioning, and breast support. Sometimes, it is more convenient to use a separate arm board that can be attached directly to the treatment couch (Fig. 15). The perpendicular support provides a hand grasp that can be adjusted to the proper height and assists the patient in holding their arm in a comfortable position away from the treatment field.

Another useful device is the breast bridge (Fig. 16), which can be placed on the patient's chest and adjusted to the skin markings, for determining separation of the tangential fields. Precise angulation of the beam portals is determined using a digital readout level. In addition, a squeeze bridge (Fig. 17) with plastic plates can be used to provide buildup when a higher surface dose is desired, or one with a wire mesh frame can be used when no additional surface dose is warranted. A beam alignment device (Fig. 18) is used to match multiple fields used in tangential breast irradiation (12). The alignment component of the device is a curved piece of aluminum with a row of nylon

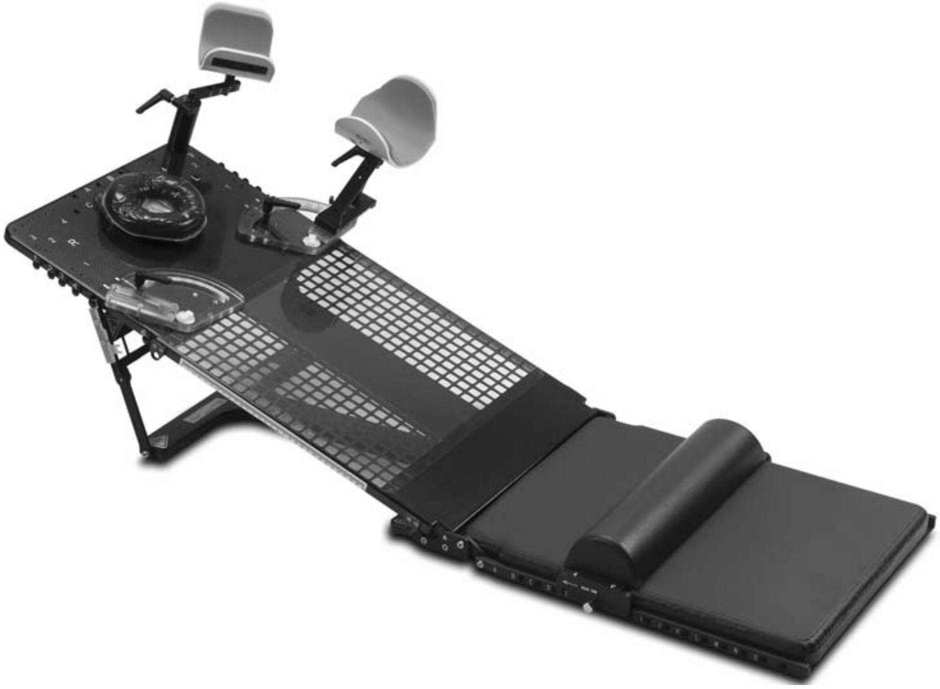


Figure 14. Tilt board or adjustable breast board. The top piece is fabricated with a hinged section that allows the sloping chest wall to be more appositional to a vertical beam. (Courtesy of MEDTEC.)

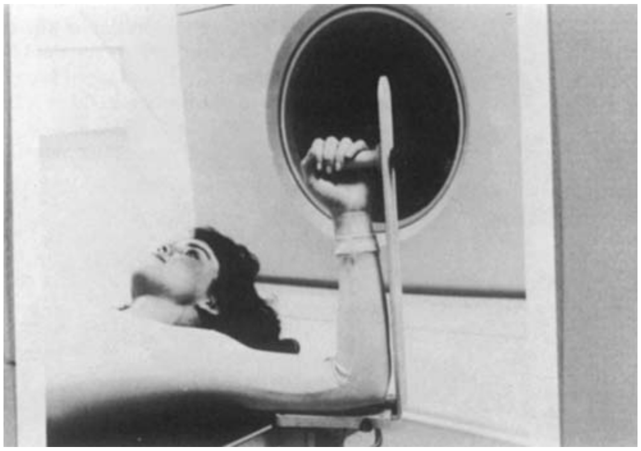


Figure 15. Patient arm support used for breast irradiation assists patients in holding their arm in a comfortable position, away from the treatment field. (Courtesy of Varian Medical Systems.)

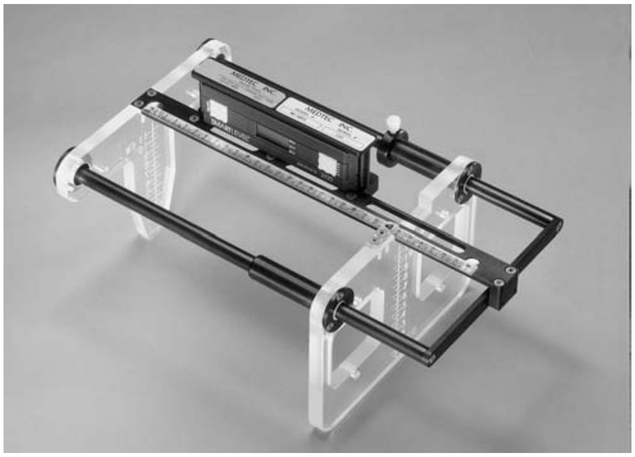


Figure 16. A Breast Bridge is used with tangential radiation fields and consists of a pair of plastic plates that can be locked at the appropriate separation determined for the individual patient. After the treatment area has been marked, the bridge is placed on the patient's chest and adjusted to the skin markings, thus determining separation of the fields. Precise angulation of the portals is determined by the digital readout level. Once the portals are set, the bridge may be removed. (Courtesy of MEDTEC.)

pins protruding from its surface. The pins have a thin inscribed line to which the light field is aligned.

In some instances, it may be advantageous to treat the cancer patient in an upright position. The treatment chair (Fig. 19) is a device that facilitates such treatments by allowing a patient to be accurately repositioned in a seated position each day. The chair provides means of stabilizing the patient by the use of hand grips, elbow holders, and a seatbelt. The back of the seat is constructed of carbon fiber and thus the radiation beam can penetrate with minimal effects, and the angle of the seat back is adjustable.

Another device used to optimize patient position is the shoulder retractor. It provides a means by which the patient's shoulders can be pulled down in a reproducible manner, as illustrated in Fig. 20. Such a device is often used in the treatment of head and neck cancer involving lateral fields.

A standard feature on most accelerator treatment couches is a Mylar window or tennis racket-type table insert. This device consists of a thin sheet of Mylar stretched over a tennis racket-type webbing material and mounted in a frame that fits into a treatment table that has removable sections. Newer table insert devices made of carbon fiber (Fig. 21) eliminate the need to "restring" such panels and minimize the "sag" that can occur with nylon string panels. Such inserts provide excellent patient support

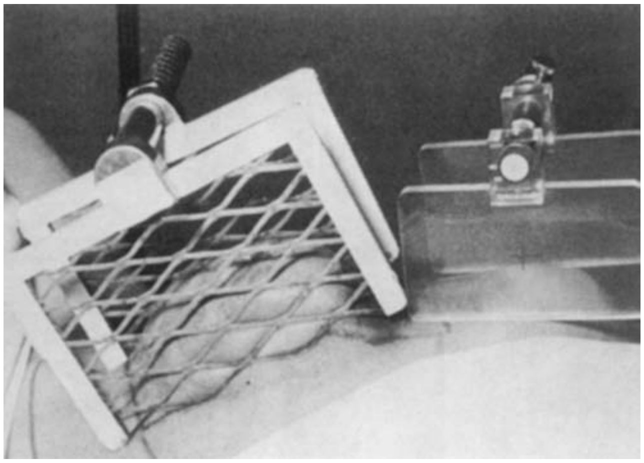


Figure 17. A squeeze bridge used for breast treatments with and without bolus. The wire mesh device is used where no additional surface dose is desired and the plastic frame device is used when increased surface dose is needed. (See Ref. 1.)

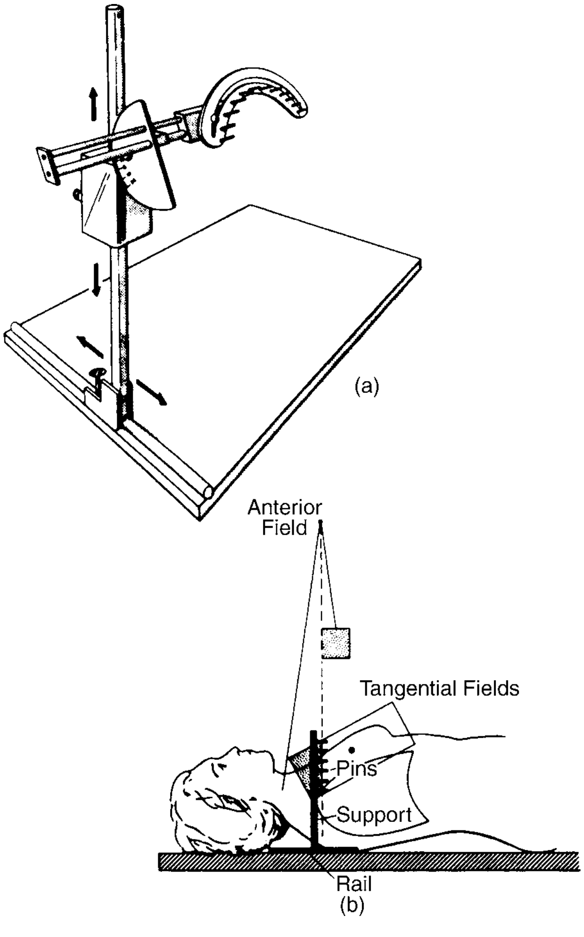


Figure 18. A beam alignment device used to implement field matching techniques. (a) Schematic of beam alignment device. (b) Example of use of device with three-field technique for breast treatment. Superior edges of tangential breast fields are coplanar and abutted to the vertical inferior edge of the anterior supraclavicular field. (See Ref. 12.)

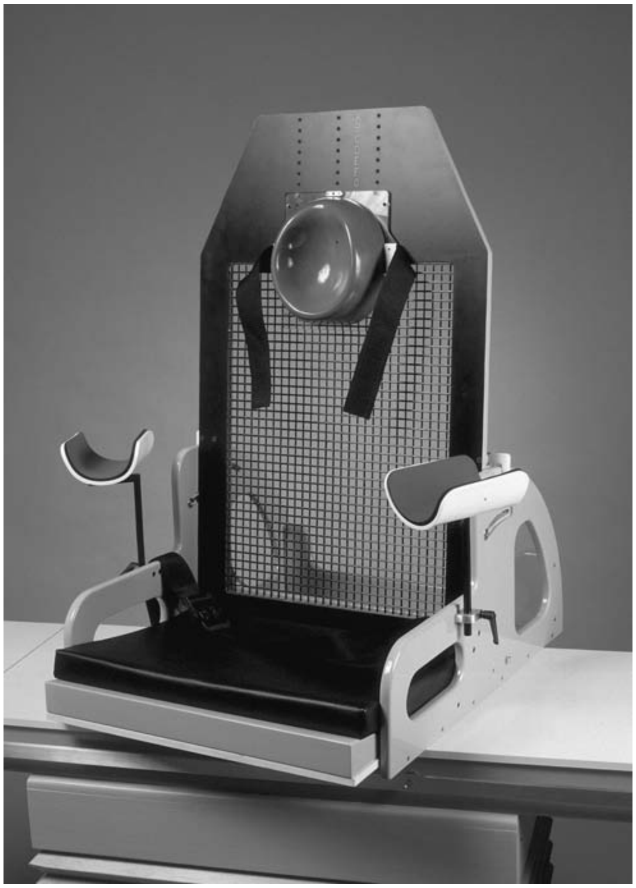


Figure 19. Treatment chair. Provides positioning and fixation for breast, lung, and thorax patients who require vertical–upright positioning; adjusts to different locking positions and can accommodates a thermoplastic mask for head fixation. (Courtesy of MEDTEC.)

with a minimum of surface buildup effect for opposing beam portals, minimum reduction of beam intensity, and good visual access to the treatment surface. In addition, most medical linacs come with clamps that can be attached to

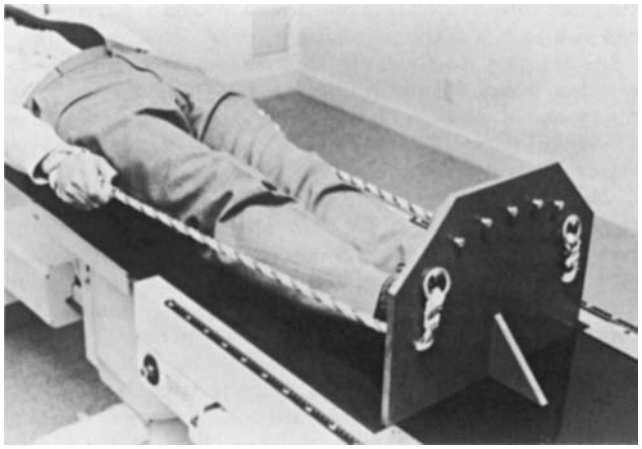


Figure 20. Shoulder retractor. Device used to pull the patient's shoulders down in a reproducible manner for treatment of the head and neck with lateral radiation fields. (Courtesy MED-TEC, Inc.)

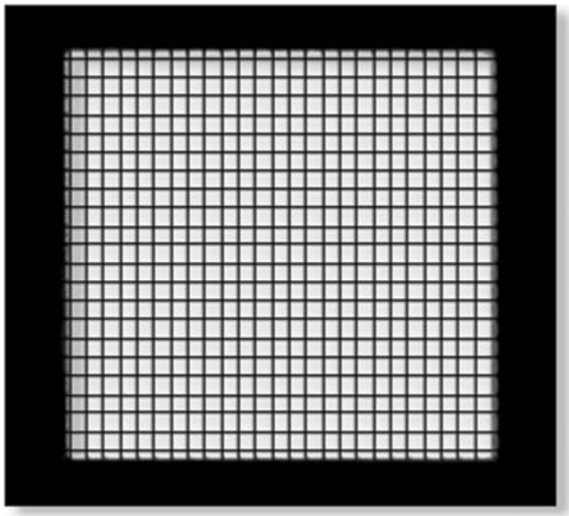


Figure 21. Treatment couch insert. Carbon fiber table and spine inserts for simulator and treatment couches. Rigid carbon fiber minimizes the "sag" that can occur with nylon string panels. (Courtesy of MEDTEC.)

the treatment couch and can be used with several different accessories, such as hand grips.

The development of 3D conformal radiation therapy (3DCRT) and more recently intensity-modulated radiation therapy (IMRT) has greatly enhanced the radiation oncologist's ability to plan and deliver very high doses that conform closely to the target volume, and falls off sharply, thus avoiding high dose to the nearby organs at risk (13,14). Both 3DCRT and IMRT invite the use of tighter margin to achieve higher dose escalation, and thus have spurred the development of new accessories and processes to better account for setup variation and organ motion that can occur during one (intra-) fraction, and between (inter-) fractions. Efforts thus far have focused on accounting for internal movement of the prostate gland and internal motion cause by respiratory function.

For example, for prostate cancer, the use of daily ultrasound imaging (Fig. 22), or daily electronic portal imaging of implanted radiopaque markers, has now become standard practice in many clinics (15,16).

Devices–methodologies used to address the problem of breathing motion in radiation treatment include: (1) gating and/or tracking and (2) breathhold devices–strategies. In gating and tracking, the state of the treatment machine is adjusted in response to a signal that is representative of a patient's breathing motion. With breathholding, the lung volume of the patient is directly immobilized prior to beam-on, and released after the beam is off. The basic components of a gating or tracking system consist of a respiration sensor whose signal is processed and evaluated by a computer for suitability to trigger, or gate, the radiation. An example of a respiratory gating system is the Real-time Position Management (RPM) system (Fig. 23) commercially available from Varian Medical Systems (Palo Alto, CA) (17,18). An example of a breath control device is the Elekta Inc. (Norcross, GA) Active Breathing Coordinator (ABC) (Fig. 24) (19). The ABC apparatus is used to suspend breathing at any predetermined position along the normal breathing cycle, or at active inspiration, and consists of a digital spirometer to measure the respiratory trace, which is in turn connected to a balloon valve. Another example is the ExacTrac system (BrainLAB AG) that combines X-ray imaging and infrared tracking that permits correlation of internal 3D tumor motion with the patient's breathing cycle (Fig. 25) (20). Automatic fusion of digitally reconstructed radio-graphic (DRR) images computed from the treatment planning CT data to the live X-rays allows any set-up error or target shift and rotation to be identified and any discrepancy compensated for via robotic table movement.

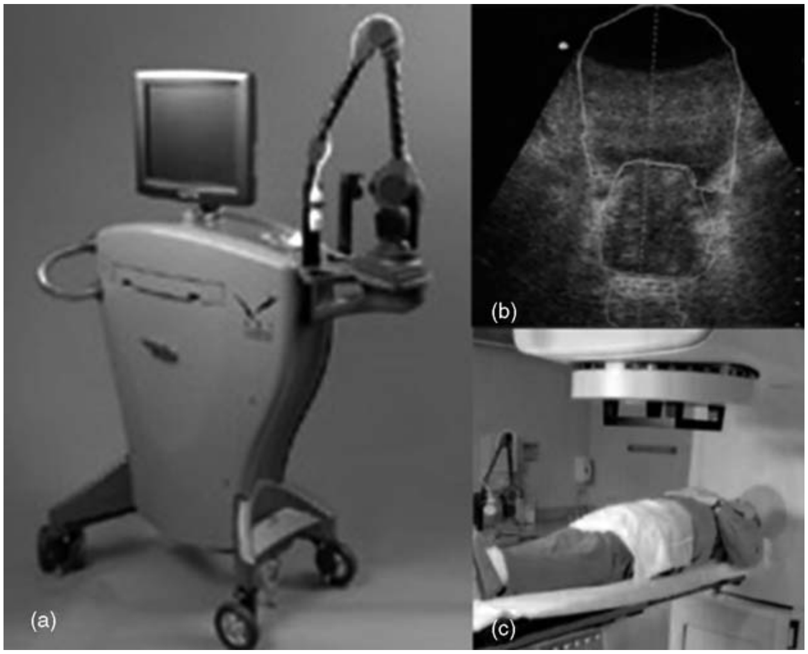
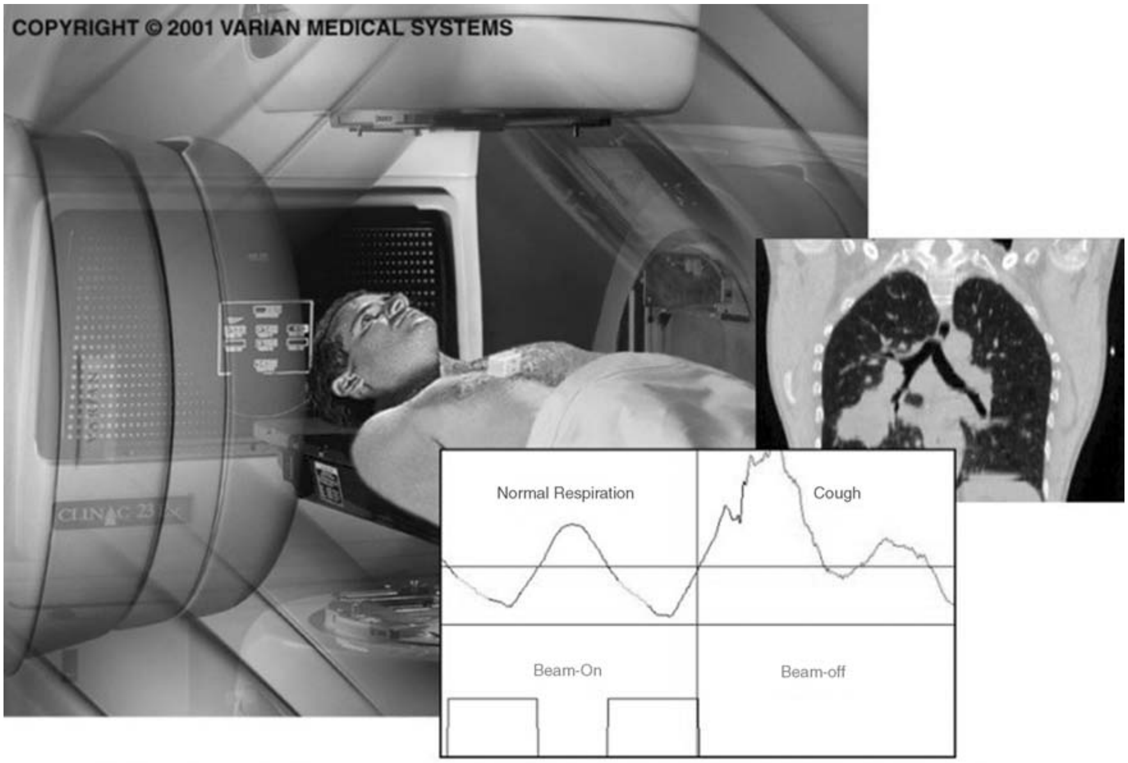


Figure 22. BAT (B-mode Acquisition and Targeting) SXi system. Targeting device that provides fast ultrasound localization of a treatment target on a daily basis. (Courtesy North American Scientific)



From both simulation and virtual simulation to treatment, the RPM Respiratory Gating system allows you to accurately monitor and compensate for tumor movement.

Figure 23. Real-time Position Management system. Device used to allow respiratory gating treatment delivery. (Courtesy of Varian Medical Systems)

For stereotactic radiosurgery or fractionated stereotactic radiotherapy, a suite of accessories is now available. Most important is a head stereotactic localizer such as the Gill–Thomas–Cosman (GTC) relocatable head ring (Fig. 26), which enables precise fixation and localization and repositioning of targets in the cranium (21). Another important device when using the newly emerging radiotherapy treatment, stereotactic body radiation therapy (SBRT), in which a high dose is delivered in either a single fraction or just a few fractions, is the frame-based body stereotactic immobilization system (22). There are several now commercially available; an example is the Elekta Stereotactic Body Frame shown in Fig. 27. This device provides a reference stereotactic coordinate system that is external to the patient's body, so that the coordinates of a target volume can be reproducibly localized during simulation and treatment. This frame has built-in reference indicators for CT or MR determination of target volume coordinates. In addition, a diaphragm control attached to the frame can be used to minimize respiratory movements. Horizontal positioning of the frame, on the CT simulator or treatment couch, is achieved using an adjustable base on the frame.

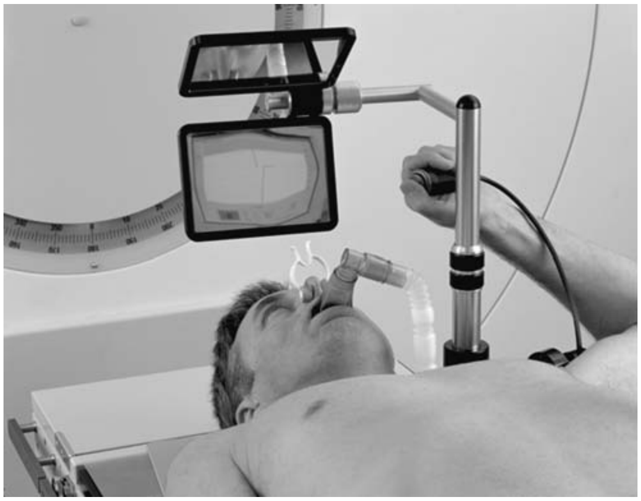


Figure 24. Active Breathing Coordinator$^{\text{TM}}$system. Device allows the radiation oncologist to pause a patient's breathing at a precisely indicated tidal volume and coordinate delivery with this pause. (Courtesy of Elekta AB)

FIELD-SHAPING, SHIELDING, AND DOSE MODIFYING DEVICES

The Lipowitz metal (Cerrobend) shielding block system introduced by Powers et al. (23) is in widespread use throughout the world. The block fabrication procedure is illustrated in Fig. 28 and more details using this form of field shaping can be found in the review article by Leavitt and Gibbs (24). Lipowitz metal consists of 13.3% tin, 50.0% bismuth, 26.7% lead, and 10.0% cadmium, and has a

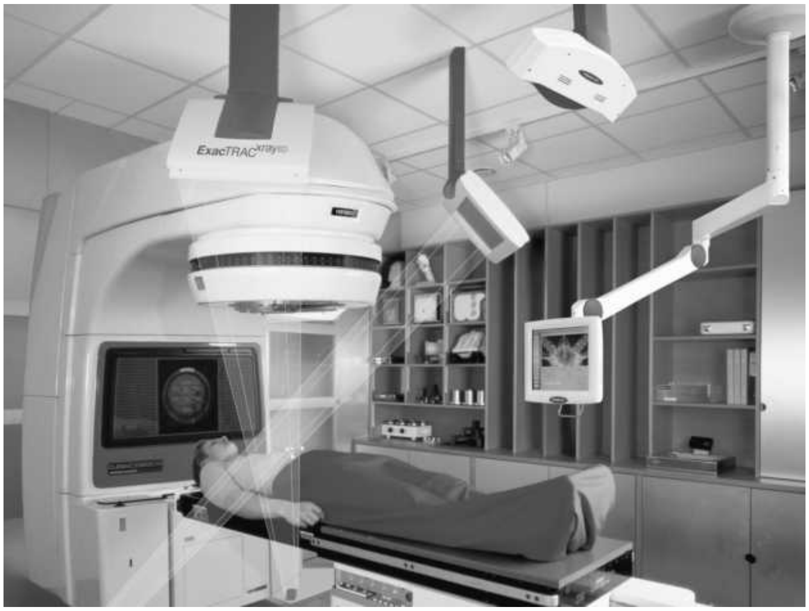


Figure 25. ExacTrac X-Ray 6D automated image-guided radiation therapy system. System consists of 2 kV X-ray units recessed into the linac floor and two ceiling-mounted amorphous silicon Flat Panel detectors integrated with a real-time infrared tracking device to enable imaging of internal structures or implanted markers for extremely accurate set-up of the target volume's planned isocenter position. (Courtesy Brain LAB AG.)

physical density at 20 °C of 9.4 g·cm^{-3} as compared with 11.3 g·cm^{3} for lead. A simulation radiograph is obtained with the patient in the treatment position and the desired treatment field aperture is drawn on the radiograph by the radiation oncologist. The marked radiograph is then used as a template for cutting a foam mold with a "hot-wire" cutting device in which molten Lipowitz alloy is poured and then cooled to form a shielding block. Computer-controlled adaptations of the hot-wire cutting technique have evolved as an adjunct to 3D treatment planning in which the treatment field shape is defined based on beam's eye-view displays. The shaped field coordinates are transferred directly to the computer-controlled blockcutting system, thereby eliminating potential errors in manual tracing, magnification, or image reversal. The other steps in the block forming and verification process remain similar to the manual procedure.

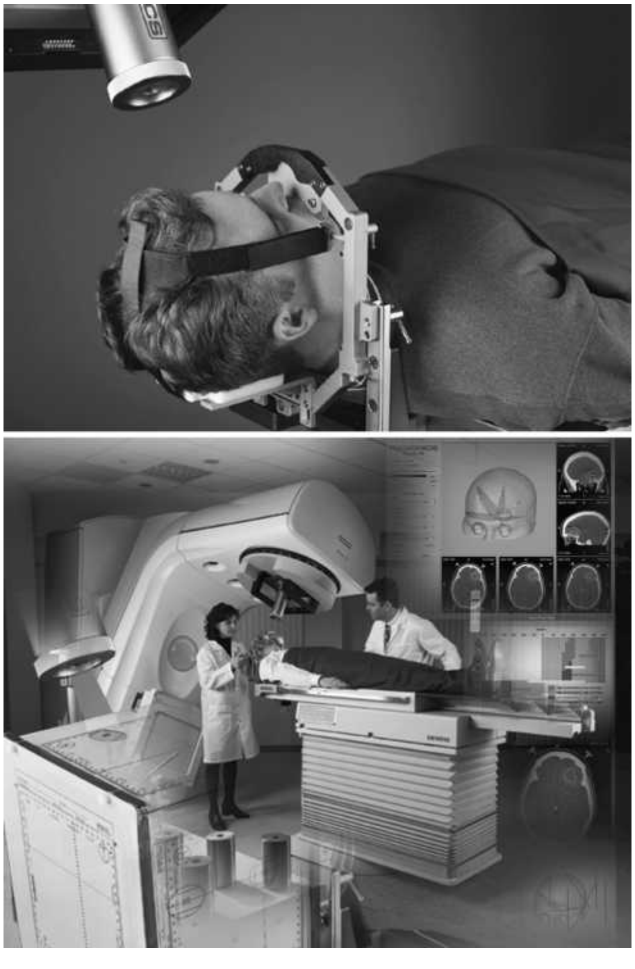


Figure 26. Gill–Thomas–Cosman relocatable head ring. Device used for stereotactic radiosurgery or fractionated stereotactic radiotherapy that enables precise fixation and localization and repositioning of targets in the cranium. Copyright © 2005 Radionics. All rights reserved. Reprinted with the permission of Radionics, a division of Tyco Healthcare Group LP.

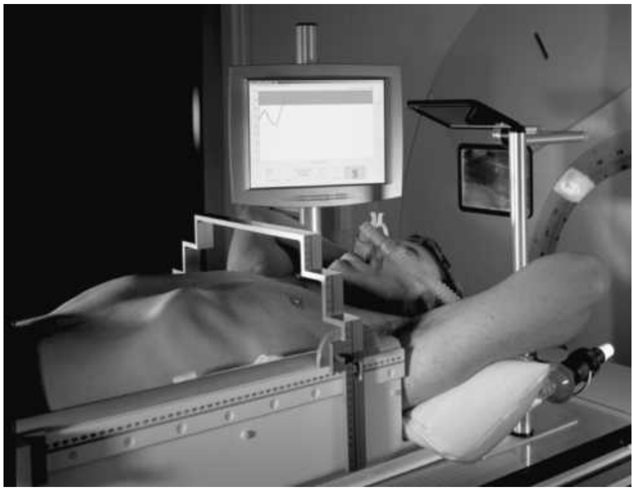


Figure 27. Elekta Stereotactic Body Frame®. Used for stereotactic body radiation therapy (SBRT), in which a high dose is delivered in either a single fraction or just a few fractions. (Courtesy of Elekta AB.)

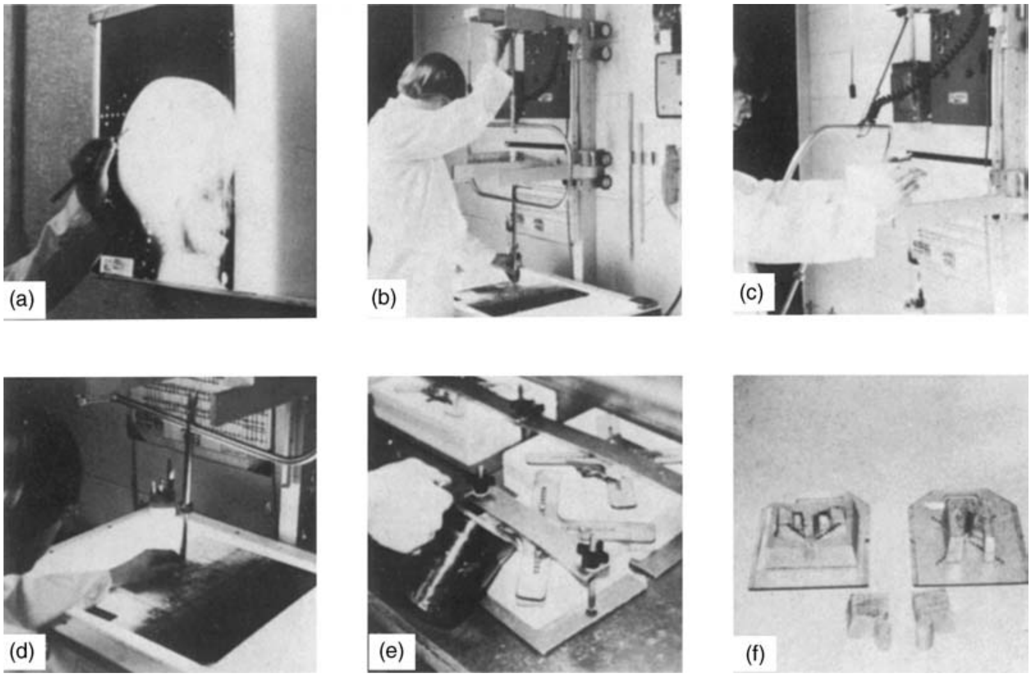


Figure 28. Shielding block for external photon beam irradiation. The design and fabrication technique is described as follows. (a) Physician defines the treatment volume on the X-ray simulator radiograph. (b) Physics technician adjusts SSD and STD of hot-wire cutter to emulate simulator geometry. (c) Proper-thickness foam block is aligned to central axis of cutter. (d) Foam mold is cut using hot-wire cutter. Courtesy of Huestis Machine Corp. (e) Foam pieces are aligned and held in place with a special clamping device. Molten alloy is poured into the mold and allowed to harden. (f) Examples of typical shielding blocks cast using this system.

Returning to accessories used for field shaping, Fig. 29 shows a Multileaf Collimator (MLC) system. The MLCs were first introduced in Japan in the 1960's (25) and have now gained widespread acceptance, and have replaced alloy blocking as the standard-of-practice for field shaping in modern radiation therapy clinics. All medical linac manufacturers now provide MLC systems. The leaves are typically carried on two opposed carriages that transport the leaves in unison. The leaves (under computer control) can be individually positioned to serve either as a block replacement, or as a means to provide IMRT treatment delivery.

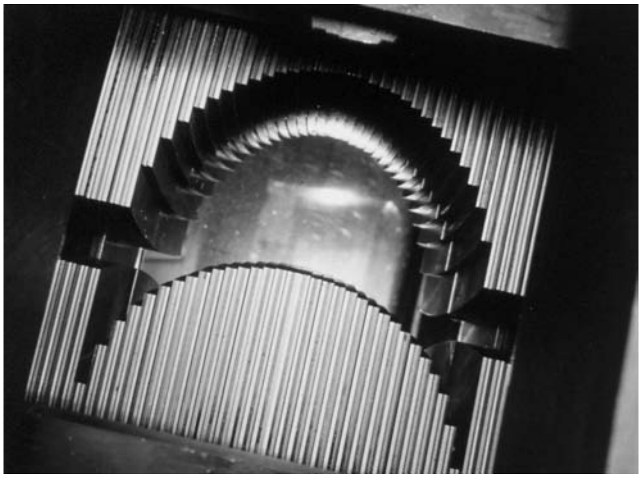


Figure 29. Multileaf Collimation (MLC) system. Computer controlled system used to shape beam aperture and also to deliver IMRT. (Courtesy of Elekta AB.)

Fig. 30 shows a commercial binary multi-leaf collimator system (called "MIMiC") designed and built by the NOMOS Corporation and incorporated into their serial tomotherapy system, known as Peacock, for planning and rotational delivery of IMRT treatments (26). This device can be attached to a conventional linac to deliver IMRT by rapidly moving leaves in or out of a slit field. Like a CT unit, the radiation source and the collimator continuously revolve around the patient.

Other more conventional devices used to achieve a desired dose distribution and to correct for perturbing influences, such as patient shape, include bolus, wedges, and compensating filters. Bolus is a tissue-equivalent material used to smooth an irregular surface, increase the dose to the patient's surface region, or sometimes to fill external air cavities, such as the ear canal or nasal passage. Bolus should have an electron density and atomic number similar to that of tissue or water. Examples of bolus materials include slabs of paraffin wax, rice bags filled with soda ash, gauze coated with Vaseline, and synthetic-based substances, such as Superflab (Fig. 31).

Wedge filters (Fig. 32) are typically made of a dense material, such as brass or steel, and are mounted on a frame that can be inserted into the beam at a specified

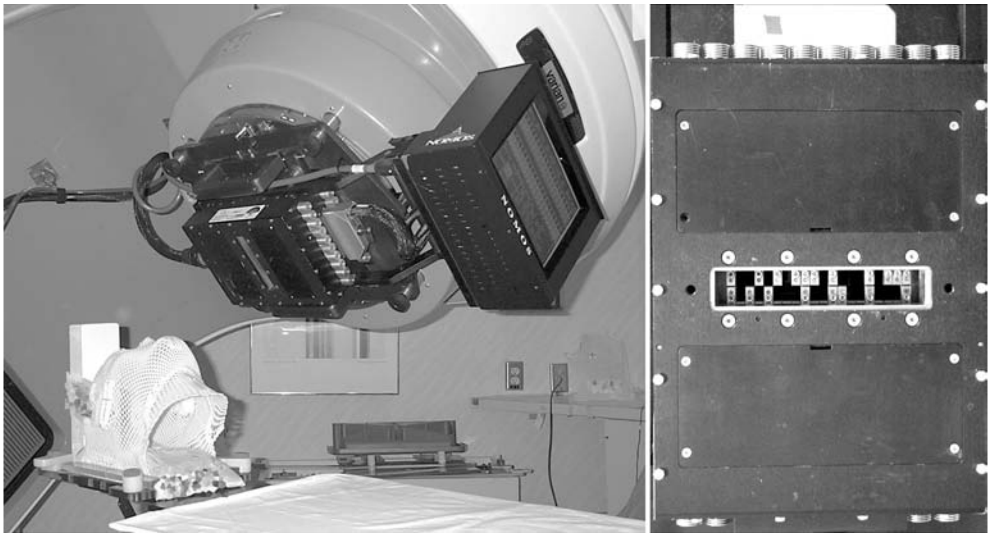


Figure 30. Binary multileaf collimator system (called MIMiC) designed and built by the NOMOS Corporation and incorporated into their serial tomotherapy system, known as Peacock for planning and rotational delivery of IMRT treatments. This device can be attached to a conventional linac to deliver IMRT by rapidly moving leaves in or out of a slit field. Like a CT unit, the radiation source and the collimator continuously revolve around the patient.

distance from the source and cause the dose distribution at a specified depth to be angled to a desired amount relative to the incident beam direction. Modern linacs have replaced physical wedges with software control of independent jaws that provide what is called dynamic wedging (27, 28). Currently, there are two versions of dynamic wedging, the Varian Enhanced Dynamic Wedge (EDW), and the Siemens Virtual Wedge (VW). The desired wedge angle is achieved by moving a collimating jaw while the beam is on, thereby shrinking the field during treatment.

A compensating filter (Fig. 33) is a beam modifier that is used to counteract the effect of air gaps caused by the patient's topography while still preserving the skin-sparing characteristic of megavoltage photon beams (29). To do this, the compensating filter is placed in the beam some distance away from the patient's skin surface. This requires that the lateral dimensions of the filter be reduced and causes the scatter radiation conditions to be altered, complicating the relationship between the thickness of the compensator along a ray and the amount of tissue deficit to be compensated.

Other accessories include shields designed specifically to protect certain organs at risk. For example, in the treatment of Hodgkin's disease and other malignant lymphomas, in which the inguinal and femoral lymph nodes are frequently irradiated, a testicular shield (Fig. 34) is frequently used. This device is typically constructed with lead and is designed to reduce scatter radiation to the testicles. Another example is the eye shields shown in

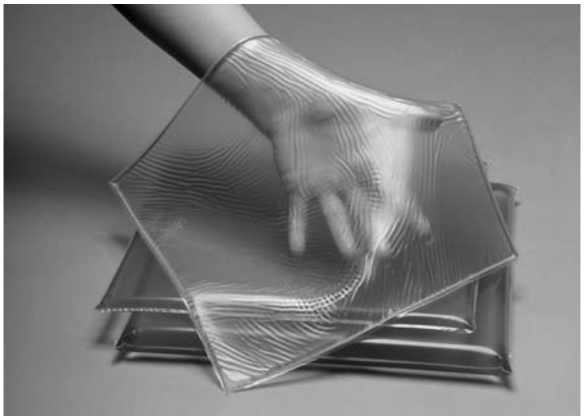


Figure 31. Example of bolus material used in radiation treatment to smooth the patient's irregular surface or to increase the dose to the surface region. (Courtesy of MEDTEC.)

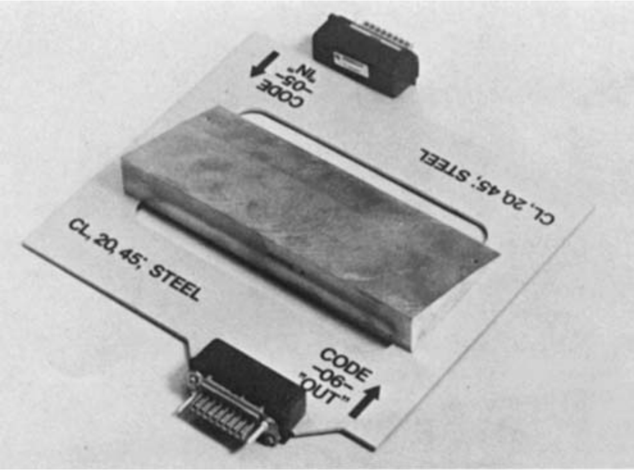


Figure 32. Wedge filter used on medical linear accelerator to shape dose distribution. (Courtesy Varian Medical Systems.)

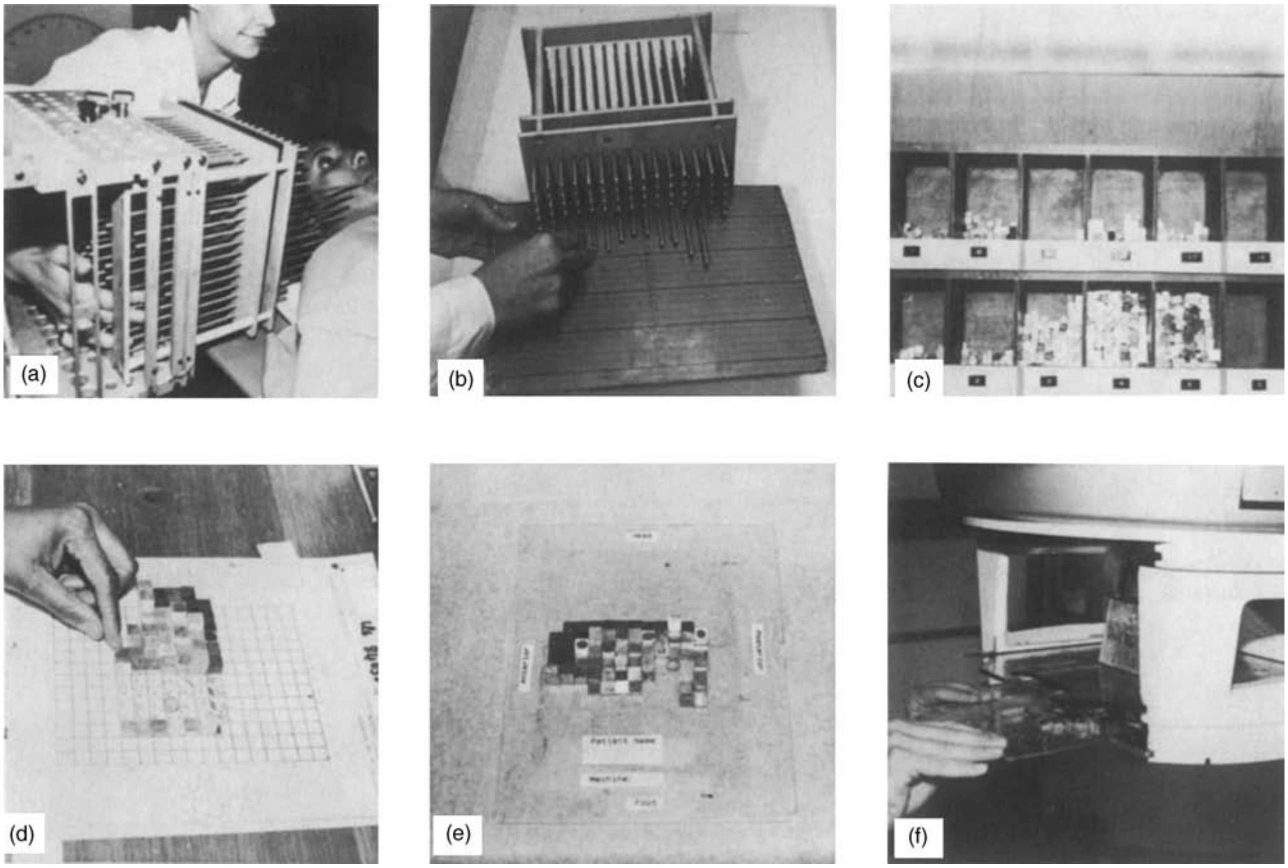


Figure 33. Compensating filter. Composite photograph illustrating design and fabrication procedure: (a) Push rod device is attached to the head of the simulator directly over the patient in the treatment position. The rods are lowered one by one, until they are in contact with the patient's surface. (b) Tissue deficits are determined from measurement of each rod length. (c) Aluminum and brass blocks, which are numbered to correspond to the tissue deficits determined, are selected from storage bins. (d) Blocks are attached on the plastic tray according to the pattern specified by the calculations. (e) Completed filter. (f) Compensating filter positioned on treatment machine.

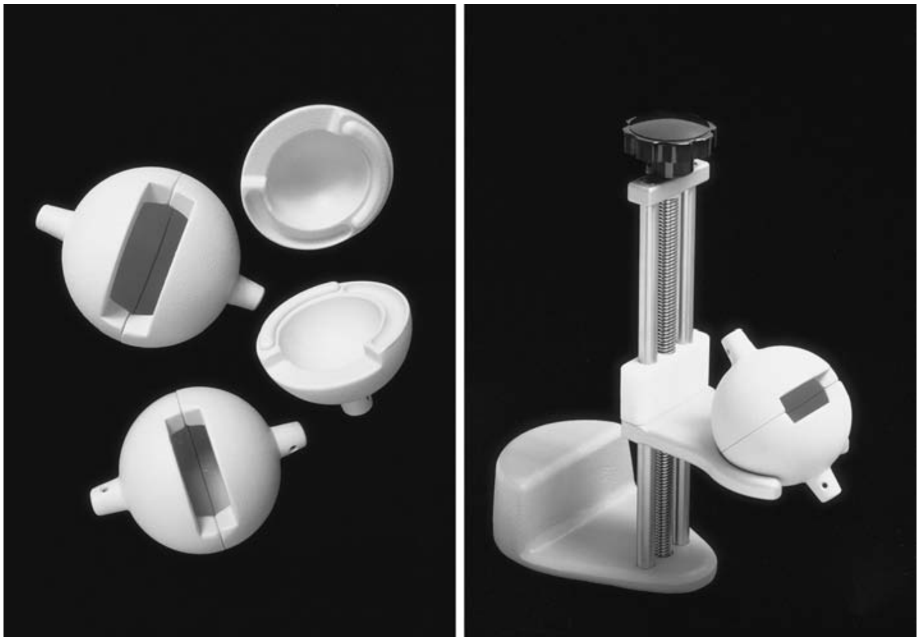


Figure 34. Testicular Shield. Device used to position and shield the gonadal area from scattered radiation. (Courtesy of MEDTEC.)

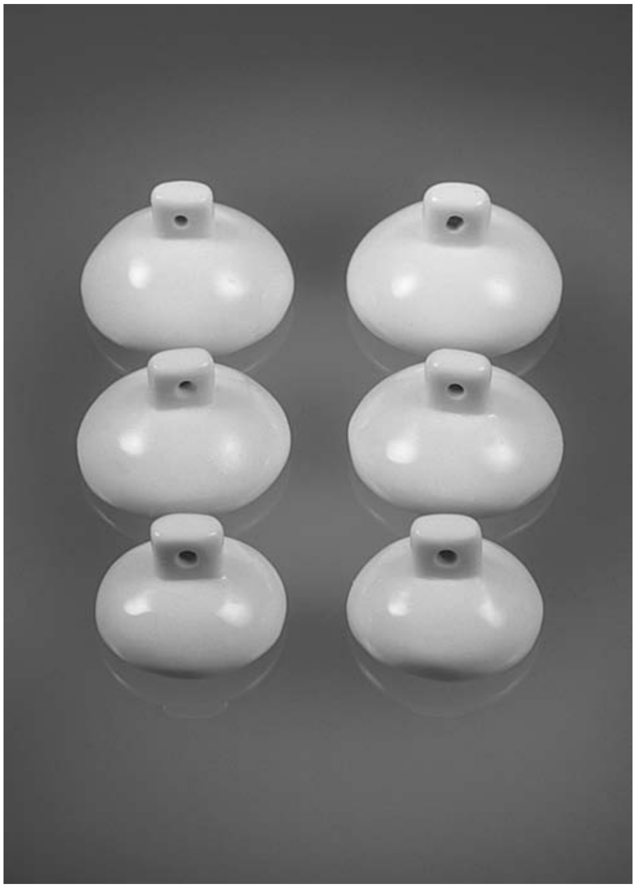


Figure 35. Tungsten eye shields provide protection of the ocular structure for electrons up to 9 MeV. Each eye shield is coated with a 2 mm minimum thickness of dental acrylic on the beam entrance of the shield to reduce electron backscatter to an acceptable level. (Courtesy of MEDTEC.)

Fig. 35 that provide protection of the ocular structure for electrons up to 9 MeV. Each eye shield is made of tungsten and coated with a 2 mm minimum thickness of dental acrylic on the beam entrance of the shield to reduce electron backscatter to an acceptable level.

TREATMENT VERIFICATION AND QUALITY ASSURANCE DEVICES

Radiographic film (port films) using film cassettes with lead or copper filters which improve the radiographic contrast of the port films are typically used to verify patient isocenter position and for portal shape (Fig. 36). Devices to support the port film cassettes and to help insure that the film plane is orthogonal to the beam direction and close to the patient's surface from which the beam exits are also important accessories (Fig. 37). In addition, radiopaque graticules (Fig. 38) that can be inserted into the treatment beam are invaluable for the evaluation of port films (30). Typically, such devices consist of platinum or tungsten wire embedded in plastic and molded in a frame that can be attached to the treatment machine accessory mount.

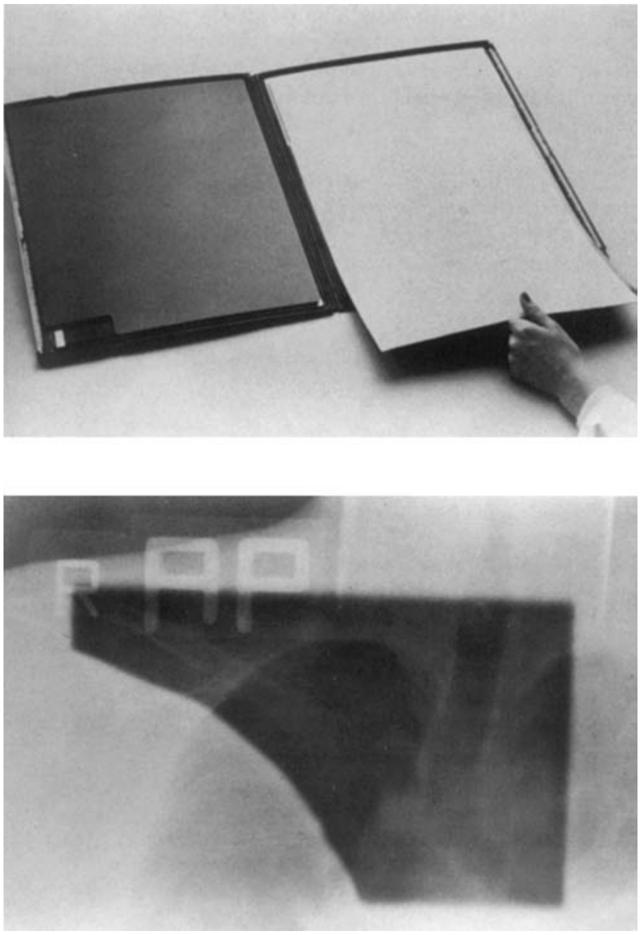


Figure 36. Radiotherapy port film cassette. Copper screens are used to improve subject contrast and improve resolution as shown in the portal localization radiograph. (Courtesy Eastman Kodak Co.)

The generally poor quality of the film images and the inconveniences of the film processing and physician reviewing procedure have spurred development of computer-aided enhancement and digital imaging techniques in radiation therapy. Computed radiography (CR) systems, such as the Kodak 2000RT CR system shown in Fig. 39, is an example of such a system that allows digital DICOM images to be distributed electronically throughout the department.

In addition, electronic portal imagers (EPID) have made great strides this past decade and such devices are poised to replace film over the next few years (31,32). More recently, linac manufacturers have integrated EPID and tomographic imaging systems on their linacs (Fig. 40) for localization of bone and soft-tissue targets and have set the stage for image-guided radiation therapy to move into routine practice (33,34). Such systems will make quantitative evaluation of immobilization and repositioning of the patient much more achievable by allowing daily imaging of the patient's treatment.

In addition to imaging verification, there is sometimes a need to verify actual dose delivered to the patient. Simple point dose verification can be achieved using TLDs, diodes,

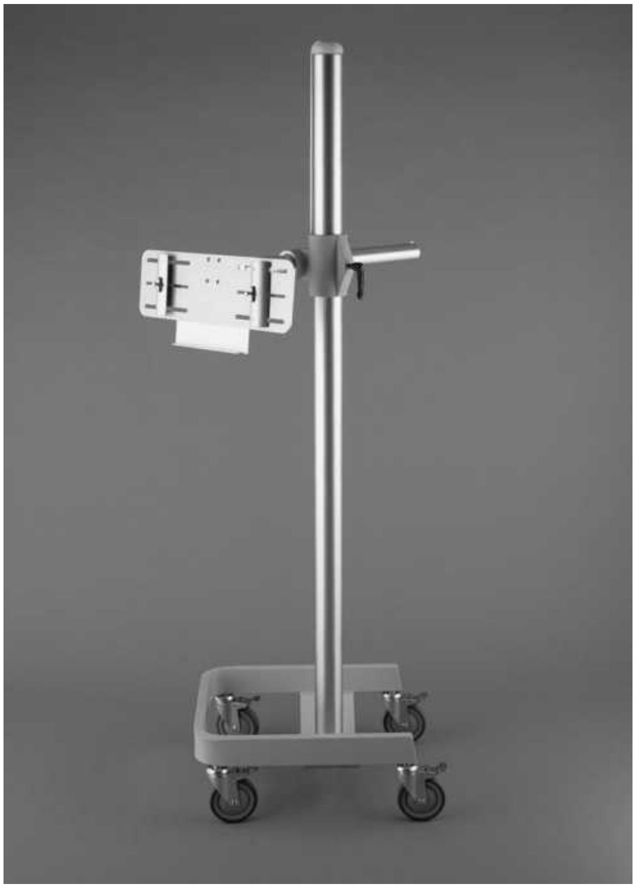


Figure 37. Port film cassette holder. Devices used to support a port film cassette behind the patient in any orientation device are especially useful for oblique treatment angles as the plane of film can be adjusted so as to produce normal incidence of the radiation field. (Courtesy of MEDTEC.)

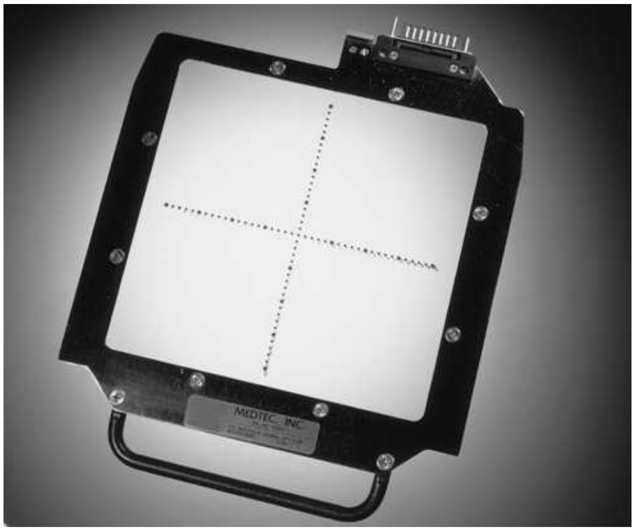


Figure 38. Example of a port film fiducial grid. Device identifies the central axis of the radiation beam and provides a scale at the calibration distance, thus providing a magnification factor for the port film. (Courtesy of MEDTEC.)

Figure 39. Kodak 2000RT CR system. Computed radiography system that allows digital DICOM images to be distributed electronically throughout the department. (Courtesy Eastman Kodak Co.)

or MOSFET dosimeters placed on the patient's skin surface or in body cavities. The most recent development in these types of devices is the OneDose patient dosimetry system (Fig. 41) developed by Sicel Technologies, Inc. It consists of a wireless handheld reader which interacts with self-adhesive external MOSFET dosimeters placed on the patient. To use, the therapist simply places the precalibrated dosimeters on the patient's skin in the treatment field, treats the patient, and then slides the dosimeter into the reader

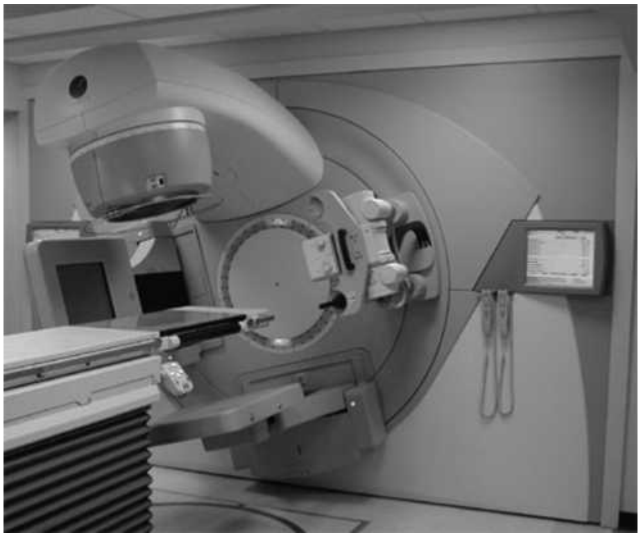


Figure 40. Elekta Synergy®system. Linac with electronic portal imager (EPID) and conebeam tomographic imaging system for localization of bone and soft-tissue targets. (Courtesy of Elekta AB.)

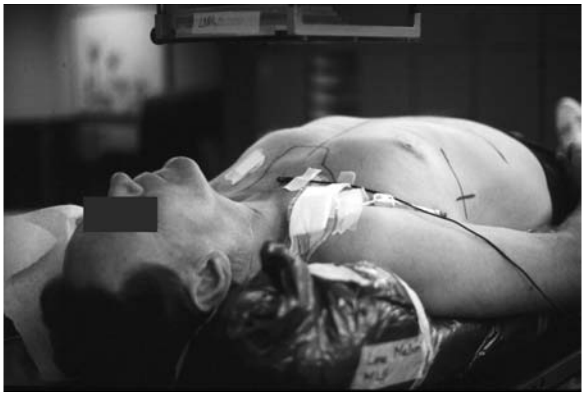

[redacted]

Figure 41. Patient dosimetry monitoring system. Semiconductor detectors are typically used and applied on the patient's surface using surgical tape. (Courtesy of MEDTEC.)

for an immediate display. The reader automatically provides a permanent record of dosage, time, and date with minimal data entry. Another dose verification device is shown in Fig. 42; because IMRT treatments presently require verification of the dose delivered and pattern for each patient, special phantoms and/or check devices have been developed to facilitate the IMRT verification measurements.

Also, there are numerous QA devices available to check the constancy of the linac's beam calibration, symmetry, and radiation-light field alignment. Generally, the QA radiation detection devices consist of several ionization chambers or semiconductors positioned in a plastic phantom that can be placed in the radiation beam (Fig. 43).

Finally, one of the latest advances is a system (still under development) that will be capable of performing continuous objective, real-time tracking of the target volume during treatment (Calypso Medical Technologies, Inc.) (35). The system is based on alternating current (ac) magnetic fields utilizing permanently implantable wireless transponders that do not require additional ionizing radiation and do not depend on subjective interpretation of images. The system is currently undergoing clinical evaluation and is not yet available for clinical use.

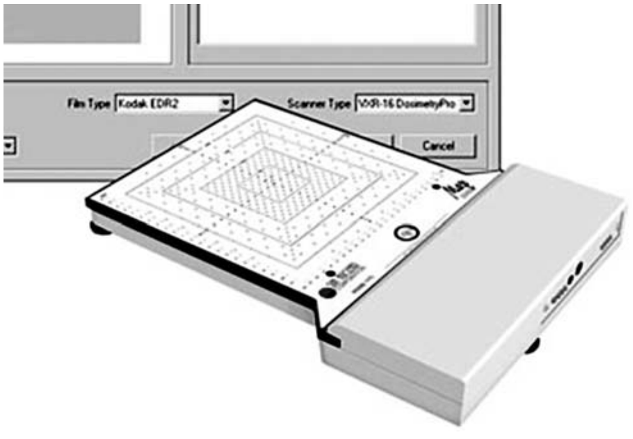


Figure 42. MapCHECK device provides 2D therapy beam measurements intended for quick and precise verification of the dose distribution resulting from an IMRT plan. (Courtesy of Sun Nuclear.)

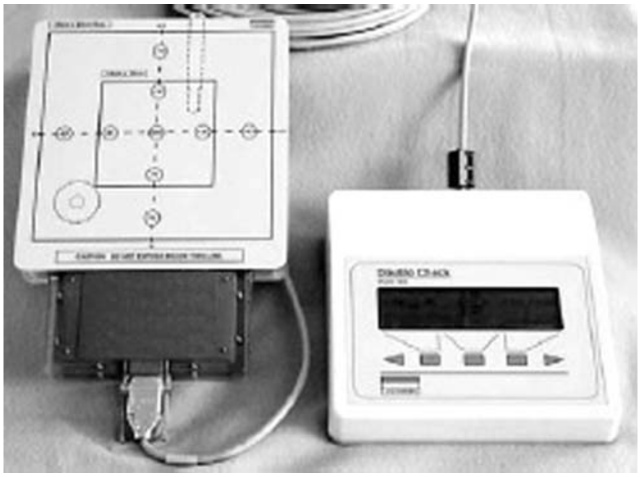


Figure 43. DoubleCheck A matrix detector, consisting of multiple ionization chambers embedded in a plastic phantom, used to check the constancy of radiation beam output , symmetry, and flatness. (Courtesy MED-TEC, Inc.)

SUMMARY

In summary, numerous radiotherapy accessories are used to aid in planning, delivering, and verifying radiation treatments. New systems continue to be developed that give the ability to more accurately position the patient and account for the internal target volume relative to the treatment machine's isocenter. Such devices enable significant reductions in the amount of normal tissue included in the irradiated volume. Finally, it should be recognized that the preferences of the individual radiation oncologist, radiation therapist, and clinical physicist still play a major role in the acceptance and use of all of the types of devices discussed in this article. The major considerations for any specific application are typically cost, speed and ease of preparation, ease of use, and effectiveness.

BIBLIOGRAPHY

1. Levitt SH, Khan FM, Potish RA, Perez CA, editors. Technological Basis of Radiation Therapy: Clinical Applications. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 1999.
2. ICRU, Report No. 24, Determination of Absorbed Dose in a Patient Irradiated by Beams of X or Gamma Rays in Radiotherapy Procedures. Washington, D.C.: International Commission on Radiation Units and Measurements. 1976.
3. Purdy JA. 3-D radiation treatment planning: a new era, in Frontiers of Radiation Therapy and Oncology. 3-D Conformal Radiotherapy: A New Era in the Irradiation of Cancer. In: Meyer JL, Purdy JA, editors. Basel: Karger; 1996. p 1–16.
4. Purdy JA. Defining our goals: volume and dose specification for 3-D conformal radiation therapy, in Frontiers of Radiation Therapy and Oncology. 3-D Conformal Radiotherapy: A New Era in the Irradiation of Cancer. In: Meyer JL, Purdy JA, editors. Basel: Karger; 1996. p 24–30.

5. Purdy JA, Klein EE, Low DA. Quality assurance and safety of new technologies for radiation oncology. Sem Radiat Oncol 1995;5(2):156–165.
6. Van Dyk J, Mah K. Simulators and CT scanners. In: Williams JR, Thwaites DI, editors. Radiotherapy Physics. New York: Oxford Medical Publications; 1993.
7. Prasad S, Pilepich MV, Perez CA. Contribution of CT to quantitative radiation therapy planning. Am J Radiol 1981;136:123.
8. Goitein M. Applications of computed tomography in radiotherapy treatment planning. Progress in Medical Radiation Physics. 1 In: Orton CG, editor. New York: Plenum; 1982. p 195–293.
9. Perez CA, et al. Design of a fully integrated three-dimensional computed tomography simulator and preliminary clinical evaluation. Init J Radiat Oncol Biol Phys 1994;30(4):887–897.
10. Mutic S, et al. Quality assurance for computed-tomography simulators and the computed-tomography-simulation process: Report of the AAPM Radiation Therapy Committee Task Group No. 68. Med Phys 2003;30(10):2762–2792.
11. Watkins DMB. Radiation Therapy Mold Technology. Toronto, Canada: Pergamon Press; 1981.
12. Buck BA, Siddon RL, Svensson GK. A beam alignment device for matching fields. Int J Radiat Oncol Biol Phys 1985;11:1939.
13. Purdy JA, et al. 3-D Conformal and Intensity Modulated Radiation Therapy: Physics and Clinical Applications. Madison, (WI): Advanced Medical Publishing, Inc; 2001. p 612.
14. Sternick ES, editor. The Theory and Practice of Intensity Modulated Radiation Therapy. Madison (WI): Advanced Medical Publishing; 1997. p 254.
15. Balter JM, et al. Measurement of prostate movement over the course of routine radiotherapy using implanted markers. Int J Radiat Oncol Biol Phys 1995;31:113–118.
16. Lattanzi J. A comparison of daily CT localization to a daily ultrasound-based system in prostate cancer. Int J Radiat Oncol Biol Phys 1999;43(4):719–725.
17. Mageras GS. Flouroscopic evaluation of diaphragmatic motion reduction with a respiratory gated radiotherapy system. J Appl Clin Med Phys 2001;2(4):191–200.
18. Vedam SS, Keall PJ, Kini VR, Mohan R. Determining parameters for respiration-gated radiotherapy. Med Phys 2001;28(10): 2139–2146.
19. Wong J, et al. The use of active breathing control (ABC) to reduce margin for breathing control. Int J Radiat Oncol Biol Phys 1999;44:911–919.
20. Verellen D, et al. Quality assurance of a system for improved target localization and patient set-up combines real-time infrared tracking and stereoscopic X-ray imaging. Radio Oncol 2003;67(1):129–141.
21. Schlegel W, et al. Computer systems and mechanical tools for stereotactically guided conformal therapy with linear accelerators. Int J Radiat Oncol Biol Phys 1992;24:781.
22. Lax I, et al. Stereotactic radiotherapy of malignancies in the abdomen: methodological aspects. Acta Oncol 1994;33:677–683.
23. Powers WE, et al. A new system of field shaping for external-beam radiation therapy. Radiology 1973;108:407–411.
24. Leavitt DD, FA Gibbs Jr. Field shaping, in Advances in Radiation Oncology Physics: Dosimetry, Treatment Planning, and Brachytherapy. In: Purdy JA, editor. New York: American Institute of Physics; 1992. pp. 500–523.
25. Takahashi S. Conformation radiotherapy: rotation techniques as applied to radiography and radiotherapy of cancer. Acta Radiol (Suppl) 1965;242:1–42.
26. Carol MP. Integrated 3-D conformal multivane intensity modulation delivery system for radiotherapy. In Proceedings of the 11th International Conference on the Use of Computers in Radiation Therapy. Madison (WI): Medical Physics Publishing; 1994.
27. Leavitt DD, et al. Dynamic wedge field techniques through computer-controlled collimator motion and dose delivery. Med Phys 1990;17:87–91.
28. Kijewski PK, Chin LM, Bjarngard BE. Wedge-shaped dose distributions by computer-controlled collimator motion. Med Phys 1978;5(5):426–429.
29. Ellis F, Hall EJ, Oliver R. A compensator for variations in tissue thickness for high energy beams. Br J Radiol 1959;32:421–422.
30. van de Geijn J, Harrington FS, Fraass B. A graticule for evaluation of megavoltage X-ray port films. Int J Radiat Oncol Biol Phys 1982;8:1999.
31. Herman MG, et al. Clinical use of electronic portal imaging: Report of AAPM RadiationTherapy Committee Task Group 58. Med Phys 2001;28(5):712–737.
32. Antonuk LE. Electronic portal imaging devices: a review and historical perspective of contemporary technologies and research. Phys Med Biol 2002;47(6):R31–R65.
33. Jaffray DA, et al. A radiographic and tomographic imaging system integrated into a medical linear accelerator for localization of bone and soft-tissue targets. Int J Radia Oncol Biol Phys 1999;45:773–789.
34. Jaffray DA, Siewerdsen JH, Wong JW, Martinez AA. Flat-panel cone-beam computed tomography for image-guided radiation therapy. Int J Radia Oncol Biol Phys 2002;53(5): 1337–1349.
35. Mate TP, et al. Principles of AC magnetic fields for objective and continuous target localization in radiation therapy (abstract). Init J Radiat Oncol Biol Phys 2004;60(1):S455.

See also RADIATION THERAPY SIMULATOR; RADIATION PROTECTION INSTRUMENTATION.