The Biomedical Engineering Handbook
Third Edition

Medical Devices and Systems

The Electrical Engineering Handbook Series

Series Editor

Richard C. Dorf

University of California, Davis

Titles Included in the Series

The Handbook of Ad Hoc Wireless Networks, Mohammad Ilyas
The Avionics Handbook, Cary R. Spitzer
The Biomedical Engineering Handbook, *Third Edition,* Joseph D. Bronzino
The Circuits and Filters Handbook, Second Edition, Wai-Kai Chen
The Communications Handbook, Second Edition, Jerry Gibson
The Computer Engineering Handbook, Vojin G. Oklobdzija
The Control Handbook, William S. Levine
The CRC Handbook of Engineering Tables, Richard C. Dorf
The Digital Signal Processing Handbook, Vijay K. Madisetti and Douglas Williams
The Electrical Engineering Handbook, *Third Edition,* Richard C. Dorf
The Electric Power Engineering Handbook, Leo L. Grigsby
The Electronics Handbook, *Second Edition,* Jerry C. Whitaker
The Engineering Handbook, Third Edition, Richard C. Dorf
The Handbook of Formulas and Tables for Signal Processing, Alexander D. Poularikas
The Handbook of Nanoscience, Engineering, and Technology, William A. Goddard, III, Donald W. Brenner, Sergey E. Lyshevski, and Gerald J. Iafrate
The Handbook of Optical Communication Networks, Mohammad Ilyas and Hussein T. Mouftah
The Industrial Electronics Handbook, J. David Irwin
The Measurement, Instrumentation, and Sensors Handbook, John G. Webster
The Mechanical Systems Design Handbook, Osita D.I. Nwokah and Yidirim Hurmuzlu
The Mechatronics Handbook, Robert H. Bishop
The Mobile Communications Handbook, *Second Edition*, Jerry D. Gibson
The Ocean Engineering Handbook, Ferial El-Hawary
The RF and Microwave Handbook, Mike Golio
The Technology Management Handbook, Richard C. Dorf
The Transforms and Applications Handbook, *Second Edition,* Alexander D. Poularikas
The VLSI Handbook, Wai-Kai Chen

The Biomedical Engineering Handbook

Third Edition

Edited by

Joseph D. Bronzino

Biomedical Engineering Fundamentals

Medical Devices and Systems

Tissue Engineering and Artificial Organs

The Biomedical Engineering Handbook

Third Edition

Medical Devices and Systems

Edited by

Joseph D. Bronzino

Trinity College
Hartford, Connecticut, U.S.A.

CRC Taylor & Francis
Taylor & Francis Group
Boca Raton London New York

A CRC title, part of the Taylor & Francis imprint, a member of the Taylor & Francis Group, the academic division of T&F Informa plc.

Published in 2006 by
CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

No claim to original U.S. Government works
Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-10: 0-8493-2122-0 (Hardcover)
International Standard Book Number-13: 978-0-8493-2122-1 (Hardcover)
Library of Congress Card Number 2005056892

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Medical devices and systems / edited by Joseph D. Bronzino.
p. cm. -- (The electrical engineering handbook series)
Includes bibliographical references and index.
ISBN 0-8493-2122-0
1. Medical instruments and apparatus--Handbooks, manuals, etc. I. Bronzino, Joseph D., 1937- II. Title. III. Series.

R856.15.B76 2006
610.28--dc22 2005056892

informa
Taylor & Francis Group
is the Academic Division of Informa plc.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Introduction and Preface

During the past five years since the publication of the Second Edition — a two-volume set — of the *Biomedical Engineering Handbook*, the field of biomedical engineering has continued to evolve and expand. As a result, this Third Edition consists of a three volume set, which has been significantly modified to reflect the state-of-the-field knowledge and applications in this important discipline. More specifically, this Third Edition contains a number of completely new sections, including:

- Molecular Biology
- Bionanotechnology
- Bioinformatics
- Neuroengineering
- Infrared Imaging

as well as a new section on ethics.

In addition, all of the sections that have appeared in the first and second editions have been significantly revised. Therefore, this Third Edition presents an excellent summary of the status of knowledge and activities of biomedical engineers in the beginning of the 21st century.

As such, it can serve as an excellent reference for individuals interested not only in a review of fundamental physiology, but also in quickly being brought up to speed in certain areas of biomedical engineering research. It can serve as an excellent textbook for students in areas where traditional textbooks have not yet been developed and as an excellent review of the major areas of activity in each biomedical engineering subdiscipline, such as biomechanics, biomaterials, bioinstrumentation, medical imaging, etc. Finally, it can serve as the "bible" for practicing biomedical engineering professionals by covering such topics as a historical perspective of medical technology, the role of professional societies, the ethical issues associated with medical technology, and the FDA process.

Biomedical engineering is now an important vital interdisciplinary field. Biomedical engineers are involved in virtually all aspects of developing new medical technology. They are involved in the design, development, and utilization of materials, devices (such as pacemakers, lithotripsy, etc.) and techniques (such as signal processing, artificial intelligence, etc.) for clinical research and use; and serve as members of the health care delivery team (clinical engineering, medical informatics, rehabilitation engineering, etc.) seeking new solutions for difficult health care problems confronting our society. To meet the needs of this diverse body of biomedical engineers, this handbook provides a central core of knowledge in those fields encompassed by the discipline. However, before presenting this detailed information, it is important to provide a sense of the evolution of the modern health care system and identify the diverse activities biomedical engineers perform to assist in the diagnosis and treatment of patients.

Evolution of the Modern Health Care System

Before 1900, medicine had little to offer the average citizen, since its resources consisted mainly of the physician, his education, and his "little black bag." In general, physicians seemed to be in short supply, but the shortage had rather different causes than the current crisis in the availability of health care professionals. Although the costs of obtaining medical training were relatively low, the demand for doctors' services also was very small, since many of the services provided by the physician also could be obtained from experienced amateurs in the community. The home was typically the site for treatment and recuperation, and relatives and neighbors constituted an able and willing nursing staff. Babies were delivered by midwives, and those illnesses not cured by home remedies were left to run their natural, albeit frequently fatal, course. The contrast with contemporary health care practices, in which specialized physicians and nurses located within the hospital provide critical diagnostic and treatment services, is dramatic.

The changes that have occurred within medical science originated in the rapid developments that took place in the applied sciences (chemistry, physics, engineering, microbiology, physiology, pharmacology, etc.) at the turn of the century. This process of development was characterized by intense interdisciplinary cross-fertilization, which provided an environment in which medical research was able to take giant strides in developing techniques for the diagnosis and treatment of disease. For example, in 1903, Willem Einthoven, a Dutch physiologist, devised the first electrocardiograph to measure the electrical activity of the heart. In applying discoveries in the physical sciences to the analysis of the biologic process, he initiated a new age in both cardiovascular medicine and electrical measurement techniques.

New discoveries in medical sciences followed one another like intermediates in a chain reaction. However, the most significant innovation for clinical medicine was the development of x-rays. These "new kinds of rays," as their discoverer W.K. Roentgen described them in 1895, opened the "inner man" to medical inspection. Initially, x-rays were used to diagnose bone fractures and dislocations, and in the process, x-ray machines became commonplace in most urban hospitals. Separate departments of radiology were established, and their influence spread to other departments throughout the hospital. By the 1930s, x-ray visualization of practically all organ systems of the body had been made possible through the use of barium salts and a wide variety of radiopaque materials.

X-ray technology gave physicians a powerful tool that, for the first time, permitted accurate diagnosis of a wide variety of diseases and injuries. Moreover, since x-ray machines were too cumbersome and expensive for local doctors and clinics, they had to be placed in health care centers or hospitals. Once there, x-ray technology essentially triggered the transformation of the hospital from a passive receptacle for the sick to an active curative institution for all members of society.

For economic reasons, the centralization of health care services became essential because of many other important technological innovations appearing on the medical scene. However, hospitals remained institutions to dread, and it was not until the introduction of sulfanilamide in the mid-1930s and penicillin in the early 1940s that the main danger of hospitalization, that is, cross-infection among patients, was significantly reduced. With these new drugs in their arsenals, surgeons were able to perform their operations without prohibitive morbidity and mortality due to infection. Furthermore, even though the different blood groups and their incompatibility were discovered in 1900 and sodium citrate was used in 1913 to prevent clotting, full development of blood banks was not practical until the 1930s, when technology provided adequate refrigeration. Until that time, "fresh" donors were bled and the blood transfused while it was still warm.

Once these surgical suites were established, the employment of specifically designed pieces of medical technology assisted in further advancing the development of complex surgical procedures. For example, the Drinker respirator was introduced in 1927 and the first heart-lung bypass in 1939. By the 1940s, medical procedures heavily dependent on medical technology, such as cardiac catheterization and angiography (the use of a cannula threaded through an arm vein and into the heart with the injection of radiopaque dye) for the x-ray visualization of congenital and acquired heart disease (mainly valve disorders due to rheumatic fever) became possible, and a new era of cardiac and vascular surgery was established.

Following World War II, technological advances were spurred on by efforts to develop superior weapon systems and establish habitats in space and on the ocean floor. As a by-product of these efforts, the development of medical devices accelerated and the medical profession benefited greatly from this rapid surge of technological finds. Consider the following examples:

1. Advances in solid-state electronics made it possible to map the subtle behavior of the fundamental unit of the central nervous system — the neuron — as well as to monitor the various physiological parameters, such as the electrocardiogram, of patients in intensive care units.
2. New prosthetic devices became a goal of engineers involved in providing the disabled with tools to improve their quality of life.
3. Nuclear medicine — an outgrowth of the atomic age — emerged as a powerful and effective approach in detecting and treating specific physiologic abnormalities.
4. Diagnostic ultrasound based on sonar technology became so widely accepted that ultrasonic studies are now part of the routine diagnostic workup in many medical specialties.
5. "Spare parts" surgery also became commonplace. Technologists were encouraged to provide cardiac assist devices, such as artificial heart valves and artificial blood vessels, and the artificial heart program was launched to develop a replacement for a defective or diseased human heart.
6. Advances in materials have made the development of disposable medical devices, such as needles and thermometers, as well as implantable drug delivery systems, a reality.
7. Computers similar to those developed to control the flight plans of the *Apollo* capsule were used to store, process, and cross-check medical records, to monitor patient status in intensive care units, and to provide sophisticated statistical diagnoses of potential diseases correlated with specific sets of patient symptoms.
8. Development of the first computer-based medical instrument, the computerized axial tomography scanner, revolutionized clinical approaches to noninvasive diagnostic imaging procedures, which now include magnetic resonance imaging and positron emission tomography as well.
9. A wide variety of new cardiovascular technologies including implantable defibrillators and chemically treated stents were developed.
10. Neuronal pacing systems were used to detect and prevent epileptic seizures.
11. Artificial organs and tissue have been created.
12. The completion of the genome project has stimulated the search for new biological markers and personalized medicine.

The impact of these discoveries and many others has been profound. The health care system of today consists of technologically sophisticated clinical staff operating primarily in modern hospitals designed to accommodate the new medical technology. This evolutionary process continues, with advances in the physical sciences such as materials and nanotechnology, and in the life sciences such as molecular biology, the genome project and artificial organs. These advances have altered and will continue to alter the very nature of the health care delivery system itself.

Biomedical Engineering: A Definition

Bioengineering is usually defined as a basic research-oriented activity closely related to biotechnology and genetic engineering, that is, the modification of animal or plant cells, or parts of cells, to improve plants or animals or to develop new microorganisms for beneficial ends. In the food industry, for example, this has meant the improvement of strains of yeast for fermentation. In agriculture, bioengineers may be concerned with the improvement of crop yields by treatment of plants with organisms to reduce frost damage. It is clear that bioengineers of the future will have a tremendous impact on the qualities of human life. The potential of this specialty is difficult to imagine. Consider the following activities of bioengineers:

- Development of improved species of plants and animals for food production
- Invention of new medical diagnostic tests for diseases

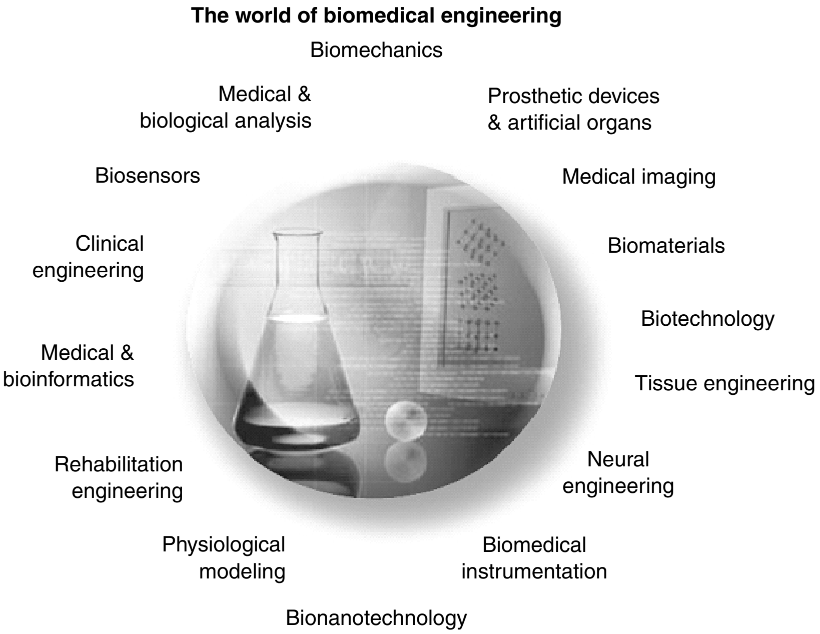


FIGURE 1 The World of Biomedical Engineering.

- Production of synthetic vaccines from clone cells
- Bioenvironmental engineering to protect human, animal, and plant life from toxicants and pollutants
- Study of protein–surface interactions
- Modeling of the growth kinetics of yeast and hybridoma cells
- Research in immobilized enzyme technology
- Development of therapeutic proteins and monoclonal antibodies

Biomedical engineers, on the other hand, apply electrical, mechanical, chemical, optical, and other engineering principles to understand, modify, or control biologic (i.e., human and animal) systems, as well as design and manufacture products that can monitor physiologic functions and assist in the diagnosis and treatment of patients. When biomedical engineers work within a hospital or clinic, they are more properly called clinical engineers.

Activities of Biomedical Engineers

The breadth of activity of biomedical engineers is now significant. The field has moved from being concerned primarily with the development of medical instruments in the 1950s and 1960s to include a more wide-ranging set of activities. As illustrated below, the field of biomedical engineering now includes many new career areas (see Figure 1), each of which is presented in this handbook. These areas include:

- Application of engineering system analysis (physiologic modeling, simulation, and control) to biologic problems
- Detection, measurement, and monitoring of physiologic signals (i.e., biosensors and biomedical instrumentation)
- Diagnostic interpretation via signal-processing techniques of bioelectric data
- Therapeutic and rehabilitation procedures and devices (rehabilitation engineering)
- Devices for replacement or augmentation of bodily functions (*artificial organs*)

- Computer analysis of patient-related data and clinical decision making (i.e., medical informatics and artificial intelligence)
- Medical imaging, that is, the graphic display of anatomic detail or physiologic function
- The creation of new biologic products (i.e., *biotechnology* and *tissue engineering*)
- The development of new materials to be used within the body (biomaterials)

Typical pursuits of biomedical engineers, therefore, include:

- Research in new materials for implanted artificial organs
- Development of new diagnostic instruments for blood analysis
- Computer modeling of the function of the human heart
- Writing software for analysis of medical research data
- Analysis of medical device hazards for safety and efficacy
- Development of new diagnostic imaging systems
- Design of telemetry systems for patient monitoring
- Design of biomedical sensors for measurement of human physiologic systems variables
- Development of expert systems for diagnosis of disease
- Design of closed-loop control systems for drug administration
- Modeling of the physiological systems of the human body
- Design of instrumentation for sports medicine
- Development of new dental materials
- Design of communication aids for the handicapped
- Study of pulmonary fluid dynamics
- Study of the biomechanics of the human body
- Development of material to be used as replacement for human skin

Biomedical engineering, then, is an interdisciplinary branch of engineering that ranges from theoretical, nonexperimental undertakings to state-of-the-art applications. It can encompass research, development, implementation, and operation. Accordingly, like medical practice itself, it is unlikely that any single person can acquire expertise that encompasses the entire field. Yet, because of the interdisciplinary nature of this activity, there is considerable interplay and overlapping of interest and effort between them. For example, biomedical engineers engaged in the development of biosensors may interact with those interested in prosthetic devices to develop a means to detect and use the same bioelectric signal to power a prosthetic device. Those engaged in automating the clinical chemistry laboratory may collaborate with those developing expert systems to assist clinicians in making decisions based on specific laboratory data. The possibilities are endless.

Perhaps a greater potential benefit occurring from the use of biomedical engineering is identification of the problems and needs of our present health care system that can be solved using existing engineering technology and systems methodology. Consequently, the field of biomedical engineering offers hope in the continuing battle to provide high-quality care at a reasonable cost. If properly directed toward solving problems related to preventive medical approaches, ambulatory care services, and the like, biomedical engineers can provide the tools and techniques to make our health care system more effective and efficient; and in the process, improve the quality of life for all.

Joseph D. Bronzino
Editor-in-Chief

Editor-in-Chief

Joseph D. Bronzino received the B.S.E.E. degree from Worcester Polytechnic Institute, Worcester, MA, in 1959, the M.S.E.E. degree from the Naval Postgraduate School, Monterey, CA, in 1961, and the Ph.D. degree in electrical engineering from Worcester Polytechnic Institute in 1968. He is presently the Vernon Roosa Professor of Applied Science, an endowed chair at Trinity College, Hartford, CT and President of the Biomedical Engineering Alliance and Consortium (BEACON) which is a nonprofit organization consisting of academic and medical institutions as well as corporations dedicated to the development and commercialization of new medical technologies (for details visit www.beaconalliance.org).

He is the author of over 200 articles and 11 books including the following: *Technology for Patient Care* (C.V. Mosby, 1977), *Computer Applications for Patient Care* (Addison-Wesley, 1982), *Biomedical Engineering: Basic Concepts and Instrumentation* (PWS Publishing Co., 1986), *Expert Systems: Basic Concepts* (Research Foundation of State University of New York, 1989), *Medical Technology and Society: An Interdisciplinary Perspective* (MIT Press and McGraw-Hill, 1990), *Management of Medical Technology* (Butterworth/Heinemann, 1992), *The Biomedical Engineering Handbook* (CRC Press, 1st ed., 1995; 2nd ed., 2000; Taylor & Francis, 3rd ed., 2005), *Introduction to Biomedical Engineering* (Academic Press, 1st ed., 1999; 2nd ed., 2005).

Dr. Bronzino is a fellow of IEEE and the American Institute of Medical and Biological Engineering (AIMBE), an honorary member of the Italian Society of Experimental Biology, past chairman of the Biomedical Engineering Division of the American Society for Engineering Education (ASEE), a charter member and presently vice president of the Connecticut Academy of Science and Engineering (CASE), a charter member of the American College of Clinical Engineering (ACCE) and the Association for the Advancement of Medical Instrumentation (AAMI), past president of the IEEE-Engineering in Medicine and Biology Society (EMBS), past chairman of the IEEE Health Care Engineering Policy Committee (HCEPC), past chairman of the IEEE Technical Policy Council in Washington, DC, and presently Editor-in-Chief of Elsevier's BME Book Series and Taylor & Francis' *Biomedical Engineering Handbook.*

Dr. Bronzino is also the recipient of the Millennium Award from IEEE/EMBS in 2000 and the Goddard Award from Worcester Polytechnic Institute for Professional Achievement in June 2004.

Contributors

Joseph Adam
Premise Development Corporation
Hartford, Connecticut

P.D. Ahlgren
Ville Marie Multidisciplinary Breast and Oncology Center
St. Mary's Hospital
McGill University
Montreal, Quebec, Canada
and
London Cancer Centre
London, Ontario
Canada

William C. Amalu
Pacific Chiropractic and Research Center
Redwood City, California

Kurt Ammer
Ludwig Boltzmann Research Institute for Physical Diagnostics
Vienna, Austria
and
Medical Imaging Research Group
School of Computing
University of Glamorgan
Pontypridd, Wales
United Kingdom

Dennis D. Autio
Dybonics, Inc.
Portland, Oregon

Raymond Balcerak
Defense Advanced Research Projects Agency
Arlington, Virginia

D.C. Barber
University of Sheffield
Sheffield, United Kingdom

Khosrow Behbehani
The University of Texas at Arlington
Arlington, Texas
and
The University of Texas
Southwestern Medical Center
Dallas, Texas

N. Belliveau
Ville Marie Multidisciplinary Breast and Oncology Center
St. Mary's Hospital
McGill University
Montreal, Quebec, Canada
and
London Cancer Centre
London, Ontario, Canada

Anna M. Bianchi
St. Raffaele Hospital
Milan, Italy

Carol J. Bickford
American Nurses Association
Washington, D.C.

Jeffrey S. Blair
IBM Health Care Solutions
Atlanta, Georgia

G. Faye Boudreaux-Bartels
University of Rhode Island
Kingston, Rhode Island

Bruce R. Bowman
EdenTec Corporation
Eden Prairie, Minnesota

Joseph D. Bronzino
Trinity College
Biomedical Engineering Alliance and Consortium (BEACON)
Harford, Connecticut

Mark E. Bruley
ECRI
Plymouth Meeting, Pennsylvania

Richard P. Buck
University of North Carolina
Chapel Hill, North Carolina

P. Buddharaju
Department of Computer Science
University of Houston
Houston, Texas

Thomas F. Budinger
University of California-Berkeley
Berkeley, California

Robert D. Butterfield
IVAC Corporation
San Diego, California

Joseph P. Cammarota
Naval Air Warfare Center
Aircraft Division
Warminster, Pennsylvania

Paul Campbell
Institute of Medical Science
and Technology
Universities of St. Andrews
and Dundee
and
Ninewells Hospital
Dundee, United Kingdom

Ewart R. Carson
City University
London, United Kingdom

Sergio Cerutti
Polytechnic University
Milan, Italy

A. Enis Çetin
Bilkent University
Ankara, Turkey

Christopher S. Chen
Department of Bioengineering
Department of Physiology
University of Pennsylvania
Philadelphia, Pennsylvania

Wei Chen
Center for Magnetic Resonance
Research
and
The University of Minnesota
Medical School
Minneapolis, Minnesota

Victor Chernomordik
Laboratory of Integrative and
Medical Biophysics
National Institute of Child Health
and Human Development
Bethesda, Maryland

David A. Chesler
Massachusetts General Hospital
Harvard University Medical
School
Boston, Massachusetts

Vivian H. Coates
ECRI
Plymouth Meeting, Pennsylvania

Arnon Cohen
Ben-Gurion University
Be'er Sheva, Israel

Steven Conolly
Stanford University
Stanford, California

Derek G. Cramp
City University
London, United Kingdom

Barbara Y. Croft
National Institutes of Health
Kensington, Maryland

David D. Cunningham
Abbott Diagnostics
Process Engineering
Abbott Park, Illinois

Ian A. Cunningham
Victoria Hospital
The John P. Roberts Research
Institute
and
The University of Western Ontario
London, Ontario, Canada

Yadin David
Texas Children's Hospital
Houston, Texas

Connie White Delaney
School of Nursing and Medical
School
The University of Minnesota
Minneapolis, Minnesota

Mary Diakides
Advanced Concepts Analysis, Inc.
Falls Church, Virginia

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

C. Drews-Peszynski
Technical University of Lodz
Lodz, Poland

Ronald G. Driggers
U.S. Army Communications and
Electronics Research,
Development and Engineering
Center (CERDEC)
Night Vision and Electronic
Sensors Directorate
Fort Belvoir, Virginia

Gary Drzewiecki
Rutgers University
Piscataway, New Jersey

Edwin G. Duffin
Medtronic, Inc.
Minneapolis, Minnesota

Jeffrey L. Eggleston
Valleylab, Inc.
Boulder, Colorado

Robert L. Elliott
Elliott-Elliott-Head Breast Cancer
Research and Treatment Center
Baton Rouge, Louisiana

K. Whittaker Ferrara
Riverside Research Institute
New York, New York

J. Michael Fitzmaurice
Agency for Healthcare Research
and Quality
Rockville, Maryland

Ross Flewelling
Nellcor Incorporation
Pleasant, California

Michael Forde
Medtronic, Inc.
Minneapolis, Minnesota

Amir H. Gandjbakhche
Laboratory of Integrative and
Medical Biophysics
National Institute of Child Health
and Human Development
Bethesda, Maryland

Israel Gannot
Laboratory of Integrative and
Medical Biophysics
National Institute of Child Health
and Human Development
Bethesda, Maryland

Leslie A. Geddes
Purdue University
West Lafayette, Indiana

Richard L. Goldberg
University of North Carolina
Chapel Hill, North Carolina

Boris Gramatikov
Johns Hopkins School
of Medicine
Baltimore, Maryland

Barton M. Gratt
School of Dentistry
University of Washington
Seattle, Washington

Walter Greenleaf
Greenleaf Medical
Palo Alto, California

Michael W. Grenn
U.S. Army Communications and
Electronics Research,
Development and Engineering
Center (CERDEC)
Night Vision and Electronic
Sensors Directorate
Fort Belvoir, Virginia

Eliot B. Grigg
Department of Plastic Surgery
Dartmouth-Hitchcock Medical
Center
Lebanon, New Hampshire

Warren S. Grundfest
Department of Bioengineering
and Electrical Engineering
Henry Samueli School of
Engineering and Applied
Science
and
Department of Surgery
David Geffen School
of Medicine
University of California
Los Angeles, California

Michael L. Gullikson
Texas Children's Hospital
Houston, Texas

Moinuddin Hassan
Laboratory of Integrative and
Medical Biophysics
National Institute of Child Health
and Human Development
Bethesda, Maryland

David Hattery
Laboratory of Integrative and
Medical Biophysics
National Institute of Child Health
and Human Development
Bethesda, Maryland

Jonathan F. Head
Elliott-Elliott-Head Breast Cancer
Research and Treatment Center
Baton Rouge, Louisiana

William B. Hobbins
Women's Breast Health Center
Madison, Wisconsin

Stuart Horn
U.S. Army Communications and
Electronics Research,
Development and Engineering
Center (CERDEC)
Night Vision and Electronic
Sensors Directorate
Fort Belvoir, Virginia

Xiaoping Hu
Center for Magnetic Resonance
Research
and
The University of Minnesota
Medical School
Minneapolis, Minnesota

T. Jakubowska
Technical University of Lodz
Lodz, Poland

G. Allan Johnson
Duke University Medical Center
Durham, North Carolina

Bryan F. Jones
Medical Imaging Research Group
School of Computing
University of Glamorgan
Pontypridd, Wales
United Kingdom

Thomas M. Judd
Kaiser Permanente
Atlanta, Georgia

Millard M. Judy
Baylor Research Institute and
MicroBioMed Corp.
Dallas, Texas

Philip F. Judy
Brigham and Women's Hospital
Harvard University Medical
School
Boston, Massachusetts

G.J.L. Kaw
Department of Diagnostic
Radiology
Tan Tock Seng Hospital
Singapore

J.R. Keyserlingk
Ville Marie Multidisciplinary
Breast and Oncology Center
St. Mary's Hospital
McGill University
Montreal, Quebec, Canada
and
London Cancer Centre
London, Ontario
Canada

C. Everett Koop
Department of Plastic Surgery
Dartmouth-Hitchcock Medical
Center
Lebanon, New Hampshire

Hayrettin Köymen
Bilkent University
Ankara, Turkey

Luis G. Kun
IRMC/National Defense
University
Washington, D.C.

Phani Teja Kuruganti
RF and Microwave Systems Group
Oak Ridge National Laboratory
Oak Ridge, Tennessee

Kenneth K. Kwong
Massachusetts General Hospital
Harvard University Medical
School
Boston, Massachusetts

Z.R. Li
South China Normal University
Guangzhou, China

Richard F. Little
National Institutes of Health
Bethesda, Maryland

Chung-Chiun Liu
Electronics Design Center and Edison Sensor Technology Center
Case Western Reserve University
Cleveland, Ohio

Zhongqi Liu
TTM Management Group
Beijing, China

Jasper Lupo
Applied Research Associates, Inc.
Falls Church, Virginia

Albert Macovski
Stanford University
Stanford, California

Luca T. Mainardi
Polytechnic University
Milan, Italy

C. Manohar
Department of Electrical & Computer Engineering
University of Houston
Houston, Texas

Joseph P. McClain
Walter Reed Army Medical Center
Washington, D.C.

Kathleen A. McCormick
SAIC
Falls Church, Virginia

Dennis McGrath
Department of Plastic Surgery
Dartmouth-Hitchcock Medical Center
Lebanon, New Hampshire

Susan McGrath
Department of Plastic Surgery
Dartmouth-Hitchcock Medical Center
Lebanon, New Hampshire

Matthew F. McKnight
Department of Plastic Surgery
Dartmouth-Hitchcock Medical Center
Lebanon, New Hampshire

Yitzhak Mendelson
Worcester Polytechnic Institute
Worcester, Massachusetts

James B. Mercer
University of Tromsø
Tromsø, Norway

Arcangelo Merla
Department of Clinical Sciences and Bioimaging
University "G.d'Annunzio"
and
Institute for Advanced Biomedical Technology
Foundation "G.d'Annunzio"
and
Istituto Nazionale Fisica della Materia
Coordinated Group of Chieti
Chieti-Pescara, Italy

Evangelia Micheli-Tzanakou
Rutgers Unversity
Piscataway, New Jersey

Robert L. Morris
Dybonics, Inc.
Portland, Oregon

Jack G. Mottley
University of Rochester
Rochester, New York

Robin Murray
University of Rhode Island
Kingston, Rhode Island

Joachim H. Nagel
University of Stuttgart
Stuttgart, Germany

Michael R. Neuman
Michigan Technological University
Houghton, Michigan

E.Y.K. Ng
College of Engineering
School of Mechanical and Production Engineering
Nanyang Technological University
Singapore

Paul Norton
U.S. Army Communications and Electronics Research, Development and Engineering Center (CERDEC)
Night Vision and Electronic Sensors Directorate
Fort Belvoir, Virginia

Antoni Nowakowski
Department of Biomedical Engineering,
Gdansk University of Technology
Narutowicza
Gdansk, Poland

Banu Onaral
Drexel University
Philadelphia, Pennsylvania

David D. Pascoe
Auburn University
Auburn, Alabama

Maqbool Patel
Center for Magnetic Resonance Research
and
The University of Minnesota Medical School
Minneapolis, Minnesota

Robert Patterson
The University of Minnesota
Minneapolis, Minnesota

Jeffrey L. Paul
Defense Advanced Research Projects Agency
Arlington, Virginia

A. William Paulsen
Emory University
Atlanta, Georgia

John Pauly
Stanford University
Stanford, California

I. Pavlidis
Department of Computer Science
University of Houston
Houston, Texas

P. Hunter Peckham
Case Western Reserve University
Cleveland, Ohio

Joseph G. Pellegrino
U.S. Army Communications and Electronics Research, Development and Engineering Center (CERDEC)
Night Vision and Electronic Sensors Directorate
Fort Belvoir, Virginia

Philip Perconti
U.S. Army Communications and Electronics Research, Development and Engineering Center (CERDEC)
Night Vision and Electronic Sensors Directorate
Fort Belvoir, Virginia

Athina P. Petropulu
Drexel University
Philadelphia, Pennsylvania

Tom Piantanida
Greenleaf Medical
Palo Alto, California

T. Allan Pryor
University of Utah
Salt Lake City, Utah

Ram C. Purohit
Auburn University
Auburn, Alabama

Hairong Qi
ECE Department
The University of Tennessee
Knoxville, Tennessee

Pat Ridgely
Medtronic, Inc.
Minneapolis, Minnesota

E. Francis Ring
Medical Imaging Research Group
School of Computing
University of Glamorgan
Pontypridd, Wales
United Kingdom

Richard L. Roa
Baylor University Medical Center
Dallas, Texas

Peter Robbie
Department of Plastic Surgery
Dartmouth-Hitchcock Medical Center
Lebanon, New Hampshire

Gian Luca Romani
Department of Clinical Sciences and Bioimaging
University "G. d'Annunzio"
and
Institute for Advanced Biomedical Technology
Foundation "G.d'Annunzio"
and
Istituto Nazionale Fisica della Materia
Coordinated Group of Chieti
Chieti-Pescara, Italy

Joseph M. Rosen
Department of Plastic Surgery
Dartmouth-Hitchcock Medical Center
Lebanon, New Hampshire

Eric Rosow
Hartford Hospital
and
Premise Development Corporation
Hartford, Connecticut

Subrata Saha
Clemson University
Clemson, South Carolina

John Schenck
General Electric Corporate Research and Development Center
Schenectady, New York

Edward Schuck
EdenTec Corporation
Eden Prairie, Minnesota

Joyce Sensmeier
HIMSS
Chicago, Illinois

David Sherman
Johns Hopkins School of Medicine
Baltimore, Maryland

Robert E. Shroy, Jr.
Picker International
Highland Heights, Ohio

Stephen W. Smith
Duke University
Durham, North Carolina

Nathan J. Sniadecki
Department of Bioengineering
University of Pennsylvania
Philadelphia, Pennsylvania

Wesley E. Snyder
ECE Department
North Carolina State University
Raleigh, North Carolina

Orhan Soykan
Corporate Science and Technology
Medtronic, Inc.
and
Department of Biomedical Engineering
Michigan Technological University
Houghton, Michigan

Primoz Strojnik
Case Western Reserve University
Cleveland, Ohio

M. Strzelecki
Technical University of Lodz
Lodz, Poland

Ron Summers
Loughborough University
Leicestershire, United Kingdom

Christopher Swift
Department of Plastic Surgery
Dartmouth-Hitchcock Medical Center
Lebanon, New Hampshire

Willis A. Tacker
Purdue University
West Lafayette, Indiana

Nitish V. Thakor
Johns Hopkins School of Medicine
Baltimore, Maryland

Roderick Thomas
Faculty of Applied Design and Engineering
Swansea Institute of Technology
Swansea, United Kingdom

P. Tsiamyrtzis
Department of Statistics
University of Economics and Business Athens
Athens, Greece

Benjamin M.W. Tsui
University of North Carolina
Chapel Hill, North Carolina

Tracy A. Turner
Private Practice
Minneapolis, Minnesota

Kamil Ugurbil
Center for Magnetic Resonance Research
and
The University of Minnesota Medical School
Minneapolis, Minnesota

Michael S. Van Lysel
University of Wisconsin
Madison, Wisconsin

Henry F. VanBrocklin
University of California-Berkeley
Berkeley, California

Jay Vizgaitis
U.S. Army Communications and Electronics Research, Development and Engineering Center (CERDEC)
Night Vision and Electronic Sensors Directorate
Fort Belvoir, Virginia

Abby Vogel
Laboratory of Integrative and Medical Biophysics
National Institute of Child Health and Human Development
Bethesda, Maryland

Wolf W. von Maltzahn
Rensselaer Polytechnic Institute
Troy, New York

Gregory I. Voss
IVAC Corporation
San Diego, California

Alvin Wald
Columbia University
New York, New York

Chen Wang
TTM International
Houston, Texas

Lois de Weerd
University Hospital of North Norway
Tromsø, Norway

Wang Wei
Radiology Department
Beijing You An Hospital
Beijing, China

B. Wiecek
Technical University of Lodz
Lodz, Poland

M. Wysocki
Technical University of Lodz
Lodz, Poland

Martin J. Yaffe
University of Toronto
Toronto, Ontario, Canada

Robert Yarchoan
HIV and AIDS Malignancy Branch
Center for Cancer Research
National Cancer Institute (NCI)
Bethesda, Maryland

M. Yassa
Ville Marie Multidisciplinary Breast and Oncology Center
St. Mary's Hospital
McGill University
Montreal, Quebec, Canada
and
London Cancer Centre
London, Ontario, Canada

Christopher M. Yip
Departments of Chemical Engineering and Applied Chemistry
Department of Biochemistry
Institute of Biomaterials and Biomedical Engineering
University of Toronto
Toronto, Ontario, Canada

E. Yu
Ville Marie Multidisciplinary Breast and Oncology Center
St. Mary's Hospital
McGill University
Montreal, Quebec, Canada
and
London Cancer Centre
London, Ontario, Canada

Wen Yu
Shanghai RuiJin Hospital
Shanghai, China

Yune Yuan
Institute of Basic Medical Science
China Army General Hospital
Beijing, China

Jason Zeibel
U.S. Army Communications and Electronics Research, Development and Engineering Center (CERDEC)
Night Vision and Electronic Sensors Directorate
Fort Belvoir, Virginia

Yi Zeng
Central Disease Control of China
Beijing, China

Xiaohong Zhou
Duke University Medical Center
Durham, North Carolina

Yulin Zhou
Shanghai RuiJin Hospital
Shanghai, China

Contents

SECTION I Biomedical Signal Analysis

Banu Onaral

SECTION II Imaging

Warren S. Grundfest

SECTION III Infrared Imaging

Nicholas A. Diakides

SECTION IV Medical Informatics

Luis G. Kun

SECTION V Biomedical Sensors

Michael R. Neuman

SECTION VI Medical Instruments and Devices

Wolf W. von Maltzahn

SECTION VII Clinical Engineering

Yadin David

SECTION VIII Ethical Issues Associated with the Use of Medical Technology

Subrata Saha and Joseph D. Bronzino

30

Infrared Imaging for Tissue Characterization and Function

Moinuddin Hassan
Victor Chernomordik
Abby Vogel
David Hattery
Israel Gannot
Richard F. Little
Robert Yarchoan
Amir H. Gandjbakhche
National Institutes of Health

Noninvasive imaging techniques are emerging into the forefront of medical diagnostics and treatment monitoring. Both near- and mid-infrared imaging techniques have provided invaluable information in the clinical setting.

Near-infrared imaging in the spectrum of 700 to 1100 nm has been used to functionally monitor diseases processes including cancer and lymph node detection and optical biopsies. Spectroscopic imaging modalities have been shown to improve the diagnosis of tumors and add new knowledge about the physiological properties of the tumor and surrounding tissues. Particular emphasis should be placed on identifying markers that predict the risk of precancerous lesions progressing to invasive cancers, thereby providing new opportunities for cancer prevention. This might be accomplished through the use of markers as contrast agents for imaging using conventional techniques or through refinements of newer technologies such as MRI or PET scanning. The spectroscopic power of light, along with the revolution in molecular characterization of disease processes has created a huge potential for *in vivo* optical imaging and spectroscopy.

In the infrared thermal waveband, information about blood circulation, local metabolism, sweat gland malfunction, inflammation, and healing can be extracted. Infrared thermal imaging has been increasingly used for detection of cancers. As this field evolves, abnormalities or changes in infrared images could be able to provide invaluable information to physicians caring for patients with a variety of disorders. The current status of modern infrared imaging is that of a first line supplement to both clinical exams and current imaging methods. Using infrared imaging to detect breast pathology is based on the principle that both metabolic and vascular activity in the tissue surrounding a new and developing tumor is usually higher than in normal tissue. Early cancer growth is dependent on increasing blood circulation by creating new blood vessels (angiogenesis). This process results in regional variations that can often be detected by infrared imaging.

Section 30.1 discusses near-infrared (NIR) imaging and its applications in imaging biological tissues. Infrared thermal imaging techniques, calibration and a current clinical trial of Kaposi's sarcoma are described in Section 30.2.

30.1 Near-Infrared Quantitative Imaging of Deep Tissue Structure

In vivo optical imaging has traditionally been limited to superficial tissue surfaces, directly or endoscopically accessible, and to tissues with a biological window (e.g., along the optical axis of the eye). These methods are based on geometric optics. Most tissues scatter light so strongly, however, that for geometric optics-based equipment to work, special techniques are needed to remove multiply scattered light (such as pinholes in confocal imaging or interferometry in optical coherence microscopies). Even with these special designs, high resolution optical imaging fails at depths of more than 1 mm below the tissue surface.

Collimated visible or infrared (IR) light impinging upon thick tissue is scattered many times in a distance of ∼1 mm, so the analysis of light-tissue interactions requires theories based on the diffusive nature of light propagation. In contrast to x-ray and Positron Emission Tomography (PET), a complex underlying theoretical picture is needed to describe photon paths as a function of scattering and absorption properties of the tissue.

Approximately a decade ago, a new field called "Photon Migration" was born that seeks to characterize the statistical physics of photon motion through turbid tissues. The goal has been to image macroscopic structures in 3D at greater depths within tissues and to provide reliable pathlength estimations for noninvasive spectral analysis of tissue changes. Although geometrical optics fails to describe light propagation under these conditions, the statistical physics of strong, multiply scattered light provides powerful approaches to macroscopic imaging and subsurface detection and characterization. Techniques using visible and NIR light offer a variety of functional imaging modalities, in addition to density imaging, while avoiding ionizing radiation hazards.

In Section 30.1.1, optical properties of biological tissue will be discussed. Section 30.1.2 is devoted to differing methods of measurements. Theoretical models for spectroscopy and imaging are discussed in Section 30.1.3. In Sections 30.1.4 and 30.1.5, two studies on breast imaging and the use of exogenous fluorescent markers will be presented as examples of NIR spectroscopy. Finally, the future direction of the field will be discussed in Section 30.1.6.

30.1.1 Optical Properties of Biological Tissue

The difficulty of tissue optics is to define optical coefficients of tissue physiology and quantify their changes to differentiate structures and functional status *in vivo*. Light-tissue interactions dictate the way that these parameters are defined. The two main approaches are the wave and particle descriptions of light propagation. The first leads to the use of Maxwell's equations, and therefore quantifies the spatially varying permittivity as a measurable quantity. For simplistic and historic reasons, the particle interpretation of light has been mostly used (see section on models of photon migration). In photon transport theory, one considers the behavior of discrete photons as they move through the tissue. This motion is characterized by absorption and scattering, and when interfaces (e.g., layers) are involved, refraction. The absorption

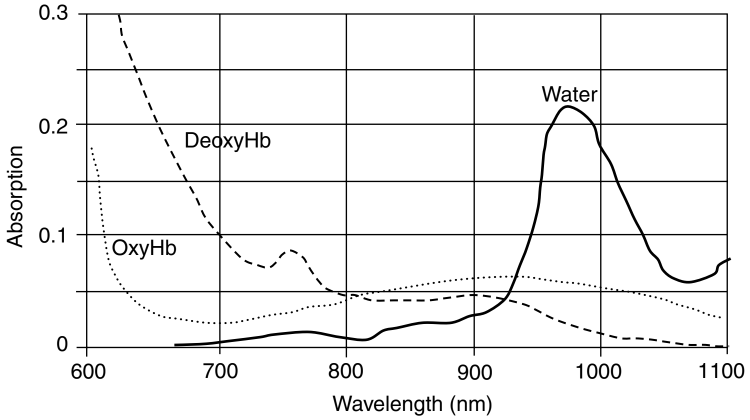


FIGURE 30.1 Absorption spectra of the three major components of tissue in the NIR region; oxy-hemoglobin, deoxy-hemoglobin and water.

coefficient, $\mu_a(\text{mm}^{-1})$, represents the inverse mean pathlength of a photon before absorption. $1/\mu_a$ is the distance in a medium where intensity is attenuated by a factor of 1/e (Beer's Lambert Law). Absorption in tissue is strongly wavelength dependent and is due to chromophores and water. Among the chromophores in tissue, the dominant component is the hemoglobin in blood. In Figure 30.1, hemoglobin absorption is devided in to oxy- and deoxy-hemoglobin. As seen in this figure, in the visible range (600–700 nm), the blood absorption is relatively high compared to absorption in the NIR. By contrast, water absorption is low in the visible and NIR regions and increases rapidly above approximately 950 nm. Thus, for greatest penetration of light in tissue, wavelengths in the 650–950 nm spectrum are used most often. This region of the light spectrum is called "the therapeutic window." One should note that different spectra of chromophores allow one to separate the contribution of varying functional species in tissue (e.g., quantification of oxy- and deoxy-hemoglobin to study tissue oxygenation).

Similarly, scattering is characterized by a coefficient, μ_s, which is the inverse mean free path of photons between scattering events. The average size of the scattered photons in tissue, in proportion to the wavelength of the light, places the scattering in the Mie region. In the Mie region, a scattering event does not result in isotropic scattering angles [1,2]. Instead, the scattering in tissue is biased in the forward direction.

For example, by studying the development of neonatal skin, Saidi et al. [3] were able to show that the principal sources of anisotropic scattering in muscle are collagen fibers. The fibers were determined to have a mean diameter of 2.2 μm. In addition to the Mie scattering from the fibers, there is isotropic Rayleigh scattering due to the presence of much smaller scatterers such as organelles in cells.

Anisotropic scattering is quantified in a coefficient, g, which is defined as the mean cosine of the scattering angle, where $p(\theta)$ is the probability of a particular scattering angle,

$$g = \langle \cos(\theta) \rangle = \frac{\int_0^{\pi} p(\theta)\cos(\theta)\sin(\theta)\mathrm{d}\theta}{\int_0^{\pi} p(\theta)\sin(\theta)\mathrm{d}\theta} \tag{30.1}$$

For isotropic scattering, $g = 0$. For complete forward scattering, $g = 1$, and for complete back scattering, $g = -1$. In tissue, g is typically 0.7 to 0.98 [3–5].

Likewise, different tissue types have differing scattering properties which are also wavelength dependent. The scattering coefficients of many soft tissues have been measured at a variety of optical wavelengths, and are within the range 10 to 100 mm^{-1}. In comparison to absorption, however, scattering changes, as a function of wavelength, are more gradual and have smaller extremes. Abnormal tissues such as tumors, fibro-adenomas, and cysts all have scattering properties that are different from normal tissue [6,7]. Thus, the scattering coefficient of an inclusion may also be an important clue to disease diagnoses.

Theories of photon migration are often based on isotropic scattering. Therefore, one must find the appropriate scaling relationships that will allow use of an isotropic scattering model. For the case of diffusion-like models (e.g., see Reference 8), it has been shown that one may use an isotropic scattering model with a corrected scattering coefficient, μ_s', and obtain equivalent results where:

$$\mu_s' = \mu_s(1 - g) \tag{30.2}$$

The corrected scattering coefficient is smaller than the actual scattering which corresponds to a greater distance between isotropic scattering events than would occur with anisotropic scattering. For this reason, μ_s' is typically called the transport-corrected scattering coefficient.

There are instances in which the spectroscopic signatures will not be sufficient for detection of disease. This can occur when the specific disease results in only very small changes to the tissue's scattering and absorption properties, or when the scattering and absorption properties are not unique to the disease. Although it is not clear what the limits of detectability are in relationship to diseased tissue properties, it is clear that there will be cases for which optical techniques based on elastic absorption are inadequate. In such cases, another source of optical contrast, such as fluorescence, will be required to detect and locate the disease. Presence of fluorescent molecules in tissues can provide useful contrast mechanisms. Concentration of these endogenous fluorophores in the body can be related to functional and metabolic activities, and therefore to the disease processes. For example, the concentrations of fluorescent molecules such as collagen and NADH have been used to differentiate between normal and abnormal tissue [9].

Advances in the molecular biology of disease processes, new immunohistopathological techniques, and the development of fluorescently-labeled cell surface markers have led to a revolution in specific molecular diagnosis of disease by histopathology, as well as in research on molecular origins of disease processes (e.g., using fluorescence microscopy in cell biology). As a result, an exceptional level of specificity is now possible due to the advances in the design of exogenous markers. Molecules can now be tailor-made to bind only to specific receptor sites in the body. These receptor sites may be antibodies or other biologically interesting molecules. Fluorophores may be bound to these engineered molecules and injected into the body, where they will preferentially concentrate at specific sites of interest [10,11].

Furthermore, fluorescence may be used as a probe to measure environmental conditions in a particular locality by capitalizing on changes in fluorophore lifetimes [12,13]. Each fluorophore has a characteristic lifetime that quantifies the probability of a specific time delay between fluorophore excitation and emission. In practice, this lifetime may be modified by specific environmental factors such as temperature, pH, and concentrations of substances such as oxygen. In these cases, it is possible to quantify local concentrations of specific substances or specific environmental conditions by measuring the lifetime of fluorophores at the site. Whereas conventional fluorescence imaging is very sensitive to non-uniform fluorophore transport and distribution (e.g., blood does not transport molecules equally to all parts of the body), fluorescence lifetime imaging is insensitive to transport non-uniformity as long as a detectable quantity of fluorophores is present in the site of interest. Throughout the following sections, experimental techniques and differing models used to quantify these sources of optical contrast will be presented.

30.1.2 Measurable Quantities and Experimental Techniques

Three classes of measurable quantities prove to be of interest in transforming results of remote sensing measurements in tissue into useful physical information. The first is the spatial distribution of light or the intensity profile generated by photons re-emitted through a surface and measured as a function of the radial distance from the source and the detector when the medium is continually irradiated by a point source (often a laser). This type of measurement is called continuous wave (CW). The intensity, nominally, does not vary in time. The second class is the temporal response to a very short pulse (~picosecond) of photons impinging on the surface of the tissue. This technique is called time-resolved and the temporal response is known as the time-of-flight (TOF). The third class is the frequency-domain technique in which an intensity-modulated laser beam illuminates the tissue. In this case, the measured outputs are the AC modulation amplitude and the phase shift of the detected signal. These techniques could be implemented in geometries with different arrangements of source(s) and detector(s); (a) in the reflection mode, source(s) and detector(s) are placed at the same side of the tissue; (b) in the transmission mode, source(s) and detector(s) are located on opposite sides of the tissue. In the latter, the source(s) and detector(s) can move in tandem while scanning the tissue surface and detectors with lateral offsets also can be used; and (c) tomographic sampling often uses multiple sources and detectors placed around the circumference of the target tissue.

For CW measurements, the instrumentation is simple and requires only a set of light sources and detectors. In this technique, the only measurable quantity is the intensity of light, and, due to multiple scattering, strong pathlength dispersion occurs which results in a loss of localization and resolution. Hence, this technique is widely used for spectroscopic measurements of bulk tissue properties in which the tissue is considered to be homogeneous [14,15]. However, CW techniques for imaging abnormal targets that use only the coherent portion of light, and thereby reject photons with long paths, have also been investigated. Using the transillumination geometry, collimated detection is used to isolate un-scattered photons [16–18]. Spatial filtering has been proposed which employs a lens to produce the Fourier spectrum of the spatial distribution of light from which the high-order frequencies are removed. The resulting image is formed using only the photons with angles close to normal [19]. Polarization discrimination has been used to select those photons which undergo few scattering events and therefore preserve a fraction of their initial polarization state, as opposed to those photons which experience multiple scattering resulting in complete randomization of their initial polarization state [20]. Several investigators have used heterodyne detection which involves measuring the beat frequency generated by the spatial and temporal combination of a light beam and a frequency modulated reference beam. Constructive interference occurs only for the coherent portion of the light [20–22]. However, the potential of direct imaging using CW techniques in very thick tissue (e.g., breast) has not been established. On the other hand, use of models of photon migration implemented in inverse method based on backprojection techniques has shown promising results. For example, Phillips Medical has used 256 optical fibers placed at the periphery of a white conical shaped vessel. The area of interest, in this case the breast, is suspended in the vessel, and surrounded by a matching fluid. Three CW laser diodes sequentially illuminate the breast using one fiber. The detection is done simultaneously by 255 fibers. It is now clear that CW imaging cannot provide direct images with clinically acceptable resolution in thick tissue. Attempts are underway to devise inverse algorithms to separate the effects of scattering and absorption and therefore use this technique for quantitative spectroscopy as proposed by Phillips [23]. However, until now, clinical application of CW techniques in imaging has been limited by the mixture of scattering and absorption of light in the detected signal. To overcome this problem, time-dependent measurement techniques have been investigated.

Time-domain techniques involve the temporal resolution of photons traveling inside the tissue. The basic idea is that photons with smaller pathlengths are those that arrive earlier to the detector. In order to discriminate between un-scattered or less scattered light and the majority of the photons, which experience a large number of multiple scattering, subnanosecond resolution is needed. This short time gating of an imaging system requires the use of a variety of techniques involving ultra-fast phenomena and/or fast detection systems. Ultra-fast shuttering is performed using the Kerr effect. The birefringence in the Kerr cell, placed between two crossed polarizers, is induced using very short pulses. Transmitted light through the Kerr cell is recorded, and temporal resolution of a few picoseconds is achieved [19]. When an impulse of light (~picoseconds or hundreds of femtoseconds) is launched at the tissue surface, the whole temporal distribution of photon intensity can be recorded by a streak camera. The streak camera can achieve temporal resolution on the order of few picoseconds up to several nanosececonds detection time. This detection system has been widely used to assess the performance of breast imaging and neonatal brain activity [24,25]. The time of flight recorded by the streak camera is the convolution of the pulsed laser source (in practice with a finite width) and the actual Temporal Point Spread Function (TPSF) of the diffuse photons. Instead of using very short pulse lasers (e.g., Ti–Sapphire lasers), the advent of pulse diode lasers with relatively larger pulse widths (100 to 400 psec) have reduced the cost of time-domain imaging. However, deconvolution of the incoming pulse and the detected TPSF have been a greater issue. Along with diode laser sources, several groups have also used time-correlated single photon counting with photomultipliers for recording the TPSF [26,27]. Fast time gating is also obtained by using Stimulated Raman Scattering. This phenomenon is a nonlinear Raman interaction in some materials such as hydrogen gas involving the amplification of photons with Stokes shift by a higher energy pump beam. The system operates by amplifying only the earliest arriving photons [28]. Less widely used techniques such as second-harmonic generation [29], parametric amplification [30] and a variety of others have been proposed for time-domain (see an excellent review in Reference 31).

For frequency-domain measurements, the requirement is to measure the DC amplitude, the AC amplitude, and the phase shift of the photon density wave. For this purpose a CW light source is modulated with a given frequency (∼100 MHz). Lock-in Amplifiers and phase sensitive CCD camera have been used to record the amplitude and phase [32,33]. Multiple sources at different wavelengths can be modulated with a single frequency or multiple frequencies [6,34]. In the latter case a network analyzer is used to produce modulation swept from several hundreds of MHz to up to 1 GHz.

30.1.3 Models of Photon Migration in Tissue

Photon Migration theories in biomedical optics have been borrowed from other fields such as astrophysics, atmospheric science, and specifically from nuclear reactor engineering [35,36]. The common properties of these physical media and biological tissues are their characterization by elements of randomness in both space and time. Because of many difficulties surrounding the development of a theory based on a detailed picture of the microscopic processes involved in the interaction of light and matter, investigations are often based on statistical theories. These can take a variety of forms, ranging from quite detailed multiple-scattering theories [36] to transport theory [37]. However, the most widely used theory is the time-dependent diffusion approximation to the transport equation:

$$\vec{\nabla} \cdot (D\vec{\nabla}\Phi(\vec{r}, t)) - \mu_a \Phi(\vec{r}, t) = \frac{1}{c}\frac{\partial \Phi(\vec{r}, t)}{\partial t} - S(\vec{r}, t) \tag{30.3}$$

where $\vec{r}$ and t are spatial and temporal variables, c is the speed of light in tissue, and D is the diffusion coefficient related to the absorption and scattering coefficients as follows:

$$D = \frac{1}{3[\mu_a + \mu_s']} \tag{30.4}$$

The quantity $\Phi(\vec{r}, t)$ is called the fluence, defined as the power incident on an infinitesimal volume element divided by its area. Note that the equation does not incorporate any angular dependence, therefore assuming an isotropic scattering. However, for the use of the diffusion theory for anisotropic scattering, the diffusion coefficient is expressed in terms of the transport-corrected scattering coefficient. $S(\vec{r}, t)$ is the source term. The gradient of fluence, $J(\vec{r}, t)$, at the tissue surface is the measured flux of photons by the detector:

$$J(\vec{r}, t) = -D\vec{\nabla}\Phi(\vec{r}, t) \tag{30.5}$$

For CW measurements, the time-dependence of the flux vanishes, and the source term can be seen as the power impinging in its area. For time-resolved measurements, the source term is a Dirac delta function describing a very short photon impulse. Equation 30.3 has been solved analytically for different types of measurements such as reflection and transmission modes assuming that the optical properties remain invariant through the tissue. To incorporate the finite boundaries, the method of images has been used. In the simplest case, the boundary has been assumed to be perfectly absorbing which does not take into account the difference between indices of refraction at the tissue–air interface. For semi-infinite and transillumination geometries, a set of theoretical expressions has been obtained for time-resolved measurements [38].

The diffusion approximation equation in the frequency-domain is the Fourier transformation of the time-domain with respect to time. Fourier transformation applied to the time-dependent diffusion equation leads to a new equation:

$$\vec{\nabla} \cdot (D\vec{\nabla}\Phi(\vec{r},\omega)) - \left[\mu_{\mathrm{a}} + \frac{\mathrm{i}\omega}{c}\right]\Phi(\vec{r},\omega) + S(\vec{r},\omega) = 0 \tag{30.6}$$

Here the time variable is replaced by the frequency ω. This frequency is the modulation angular frequency of the source. In this model, the fluence can be seen as a complex number describing the amplitude and phase of the photon density wave, dumped with a DC component:

$$\Phi(\vec{r},\omega) = \Phi_{\mathrm{AC}}(\vec{r},\omega) + \Phi_{\mathrm{DC}}(\vec{r},0) = I_{\mathrm{AC}}\exp(\mathrm{i}\theta) + \Phi_{\mathrm{DC}}(\vec{r},0) \tag{30.7}$$

In the RHS of Equation 30.7, the quantity θ is the phase shift of the diffusing wave. For a nonabsorbing medium, its wavelength is:

$$\lambda = 2\pi\sqrt{\frac{2c}{3\mu_{\mathrm{s}}'\omega}} \tag{30.8}$$

Likewise in the time-domain, Equation 30.3 has an analytical solution for the case that the tissue is considered homogeneous. The analytical solution permits one to deduce the optical properties in a spectroscopic setting.

For imaging, where the goal is to distinguish between structures in tissue, the diffusion coefficient and the absorption coefficient in Equation 30.3 and Equation 30.6 become spatial-dependent and are replaced by $D(r)$ and $\mu_{\mathrm{a}}(r)$. For the cases that an abnormal region is embedded in otherwise homogeneous tissue, perturbation methods based on Born approximation or Rytov approximation have been used (see excellent review in Reference 39). However, for the cases that the goal is to reconstruct the spectroscopic signatures inside the tissue, no analytical solution exists. For these cases, inverse algorithms are devised to map the spatially varying optical properties. Numerical methods such as finite-element or finite-difference methods have been used to reconstruct images of breast, brain, and muscle [40–42]. Furthermore, in those cases that structural heterogeneity exists, a priori information from other image modalities such as MRI can be used. An example is given in Figure 30.2. Combining MRI and NIR imaging, rat cranium functional imaging during changes in inhaled oxygen concentration was studied [43]. Figure 30.2a,b correspond to the MRI image and the corresponding constructed finite-element mesh. Figure 30.2c,d correspond to the oxygen map of the brain with and without incorporation of MRI geometry and constraints.

The use of MRI images has improved dramatically the resolution of the oxygen map. The use of optical functional imaging in conjunction with other imaging modalities has opened new possibilities in imaging and treating diseases at the bedside.

The second theoretical framework used in tissue optics is the random walk theory (RWT) on a lattice developed at the National Institutes of Health [44,45] and historically precedes the use of the diffusion approximation theory. It has been shown that RWT may be used to derive an analytical solution for the distribution of photon path-lengths in turbid media such as tissue [44]. RWT models the diffusion-like motion of photons in turbid media in a probabilistic manner. Using RWT, an expression may be derived for the probability of a photon arriving at any point and time given a specific starting point and time.

Tissue may be modeled as a 3D cubic lattice containing a finite inclusion, or region of interest, as shown in Figure 30.3. The medium has an absorbing boundary corresponding to the tissue surface, and the lattice spacing is proportional to the mean photon scattering distance, $1/\mu_{\mathrm{s}}'$. The behavior of photons in the RWT model is described by three dimensionless parameters, ρ, n, μ, which are respectively the radial distance, the number of steps, and the probability of absorption per lattice step. In the RWT model, photons may move to one of the six nearest neighboring lattice points, each with probability 1/6. If the number of steps, n, taken by a photon traveling between two points on the lattice is known, then the length of the photon's path is also known.

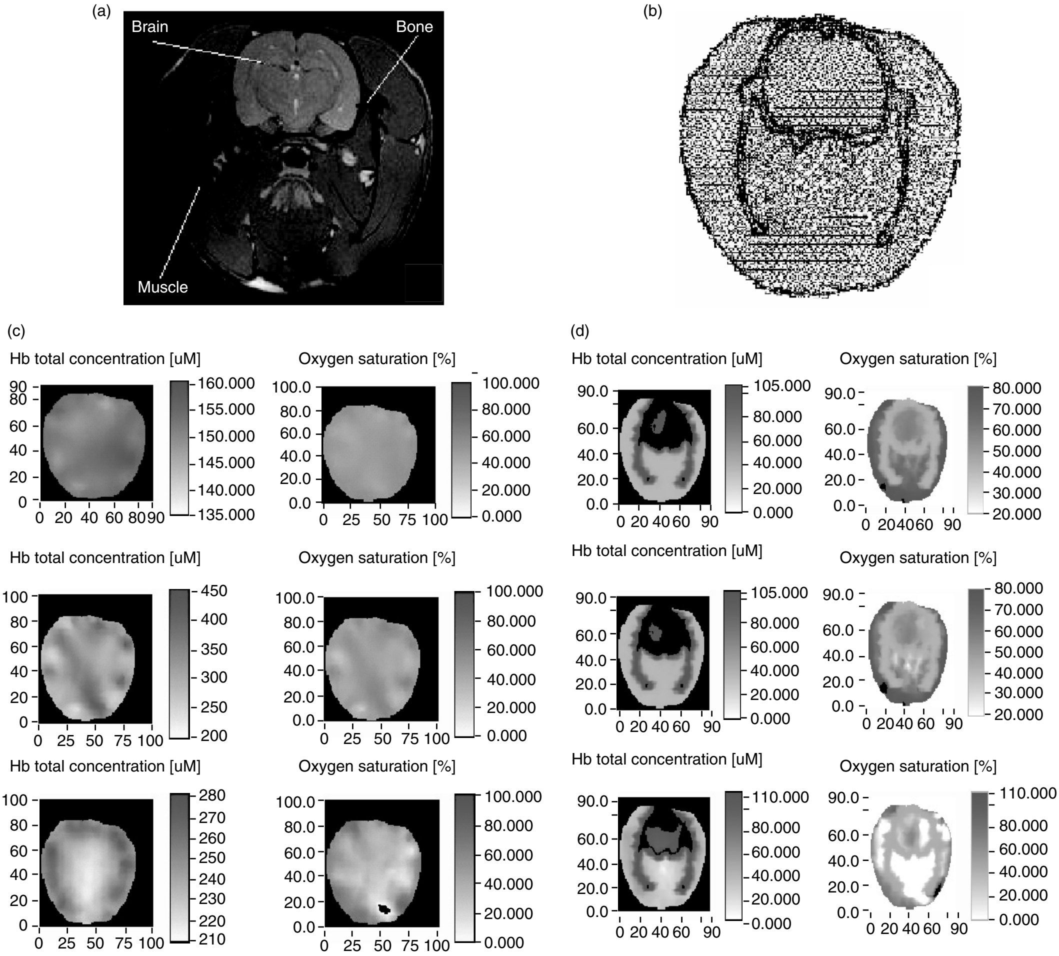


FIGURE 30.2 (See color inset following page 29-16.) Functional imaging of rat cranium during changes in inhaled oxygen concentration: (a) MRI image; (b) creation of the mesh to distinguish different compartments in the brain; (c) map of hemoglobin concentration and oxygen saturation of the rat brain without structural constraints from MRI; (d) same as (c) with structural constraints including tissue heterogeneity. In (c) and (d) the rows from top correspond to 13, 8, and 0% (after death) oxygen inhaled. (Courtesy of Dartmouth College.)

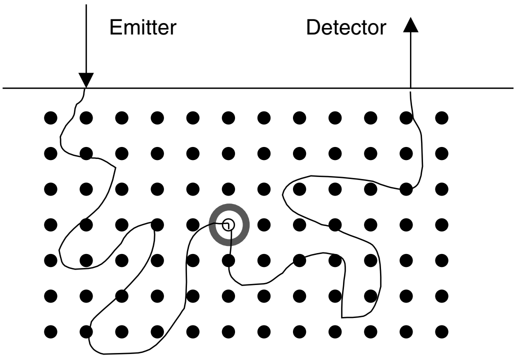


FIGURE 30.3 2D random walk lattice showing representative photon paths from an emitter to a specific site and then to a detector.

Random walk theory is useful in predicting the probability distribution of photon path lengths over distances of at least five mean photon scattering distances. The derivation of these probability distributions is described in papers [44,45]. For simplicity in this derivation, the tissue–air interface is considered to be perfectly absorbing; a photon arriving at this interface is counted as arriving at a detector on the tissue surface. The derivation uses the Central Limit Theorem and a Gaussian distribution around lattice points to obtain a closed-form solution that is independent of the lattice structure.

The dimensionless RWT parameters, ρ, n, and μ, described above, may be transformed to actual parameters, in part, by using time, t, the speed of light in tissue, c, and distance traveled, r, as follows:

$$\rho \to \frac{r\mu_s'}{\sqrt{2}}, \qquad n \to \mu_s' ct, \quad \mu \to \frac{\mu_a}{\mu_s'} \tag{30.9}$$

As stated previously, scattering in tissue is highly anisotropic. Therefore, one must find the appropriate scaling relationships that will allow the use of an isotropic scattering model such as RWT. Like diffusion theory, for RWT [46], it has been shown that one may use an isotropic scattering model with a corrected scattering coefficient, μ_s', and obtain equivalent results. The corrected scattering coefficient is smaller than the actual scattering that corresponds to a greater distance between isotropic scattering events than would occur with anisotropic scattering. RWT has been used to show how one would transition from the use of μ_s to μ_s' as the distance under considerable increases [47].

As an example, for a homogeneous slab into which a photon has been inserted, the probability, P, of a photon arriving at a point ρ after n steps is [48]:

$$P(n,\rho) = \frac{\sqrt{3}}{2}\left[\frac{1}{2\pi(n-2)}\right]^{3/2} \mathrm{e}^{-3\rho^2/2(n-2)} \sum_{k=-\infty}^{\infty}\left[\mathrm{e}^{-3[(2k+1)L-2]^2/2(n-2)} - \mathrm{e}^{-3[(2k+1)L]^2/2(n-2)}\right]\mathrm{e}^{-n\mu} \tag{30.10}$$

where L is the thickness of the slab. The method of images has been used to take into account the two boundaries of the slab. Plotting Equation 30.10 yields a photon arrival curve as shown in Figure 30.4; Monte Carlo simulation data are overlaid. In the next two sections the use of RWT for imaging will be presented.

30.1.4 RWT Applied to Quantitative Spectroscopy of the Breast

One important and yet extremely challenging areas to apply diffuse optical imaging of deep tissues is the human breast (see review article of Hawrysz and Sevick-Muraca [49]). It is clear that any new imaging

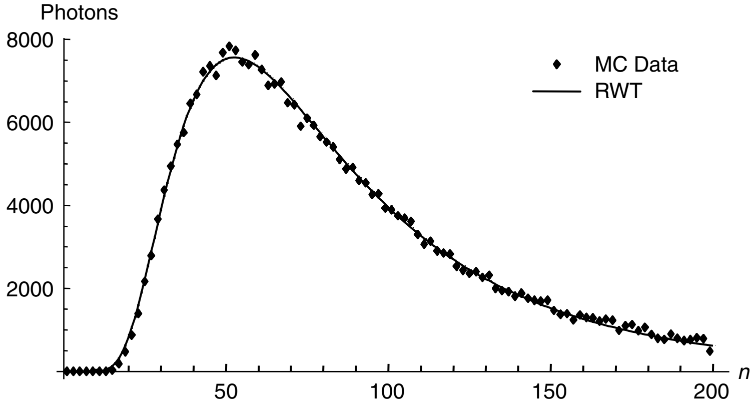


FIGURE 30.4 RWT prediction and Monte Carlo simulation results for transillumination of a 15-mm thick slab with scattering 1/mm and 109 photons.

or spectroscopic modalities that can improve the diagnosis of breast tumors or can add new knowledge about the physiological properties of the breast and surrounding tissues will have a great significance in medicine.

Conventional transillumination using continuous wave (CW) light was used for breast screening several decades ago [50]. However, because of the high scattering properties of tissue, this method resulted in poor resolution. In the late 1980s, time-resolved imaging techniques were proposed to enhance spatial resolution by detecting photons with very short time-of-flight within the tissue. In this technique, a very short pulse, of ~picosecond duration, impinges upon the tissue. Photons experience dispersion in their pathlengths, resulting in temporal dispersion in their time-of-flight (TOF).

To evaluate the performance of time-resolved transillumination techniques, RWT on a lattice was used. The analysis of breast transillumination was based on the calculation of the point spread function (PSF) of time resolved photons as they visit differing sites at different planes inside a finite slab of thickness L. The PSF [51], is defined as the probability that a photon inserted into the tissue visits a given site, is detected at the nth step (i.e., a given time), and has the following rather complicated analytical expression:

$$W_n(\mathbf{s},\mathbf{r},\mathbf{r}_0) = \sum_{l=0}^{n} p_l(\mathbf{r},\mathbf{s})p_{n-l}(\mathbf{s},\mathbf{r}_0) = \frac{9}{16\pi^{5/2}n^{3/2}}\sum_{k=-\infty}^{\infty}\sum_{m=-\infty}^{\infty}\{F_n[\alpha_+(k),\beta_+(m,p)] \qquad (30.11)$$

$$+ F_n[\alpha_-(k),\beta_-(m,p)] - F_n[\alpha_+(k),\beta_-(m,p)] - F_n[\alpha_-(k),\beta_+(m,p)]\}$$

$$F_n(a,b) = \left(\frac{1}{a}+\frac{1}{b}\right)\exp\left[-\frac{(a+b)^2}{n}\right] \qquad (30.12)$$

$$\alpha_\pm(k) = \left\{\frac{3}{2}\left[s_1^2 + (s_3 + 2kN \pm 1)^2\right]\right\}^{\frac{1}{2}} \qquad (30.13)$$

$$\beta_\pm(k,\rho) = \left\{\frac{3}{2}\left[(\rho - s_1)^2 + (N - s_3 + 2kN \pm 1)^2\right]\right\}^{\frac{1}{2}} \qquad (30.14)$$

where $N = (\mu_s'/\sqrt{2}) + 1$ is dimensionless RWT thickness of the slabs, $\bar{s}(s_1, s_2, s_3)$ are the dimensionless coordinates (see Equation 30.9) of any location for which the PSF is calculated. Evaluation of time-resolved imaging showed that strong scattering properties of tissues prevent direct imaging of abnormalities [52]. Hence, devising theoretical constructs to separate the effects of the scattering from the absorption was proposed, thus allowing one to map the optical coefficients as spectroscopic signatures of an abnormal tissue embedded in thick, otherwise normal tissue. In this method, accurate quantification of the size and optical properties of the target becomes a critical requirement for the use of optical imaging at the bedside. RWT on a lattice has been used to analyze the time-dependent contrast observed in time-resolved transillumination experiments and deduce the size and optical properties of the target and the surrounding tissue from these contrasts. For the theoretical construction of contrast functions, two quantities are needed. First, the set of functions [51] defined previously. Second, the set of functions [53] defined as the probability that a photon is detected at the nth step (i.e., time) in a homogeneous medium (Equation 30.10)[48].

To relate the contrast of the light intensity to the optical properties and location of abnormal targets in the tissue, one can take advantage of some features of the theoretical framework. One feature is that the early time response is most dependent on scattering perturbations, whereas the late time behavior is most dependent on absorptive perturbations, thus allowing one to separate the influence of scattering and absorption perturbations on the observed image contrast. Increased scattering in the abnormal target is modeled as a time delay. Moreover, it was shown that the scattering contrast is proportional to the time-derivative of the PSF, $\mathrm{d}W_n/\mathrm{d}n$, divided by P_n [53]. The second interesting feature in RWT methodology assumes that the contrast from scattering inside the inclusion is proportional to the cross-section of the target (in the z direction) [51,53], instead of depending on its volume as modeled in the perturbation analysis [54].

Several research groups intend to implement their theoretical expressions into general inverse algorithms for optical tomography, that is, to reconstruct three-dimensional maps of spatial distributions of tissue optical characteristics [49], and thereby quantify optical characteristics, positions and sizes of abnormalities. Unlike these approaches, method is a multi-step analysis of the collected data. From images observed at differing flight times, we construct the time-dependent contrast functions, fit our theoretical expressions, and compute the optical properties of the background, and those of the abnormality along with its size. The outline of data analysis is given in Reference 55.

By utilizing the method for different wavelengths, one can obtain diagnostic information (e.g., estimates of blood oxygenation of the tumor) for corresponding absorption coefficients that no other imaging modality can provide directly. Several research groups have already successfully used multi-wavelength measurements using frequency-domain techniques, to calculate physiological parameters (oxygenation, lipid, water) of breast tumors (diagnosed with other modalities) and normal tissue [56].

Researchers at Physikalisch-Technische-Bundesanstalt (PTB) of Berlin have designed a clinically practical optical imaging system, capable of implementing time-resolved *in vivo* measurements on the human breast [27]. The breast is slightly compressed between two plates. A scan of the whole breast takes but a few minutes and can be done in mediolateral and craniocaudal geometries. The first goal is to quantify the optical parameters at several wavelengths and thereby estimate blood oxygen saturation of the tumor and surrounding tissue under the usual assumption that the chromophores contributing to absorption are oxy- and deoxy-hemoglobin and water. As an example, two sets of data, obtained at two wavelengths ($\lambda = 670$ and 785 nm), for a patient (84-year-old) with invasive ductal carcinoma, were analyzed. Though the images exhibit poor resolution, the tumor can be easily seen in the optical image shown in Figure 30.5a. In this figure, the image is obtained from reciprocal values of the total integrals of the distributions of times of flight of photons, normalized to a selected "bulk" area. The tumor center is located at $x = -5$, $y = 0.25$ mm.

The best spatial resolution is observed, as expected, for shorter time-delays allowing one to determine the position of the tumor center on the 2-D image (transverse coordinates) with accuracy ~2.5 mm. After preliminary data processing that includes filtering and deconvolution of the raw time-resolved data, we created linear contrast scans passing through the tumor center and analyzed these scans, using our algorithm. It is striking that one observes similar linear dependence of the contrast amplitude on the derivative of PSF ($\lambda = 670$ nm), as expected in the model (see Figure 30.5b). The slope of this linear dependence was used, to estimate the amplitude of the scattering perturbation [55].

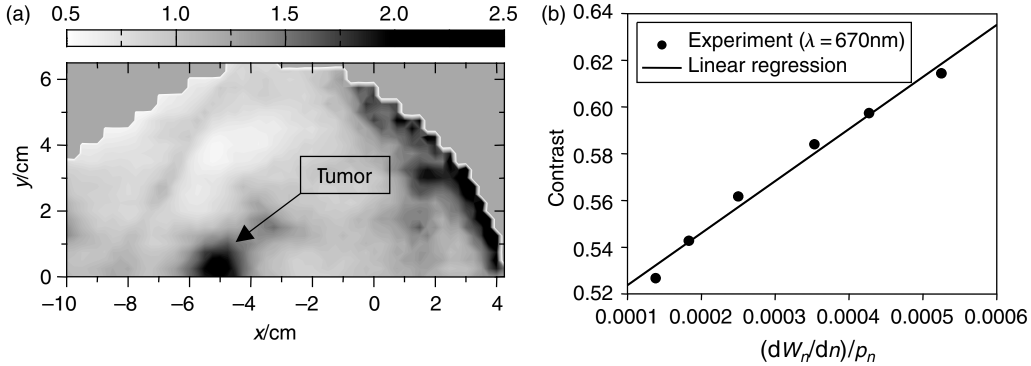


FIGURE 30.5 (a) 2-D optical image of the breast with the tumor. (Courtesy of Physikalisch-Technische-Bundesanstalt, Berlin.) (b) Contrast obtained from linear scan through the tumor plotted vs. the derivative of PSF. From the linear regression the scattering coefficient of the tumor is deduced.

TABLE 30.1 Optical Parameters of Tumor and Background Breast Tissue

Unknown coefficients	Reconstructed values (mm^{-1})	
	$\lambda = 670$ nm	$\lambda = 785$ nm
Absorption (background)	0.0029 $^{-1}$	0.0024 $^{-1}$
Scattering (background)	1.20 $^{-1}$	1.10 $^{-1}$
Absorption (tumor)	0.0071 $^{-1}$	0.0042 $^{-1}$
Scattering (tumor)	1.76 $^{-1}$	1.6 $^{-1}$

Dimensions and values of optical characteristics of the tumor and surrounding tissues were then reconstructed for both wavelengths. Results show that the tumor had larger absorption and scattering than the background. Estimated parameters are presented in Table 30.1.

Both absorption and scattering coefficients of the tumor and background all proved to be larger at the red wavelength (670 nm). Comparison of the absorption in the red and near infrared range is used to estimate blood oxygen saturation of the tumor and background tissue. Preliminary results of the analysis gave evidence that the tumor tissue is in a slightly deoxygenated state with higher blood volume, compared to surrounding tissue.

The spectroscopic power of optical imaging, along with the ability to quantify physiological parameters of human breast, have opened a new opportunity for assessing metabolic and physiological activities of the human breast during treatment.

30.1.5 Quantitative Fluorescence Imaging and Spectroscopy

As mentioned in Section 30.1.1, advances in the molecular biology of disease processes, new immunohistopathological techniques, and the development of specific fluorescently-labeled cell surface markers have led a revolution in research on the molecular origins of disease processes. On the other hand, reliable, sensitive, and specific, non-invasive techniques are needed for *in vivo* determinations of abnormalities within tissue. If successfully developed, noninvasive "optical biopsies" may replace conventional surgical biopsies and provide the advantages of smaller sampling errors, reduction in cost and time for diagnosis resulting in easier integration of diagnosis and therapy by following the progression of disease or regression in response to therapy. Clinically practical fluorescence imaging techniques must meet several requirements. First, the pathology under investigation must lie above a depth where the attenuation of the signal results in a poor signal-to-noise ratio and resolvability. Second, the specificity of the marker must be high enough that one can clearly distinguish between normal and abnormal lesions. Finally, one must have a robust image reconstruction algorithm which enables one to quantify the fluorophore concentration at a given depth.

The choices of projects in this area of research are dictated by the importance of the problem, and the impact of the solution on health care. Below, the rationale of two projects, are described that National Institutes of Health are pursuing.

Sjogren's Syndrome (SS) has been chosen as an appropriate test case for developing a noninvasive optical biopsy based on 3-D localization of exogenous specific fluorescent labels. SS is an autoimmune disease affecting minor salivary glands which are near (0.5 to 3.0 mm below) the oral mucosal surface [57]. Therefore the target pathology is relatively accessible to noninvasive optical imaging. The hydraulic conductivity of the oral mucosa is relatively high, which along with the relatively superficial location of the minor salivary glands, makes topical application and significant labeling of diseased glands with large fluorescent molecules easy to accomplish. Fluorescence ligands (e.g., fluorescent antibodies specific to CD4^{+} T cell-activated lymphocytes infiltrating the salivary glands) are expected to bind specifically to the atypical cells in the tissue, providing high contrast and a quantitative relationship to their concentration (and therefore to the stage of the disease process). The major symptoms (dry eyes and dry mouth due to decreased tear and saliva secretion) are the result of progressive immune-mediated dysfunction of the lacrimal and salivary glands. Currently, diagnosis is made by excisional biopsies of the minor salivary glands in the lower lip. This exam, though considered the best criterion for diagnosis, involves a surgical procedure under local anesthesia followed by postoperative discomfort (swelling, pain) and frequently a temporary loss of sensation at the lower lip biopsy site. Additionally, biopsy is inherently subject to sampling errors and the preparation of histopathological slides is time consuming, complicated, expensive, and requires the skills of several professionals (dentist, pathologist, and laboratory technician). Thus, there is a clear need for a noninvasive diagnostic procedure which reflects the underlying gland pathology and has good specificity. A quantitative, noninvasive assay would also allow repetition of the test to monitor disease progression and the effect of treatment. However, the quantification of fluorophore concentration within the tissue from surface images requires determining the intensities of different fluorophore sources, as a function of depth and transverse distance and predicting the 3-D distribution of fluorophores within the tissue from a series of images [58].

The second project involves the lymphatic imaging-sentinel node detection. The stage of cancer at initial diagnosis often defines prognosis and determines treatment options. As part of the staging procedure of melanoma and breast cancer, multiple lymph nodes are surgically removed from the primary lymphatic draining site and examined histologically for the presence of malignant cells. Because it is not obvious which nodes to remove at the time of resection of the primary tumor, standard practice involves dissection of as many lymph nodes as feasible. Since such extensive removal of lymphatic tissue frequently results in compromised lymphatic drainage in the examined axilla, alternatives have been sought to define the stage at the time of primary resection. A recent advance in lymph node interrogation has been the localization and removal of the "sentinel" node. Although there are multiple lymphatic channels available for trafficking from the primary tumor, the assumption was made that the anatomic location of the primary tumor in a given individual drains into lymphatic channels in an orderly and reproducible fashion. If that is in fact the case, then there is a pattern by which lymphatic drainage occurs. Thus, it would be expected that malignant cells from a primary tumor site would course from the nearest and possibly most superficial node into deeper and more distant lymphatic channels to ultimately arrive in the thoracic duct, whereupon malignant cells would gain access to venous circulation. The sentinel node is defined as the first drainage node in a network of nodes that drain the primary cancer. Considerable evidence has accrued validating the clinical utility of staging breast cancer by locating and removing the sentinel node at the time of resection of the primary tumor. Currently, the primary tumor is injected with a radionucleotide one day prior to removal of the primary tumor. Then, just before surgery, it is injected with visible dye. The surgeon localizes crudely the location of the sentinel node using a hand-held radionucleotide detector, followed by a search for visible concentrations of the injected dye. The method requires expensive equipment and also presents the patient and hospital personnel with the risk of exposure to ionizing radiation. As an alternative to the radionucleotide, we are investigating the use of IR-dependent fluorescent detection methods to determine the location of sentinel node(s).

For *in vivo* fluorescent imaging, a complicating factor is the strong attenuation of light as it passes through tissue. This attenuation deteriorates the signal-to-noise ratio of detected photons. Fortunately, development of fluorescent dyes (such as porphyrin and cyanine) that excite and re-emit in the "biological window" at NIR wavelengths, where scattering and absorption coefficients are relatively low, have provided new possibilities for deep fluorescence imaging in tissue. The theoretical complication occurs at depths greater than 1 mm where photons in most tissues enter a diffusion-like state with a large dispersion in their path-lengths. Indeed, the fluorescent intensity of light detected from deep tissue structures depends not only on the location, size, concentration, and intrinsic characteristics (e.g., lifetime, quantum efficiency) of the fluorophores, but also on the scattering and absorption coefficients of the tissue at both the excitation and emission wavelengths. Hence, in order to extract intrinsic characteristics of fluorophores within tissue, it is necessary to describe the statistics of photon pathlengths which depend on all these differing parameters.

Obviously, the modeling of fluorescent light propagation depends on the kinds of experiments that one plans to perform. For example, for frequency-domain measurements, Patterson and Pogue [59] used the diffusion approximation of the transport equation to express their results in terms of a product of two Green's function propagators multiplied by a term that describes the probability of emission of a fluorescent photon at the site. One Green's function describes the movement of an incident photon to the fluorophore, and the other describes movement of the emitted photon to the detector. In this representation, the amount of light emitted at the site of the fluorophore is directly proportional to the total amount of light impinging on the fluorophore, with no account for the variability in the number of visits by a photon before an exciting transformation. Since a transformation on an early visit to the site precludes a transformation on all later visits, this results in an overestimation of the number of photons which have a fluorescence transformation at a particular site. This overestimation is important when fluorescent absorption properties are spatially inhomogeneous and largest at later arrival times. RWT has been used to allow for this spatial inhomogeneity by introducing the multiple-passage probabilities concept, thus rendering the model more physically plausible [60]. Another incentive to devise a general theory of diffuse fluorescence photon migration is the capability to quantify local changes in fluorescence lifetime. By selecting fluorophore probes with known lifetime dependence on specific environmental variables, lifetime imaging enables one to localize and quantify such metabolic parameters as temperature and pH, as well as changes in local molecular concentrations *in vivo*.

In the probabilistic RWT model, the description of a photon path may be divided into three parts: the path from the photon source to a localized, fluorescing target; the interaction of the photon with the fluorophore; and finally, the path of the fluorescently emitted photon to a detector. Each part of the photon path may be described by a probability: first, the probability that an incident photon will arrive at the fluorophore site; second, the probability that the photon has a reactive encounter with the fluorophore and the corresponding photon transit delay, which is dependent on the lifetime of the fluorophore and the probability of the fluorophore emitting a photon; and third, the probability that the photon emitted by the fluorophore travels from the reaction site to the detector. Each of these three sequences is governed by a stochastic process. The mathematical description of the three processes is extremely complicated. The complete solution for the probability of fluorescence photon arrival at the detector is [61]:

$$\hat{\gamma}(r,s,r_0) = [\eta\Phi\hat{p}'_{\xi}(r \mid s)\hat{p}_{\xi}(s \mid r_0)] \times \Bigg[\langle\Delta n\rangle(1-\eta)[\exp(\xi)-1] + \{\eta\langle\Delta n\rangle[\exp(\xi)-1]+1\}\Bigg\{1+\Bigg[(1/8)(3/\pi)^{3/2}\sum_{j=1}^{\infty}\exp(-2j\xi)/j^{3/2}\Bigg]\Bigg\}\Bigg]^{-1} \tag{30.15}$$

where η is the probability of fluorescent absorption of an excitation wavelength photon, Φ is the quantum efficiency of the fluorophore which is the probability that an excited fluorophore will emit a photon at the emission wavelength, $\langle\Delta n\rangle$ is the mean number of steps the photon would have taken had the photon not been exciting the fluorophore (which corresponds to the fluorophore lifetime in random walk parameters) and ξ is a transform variable corresponding to the discrete analog of the Laplace transform and may be considered analogous to frequency. The probability of a photon going from the excitation source to the fluorophore site is $\hat{p}_{\xi}(s \mid r_0)$, and the probability of a fluorescent photon going from the fluorophore site to the detector is $\hat{p}'_{\xi}(r \mid s)$; the prime indicates that the wavelength of the photon has changed and therefore the optical properties of the tissue may be different. In practice, this solution is difficult to work with, so some simplifying assumptions are desired. With some simplification the result in the frequency domain is:

$$\hat{\gamma}(r,s,r_0) = \eta\Phi\{\hat{p}'_{\xi}(r \mid s)\hat{p}_{\xi}(s \mid r_0) - \xi\langle\Delta n\rangle\hat{p}'_{\xi}(r \mid s)\hat{p}_{\xi}(s \mid r_0)\} \tag{30.16}$$

The inverse Laplace transform of this equation gives the diffuse fluorescent intensity in the time-domain, and the integral of the latter over time leads to CW measurements. The accuracy of such cumbersome equations is tested in well-defined phantoms and fluorophores embedded in *ex vivo* tissue. In Figure 30.6, a line scan of fluorescent intensity collected from 500 μm^3 fluorescent dye (Molecular Probe, far red microspheres: 690 nm excitation; 720 nm emission), embedded in 10.4 mm porcine tissue

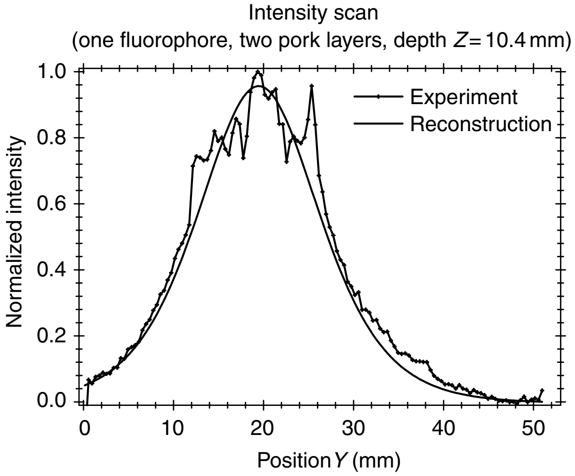


FIGURE 30.6 Intensity scan of a fluorophore 10.4 mm below the tissue surface.

with a lot of heterogeneity (e.g., fat), are presented. The dashed line is the corresponding RWT fit. The inverse algorithm written in C++ was able to construct the depth of the fluorophore with 100% accuracy. Knowing the heterogeneity of the tissue (seen in the intensity profile) this method presents huge potential to interrogate tissue structures deeply embedded in tissue for which specific fluorescent labeling such as antibodies for cell surfaces exists.

30.1.6 Future Directions

A clinically useful optical imaging device requires multidisciplinary and multi-step approaches. At the desk, one devises quantitative theories, and develop methodologies applicable to *in vivo* quantitative tissue spectroscopy and tomographic imaging in different imaging geometries (i.e., transmission or reflection), different types of measurements (e.g., steady-state or time-resolved). Effects of different optical sources of contrast such as endogenous or exogenous fluorescent labels, variations in absorption (e.g., hemoglobin or chromophore concentration) and scattering should be incorporated in the model. At the bench, one designs and conducts experiments on tissue-like phantoms and runs computer simulations to validate the theoretical findings. If successful, one tries to bring the imaging or spectroscopic device to the bedside. For this task, one must foster strong collaborations with physicians who can help to identify physiological sites where optical techniques may be clinically practical and can offer new diagnostic knowledge and less morbidity over existing methods. An important intermediate step is the use of animal models for preclinical studies. Overall, this is a complicated path. However, the spectroscopic power of light, along with the revolution in molecular characterization of disease processes has created a huge potential for *in vivo* optical imaging and spectroscopy. Maybe the twenty-first century will be the second "*siècle des lumieres.*"

30.2 Infrared Thermal Monitoring of Disease Processes: Clinical Study

The relationship between a change in body temperature and health status has been of interest to physicians since Hippocrates stated "should one part of the body be hotter or colder than the rest, then disease is present in that part." Thermography provides a visual display of the surface temperature of the skin. Skin temperature recorded by an infrared scanner is the resultant balance of thermal transport within the tissues and transport to the environment. In medical applications, thermal images of human skin contain a large amount of clinical information that can help to detect numerous pathological conditions ranging from cancer to emotional disorders. For the clinical assessment of cancer, physicians need to determine the activity of the tumor and its location, extent, and its response to therapy. All of these factors make it possible for tumors to be examined using thermography. Advantages to using this method are that it is completely nonionizing, safe, and can be repeated as often as required without exposing the patient to risk. Unfortunately, the skin temperature distribution is misinterpreted in many cases, because a high skin temperature does not always indicate a tumor. Therefore, thermography requires extensive education about how to interpret the temperature distribution patterns as well as additional research to clarify various diseases based on skin temperature.

Before applying the thermal technique in the clinical setting, it is important to consider how to avoid possible error in the results. Before the examination, the body should attain thermal equilibrium with its environment. A patient should be unclothed for at least 20 min in a controlled environment at a temperature of approximately 22°C. Under such clinical conditions, thermograms will show only average temperature patterns over an interval of time. The evaluation of surface temperature by infrared techniques requires wavelength and emissive properties of the surface (emissivity) to be examined over the range of wavelengths to which the detector is sensitive. In addition, a thermal camera should be calibrated with a known temperature reference source to standardize clinical data.

Before discussing a specific clinical application of thermography, an accurate technique for measuring emissivity is presented in Section 30.2.1. In Section 30.2.2, a procedure for temperature calibration of an infrared detector is discussed. The clinical applications of thermography with Kaposi's sarcoma are detailed in Section 30.2.3.

30.2.1 Emissivity Corrected Temperature

Emissivity is described as a radiative property of the skin. It is a measure of how well a body can radiate energy compared to a black body. Knowledge of emissivity is important when measuring skin temperature with an infrared detector system at different ambient radiation temperatures. Currently, different spectral band infrared detector systems are used in clinical studies such as 3–5 and 8–14 μm. It is well known that the emissivity of the skin varies according to the spectral range. The skin emits infrared radiation mainly between 2–20 μm with maximum emission at a wavelength around 10 μm [62]. Jones [63] showed with an InSb detector that only 2% of the radiation emitted from a thermal black body at 30°C was within the 3–5 μm spectral range; the wider spectral response of HgCdTe detector (8–14 μm) corresponded to 40–50% of this black body radiation.

Many investigators have reported on the values for emissivity of skin *in vivo*, measured in different spectral bands with different techniques. Hardy [64] and Stekettee [65] showed that the spectral emissivity of skin was independent of wavelength (λ) when $\lambda > 2$ μm. These results contradicted those obtained by Elam et al. [66]. Watmough and Oliver [67] pointed out that emissivity lies within 0.98–1 and was not less than 0.95 for a wavelength range of 2–5 μm. Patil and Williams [68] reported that the average emissivity of normal breast skin was 0.99 ± 0.045, 0.972 ± 0.041, and 0.975 ± 0.043 within the ranges 4–6, 6–18, and 4–18 μm respectively. Steketee [65] indicated that the average emissivity value of skin was 0.98 ± 0.01 within the range 3–14 μm. It is important to know the precise value of emissivity because an emissivity difference of 0.945–0.98 may cause an error of skin temperature of 0.6°C [64].

There is considerable diversity in the reported values of skin emissivity even in the same spectral band. The inconsistencies among reported values could be due to unreliable and inadequate theories and techniques employed for measuring skin emissivity. Togawa [69] proposed a technique in which the emissivity was calculated by measuring the temperature upon a transient stepwise change in ambient radiation temperature [69,70] surrounding an object surface as shown in Figure 30.7.

The average emissivity for the 12 normal subjects measured by a radiometer and infrared camera are presented in Table 30.2. The emissivity values were found to be significantly different between the 3–5 and 8–14 μm spectral bands ($p < .001$). An example of a set of images obtained during measurement using an infrared camera (3–5 μm band) on the forearm of a healthy male subject is shown in Figure 30.8.

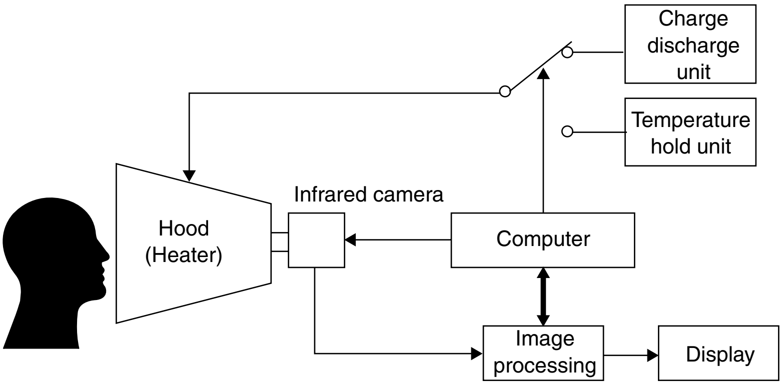


FIGURE 30.7 Schematic diagram of the emissivity measurement system [70].

TABLE 30.2 Emissivity Values

Average normal forearm skin of 12 subjects	
Infrared camera (3–5 μm)	0.958 ± 0.002
Radiometer (8–14 μm)	0.973 ± 0.0003

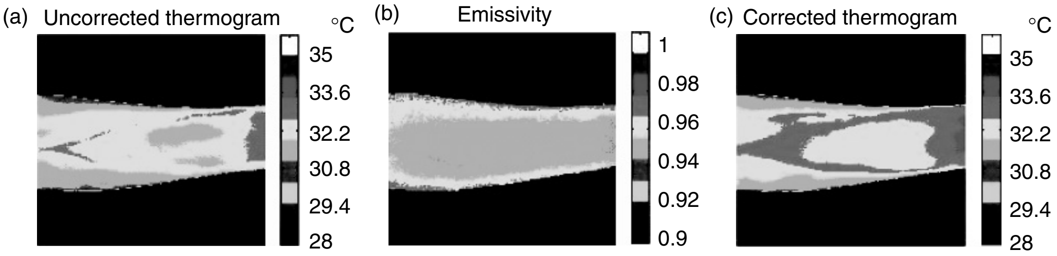


FIGURE 30.8 (See color insert.) An example of images obtained from the forearm of a normal healthy male subject. (a) Original thermogram; (b) emissivity image; (c) thermogram corrected by emissivity.

An accurate value of emissivity is important, because an incorrect value of emissivity can lead to a temperature error in radiometric thermometry especially when the ambient radiation temperature varies widely. The extent to which skin emissivity depends on the spectral range of the infrared detectors is demonstrated in Table 30.2, which shows emissivity values measured at 0.958 ± 0.002 and 0.973 ± 0.003 by an infrared detector with spectral bands of 3–5 and 8–14 μm respectively. These results can give skin temperatures that differ by 0.2°C at a room temperature of 22°C. Therefore, it is necessary to consider the wavelength dependence of emissivity, when high precision temperature measurements are required.

Emissivity not only depends on wavelength but is also influenced by surface quality, moisture on the skin surface, etc. In the infrared region of 3 to 50 μm, the emissivity of most nonmetallic substances is higher for a rough surface than a smooth one [71]. The presence of water also increases the value of emissivity [72]. These influences may account for the variation in results.

30.2.2 Temperature Calibration

In infrared thermography, any radiator is suitable as a temperature reference if its emissivity is known and constant within a given range of wavelengths. Currently, many different commercial blackbody calibrators are available to be used as temperature reference sources. A practical and simple blackbody radiator with a

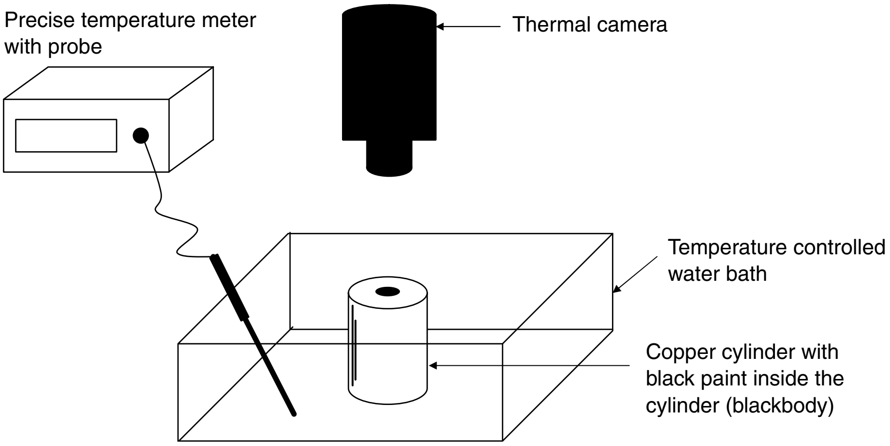

Precise temperature meter with probe
Thermal camera
Temperature controlled water bath
Copper cylinder with black paint inside the cylinder (blackbody)

FIGURE 30.9 Schematic diagram of temperature calibration system.

known temperature and measurement system is illustrated in Figure 30.9. The system consists of a hollow copper cylinder, a temperature controlled water bath and a precise temperature meter with probe. The height of the cylinder is 15 cm and the diameter is 7.5 cm. The cylinder is closed except for a hole in the center of the upper end which is 2 cm in diameter. To make the blackbody radiator, the inner surface of the cylinder is coated with black paint (3M Velvet Coating no. 2010) with emissivity of 0.93. Before the calibration, $\frac{3}{4}$ of the cylinder is placed vertically in the water and the thermal camera is placed on the top of the cylinder in a vertical direction with a distance of focus length between the surface of the hole and the camera. The water temperature ranges from 18 to 45°C by increments of 2°C. This range was selected since human temperature generally varies from 22 to 42°C in clinical studies. After setting the water temperature, the thermal camera measures the surface temperature of the hole while the temperature meter with probe measures the water temperature. The temperature of the camera is calibrated according to the temperature reading of the temperature meter.

30.2.3 Clinical Study: Kaposi's Sarcoma

The oncology community is testing a number of novel targeted approaches such as antiangiogenic, antivascular, immuno- and gene therapies for use against a variety of cancers. To monitor such therapies, it is desirable to establish techniques to assess tumor vasculature and changes with therapy [73]. Currently, several imaging techniques such as dynamic contrast-enhanced magnetic resonance (MR) imaging [74–76], positron emission tomography (PET) [77–79], computed tomography (CT) [80–83], color Doppler ultrasound (US) [84,85], and fluorescence imaging [86,87] have been used in angiogenesis-related research. With regard to monitoring vasculature, it is desirable to develop and assess noninvasive and quantitative techniques that can not only monitor structural changes, but can also assess the functional characteristics or the metabolic status of the tumor. There are currently no standard noninvasive techniques to assess parameters of angiogenesis in lesions of interest and to monitor changes in these parameters with therapy. For antiangiogenic therapies, factors associated with blood flow are of particular interest.

Kaposi's sarcoma (KS) is a highly vascular tumor that occurs frequently among people infected with acquired immunodeficiency syndrome (AIDS). During the first decade of the AIDS epidemic, 15 to 20% of AIDS patients developed this type of tumor [88]. Patients with KS often display skin and oral lesions. In addition, KS frequently involves lymph nodes and visceral organs [89]. KS is an angio-proliferative disease characterized by angiogenesis, endothelial spindle-cell growth (KS cell growth), inflammatory-cell infiltration and edema [90]. A gamma herpesvirus called Kaposi's sarcoma associated herpesvirus (KSHV) or human herpesvirus type 8 (HHV-8) is an essential factor in the pathogenesis of KS [91]. Cutaneous

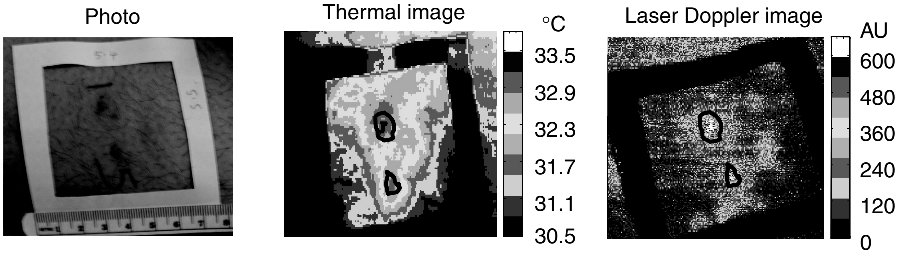


FIGURE 30.10 (See color insert.) Typical multi-modality images obtained from a patient with KS lesion. The number "1" and "5" in the visual image were written on the skin to identify the lesions for tumor measurement. The solid line in the thermal and LDI demarks the border of the visible KS lesion. Shown is a representative patient from the study reported in Reference 95.

KS lesions are easily accessible for noninvasive techniques that involve imaging of tumor vasculature, and they may thus represent a tumor model in which to assess certain parameters of angiogenesis [92,93].

Recently, two such potential noninvasive imaging techniques, infrared thermal imaging (thermography) and laser Doppler imaging (LDI) have been used to monitor patients undergoing an experimental anti-KS therapy [94,95]. Thermography graphically depicts temperature gradients over a given body surface area at a given time. It is used to study biological thermoregulatory abnormalities that directly or indirectly influence skin temperature [96–100]. However, skin temperature is only an indirect measure of skin blood flow, and the superficial thermal signature of skin is also related to local metabolism. Thus, this approach is best used in conjunction with other techniques. LDI can more directly measure the net blood velocity of small blood vessels in tissue, which generally increases as blood supply increases during angiogenesis [101,102]. Thermal patterns were recorded using an infrared camera with a uniform sensitivity in the wavelength range of 8 to 12 μm and LDI images were acquired by scanning the lesion area of the KS patients at two wavelengths, 690 and 780 nm.

An example of the images obtained from a typical KS lesion using different modalities is shown in Figure 30.10 [95]. As can be seen in the thermal image, the temperature of the lesion was approximately 2°C higher than that of the normal tissue adjacent to the lesion. Interestingly, in a number of lesions, the area of increased temperature extended beyond the lesion edges as assessed by visual inspection or palpation [95]. This may reflect relatively deep involvement of the tumor in areas underlying normal skin. However, the thermal signature of the skin not only reflects superficial vascularity, but also deep tissue metabolic activity. In the LDI image of the same lesion, there was increased blood flow in the area of the lesion as compared to the surrounding tissue, with a maximum increase of over 600 AU (arbitrary units). Unlike the thermal image, the increased blood velocity extended only slightly beyond the area of this visible lesion, possibly because the tumor process leading to the increased temperature was too deep to be detected by LDI. Both of these techniques were used successfully to visualize KS lesions [95], and although each measures an independent parameter (temperature or blood velocity), there was a strong correlation in a group of 16 patients studied by both techniques (Figure 30.11) [95]. However, there were some differences in individual lesions since LDI measured blood flow distribution in the superficial layer of the skin of the lesion, whereas the thermal signature provided a combined response of superficial vascularity and metabolic activities of deep tissue.

In patients treated with an anti-KS therapy, there was a substantial decrease in temperature and blood velocity during the initial 18-week treatment period as shown in Figure 30.12 [95]. The changes in these two parameters were generally greater than those assessed by either measurement of tumor size or palpation. In fact, there was no statistically significant decrease in tumor size overall. These results suggest that thermography and LDI may be relatively more sensitive in assessing the response of therapy in KS than conventional approaches. Assessing responses to KS therapy is now generally performed by visual measuring and palpating the numerous lesions and using rather complex response criteria. However, the current tools are rather cumbersome and often subject to observer variation, complicating the assessment

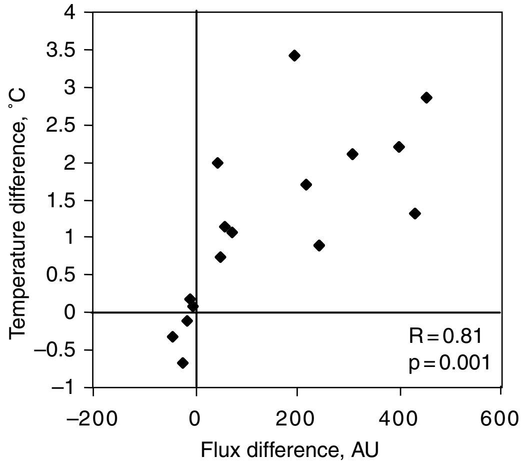


FIGURE 30.11 Relationship between the difference in temperature and flux of the lesion and surrounding area of the lesion of each subject. A positive correlation was observed between these two methods ($R = 0.8, p < .001$). (Taken from Hassan et al., TCRT, 3, 451–457, 2004. With permission.)

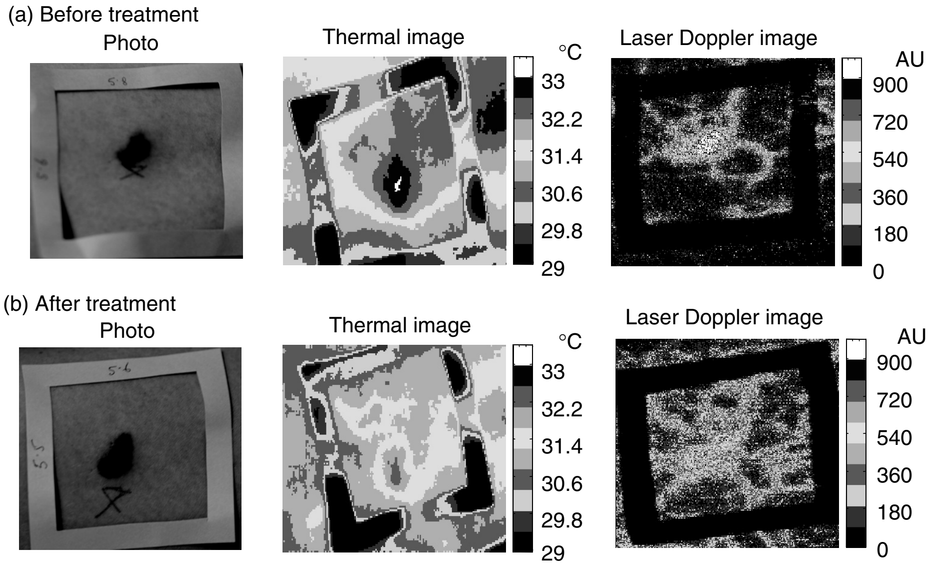


FIGURE 30.12 (See color insert.) Typical example of lesion obtained from a subject with KS (a) before, and (b) after the treatment. Improvement after the treatment can be assessed by the thermal or LDI images after 18 weeks. Shown is a patient from the clinical trial reported in Reference 95.

of new therapies. The techniques described here, possibly combined with other techniques to assess vasculature and vessel function, have the potential of being more quantitative, sensitive, and reproducible than established techniques. Moreover, it is possible that they may show a response to therapy sooner than conventional than conventional means of tumor assessment.

Acknowledgments

Special thanks go to Dr. Herbert Rinneberg (Physikalisch-Technische-Bundesanstalt, Berlin) and Dr. Brian Pogue (Dartmouth College) for providing the optical images. The authors also wish to express their thanks to Dr. Tatsuo Togawa, a former professor of the Institutes of Biomaterials and Bioengineering, Tokyo Medical and Dental University, Tokyo, Japan for his valuable suggestions and allowing emissivity measurements to be performed in his lab. The authors also like to thanks Kathleen Wyvill and Karen Aleman for their help.

References

[1] M. Born and E. Wolf, *Principles in Optics*, 7th ed. Cambridge: Cambridge University Press, 1999.

[2] A.T. Young, "Rayleigh scattering," *Phys. Today*, 42, 1982.

[3] I.S. Saidi, S.L. Jacques, and F.K. Tittel, "Mie and Rayleigh modeling of visible-light scattering in neonatal skin," *Appl. Opt.*, 34, 7410, 1995.

[4] M.J.C. Van Gemert, S.L. Jacques, H.J.C.M. Sterenberg, and W.M. Star, "Skin optics," *IEEE Trans.*, 36, 1146, 1989.

[5] R. Marchesini, A. Bertoni, S. Andreola, E. Melloni, and A. Sicherolli, "Extinction and absorption coefficients and scattering phase functions of human tissues *in vitro*," *Appl. Opt.*, 28, 2318, 1989.

[6] J. Fishkin, O. Coquoz, E. Anderson, M. Brenner, and B. Tromberg, "Frequency-domain photon migration measurements of normal and malignant tissue optical properties in a human subject," *Appl. Opt.*, 36, 10, 1997.

[7] T.L. Troy, D.L. Page, and E.M. Sevick-Muraca, "Optical properties or normal and diseased breast tissues: prognosis for optical mammography," *J. Biomed. Opt.*, 1, 342, 1996.

[8] A.H. Gandjbakhche, R.F. Bonner, and R. Nossal, "Scaling relationships for anisotropic random walks," *J. Stat. Phys.*, 69, 35, 1992.

[9] G.A. Wagnieres, W.M. Star, and B.C. Wilson, "*In vivo* fluorescence spectroscopy and imaging for oncological applications," *Photochem. Photobiol.*, 68, 603, 1998.

[10] R. Weissleder, "A clearer vision for *in vivo* imaging," *Nat. Biotechnol.*, 19, 316, 2001.

[11] V.F. Kamalov, I.A. Struganova, and K. Yoshihara, "Temperature dependent radiative lifetime of J-aggregates," *J. Phys. Chem.*, 100, 8640, 1996.

[12] S. Mordon, J.M. Devoisselle, and V. Maunoury, "*In vivo* pH measurement and imaging of a pH-sensitive fluorescent probe (5–6 carboxyfluorescein): instrumental and experimental studies," *Photochem. Photobiol.*, 60, 274, 1994.

[13] C.L. Hutchinson, J.R. Lakowicz, and E.M. Sevick-Muraca, "Fluorescence lifetime-based sensing in tissues: a computational study," *Biophys. J.*, 68, 1574, 1995.

[14] F.F. Jobsis, "Noninvasive infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters," *Science*, 198, 1264, 1977.

[15] T.J. Farrell, M.S. Patterson, and B. Wilson, "A diffusion theory model of spatially resolved, steady-state diffuse reflectance for the noninvasive determination of tissue optical properties *in vivo*," *Med. Phys.*, 9, 879, 1992.

[16] P.C. Jackson, P.H. Stevens, J.H. Smith, D. Kear, H. Key, and P. N. T. Wells, "Imaging mammalian tissues and organs using laser collimated transillumination," *J. Biomed. Eng.*, 6, 70, 1987.

[17] G. Jarry, S. Ghesquiere, J.M. Maarek, F. Fraysse, S. Debray, M.-H. Bui, and D. Laurent, "Imaging mammalian tissues and organs using laser collimated transillumination," *J. Biomed. Eng.*, 6, 70, 1984.

[18] M. Kaneko, M. Hatakeyama, P. He, Y. Nakajima, H. Isoda, M. Takai, T. Okawada, M. Asumi, T. Kato, S. Goto "Construction of a laser transmission photo-scanner: pre-clinical investigation," *Radiat. Med.*, 7, 129, 1989.

[19] L. Wang, P.P. Ho, C. Liu, G. Zhang, and R.R. Alfano, "Ballistic 2-D imaging through scattering walls using an ultrafast optical Kerr gate," *Science*, 253, 769, 1991.

[20] A. Schmitt, R. Corey, and P. Saulnier, "Imaging through random media by use of low-coherence optical heterodyning," *Opt. Lett.*, 20, 404, 1995.

[21] H. Inaba, M. Toida, and T. Ichmua, "Optical computer-assisted tomography realized by coherent detection imaging incorporating laser heterodyne method for biomedical applications," *SPIE Proc.*, 1399, 108, 1990.

[22] H. Inaba, "Coherent detection imaging for medical laser tomography," In *Medical Optical Tomography: Functional Imaging and Monitoring*, Muller, G., ed. p. 317, 1993.

[23] S.B. Colak, D.G. Papaioannou, G.W. T'Hoooft, M.B. van der Mark, H. Schomberg, J.C.J. Paasschens, J.B.M. Melissen, and N.A.A.J. van Austen, "Tomographic image reconstruction from optical projections in light diffusing media," *Appl. Opt.*, 36, 180, 1997.

[24] J.C. Hebden, D.J. Hall, M. Firbank, and D.T. Delpry, "Time-resolved optical imaging of a solid tissue-equivalent phantom," *Appl. Opt.*, 34, 8038, 1995.

[25] J.C. Hebden, "Evaluating the spatial resolution performance of a time-resolved optical imaging system," *Med. Phys.*, 19, 1081, 1992.

[26] R. Cubeddu, A. Pifferi, P. Taroni, A. Torriceli, and G. Valentini, "Time-resolved imaging on a realistic tissue phantom: us' and ua images versus time-integrated images," *Appl. Opt.*, 35, 4533, 1996.

[27] D. Grosenick, H. Wabnitz, H. Rinneberg, K.T. Moesta, and P. Schleg, "Development of a time-domain optical mammograph and first *in-vivo* application," *Appl. Opt.*, 38, 2927, 1999.

[28] M. Bashkansky, C. Adler, and J. Reinties, "Coherently amplified Raman polarization gate for imaging through scattering media," *Opt. Lett.*, 19, 350, 1994.

[29] K.M. Yoo, Q. Xing, and R.R. Alfano, "Imaging objects hidden in highly scattering media using femtosecond second-harmonic-generation cross-correlation time gating," *Opt. Lett.*, 16, 1019, 1991.

[30] G.W. Faris and M. Banks, "Upconverting time gate for imaging through highly scattering media," *Opt. Lett.*, 19, 1813, 1994.

[31] J.C. Hebden, S.R. Arridge, and D.T. Delpry, "Optical imaging in medicine I: experimental techniques," *Phys. Med. Biol.*, 42, 825, 1997.

[32] J.R. Lakowitz and K. Brendt, "Frequency domain measurements of photon migration in tissues," *Chem. Phys. Lett.*, 166, 246, 1990.

[33] M.A. Franceschini, K.T. Moesta, S. Fantini, G. Gaida, E. Gratton, H. Jess, W.W. Mantulin, M. Seeber, P.M. Schlag, and M. Kaschke, "Frequency-domain techniques enhance optical mammography: initial clinical results," *Proc. Natl Acad. Sci., Med. Sci.*, 94, 6468, 1997.

[34] B. Tromberg, O. Coquoz, J.B. Fishkin, T. Pham, E. Anderson, J. Butler, M. Cahn, J.D. Gross, V. Venugopalan, and D. Pham, "Non-invasive measurements of breast tissue optical properties using frequency-domain photon migration," *Philos. Trans. R. Soc. Lond. Ser. B*, 352, 661, 1997.

[35] J.J. Duderstadt and L.J. Hamilton, *Nuclear Reactor Analysis.* New York: Wiley, 1976.

[36] K.M. Case and P.F. Zweifel, *Linear Transport Theory*. Reading: Addison Wesley, 1967.

[37] A. Ishimaru, *Wave Propogation and Scattering in Random Media.* New York: Academic Press, 1978.

[38] M.S. Patterson, B. Chance, and B. Wilson, "Time resolved reflectance and transmittance for the non-invasive measurement of tissue optical properties," *Appl. Opt.*, 28, 2331, 1989.

[39] S.R. Arridge and J.C. Hebden, "Optical imaging in medicine: II. Modelling and reconstruction," *Phys. Med. Biol.*, 42, 841, 1997.

[40] S.R. Nioka, M. Miwa, S. Orel, M. Schnall, M. Haida, S. Zhao, and B. Chance, "Optical imaging of human breast cancer," *Adv. Exp. Med. Biol.*, 361, 171, 1994.

[41] S. Fantini, S.A. Walker, M.A. Franceschini, M. Kaschke, P.M. Schlag, and K.T. Moesta, "Assessment of the size, position, and optical properties of breast tumors *in vivo* by noninvasive optical methods," *Appl. Opt.*, 37, 1982, 1998.

[42] M. Maris, E. Gratton, J. Maier, W. Mantulin, and B. Chance, "Functional near-infrared imaging of deoxygenated haemoglobin during exercise of the finger extensor muscles using the frequency-domain techniques," *Bioimaging*, 2, 174, 1994.

[43] B.W. Pogue and K.D. Paulsen, "High-resolution near-infrared tomographic imaging simulations of the rat cranium by use of a priori magnetic resonance imaging structural information," *Opt. Lett.*, 23, 1716, 1998.

[44] R.F. Bonner, R. Nossal, S. Havlin, and G.H. Weiss, "Model for photon migration in turbid biological media," *J. Opt. Soc. Am. A*, 4, 423, 1987.

[45] A.H. Gandjbakhche and G.H. Weiss, "Random walk and diffusion-like models of photon migration in turbid media," *Progress in Optics*, Wolf, E., ed. Elsevier Science B.V., vol. XXXIV, p. 333, 1995.

[46] A.H. Gandjbakhche, R. Nossal, and R.F. Bonner, "Scaling relationships for theories of anisotropic random walks applied to tissue optics," *Appl. Opt.*, 32, 504, 1993.

[47] V. Chernomordik, R. Nossal, and A.H. Gandjbakhche, "Point spread functions of photons in time-resolved transillumination experiments using simple scaling arguments," *Med. Phys.*, 23, 1857, 1996.

[48] A.H. Gandjbakhche, G.H. Weiss, R.F. Bonner, and R. Nossal, "Photon path-length distributions for transmission through optically turbid slabs," *Phys. Rev. E*, 48, 810, 1993.

[49] D.J. Hawrysz and E.M. Sevick-Muraca, "Developments toward diagnostic breast cancer imaging using near-infrared optical measurements and fluorescent contract agents," *Neoplasia*, 2, 388, 2000.

[50] M. Cutler, "Transillumination as an aid in the diagnosis of breast lesions," *Surg. Gynecol. Obstet.*, 48, 721, 1929.

[51] A. H. Gandjbakhche, V. Chernomordik et al., "Time-dependent contract functions for quantitative imaging in time-resolved transillumination experiments," *Appl. Opt.*, 37, 1973, 1998.

[52] A.H. Gandjbakhche, R. Nossal, and R.F. Bonner, "Resolution limits for optical transillumination of abnormalities deeply embedded in tissues," *Med. Phys.*, 21, 185, 1994.

[53] V. Chernomordik, D. Hattery, A. Pifferi, P. Taroni, A. Torricelli, G. Valentini, R. Cubeddu, and A.H. Gandjbakhche, "A random walk methodology for quantification of the optical characteristics of abnormalities embedded within tissue-like phantoms," *Opt. Lett.*, 25, 951, 2000.

[54] M. Morin, S. Verreault, A. Mailloux, J. Frechette, S. Chatigny, Y. Painchaud, and P. Beaudry, "Inclusion characterization in a scattering slab with time-resolved transmittance measurements: perturbation analysis," *Appl. Opt.*, 39, 2840–2852, 2000.

[55] V. Chernomordik, D.W. Hattery, D. Grosenick, H. Wabnitz, H. Rinneberg, K.T. Moesta, P.M. Schlag, and A.H. Gandjbakhche, "Quantification of optical properties of a breast tumor using random walk theory," *J. Biomed. Opt.*, 7, 80–87, 2002.

[56] A.P. Gibson, J.C. Hebden, and S.R. Arridge, "Recent advances in diffuse optical imaging," *Phys. Med. Biol.*, 50, R1–R43, 2005.

[57] R.I. Fox, "Treatment of patient with Sjogren syndrome," *Rhem. Dis. Clin. North Amer.*, 18, 699–709, 1992.

[58] V. Chernomordik, D. Hattery, I. Gannot, and A.H. Gandjbakhche, "Inverse method 3D reconstruction of localized in-vivo fluorescence. Application to Sjogren syndrome," *IEEE J. Select Topics in Quant. Elec.*, 5, 930, 1999.

[59] M.S. Patterson and B.W. Pogue, "Mathematical model for time-resolved and frequency-domain fluorescence spectroscopy in biological tissue," *Appl. Opt.*, 33, 1963, 1994.

[60] A.H. Gandjbakhche, R.F. Bonner, R. Nossal, and G.H. Weiss, "Effects on multiple passage probabilities on fluorescence signals from biological media," *Appl. Opt.*, 36, 4613, 1997.

[61] D. Hattery, V. Chernomordik, M. Loew, I. Gannot, and A.H. Gandjbakhche, "Analytical solutions for time-resolved fluorescence lifetime imaging in a turbid medium such as tissue," *JOSA(A)*, 18, 1523, 2001.

[62] E. Samuel, "Thermography — some clinical applications," *Biomed. Eng.*, 4, 15–19, 1969.

[63] C.H. Jones, "Physical aspects of thermography in relation to clinical techniques," *Bibl. Radiol.*, 6, 1–8, 1975.

[64] J. Hardy, "The radiation power of human skin in the infrared," *Am. J. Physiol.*, 127, 454–462, 1939.

[65] J. Steketee, "Spectral emissivity of skin and pericardium," *Phys. Med. Biol.*, 18, 686–694, 1973.

[66] R. Elam, D. Goodwin, and K. Williams, "Optical properties of human epidermics," *Nature*, 198, 1001–1002, 1963.

[67] D.J. Watmough and R. Oliver, "Emissivity of human skin in the waveband between 2micra and 6micra," *Nature*, 219, 622–624, 1968.

[68] K.D. Patil and K.L. Willaiam, "Spectral study of human radiation. Non-ionizing radiation," *Non-Ionizing Radiation*, 1, 39–44, 1969.

[69] T. Togawa, "Non-contact skin emissivity: measurement from reflectance using step change in ambient radiation temperature," *Clin. Phys. Physiol. Meas.*, 10, 39–48, 1989.

[70] M. Hassan and T. Togawa, "Observation of skin thermal inertia distribution during reactive hyperaemia using a single-hood measurement system," *Physiol. Meas.*, 22, 187–200, 2001.

[71] W.H. McAdams, *Heat Transmission*. New York: McGraw Hill, p. 472, 1954.

[72] H.T. Hammel, J.D. Hardy, and D. Murgatroyd, "Spectral transmittance and reflectance of excised human skin," *J. Appl. Physiol.*, 9, 257–264, 1956.

[73] D.M. McDonald and P.L. Choyke, "Imaging of angiogenesis: from microscope to clinic," *Nat. Med.*, 9, 713–725, 2003.

[74] J.S. Taylor, P.S. Tofts, R. Port, J.L. Evelhoch, M. Knopp, W.E. Reddick, V.M. Runge, and N. Mayr, "MR imaging of tumor microcirculation: promise for the new millennium," *J. Magn. Reson. Imaging*, 10, 903–907, 1999.

[75] K.L. Verstraete, Y. De Deene, H. Roels, A. Dierick, D. Uyttendaele, and M. Kunnen, "Benign and malignant musculoskeletal lesions: dynamic contrast-enhanced MR imaging—parametric 'first-pass' images depict tissue vascularization and perfusion," *Radiology*, 192, 835–843, 1994.

[76] L.D. Buadu, J. Murakami, S. Murayama, N. Hashiguchi, S. Sakai, K. Masuda, S. Toyoshima, S. Kuroki, and S. Ohno, "Breast lesions: correlation of contrast medium enhancement patterns on MR images with histopathologic findings and tumor angiogenesis," *Radiology*, 200, 639–649, 1996.

[77] A. Fredriksson and S. Stone-Elander, "PET screening of anticancer drugs. A faster route to drug/target evaluations *in vivo*," *Meth. Mol. Med.*, 85, 279–294, 2003.

[78] G. Jerusalem, R. Hustinx, Y. Beguin, and G. Fillet, "The value of positron emission tomography (PET) imaging in disease staging and therapy assessment," *Ann. Oncol.*, 13, 227–234, 2002.

[79] H.C. Steinert, M. Hauser, F. Allemann, H. Engel, T. Berthold, G.K. von Schulthess, and W. Weder, "Non-small cell lung cancer: nodal staging with FDG PET versus CT with correlative lymph node mapping and sampling," *Radiology*, 202, 441–446, 1997.

[80] S. D. Rockoff, "The evolving role of computerized tomography in radiation oncology," *Cancer*, 39, 694–696, 1977.

[81] K.D. Hopper, K. Singapuri, and A. Finkel, "Body CT and oncologic imaging," *Radiology*, 215, 27–40, 2000.

[82] K.A. Miles, M. Hayball, and A.K. Dixon, "Colour perfusion imaging: a new application of computed tomography," *Lancet*, 337, 643–645, 1991.

[83] K.A. Miles, C. Charnsangavej, F.T. Lee, E.K. Fishman, K. Horton, and T.Y. Lee, "Application of CT in the investigation of angiogenesis in oncology," *Acad. Radiol.*, 7, 840–850, 2000.

[84] N. Ferrara, "Role of vascular endothelial growth factor in physiologic and pathologic angiogenesis: therapeutic implications," *Semin. Oncol.*, 29, 10–14, 2002.

[85] D.E. Goertz, D.A. Christopher, J.L. Yu, R.S. Kerbel, P.N. Burns, and F.S. Foster, "High-frequency color flow imaging of the microcirculation," *Ultrasound Med. Biol.*, 26, 63–71, 2000.

[86] E.M. Gill, G.M. Palmer, and N. Ramanujam, "Steady-state fluorescence imaging of neoplasia," *Meth. Enzymol.*, 361, 452–481, 2003.

[87] K. Svanberg, I. Wang, S. Colleen, I. Idvall, C. Ingvar, R. Rydell, D. Jocham, H. Diddens, S. Bown, G. Gregory, S. Montan, S. Andersson-Engels, and S. Svanberg, "Clinical multi-colour fluorescence imaging of malignant tumours — initial experience," *Acta Radiol.*, 39, 2–9, 1998.

[88] V. Beral, T.A. Peterman, R.L. Berkelman, and H.W. Jaffe, "Kaposi's sarcoma among persons with AIDS: a sexually transmitted infection?" *Lancet*, 335, 123–128, 1990.

[89] B.A. Biggs, S.M. Crowe, C.R. Lucas, M. Ralston, I.L. Thompson, and K. J. Hardy, "AIDS related Kaposi's sarcoma presenting as ulcerative colitis and complicated by toxic megacolon," *Gut*, 28, 1302–1306, 1987.

[90] E. Cornali, C. Zietz, R. Benelli, W. Weninger, L. Masiello, G. Breier, E. Tschachler, A. Albini, and M. Sturzl, "Vascular endothelial growth factor regulates angiogenesis and vascular permeability in Kaposi's sarcoma," *Am. J. Pathol.*, 149, 1851–1869, 1996.

[91] Y. Chang, E. Cesarman, M.S. Pessin, F. Lee, J. Culpepper, D.M. Knowles, and P.S. Moore, "Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma," *Science*, 266, 1865–1869, 1994.

[92] R. Yarchoan, "Therapy for Kaposi's sarcoma: recent advances and experimental approaches," *J. Acquir. Immune Defic. Syndr.*, 21, S66–73, 1999.

[93] R.F. Little, K.M. Wyvill, J.M. Pluda, L. Welles, V. Marshall, W.D. Figg, F.M. Newcomb, G. Tosato, E. Feigal, S.M. Steinberg, D. Whitby, J.J. Goedert, and R. Yarchoan, "Activity of thalidomide in AIDS-related Kaposi's sarcoma," *J. Clin. Oncol.*, 18, 2593–2602, 2000.

[94] M. Hassan, D. Hattery, V. Chernomordik, K. Aleman, K. Wyvill, F. Merced, R.F. Little, R. Yarchoan, and A. Gandjbakhche, "Non-invasive multi-modality technique to study angiogenesis associated with Kaposi's sarcoma," *Proceedings of EMBS BMES*, pp. 1139–1140, 2002.

[95] M. Hassan, R.F. Little, A. Vogel, K. Aleman, K. Wyvill, R. Yarchoan, and A. Gandjbakhche, "Quantitative assessment of tumor vasculature and response to therapy in Kaposi's sarcoma using functional noninvasive imaging," *TCRT*, 3, 451–458, 2004.

[96] C. Maxwell-Cade, "Principles and practice of clinical thermography," *Radiography*, 34, 23–34, 1968.

[97] J.F. Head and R.L. Elliott, "Infrared imaging: making progress in fulfilling its medical promise," *IEEE Eng. Med. Biol. Mag.*, 21, 80–85, 2002.

[98] S. Bornmyr and H. Svensson, "Thermography and laser-Doppler flowmetry for monitoring changes in finger skin blood flow upon cigarette smoking," *Clin. Physiol.*, 11, 135–141, 1991.

[99] K. Usuki, T. Kanekura, K. Aradono, and T. Kanzaki, "Effects of nicotine on peripheral cutaneous blood flow and skin temperature," *J. Dermatol. Sci.*, 16, 173–181, 1998.

[100] M. Anbar, "Clinical thermal imaging today," *IEEE Eng. Med. Biol. Mag.*, 17, 25–33, 1998.

[101] J. Sorensen, M. Bengtsson, E.L. Malmqvist, G. Nilsson, and F. Sjoberg, "Laser Doppler perfusion imager (LDPI) — for the assessment of skin blood flow changes following sympathetic blocks," *Acta Anaesthesiol. Scand.*, 40, 1145–1148, 1996.

[102] A. Rivard, J.E. Fabre, M. Silver, D. Chen, T. Murohara, M. Kearney, M. Magner, T. Asahara, and J.M. Isner, "Age-dependent impairment of angiogenesis," *Circulation*, 99, 111–120, 1999.

31

Thermal Imaging in Diseases of the Skeletal and Neuromuscular Systems

E. Francis Ring
University of Glamorgan

Kurt Ammer
Ludwig Boltzmann Research Institute for Physical Diagnostics and University of Glamorgan

31.1 Introduction

Clinical medicine has made considerable advances over the last century. The introduction of imaging modalities has widened the ability of physicians to locate and understand the extent and activity of a disease. Conventional radiography has dramatically improved, beyond the mere demonstration of bone and calcified tissue. Computed tomography ultrasound, positron emission tomography, and magnetic resonance imaging are now available for medical diagnostics.

Infrared imaging has also added to this range of imaging procedures. It is often misunderstood, or not been used due to lack of knowledge of thermal physiology and the relationship between temperature and disease.

In Rheumatology, disease assessment remains complex. There are a number of indices used, which testify to the absence of any single parameter for routine investigation. Most indices used are subjective.

Objective assessments are of special value, but may be more limited due to their invasive nature. Infrared imaging is noninvasive, and with modern technology has proved to be reliable and useful in rheumatology.

From early times physicians have used the cardinal signs of inflammation, that is, pain, swelling, heat, redness, and loss of function. When a joint is acutely inflamed, the increase in heat can be readily detected by touch. However, subtle changes in joint surface temperature occur and increase and decrease in temperature can have a direct expression of reduction or exacerbation of inflammation.

31.2 Inflammation

Inflammation is a complex phenomenon, which may be triggered by various forms of tissue injury. A series of cellular and chemical changes take place that are initially destructive to the surrounding tissue. Under normal circumstances the process terminates when healing takes place, and scar tissue may then be formed.

A classical series of events take place in the affected tissues. First, a brief arteriolar constriction occurs, followed by a prolonged dilatation of arterioles, capillaries, and venules. The initial increased blood flow caused by the blood vessel dilation becomes sluggish and leucocytes gather at the vascular endothelium. Increased permeability to plasma proteins causes exudates to form, which is slowly absorbed by the lymphatic system. Fibrinogen, left from the reabsorption partly polymerizes to fibrin. The increased permeability in inflammation is attributed to the action of a number of mediators, including histamines, kinins, and prostaglandins. The final process is manifest as swelling caused by the exudates, redness, and increased heat in the affected area resulting from the vasodilation, and increased blood flow. Loss of function and pain accompany these visible signs.

Increase in temperature and local vascularity can be demonstrated by some radionuclide procedures. In most cases, the isotope is administered intravenously and the resulting uptake is imaged or counted with a gamma camera. Superficial increases in blood flow can also be shown by laser doppler imaging although the response time may be slow. Thermal imaging, based on infrared emission from the skin is both fast and noninvasive.

This means that it is a technique that is suitable for repeated assessment, and especially useful in clinical trials of treatment whether by drugs, physical therapy, or surgery.

Intra-articular injection, particularly to administer corticosteroids came into use in the middle of the last century. Horvath and Hollander in 1949 [1] used intra-articular thermocouples to monitor the reduction in joint inflammation and synovitis following treatment. This method of assessment while useful to provide objective evidence of anti-inflammatory treatment was not universally used for obvious ethical reasons.

The availability of noncontact temperature measurement for infrared radiometry was a logical progression. Studies in a number of centers were made throughout the 1960s to establish the best analogs of corticosteroids and their effective dose. Work by Collins and Cosh in 1970 [2] and Ring and Collins 1970 [3] showed that the surface temperature of an arthritic joint was related to the intra-articular joint, and to other biochemical markers of inflammation obtained from the exudates. In a series of experiments with different analogues of prednisolone (all corticosteroids), the temperature measured by thermal imaging in groups of patients can be used to determine the duration and degree of reduction in inflammation [4,5].

At this time, a thermal challenge test for inflamed knees was being used in Bath, based on the application of a standard ice pack to the joint. This form of treatment is still used, and results in a marked decrease of joint temperature, although the effect may be transient.

The speed of temperature recovery after an ice pack of 1 kg of crushed ice to the knee for 10 min, was shown to be directly related to the synovial blood flow and inflammatory state of the joint. The mean temperature of the anterior surface of the knee joint could be measured either by infrared radiometry or by quantitative thermal imaging [6].

A number of new nonsteroid anti-inflammatory agents were introduced into rheumatology in the 1970s and 1980s. Infrared imaging was shown to be a powerful tool for the clinical testing of these drugs, using temperature changes in the affected joints as an objective marker. The technique had been successfully used on animal models of inflammation, and effectively showed that optimal dose response curves could be obtained from temperature changes at the experimental animal joints. The process with human patients suffering from acute Rheumatoid Arthritis was adapted to include a washout period for previous medication. This should be capable of relieving pain but no direct anti-inflammatory action per se. The compound used by all the pharmaceutical companies was paracetamol. It was shown by Bacon et al. [7] that small joints such as fingers and metacarpal joints increased in temperature quite rapidly while paracetamol treatment was given, even if pain was still suppressed. Larger joints, such as knees and ankles required more than one week of active anti-inflammatory treatment to register the same effect. Nevertheless, the commonly accepted protocol was to switch to the new test anti-inflammatory treatment after one week of washout with the analgesic therapy. In every case if the dose was ineffective the joint temperature was not reduced. At an effective dose, a fall in temperature was observed, first in the small joints, then later in the larger joints. Statistical studies were able to show an objective decrease in joint temperature by infrared imaging as a result of a new and successful treatment. Not all the new compounds found their way into routine medicine; a few were withdrawn as a result of undesirable side effects. The model of infrared imaging to measure the effects of a new treatment for arthritis was accepted by all the pharmaceutical companies involved and the results were published in the standard peer reviewed medical journals. More recently attention has been focused on a range of new biological agents for reducing inflammation. These also are being tested in trials that incorporate quantitative thermal imaging.

To facilitate the use and understanding of joint temperature changes, Ring and Collins [3], Collins et al. [8] devised a system for quantitation. This was based on the distribution of isotherms from a standard region of interest. The Thermal Index was calculated as the mean temperature difference from a reference temperature. The latter was determined from a large study of 600 normal subjects where the average temperature threshold for ankles, knees, hands, elbows, and shoulder were calculated. Many of the clinical trials involved the monitoring of hands, elbows, knees, and ankle joints. Normal index figure obtained from controls under the conditions described was from 1 to 2.5 on this scale. In inflammatory arthritis this figure was increased to 4–5, while in osteoarthritic joints, the increase in temperature was usually less, 3–4. In gout and infection higher values around 6–7 on this scale were recorded.

However, to determine normal values of finger joints is a very difficult task. This difficulty arises partly from the fact, that cold fingers are not necessarily a pathological finding. Tender joints showed higher temperatures than nontender joints, but a wide overlap of readings from nonsymptomatic and symptomatic joints was observed [9]. Evaluation of finger temperatures from the reference database of normal thermograms [10] of the human body might ultimately solve the problem of being able to establish a normal range for finger joint temperatures in the near future.

31.3 Paget's Disease of Bone

The early descriptions of Osteitis Deformans by Czerny [11] and Paget [12] refer to "chronic inflammation of bone." An increased skin temperature over an active site of this disease has been a frequent observation and that the increase may be around 4°C. Others have shown an increase in peripheral blood flow in almost all areas examined. Increased periosteal vascularity has been found during the active stages of the disease. The vascular bed is thought to act as an arterio-venous shunt, which may lead to high output cardiac failure. A number of studies, initially to monitor the effects of calcitonin, and later bisphosphonate therapy have been made at Bath (UK). As with the clinical trials previously mentioned, a rigorous technique is required to obtain meaningful scientific data. It was shown that the fall in temperature during calcitonin treatment was also indicated more slowly, by a fall in alkaline phosphatase, the common biochemical marker. Relapse and the need for retreatment was clearly indicated by thermal imaging. Changes in the thermal index often preceded the onset of pain and other symptoms by 2 to 3 weeks. It was also shown that the level of increased temperature over the bone was related to the degree of bone pain. Those patients who had maximal temperatures recorded at the affected bone experienced severe bone pain. Moderate pain was found in those with raised temperature, and no pain in those patients with normal temperatures. The most dramatic temperature changes were observed at the tibia, where the bone is very close to the skin surface. In a mathematical model, Ring and Davies [13] showed that the increased temperature measured over the tibia was primarily derived from osseous blood flow and not from metabolic heat. This disease is often categorized as a metabolic bone disease.

31.4 Soft Tissue Rheumatism

31.4.1 Muscle Spasm and Injury

Muscle work is the most important source for metabolic heat. Therefore, contracting muscles contribute to the temperature distribution at the body's surface of athletes [14,15]. Pathological conditions such as muscle spasms or myofascial trigger points may become visible at regions of increased temperature [16]. An anatomic study from Israel proposes in the case of the levator scapulae muscle that the frequently seen hot spot on thermograms of the tender tendon insertion on the medial angle of the scapula might be caused by an inflamed bursae and not by a taut band of muscle fibers [17].

Acute muscle injuries may also be recognized by areas of increased temperature [18] due to inflammation in the early state of trauma. However, long lasting injuries and also scars appear at hypothermic areas caused by reduced muscle contraction and therefore reduced heat production. Similar areas of decreased temperature have been found adjacent to peripheral joints with reduced range of motion due to inflammation or pain [19]. Reduced skin temperatures have been related to osteoarthritis of the hip [20] or to frozen shoulders [21,22]. The impact of muscle weakness on hypothermia in patients suffering from paresis was discussed elsewhere [23].

31.4.2 Sprains and Strains

Ligamentous injuries of the ankle [24] and meniscal tears of the knee [25] can be diagnosed by infrared thermal imaging. Stress fractures of bone may become visible in thermal images prior to typical changes in x-rays [26] Thermography provides the same diagnostic prediction as bone scans in this condition.

31.4.3 Enthesopathies

Muscle overuse or repetitive strain may lead to painful tendon insertions or where tendons are shielded by tendon sheaths or adjacent to bursae, to painful swellings. Tendovaginitis in the hand was successfully diagnosed by skin temperature measurement [27]. The acute bursitis at the tip of the elbow can be detected through an intensive hot spot adjacent to the olecranon [28]. Figure 31.1 shows an acute tendonitis of the Achilles tendon in a patient suffering from inflammatory spondylarthropathy.

31.4.3.1 Tennis Elbow

Painful muscle insertion of the extensor muscles at the elbow is associated with hot areas on a thermogram [29]. Thermal imaging can detect persistent tendon insertion problems of the elbow region in a similar way as isotope bone scanning [30]. Hot spots at the elbow have also been described as having a high association with a low threshold for pain on pressure [31]. Such hot areas have been successfully used as outcome measure for monitoring treatment [32,33]. In patients suffering from fibromyalgia, bilateral hot spots at the elbows is a common finding [34]. Figure 31.2 is the image of a patient suffering from tennis elbow with a typical hot spot in the region of tendon insertion.

31.4.3.2 Golfer Elbow

Pain due to altered tendon insertions of flexor muscles on the medial side of the elbow is usually named Golfer's elbow. Although nearly identical in pathogenesis as the tennis elbow, temperature symptoms in this condition were rarely found [35].

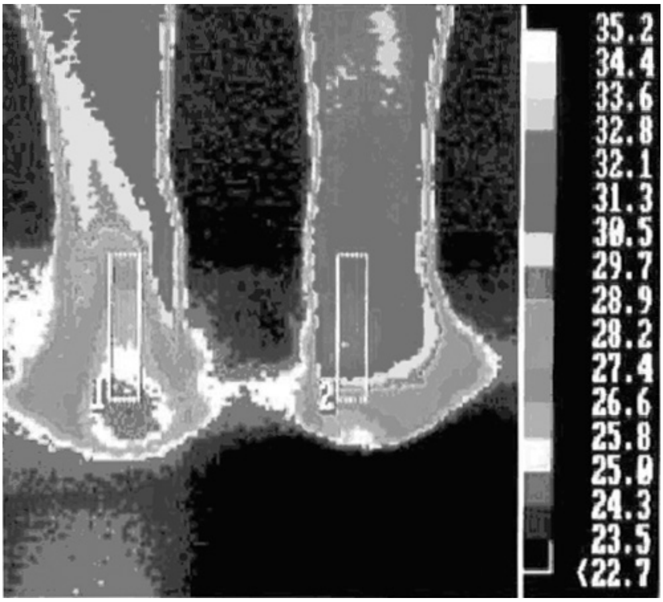


FIGURE 31.1 (See color insert following page 29-16.) Acute tendonitis of the right Achilles tendon in a patient suffering from inflammatory spondylarthropathy.

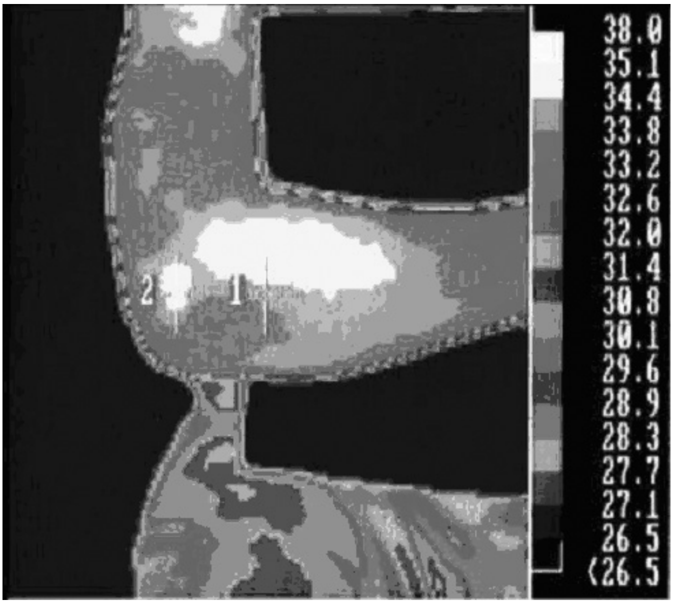


FIGURE 31.2 (See color insert.) Tennis elbow with a typical hot spot in the region of tendon insertion.

31.4.3.3 Periarthropathia of the Shoulder

The term periarthropathia includes a number of combined alterations of the periarticular tissue of the humero-scapular joint. The most frequent problems are pathologies at the insertion of the supraspinous and infraspinous muscles, often combined with impingement symptoms in the subacromial space. Long lasting insertion alteration can lead to typical changes seen on radiographs or ultrasound images, but

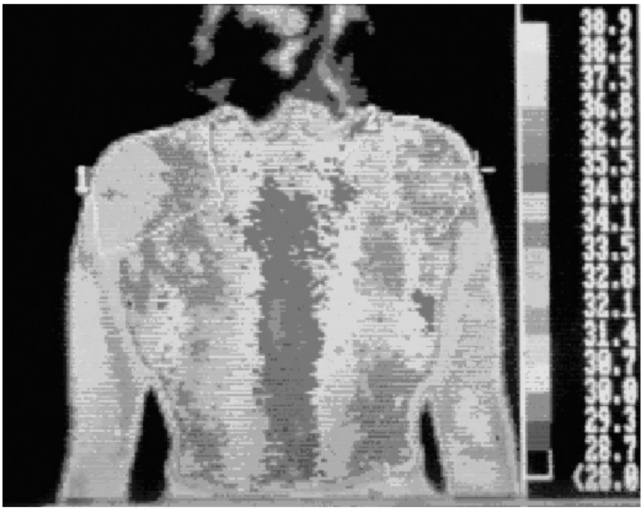


FIGURE 31.3 (See color insert.) Decreased temperature in patient with a frozen shoulder on the left-hand side.

unfortunately there are no typical temperature changes caused by the disease [22,36]. However, persistent loss in range of motion will result in hypothermia of the shoulder region [21,22,36,37]. Figure 31.3 gives an example of an area of decreased temperature over the left shoulder region in patient with restricted range of motion.

31.4.4 Fibromyalgia

The terms tender points (important for the diagnosis of fibromyalgia) and trigger points (main feature of the myofascial pain syndrome) must not be confused. Tender points and trigger points may give a similar image on the thermogram. If this is true, patients suffering from fibromyalgia may present with a high number of hot spots in typical regions of the body. A study from Italy could not find different patterns of heat distribution in patients suffering from fibromyalgia and patients with osteoarthritis of the spine [38]. However, they reported a correspondence of nonspecific hyperthermic patterns with painful muscle areas in both groups of patients. Our thermographic investigations in fibromyalgia revealed a diagnostic accuracy of 60% of hot spots for tender points [34]. The number of hot spot was greatest in fibromyalgia patients and the smallest in healthy subjects. More than 7 hot spots seem to be predictive for tenderness of more than 11 out of 18 specific sites [39]. Based on the count of hot spots, 74.2% of 252 subjects (161 fibromyalgia, 71 with widespread pain but less than 11 tender sites out of 18, and 20 healthy controls) have been correctly diagnosed. However, the intra- and inter-observer reproducibility of hot spot count is rather poor [40]. Software assisted identification of hot or cold spots based on the angular distribution around a thermal irregularity [41] might overcome that problem of poor repeatability.

31.5 Peripheral Nerves

31.5.1 Nerve Entrapment

Nerve entrapment syndromes are compression neuropathies at specific sites in human body. These sites are narrow anatomic passages where nerves are situated. The nerves are particularly prone to extrinsic or intrinsic pressure. This can result in paraesthesias such as tingling or numb feelings, pain, and ultimately in muscular weakness and atrophy.

Uematsu [42] has shown in patients with partial and full lesion of peripheral nerves that both conditions can be differentiated by their temperature reaction to the injury. The innervated area of partially lesioned nerve appears hypothermic caused by activation of sympathetic nerve fibers. Fully dissected nerves result in a total loss of sympathetic vascular control and therefore in hyperthermic skin areas.

The spinal nerves, the brachial nerve plexus, and the median nerve at the carpal tunnel are the most frequently affected nerves with compression neuropathy.

31.5.1.1 Radiculopathy

A slipped nucleus of an intervertebral disk may compress the adjacent spinal nerve or better the sensory and motor fibers of the dorsal root of the spinal nerve. This may or must not result in symptoms of compression neuropathy in the body area innervated by these fibers.

The diagnostic value of infrared thermal imaging in radiculopathies is still under debate. A review by Hoffman et al. [43] from 1991 concluded that thermal imaging should be used only for research and not in clinical routine. This statement was based on the evaluation of 28 papers selected from a total of 81 references.

The study of McCulloch et al. [44] planned and conducted at a high level of methodology, found thermography not valid. However, the applied method of recording and interpretation of thermal images was not sufficient. The chosen room temperature of 20 to 22°C might have been too low for the identification of hypothermic areas. Evaluation of thermal images was based on the criterion that at least 25% of a dermatome present with hypothermia of 1°C compared to the contralateral side. This way of interpretation might be feasible for contact thermography, but does not meet the requirements of quantitative infrared imaging.

The paper of Takahashi et al. [45] showed that the temperature deficit identified by infrared imaging is an additional sign in patients with radiculoapathy. Hypothermic areas did not correlate with sensory dermatomes and only slightly with the underlying muscles of the hypothermic area. The diagnostic sensitivity (22.9–36.1%) and the positive predictive value (25.2–37.0%) were low for both, muscular symptoms such as tenderness or weakness and for spontaneous pain and sensory loss. In contrast, high specificity (78.8–81.7%), high negative predictive values (68.5–86.2%), and a high diagnostic accuracy were obtained.

Only the papers by Kim and Cho [46] and Zhang et al. [47] found thermography of high value for the diagnosis of both lumbosacral and cervical radiculopathies. However, these studies have several methodological flaws. Although a high number of patients were reported, healthy control subjects were not mentioned in the study on lumbosacral radiculopathy. The clinical symptoms are not described and the reliability of the used thermographic diagnostic criteria remains questionable.

31.5.1.2 Thoracic Outlet Syndrome

Similar to fibromyalgia, the disease entity of the thoracic outlet syndrome (TOS) is under continuous debate [48]. Consensus exists, that various subforms related to the severity of symptoms must be differentiated. Recording thermal images during diagnostic body positions can reproducibly provoke typical temperature asymmetries in the hands of patients with suspected thoracic outlet syndrome [49,50]. Temperature readings from thermal images from patients passing that test can be reproduced by the same and by different readers with high precision [51]. The original protocol included a maneuver in which the fist was opened and closed 30 times before an image of the hand was recorded. As this test did not increase the temperature difference between index and little finger, the fist maneuver was removed from the protocol [52]. Thermal imaging can be regarded as the only technique that can objectively confirm the subjective symptoms of mild thoracic outlet syndrome. It was successfully used as outcome measure for the evaluation of treatment for this pain syndrome [53]. However, in a patient with several causes for the symptoms paraestesias and coldness of the ulnar fingers, thermography could show only a marked cooling of the little finger, but could not identify all reasons for that temperature deficit [54]. It was also difficult to differentiate between subjects whether they suffer from TOS or carpal tunnel syndrome. Only

66.3% of patients were correctly allocated to three diagnostic groups, while none of the carpal tunnel syndromes have been identified [55].

31.5.1.3 Carpal Tunnel Syndrome

Entrapment of the median nerve at the carpal tunnel is the most common compression neuropathy. A study conducted in Sweden revealed a prevalence of 14.4%; for pain, numbness, and tingling in the median nerve distribution in the hands. Prevalence of clinically diagnosed carpal tunnel syndrome (CTS) was 3.8 and 4.9% for pathological results of nerve conduction of the median nerve. Clinically and electrophysiologically confirmed CTS showed a prevalence of 2.7% [56].

The typical distribution of symptoms leads to the clinical suspect of CTS [57], which must be confirmed by nerve conduction studies. The typical electroneurographic measurements in patients with CTS show a high intra- and inter-rater reproducibility [58]. The course of nerve conduction measures for a period of 13 years in patients with and without decompression surgery was investigated and it was shown that most of the operated patients presented with less pathological conduction studies within 12 months after operation [59]. Only 2 of 61 patients who underwent a simple nerve decompression by division of the carpal ligament as therapy for CTS had pathological findings in nerve conduction studies 2 to 3 years after surgery [60].

However, nerve conduction studies are unpleasant for the patient and alternative diagnostic procedures are welcome. Liquid crystal thermography was originally used for the assessment of patients with suspected CTS [61–64]. So et al. [65] used infrared imaging for the evaluation of entrapment syndromes of the median and ulnar nerves. Based on their definition of abnormal temperature difference to the contralateral side, they found thermography without any value for assisting diagnosis and inferior to electrodiagnostic testing. Tchou reported infrared thermography of high diagnostic sensitivity and specificity in patients with unilateral CTS. He has defined various regions of interest representing mainly the innervation area of the median nerve. Abnormality was defined if more than 25% of the measured area displayed a temperature increase of at least 1°C when compared with the asymptomatic hand [66].

Ammer has compared nerve conduction studies with thermal images in patients with suspected CTS. Maximum specificity for both nerve conduction and clinical symptoms was obtained for the temperature difference between the 3rd and 4th finger at a threshold of 1°C. The best sensitivity of 69% was found if the temperature of the tip of the middle finger was by 1.2°C less than temperature of the metacarpus [67].

Hobbins [68] combined the thermal pattern with the time course of nerve injuries. He suggested the occurrence of a hypothermic dermatome in the early phase of nerve entrapment and hyperthermic dermatomes in the late phase of nerve compression. Ammer et al. [69] investigated how many patients with a distal latency of the median nerve greater than 6 msec present with a hyperthermic pattern. They reported a slight increase of the frequency of hyperthermic patterns in patients with severe CTS indicating that the entrapment of the median nerve is followed by a loss of the autonomic function in these patients.

Ammer [70] has also correlated the temperature of the index finger with the temperature of the sensory distribution of the median nerve on the dorsum of the hand and found nearly identical readings for both areas. A similar relationship was obtained for the ulnar nerve. The author concluded from these data that the temperature of the index or the little finger is highly representative for the temperature of the sensory area of the median or ulnar nerve, respectively.

Many studies on CTS have used a cold challenge to enhance the thermal contrast between affected fingers. A slow recovery rate after cold exposure is diagnostic for Raynaud's Phenomenon [71]. The coincidence of CTS and Raynaud's phenomenon was reported in the literature [72,73].

31.5.1.4 Other entrapment syndromes

No clear thermal pattern was reported for the entrapment of the ulnar nerve [65]. A pilot study for the comparison of hands from patients with TOS or entrapment of the ulnar nerve at the elbow found only 1 out of 7 patients with ulnar entrapment who presented with temperature asymmetry of the affected extremity [74]. All patients with TOS who performed provocation test during image recording showed at least in one thermogram an asymmetric temperature pattern.

31.5.2 Peripheral Nerve Paresis

Paresis is an impairment of the motor function of the nervous system. Loss of function of the sensory fibers may be associated with motor deficit, but sensory impairment is not included in the term paresis. Therefore, most of the temperature signs in paresis are related to impaired motor function.

31.5.2.1 Brachial Plexus Paresis

Injury of the brachial plexus is a severe consequence of traffic accidents and motor cyclers are most frequently affected. The loss of motor activity in the affected extremity results in paralysis, muscle atrophy, and decreased skin temperature. Nearly 0.5 to 0.9% of newborns acquire brachial plexus paresis during delivery [75]. Early recovery of the skin temperature in babies with plexus paresis precede the recovery of motor function as shown in a study from Japan [76].

31.5.2.2 Facial Nerve

The seventh cranial nerve supplies the mimic muscles of the face and an acquired deficit is often named Bell's palsy. This paresis has normally a good prognosis for full recovery. Thermal imaging was used as outcome measure in acupuncture trials for facial paresis [77,78]. Ammer et al. [79] found slight asymmetries in patients with facial paresis, in which hyperthermia of the affected side occurred more frequently than hypothermia. However, patients with apparent herpes zoster causing facial palsy presented with higher temperature differences to the contralateral side than patients with nonherpetic facial paresis [80].

31.5.2.3 Peroneal Nerve

The peroneal nerve may be affected by metabolic neuropathy in patients with metabolic disease or by compression neuropathy due to intensive pressure applied at the site of fibula head. This can result in "foot drop," an impairment in which the patient cannot raise his forefoot. The thermal image is characterized by decreased temperatures on the anterior lower leg, which might become more visible after the patient has performed some exercises [81].

31.6 Complex Regional Pain Syndrome

A temperature difference between the affected and the nonaffected limb equal or greater than 1°C is one of the diagnostic criteria of the complex regional pain syndrome (CRPS) [82]. Ammer conducted a study in patients after radius fracture treated conservatively with a plaster cast [83]. Within 2 h after plaster removal and 1 week later thermal images were recorded. After the second thermogram an x-ray image of both hands was taken. The mean temperature difference between the affected and unaffected hand was 0.6 after plaster removal and 0.63 one week later. In 21 out of 41 radiographs slight bone changes suspected of algodystropy have been found. Figure 31.4 summarizes the results with respect to the outcome of x-ray images. Figure 31.5 shows the time course of an individual patient.

It was also shown, that the temperature difference decrease during successful therapeutic intervention and temperature effect was paralleled by reduction of pain and swelling and resolution of radiologic changes [84].

Disturbance of vascular adaptation mechanism and delayed response to temperature stimuli was obtained in patients suffering from CRPS [85,86]. These alterations have been interpreted as being caused by abnormalities of the autonomic nerve system. It was suggested to use a cold challenge on the contralateral side of the injured limb for prediction and early diagnosis of CRPS. Gulevich et al. [87] confirmed the high diagnostic sensitivity and specificity of cold challenge for the CRPS. Wasner et al. [88] achieved similar results by whole body cooling or whole body warming. Most recently a Dutch study found that the asymmetry factor, which was based on histograms of temperatures from the affected and nonaffected

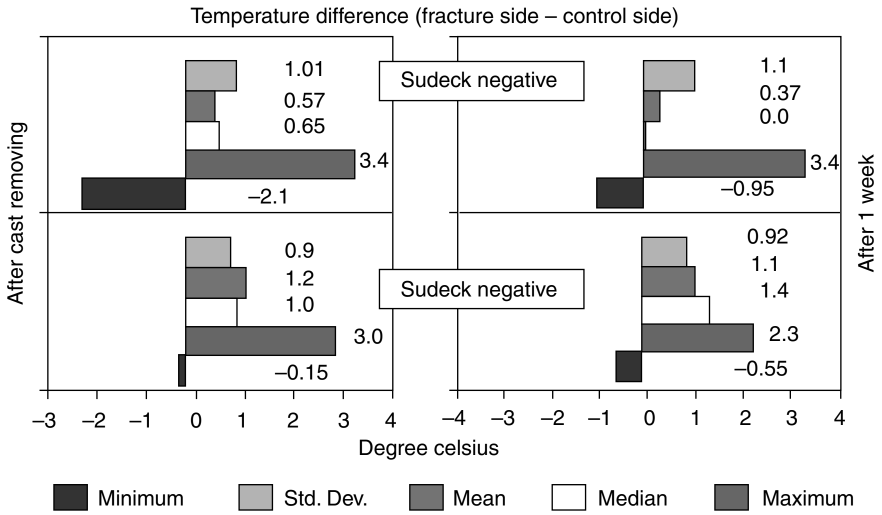


FIGURE 31.4 Diagram of temperatures obtained in patients with positive or negative x-ray images. (From Ammer, K. *Thermol. Österr.*, 1, 4, 1991. With permission.)

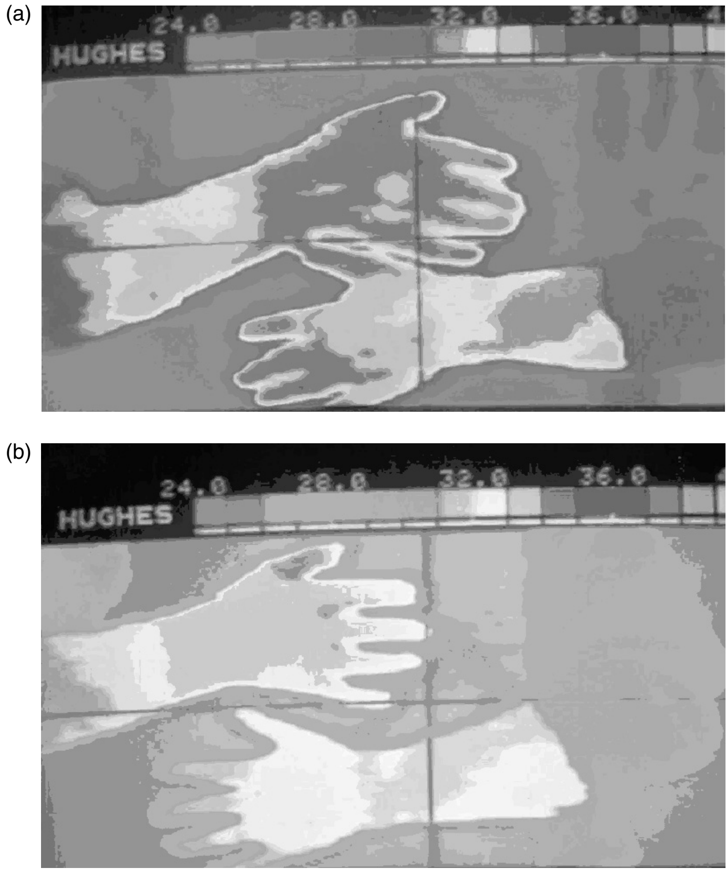


FIGURE 31.5 (See color insert.) Early CRPS after radius fracture. (a) 2 h after cast removal; (b) 1 week later.

hand had the highest diagnostic power for CRPS, while the difference of mean temperatures did not discriminate between disease and health [89].

31.7 Thermal Imaging Technique

The parameters for a reliable technique have been described in the past. Engel et al. [90] is a report published in 1978 by a European working group on Thermography in Locomotor Diseases. This paper discusses aspects of standardization including the need for adequate temperature control of the examination room and the importance of standard views used for image capture. More recently Ring and Ammer [91] described an outline of necessary considerations for good thermal imaging technique in clinical medicine. This outline has been subsequently expanded to encompass a revised set of standard views, and associated regions of interest for analysis. The latter is especially important for standardization, since the normal approach used is to select a region of interest subjectively. This means that without a defined set of reference points it is difficult for the investigator to reproduce the same region of interest on subsequent occasions. It is also even more difficult for another investigator to achieve the same, leading to unacceptable variables in the derived data. The aspects for standardization of the technique and the standard views and regions of interest recently defined are the product of a multicentered Anglo-Polish study who are pursuing the concept of a database of normal reference thermograms. The protocol can be found on a British University Research Group's website from University of Glamorgan [10].

31.7.1 Room Temperature

Room temperature is an important issue when investigating this group of diseases. Inflammatory conditions such as arthritis, are better revealed in a room temperature of 20°C, for the extremities, and may need to be at 18°C for examining the trunk. This presumes that the relative humidity will not exceed 45%, and a very low airspeed is required. At no time during preparatory cooling or during the examination should the patient be placed in a position where they can feel a draught from moving air. However in other clinical conditions where an effect from neuromuscular changes is being examined, a higher room temperature is needed to avoid forced vasoconstriction. This is usually performed at 22 to 24°C ambient. At higher temperatures, the subject may begin to sweat, and below 17°C shivering may be induced. Both of these thermoregulatory responses by the human body are undesirable for routine thermal imaging.

31.7.2 Clinical Examination

In this group of diseases, it can be particularly important that the patient receives a clinical examination in association with thermal imaging. Observations on medication, range of movement, experience of pain related to movement, or positioning may have a significant effect on the interpretation of the thermal images. Documentation of all such clinical findings should be kept on record with the images for future reference.

References

[1] Horvath, S.M. and Hollander, J.L. Intra-articular temperature as a measure of joint reaction. *J. Clin. Invest.*, 13, 615, 1949.

[2] Collins, A.J. and Cosh, J.A. Temperature and biochemical studies of joint inflammation. *Ann. Rheum. Dis.*, 29, 386, 1970.

[3] Ring, E.F.J. and Collins, A.J. Quantitative thermography. *Rheumatol. Phys. Med.*, 10, 337, 1970.

[4] Esselinckx, W. et al. Thermographic assessment of three intra-articular prednisolone analogues given in rheumatoid arthritis. *Br. J. Clin. Pharm.*, 5, 447, 1978.

[5] Bird, H.A., Ring, E.F.J., and Bacon, P.A. A thermographic and clinical comparison of three intra-articular steroid preparations in rheumatoid arthritis. *Ann. Rheum. Dis.*, 38, 36, 1979.

[6] Collins, A.J. and Ring, E.F.J. Measurement of inflammation in man and animals. *Br. J. Pharm.*, 44, 145, 1972.

[7] Bacon, P.A., Ring, E.F.J., and Collins, A.J. Thermography in the assessment of anti rheumatic agents, in *Rheumatoid Arthritis.* Gordon, J.L. and Hazleman, B.L., Eds., Elsevier/North Holland Biomedical Press, Amsterdam, 1977, p. 105.

[8] Collins, A.J. et al. Quantitation of thermography in arthritis using multi-isothermal analysis. I. The thermographic index. *Ann. Rheum. Dis.*, 33, 113, 1974.

[9] Ammer, K., Engelbert, B., and Kern, E. The determination of normal temperature values of finger joints. *Thermol. Int.*, 12, 23, 2002.

[10] Website address, Standard protocol for image capture and analysis, www.medimaging.org

[11] Czerny, V. Eine fokale Malazie des Unterschenkels. *Wien. Med. Wochenschr.*, 23, 895, 1873.

[12] Paget, J. On a form of chronic inflammation of bones. *Med. Chir. Transact.*, 60, 37, 1877.

[13] Ring, E.F.J. and Davies, J. Thermal monitoring of Paget's disease of bone. *Thermology*, 3, 167, 1990.

[14] Tauchmannova, H., Gabrhel, J., and Cibak, M. Thermographic findings in different sports, their value in the prevention of soft tissue injuries. *Thermol. Österr.* 3, 91–95, 1993.

[15] Smith, B.L, Bandler, M.K, and Goodman, P.H. Dominant forearm hyperthermia, a study of fifteen athletes. *Thermology*, 2, 25–28, 1986.

[16] Fischer, A.A. and Chang, C.H. Temperature and pressure threshold measurements in trigger points. *Thermology*, 1, 212, 1986.

[17] Menachem, A., Kaplan, O., and Dekel, S. Levator scapulae syndrome: an anatomic–clinical study. *Bull. Hosp. Jt. Dis.*, 53, 21, 1993.

[18] Schmitt, M. and Guillot, Y. Thermography and muscle injuries in sports medicine, in *Recent Advances in; Medical Thermography*, Ring, E.F.J. and Philips, J., Eds., Plenum Press, London, 1984, p. 439.

[19] Ammer, K. Low muscular activity of the lower leg in patients with a painful ankle. *Thermol. Österr.*, 5, 103, 1995.

[20] Kanie, R. Thermographic evaluation of osteoarthritis of the hip. *Biomed. Thermol.*, 15, 72, 1995.

[21] Vecchio, P.C. et al. Thermography of frozen shoulder and rotator cuff tendinitis. *Clin. Rheumatol.*, 11, 382, 1992.

[22] Ammer, K. et al. Thermography of the painful shoulder. *Eur. J. Thermol.*, 8, 93, 1998.

[23] Hobbins, W.B. and Ammer, K. Controversy: why is a paretic limb cold, high activity of the sympathetic nerve system or weakness of the muscles? *Thermol. Österr.*, 6, 42, 1996.

[24] Ring, E.F.J. and Ammer, K. Thermal imaging in sports medicine. *Sports Med. Today*, 1, 108, 1998.

[25] Gabrhel, J. and Tauchmannova, H. Wärmebilder der Kniegelenke bei jugendlichen Sportlern. *Thermol. Österr.*, 5, 92, 1995.

[26] Devereaux, M.D. et al. The diagnosis of stress fractures in athletes. *JAMA*, 252, 531, 1984.

[27] Graber, J. Tendosynovitis detection in the hand. *Verh. Dtsch. Ges. Rheumatol.*, 6, 57, 1980.

[28] Mayr, H. Thermografische Befunde bei Schmerzen am Ellbogen. *Thermol. Österr.*, 7, 5–10, 1997.

[29] Binder, A.I. et al. Thermography of tennis elbow, in *Recent Advances in Medical Thermography.* Ring, E.F.J. and Philips, J., Eds., Plenum Press, London, 1984, p. 513.

[30] Thomas, D. and Savage, J.P. Persistent tennis elbow: evaluation by infrared thermography and nuclear medicine isotope scanning. *Thermology*, 3, 132; 1989.

[31] Ammer, K. Thermal evaluation of tennis elbow, in *The Thermal Image in Medicine and Biology.* Ammer, K. and Ring, E.F.J., Eds., Uhlen Verlag, Wien, 1995, p. 214.

[32] Devereaux, M.D., Hazleman, B.L., and Thomas, P.P. Chronic lateral humeral epicondylitis — a double-blind controlled assessment of pulsed electromagnetic field therapy. *Clin. Exp. Rheumatol.* 3, 333, 1985.

[33] Ammer, K. et al. Thermographische und algometrische Kontrolle der physikalischen Therapie bei Patienten mit Epicondylopathia humeri radialis. *ThermoMed*, 11, 55–67, 1995.

[34] Ammer, K., Schartelmüller, T., and Melnizky, P. Thermography in fibromyalgia. *Biomed. Thermol.* 15, 77, 1995.

[35] Ammer, K. Only lateral, but not medial epicondylitis can be detected by thermography. *Thermol. Österr.*, 6, 105, 1996.

[36] Hirano, T. et al. Clinical study of shoulder surface temperature in patients with periarthritis scapulohumeralis (abstract). *Biomed. Thermol.*, 11, 303, 1991.

[37] Jeracitano, D. et al. Abnormal temperature control suggesting sympathetic dysfunction in the shoulder skin of patients with frozen shoulder. *Br. J. Rheumatol.*, 31, 539, 1992.

[38] Biasi, G. et al. The role computerized telethermography in the diagnosis of fibromyalgia syndrome. *Minerva Medica*, 85, 451, 1994.

[39] Ammer, K. Thermographic diagnosis of fibromyalgia. *Ann Rheum Dis. XIV European League Against Rheumatism Congress, Abstracts*, 135, 1999.

[40] Ammer, K., Engelbert, B., and Kern, E. Reproducibility of the hot spot count in patients with fibromyalgia, an intra- and inter-observer comparison. *Thermol. Int.*, 11, 143, 2001.

[41] Anbar, M. Recent technological developments in thermology and their impact on clinical applications. *Biomed. Thermol.*, 10, 270, 1990.

[42] Uematsu, S. Thermographic imaging of cutaneous sensory segment in patients with peripheral nerve injury. *J. Neurosurg.*, 62, 716–720, 1985.

[43] Hoffman, R.M., Kent, D.L., and. Deyo, R.A. Diagnostic accuracy and clinical utility of thermography for lumbar radiculopathy. A meta-analysis. *Spine*, 16, 623, 1991.

[44] McCulloch, J. et al. Thermography as a diagnostic aid in sciatica. *J. Spinal Disord.*, 6, 427, 1993.

[45] Takahashi, Y., Takahashi, K., and Moriya, H. Thermal deficit in lumbar radiculopathy. *Spine*, 19, 2443, 1994.

[46] Kim, Y.S. and Cho, Y.E. Pre- and postoperative thermographic imaging of lumbar disk herniations. *Biomed. Thermol.*, 13, 265, 1993.

[47] Zhang, H.Y., Kim, Y.S., and Cho, Y.E. Thermatomal changes in cervical disc herniations. *Yonsei Med. J.*, 40, 401, 1999.

[48] Cuetter, A.C. and Bartoszek, D.M. The thoracic outlet syndrome: controversies, overdiagnosism overtreatment and recommendations for management. *Muscle Nerve*, 12, 419, 1989.

[49] Schartelmüller, T. and Ammer, K. Thoracic outlet syndrome, in *The Thermal Image in Medicine and Biology.* Ammer, K. and Ring, E.F.J., Eds., Uhlen Verlag, Wien, 1995, p. 201.

[50] Schartelmüller, T. and Ammer, K. Infrared thermography for the diagnosis of thoracic outlet syndrome. *Thermol. Österr.*, 6, 130, 1996.

[51] Melnizky, P, Schartelmüller, T., and Ammer, K. Prüfung der intra-und interindividuellen Verläßlichkeit der Auswertung von Infrarot-Thermogrammen. *Eur. J. Thermol.*, 7, 224, 1997.

[52] Ammer, K. Thermographie der Finger nach mechanischem Provokationstest. *ThermoMed*, 17/18, 9, 2003.

[53] Schartelmüller, T., Melnizky, P., and Engelbert, B. Infrarotthermographie zur Evaluierung des Erfolges physikalischer Therapie bei Patenten mit klinischem Verdacht auf Thoracic Outlet Syndrome. *Thermol. Int.*, 9, 20, 1999.

[54] Schartelmüller, T. and Ammer, K. Zervikaler Diskusprolaps, Thoracic Outlet Syndrom oder periphere arterielle Verschlußkrankheit-ein Fallbericht. *Eur. J. Thermol.*, 7, 146, 1997.

[55] Ammer, K. Diagnosis of nerve entrapment syndromes by thermal imaging, in *Proceedings of The First Joint BMES/EMBS Conference. Serving Humanity, Advancing Technology*, October 13–16, 1999, Atlanta, GA, USA, p. 1117.

[56] Atroshi, I. et al. Prevalence of carpal tunnel syndrome in a general population. *JAMA*, 282, 153, 1999.

[57] Ammer, K., Mayr, H., and Thür, H. Self-administered diagram for diagnosing carpal tunnel syndrome. *Eur. J. Phys. Med. Rehab.*, 3, 43, 1993.

[58] Melnizky, P., Ammer, K., and Schartelmüller, T. Intra- und interindividuelle Verläßlichkeit der elektroneurographischen Untersuchung des Nervus medianus. *Österr. Z. Phys. Med. Rehab.*, 7, S83, 1996.

[59] Schartelmüller, T., Ammer, K., and Melnizky, P. Natürliche und postoperative Entwicklung elektroneurographischer Untersuchungsergebnisse des N. medianus von Patienten mit Carpaltunnelsyndrom (CTS). *Österr. Z. Phys. Med.*, 7, 183, 1997.

[60] Rosen, H.R. et al. Is surgical division of the carpal ligament sufficient in the treatment of carpal tunnel syndrome? *Chirurg*, 61, 130, 1990.

[61] Herrick, R.T. et al. Thermography as a diagnostic tool for carpal tunnel syndrome, in *Medical Thermology*, Abernathy, M. and Uematsu, S., Eds., American Academy of Thermology, 1986, p. 124.

[62] Herrick, R.T. and Herrick, S.K., Thermography in the detection of carpal tunnel syndrome and other compressive neuropathies. *J. Hand Surg.*, 12A, 943–949, 1987.

[63] Gateless, D., Gilroy, J., and Nefey, P. Thermographic evaluation of carpal tunnel syndrome during pregnancy. *Thermology*, 3, 21, 1988.

[64] Meyers, S. et al. Liquid crystal thermography, quantitative studies of abnormalities in carpal tunnel syndrome. *Neurology*, 39, 1465, 1989.

[65] So, Y.T., Olney, R.K., and Aminoff, M.J. Evaluation of thermography in the diagnosis of selected entrapment neuropathies. *Neurology*, 39, 1, 1989.

[66] Tchou, S. and Costich, J.F. Thermographic study of acute unilateral carpal tunnel syndromes. *Thermology*, 3, 249–252, 1991.

[67] Ammer, K. Thermographische Diagnose von peripheren Nervenkompressionssyndromen. *ThermoMed*, 7, 15, 1991.

[68] Hobbins, W.B. Autonomic vasomotor skin changes in pain states: significant or insignificant? *Thermol. Österr.*, 5, 5, 1995.

[69] Ammer, K. et al. The thermal image of patients suffering from carpal tunnel syndrome with a distal latency higher than 6.0 msec. *Thermol. Int.*, 9, 15, 1999.

[70] Ammer, K. and Melnizky, P. Determination of regions of interest on thermal images of the hands of patients suffering from carpal tunnel syndrome. *Thermol. Int.*, 9, 56, 1999.

[71] Ammer, K. Thermographic diagnosis of Raynaud's Phenomenon. *Skin Res. Technol.*, 2, 182, 1996.

[72] Neundörfer, B., Dietrich, B., and Braun, B. Raynaud–Phänomen beim Carpaltunnelsyndrom. *Wien. Klin. Wochenschr.*, 89, 131–133, 1977.

[73] Grassi, W. et al. Clinical diagnosis found in patients with Raynaud's phenomenon: a multicentre study. *Rheumatol. Int.*, 18, 17, 1998.

[74] Mayr, H. and Ammer, K. Thermographische Diagnose von Nervenkompressionssyndromen der oberen Extremität mit Ausnahme des Karpaltunnelsyndroms (abstract). *Thermol. Österr.*, 4, 82, 1994.

[75] Mumenthaler, M. and Schliack, H. *Läsionen periphere Nerven.* Georg Thieme Verlag, Stuttgart-New York, Auflage, 1982, p. 4.

[76] Ikegawa, S. et al. Use of thermography in the diagnosis of obstetric palsy (abstract). *Thermol. Österr.*, 7, 31, 1997.

[77] Zhang, D. et al. Preliminary observation of imaging of facial temperature along meridians. *Chen Tzu Yen Chiu*, 17, 71, 1992.

[78] Zhang, D. et al. Clinical observations on acupuncture treatment of peripheral facial paralysis aided by infra-red thermography — a preliminary report. *J. Tradit. Chin. Med.*, 11, 139, 1991.

[79] Ammer, K., Melnizky, P. and Schartelmüller, T. Thermographie bei Fazialisparese. *ThermoMed*, 13, 6–11, 1997.

[80] Schartelmüller, T., Melnizky, P., and Ammer, K. Gesichtsthermographie, Vergleich von Patienten mit Fazialisparese und akutem Herpes zoster ophthalmicus. *Eur. J. Thermol.*, 8, 65, 1998.

[81] Melnizky, P., Ammer, K., and Schartelmüller, T. Thermographische Überprüfung der Heilgymnastik bei Patienten mit Peroneusparese. *Thermol. Österr.*, 5, 97, 1995.

[82] Wilson, P.R. et al. Diagnostic algorithm for complex regional pain syndromes, in *Reflex Sympathetic Dystrophy, A Re-appraisal.* Jänig, W. and Stanton-Hicks, M., Eds., Seattle, IASP Press, 1996, p. 93.

[83] Ammer, K. Thermographie nach gipsfixierter Radiusfraktur. *Thermol. Österr.*, 1, 4, 1991.

[84] Ammer, K. Thermographische Therapieüberwachung bei M.Sudeck. *ThermoMed*, 7, 112–115, 1991.

[85] Cooke, E.D. et al. Reflex sympathetic dystrophy (algoneurodystrophy): temperature studies in the upper limb. *Br. J. Rheumatol.*, 8, 399, 1989.

[86] Herrick, A. et al. Abnormal thermoregulatory responses in patients with reflex sympathetic dystrophy syndrome. *J. Rheumatol.*, 21, 1319, 1994.

[87] Gulevich, S.J. et al. Stress infrared telethermography is useful in the diagnosis of complex regional pain syndrome, type I (formerly reflex sympathetic dystrophy). *Clin. J. Pain*, 13, 50, 1997.

[88] Wasner, G., Schattschneider, J., and Baron, R. Skin temperature side differences — a diagnostic tool for CRPS? *Pain*, 98, 19, 2002.

[89] Huygen, F.J.P.M. et al. Computer-assisted skin videothermography is a highly sensitive quality tool in the diagnosis and monitoring of complex regional pain syndrome type I. *Eur. J. Appl. Physiol.*, 91, 516, 2004.

[90] Engel, J.M. et al. Thermography in locomotor diseases, recommended procedure. Anglo-dutch thermographic society group report. *Eur. J. Rheumatol. Inflam.*, 2, 299–306, 1979.

[91] Ring, E.F.J. and Ammer, K. The technique of infra red imaging in medicine. *Thermol. Int.*, 10, 7, 2000.

32

Functional Infrared Imaging in Clinical Applications

Arcangelo Merla
Gian Luca Romani
University "G. d'Annunzio"

32.1 Introduction

Infrared imaging allows the representation of the surface thermal distribution of the human body. Several studies have been performed so far to assess the contribution that such information may provide to the clinicians. The skin temperature distribution of the human body depends on the complex relationships defining the heat exchange processes between skin tissue, inner tissue, local vasculature, and metabolic activity. All of these processes are mediated and regulated by the sympathetic and parasympathetic activity to maintain the thermal homeostasis. The presence of a disease can locally affect the heat balance or exchange processes resulting in an increase or in a decrease of the skin temperature. Such a temperature change can be better estimated with respect to the surrounding regions or the unaffected contra lateral region. But then, the disease should also effect the local control of the skin temperature. Therefore, the characteristic parameters modeling the activity of the skin thermoregulatory system can be used as diagnostic parameters. The functional infrared (fIR) Imaging — also named infrared functional imaging (fIR imaging) — is the study for diagnostic purposes, based on the modeling of the bio-heat exchange processes, of the functional properties and alterations of the human thermoregulatory system. In this chapter, we will review some of the most important recent clinical applications of the functional infrared imaging of our group.

32.2 Quantifying the Relevance and Stage of Disease with the τ Image Technique

Infrared imaging can provide diagnostic information according different possible approaches. The approach generally followed consists of the detection of significant differences between the skin thermal distributions of the two hemisoma or in the pattern recognition of specific features with respect to average healthy population [1]. The underlying hypothesis is that the skin temperature distribution, at a given time, is considered at a steady state. Of course this is a rough approximation of the reality because of the homeostasis. More valuable and quantitative information can be obtained from the study of the skin temperature dynamics in the unsteady state, where the processes involved and controlled by the thermoregulatory system can be modeled and described through their characteristic parameters [2–7]. The presence of diseases interfering with the skin thermoregulatory system can be then inferred by the analysis of its functional alterations [8–18]. To enhance the functional content of the thermoregulatory response, one needs to pass through modeling of the thermal properties and dynamics of the skin thermoregulatory system. Such a modeling can bring more quantitative and detailed diagnostic parameters with respect to the particular disease being analyzed. Merla et al. [7,17,19,20] proposed a new imaging technique, based on this approach, for the clinical study of a variety of diseases. The theory behind the technique is based on the fact that the human thermoregulatory system maintains a reasonably constant body temperature against a wide range of environmental conditions. The body uses several physiological processes to control the heat exchange with the environment. The mechanism controlling thermal emission and dermal microcirculation is driven by the sympathetic nervous system. A disease locally affecting the thermoregulatory system (i.e., traumas, lesions, vein thrombosis, varicocele, dermatitis, Raynaud's phenomenon, and scleroderma, etc.) may produce an altered sympathetic function and a change in the local metabolic rate. Local vasculature and microvasculature may be rearranged resulting in a modification of the skin temperature distribution.

Starting from a general energy balance equation, it is straightforward to demonstrate that the recovery time from any kind of thermal stress for a given region of interest depends on the region thermal parameters. A given disease may alter the normal heat capacity and the tissue/blood ratio mass density of a region. An example is given in Figure 32.1 that shows the different thermoregulatory behaviors exhibited by two

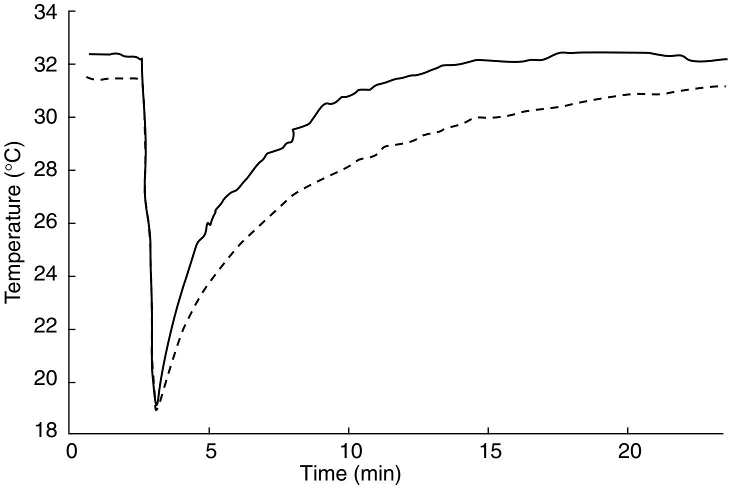


FIGURE 32.1 Muscular lesion on the left thigh abductor with hemorrhage shedding: thermal recovery curves following a cold thermal stress. The dotted line represents the recovery of a healthy area close to the damaged one. The continuous line represents the curve related to a muscular lesion region. Both recoveries exhibit exponential feature; the injured area exhibits a faster rewarming with a shorter time constant.

adjacent regions — one healthy and one affected by a muscular lesion — after local cooling applied to the skin. A controlled thermal stress applied to the region of interest and the surrounding tissue permits to study and to model the response of the region itself. The most important terms involved in the energy balance during the recovery are the heat storage in the tissue, heat clearance by blood perfusion, and convective heat exchange with the environment, as described by the following equation:

$$\frac{\partial T}{\partial t}\rho \cdot c \cdot V = hA\,(T_{\text{o}} - T) + \rho_{\text{bl}} \cdot c_{\text{bl}} \cdot w_{\text{bl}}(t) \cdot (T_{\text{bl}} - T) \tag{32.1}$$

where subscripts o and bl designate the properties of the environment and blood, respectively, while ρ is the density, c is the specific heat, V is the volume, T is the temperature, t is the time, h is the combined heat transfer coefficient between the skin and the environment, A is the surface area, and w is the blood perfusion rate.

The initial condition for (32.1) is

$$T = T_{\text{i}} \qquad \text{for } t = 0 \tag{32.2}$$

where T_{i} is the skin temperature and $t = 0$ is the time at the recovery starting.

Equation 32.1 can be easily integrated under the assumption of constant blood perfusion rate w_{bl} and blood temperature T_{bl}, yielding:

$$T(t) = \frac{W \cdot (T_{\text{bl}} - T_{\text{o}})}{W + H} + \left(T_{\text{i}} - T_{\text{o}} - \frac{W \cdot (T_{\text{bl}} - T_{\text{o}})}{W + H}\right) \cdot \text{e}^{-(W+H)\cdot t} + T_{\text{o}} \tag{32.3}$$

where

$$H = \frac{h \cdot A}{\rho \cdot c \cdot V} \qquad W = \frac{\rho_{\text{bl}} \cdot c_{\text{bl}} \cdot w_{\text{bl}}}{\rho \cdot c \cdot V} \tag{32.4}$$

The time t_{f} to reach a certain preset (final) temperature T_{f} is then given by:

$$t_{\text{f}} = -\frac{1}{W + H} \ln\left(\frac{(1 + (H/W)) \cdot (T_{\text{f}} - T_{\text{o}}) - W(T_{\text{bl}} - T_{\text{o}})}{(1 + (H/W)) \cdot (T_{\text{i}} - T_{\text{o}}) - W(T_{\text{bl}} - T_{\text{o}})}\right) \tag{32.5}$$

Equation 32.5, with the assumption of constant blood perfusion, relates the time to reach a preset temperature to the local thermal properties and to local blood perfusion.

The exponential solution described in (32.3) suggests to use the time constant τ as a characterizing parameter for the description of the recovery process after any kind of controlled thermal stress, with τ mainly determined by the local blood flow and thermal capacity of the tissue.

The fIR imaging permits an easy evaluation of τ, which can be regarded as a parameter able to discriminate areas interested by the specific disease from healthy ones.

Rather than a static imaging of the skin thermal distribution to pictorially describe the effect of the given disease, an image reporting the τ recovery time pixel to pixel can be used to characterize that disease [7,17,19,20]. Areas featuring an associated blood shedding, or an inflammatory state,or an increased blood reflux, often exhibit a faster recovery time with respect to the surroundings. Those areas then exhibit a smaller τ value. In contrast, in presence of localized calcifications, early ulcers or scleroderma, and impaired microvascular control, the involved areas show a slower recovery than the healthy surrounding areas and are therefore characterized by a longer τ time.

The reliability and value of the τ image technique rely on the good quality of the data and on their appropriate processing. While the interested reader can find a detailed description for proper materials and method for the τ image technique in Reference 17, it is worthwhile to report hereby the general algorithm for the method:

1. Subject marking (to permit movement correction of the thermal image series) and acclimation to the measurement room kept at controlled environmental conditions

2. Adequate calibration of the thermal imaging device
3. Recording of the baseline temperature dynamics for the region of interest
4. Execution of the thermal stress (usually performed through a cold or warm dry patch at controlled temperature and temperature exchange rate)
5. Recording of the thermal recovery until the complete restoration of the baseline features
6. Postprocessing movement correction of the thermal image series
7. Fitting of the pixel by pixel experimental recovery data to an exponential curve and extraction of the time constant τ for each pixel of the region of interest
8. Pixel by pixel color coding and mapping of the time constant τ values

The τ image technique has been first proposed as complementary diagnostic tool for the diagnosis of muscular lesions, Raynaud's phenomenon, and deep vein thrombosis [7,17,19,20]. In those studies, the technique correctly depicted the stages of the diseases accordingly with the gold standard evaluation techniques. A mild cold stress has been used as a thermal stress. For the muscular lesions, according to the importance of the lesion, the lower values (2–4 min) of the recovery time τ were found in agreement with the location and the severity of the trauma (Figure 32.2). The dimensions of the lesions as estimated by ultrasonography were proportionally related to those of their tracks on the τ image. In the diagnosis of Raynaud's phenomenon secondary to scleroderma greater values (18–20 min) of the recovery time τ corresponded to finger districts more affected by the disease (Figure 32.3). Clinical investigation and capillaroscopy confirmed the presence of scleroderma and the microvascular damage.

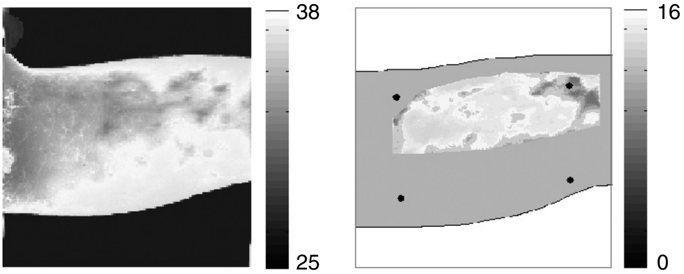


FIGURE 32.2 Second-class muscular lesion on the left leg abductor — Medial view. Left: Static thermography image. The bar shows the pixel temperature. The light gray spots indicate the presence of the trauma. Right: Time constant τ image after mild cold stress. The bar illustrates the recovery time, in minutes, for each pixel. The black spots are the markers used as position references. (From Merla et al., *IEEE Eng. Med. Biol. Magn.*, 21, 86, 2002. With permission.)

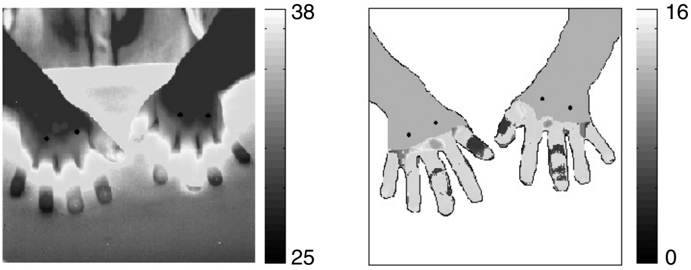


FIGURE 32.3 Raynaud's Phenomenon Secondary to Scleroderma. Left: Static thermography image. The bar shows the pixel temperature. Right: Time constant τ image after mild cold stress. The bar illustrates the recovery time, in minutes, for each pixel. The regions associated with longer recovery times identify the main damaged finger regions. (From Merla et al., *IEEE Eng. Med. Biol. Magn.*, 21, 86, 2002. With permission.)

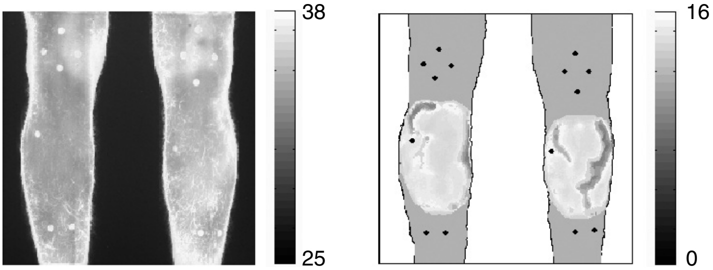


FIGURE 32.4 Bi-lateral vein thrombosis. Left: Static thermography image. The bar shows the pixel temperature. Right: Time constant τ image after mild cold stress. The bar illustrates the recovery time, in minutes, for each pixel. The areas associated with shorter recovery times identify the regions interested by the thrombosis. (From Merla et al., *IEEE Eng. Med. Biol. Magn.*, 21, 86, 2002. With permission.)

In the reported deep vein thrombosis cases, the authors found the lower values (1–3 min) of the recovery time τ in agreement with the location and the severity of the blood flow reflux according to the Echo Color Doppler findings (Figure 32.4).

The τ image technique provides useful diagnostic information and can be applied also as a follow-up tool. It is an easy and not invasive diagnostic procedure that can be successfully used in the diagnosis and monitoring of several diseases affecting the local thermoregulatory properties, both in a direct and an indirect way. The τ image technique opens new possibilities for the applications of IR imaging in the clinical field. It is worth noting that a certain amount of information is already present — but embedded — in the traditional static image (see Figure 32.2), but the interpretation is difficult and relies on the ability of the clinicians. The method is based on the assumptions of a time constant blood perfusion and blood temperature. While such assumptions are not completely correct from the physiological point of view, the experimental exponential-shaped recovery function allows such a simplification. With respect to some diseases, such as the Raynaud's phenomenon, the τ image technique may provide useful information to image the damage and quantitatively follow its time evolution.

32.3 Raynaud's Phenomenon

Raynaud's phenomenon (RP) is defined as a painful vasoconstriction — that may follow cold or emotional stress — of small arteries and arterioles of extremities, like fingers and toes. RP can be primary (PRP) or secondary Systemic Sclerosis (SSc) to scleroderma. The latter is usually associated with a connective tissues disease. RP precedes the systemic autoimmune disorders development, particularly scleroderma, by many years and it can evolve into secondary RP. The evaluation of vascular disease is crucial in order to distinguish between PRP and SSc. In PRP, episodic ischemia in response to cold exposure or to emotional stimuli is usually completely reversible: absence of tissue damage is the typical feature [21], but also mild structural changes are demonstrated [22]. In contrast, scleroderma RP shows irreversible tissue damage and severe structural changes in the finger vascular organization [2]. None of the physiological measurement techniques currently in use, but infrared imaging, is completely satisfactory in focusing primary or secondary RP [3]. The main limit of such techniques (nail fold capillary microscopy, cutaneous laser-Doppler flowmetry, and plethysmography) is the fact that they can proceed just into a partial investigation, usually assessing only one finger once. The measurement of skin temperature is an indirect method to estimate change in skin thermal properties and blood flow. Thermography, protocols [3–5, 23–27] usually include cold patch testing to evaluate the capability of the patients hands to rewarm. The pattern of the rewarming curves is usually used to depict the underlying structural diseases. Analysis of rewarming curves has been used in several studies to differentiate healthy subjects from PRP or SSc Raynaud's patients. Parameters usually considered so far are: the lag time preceding the onset of rewarming

[3–5] or to reach a preset final temperature [26]; the rate of the rewarming and the maximum temperature of recovery [27]; and the degree of temperature variation between different areas of the hands [25].

Merla et al. [14,16] proposed to model the natural response of the fingertips to exposure to a cold environment to get a diagnostic parameter derived by the physiology of such a response. The thermal recovery following a cold stress is driven by thermal exchange with the environment, transport by the incoming blood flow, conduction from adjacent tissue layers, and metabolic processes. The finger temperature is determined by the net balance of the energy input/output. The more significant contribution come from the input power due to blood perfusion and the power lost to the environment [28]:

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = -\frac{\mathrm{d}Q_{\mathrm{env}}}{\mathrm{d}t} + \frac{\mathrm{d}Q_{\mathrm{ctr}}}{\mathrm{d}t} \tag{32.6}$$

Normal finger recovery after a cold stress is reported in Figure 32.5. In absence of thermoregulatory control, fingers exchange heat only with the environment: in this case, their temperature T_{exp} follows an exponential pattern with time constant τ given by:

$$\tau = \frac{\rho \cdot c \cdot V}{h \cdot A} \tag{32.7}$$

where ρ is the mass density, c the specific heat, V the finger volume, h is the combined heat transfer coefficient between the finger and the environment, and A is the finger surface area. Thanks to the thermoregulatory control, the finger maintains its temperature T greater than T_{exp}. For a Δt time, the area of the trapezoid *ABCF* times $h \cdot A$ in Figure 32.5 computes the heat provided by the thermoregulatory system, namely ΔQ_{ctrl}. This amount summed to ΔQ_{env} yields Q, the global amount of heat stored in the finger.

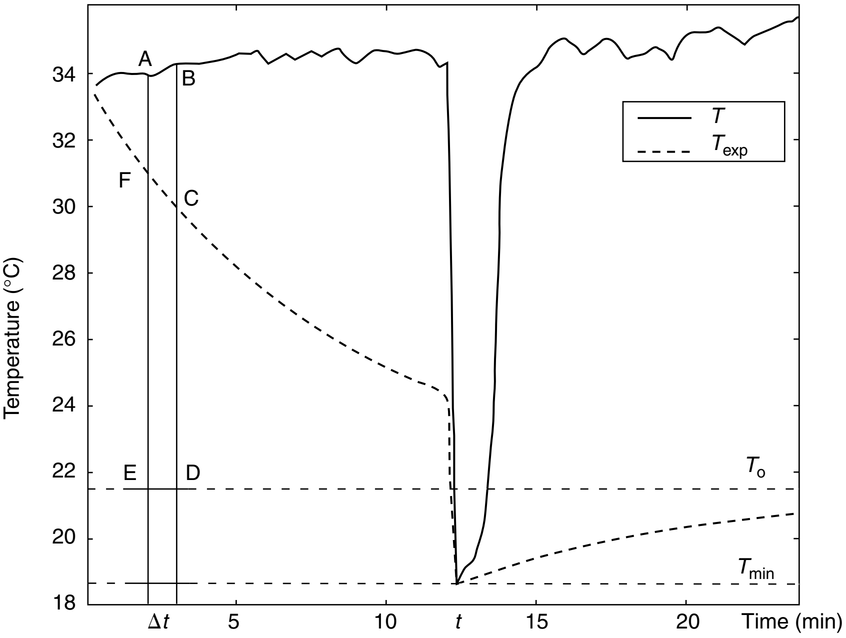


FIGURE 32.5 Experimental rewarming curves after cold stress in normal subjects. The continuous curve represents the recorded temperature finger. The outlined curve represents the exponential temperature pattern exhibited by the finger in absence of thermoregulatory control. In this case, the only heat source for the finger is the environment. (From Merla et al., *IEEE Eng. Med. Biol. Mag.*, 21, 73, 2002. With permission.)

Then, the area of the trapezoid *ABDE* is proportional to the amount *Q* of heat stored in the finger during a Δt interval. Therefore, *Q* can be computed integrating the area surrounded by the temperature curve *T* and the constant straight line T_0:

$$Q = -h \cdot A \cdot \int_{t_1}^{t_2} (T_0 - T(\varsigma))\, \mathrm{d}\varsigma \tag{32.8}$$

where the minus sign takes into account that the heat stored by the finger is counted as positive. *Q* is intrinsically related to the finger thermal capacity, according to the expression

$$\Delta Q = \rho \cdot c \cdot V \cdot \Delta T \tag{32.9}$$

Under the hypothesis of constant T_0, the numerical integration in (32.8) can be used to characterize the rewarming exhibited by a healthy or a suffering finger.

The *Q* parameter has been used in References 14 and 16 to discriminate and classify PRP, SSc, and healthy subjects on a set of 40 (20 PRP, 20 SSc), and 18 healthy volunteers, respectively. For each subject, the response to a mild cold challenge of hands in water was assessed by fIR imaging. Rewarming curves were recorded for each of the five fingers of both hands; the temperature integral *Q* was calculated along the 20 min following the cold stress. Ten subjects, randomly selected within the 18 normal ones, repeated two times and in different days the test to evaluate the repeatability of the fIR imaging findings. The repeatability test confirmed that fIR imaging and *Q* computation is a robust tool to characterize the thermal recovery of the fingers.

The grand average *Q* values provided by the first measurement was (1060.0 ± 130.5)°C min, while for the second assessment it was (1012 ± 135.1)°C min ($p > .05$, one-way ANOVA test). The grand average *Q* values for PRP, SSc, and healthy subjects' groups are shown in Figure 32.6, whereas single values obtained for each finger of all of the subjects are reported in Figure 32.7.

The results in References 14 and 16 highlight that the PRP group features low intra- and inter-individual variability whereas the SSc group displays a large variability between healthy and unhealthy fingers. *Q* values for SSc finger are generally greater than PRP ones.

The temperature integral at different finger regions yields very similar results for all fingers of the PRP group, suggesting common thermal and BF properties. SSc patients showed different thermoregulatory

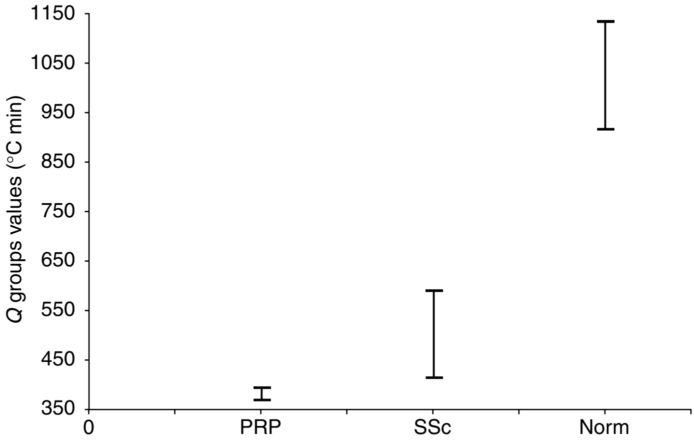


FIGURE 32.6 One-way ANOVA test applied to the *Q* parameter calculated for each group (PRP, SSc, and healthy). The *Q* parameter clearly discriminates among the three groups. (From Merla et al., *IEEE Eng. Med. Biol. Magn.*, 21, 73, 2002. With permission.)

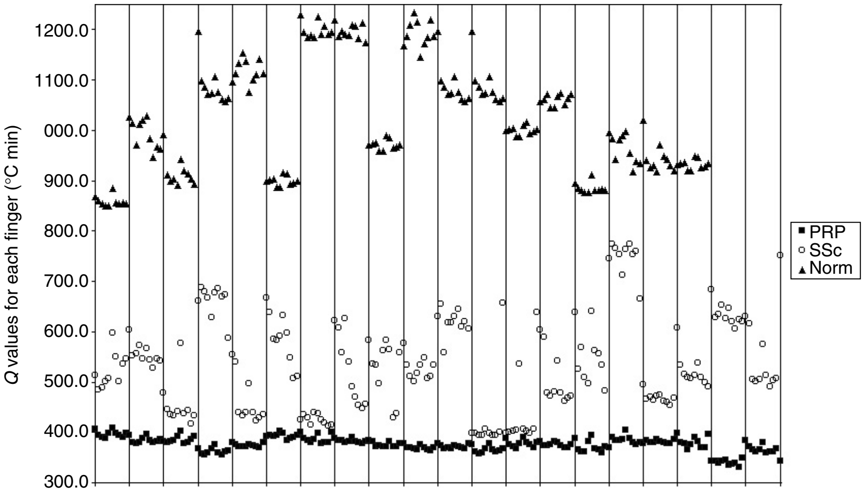


FIGURE 32.7 *Q* values calculated for each finger of each subjects. Vertical grid lines are placed to discriminate the ten fingers. PRP fingers are characterized by a strong intra- and inter-individual homogeneity. Greater mean *Q* values and greater intra- and inter-individual variations characterizes the SSc fingers. (From Merla et al., *IEEE Eng. Med. Biol. Magn.*, 21, 73, 2002. With permission.)

responses in the different segments of finger. This feature is probably due to the local modification in the tissue induced by the scleroderma. Scleroderma patients also featured a significantly different behavior across the five fingers depending on the disease involvement.

In normal and PRP groups all fingers show a homogeneous behavior and PRP fingers always exhibit a poorer recovery than normal ones. Additionally, in both groups, the rewarming always starts from the finger distal area differently from what happens in SSc patients. The sensitivity of the method in order to distinguish patients from normal is 100%. The specificity in distinguishing SSc from PRP is 95%.

The grand average *Q* clearly highlights the difference between PRP, SSc, and between normal subjects. It provides useful information about the abnormalities of their thermoregulatory finger properties. The PRP patients exhibited common features in terms of rewarming. Such behavior can be explained in terms of an equally low and constant BF in all fingers and to differences in the amount of heat exchanged with the environment [2].

Conversely, no common behavior was found for the SSc patients, since their disease determines — for each finger — very different thermal and blood perfusion properties. Scleroderma seems to increase the tissue thermal capacity with a reduced ability to exchange. As calculated from the rewarming curves, *Q* parameter seems to be particularly effective to describe the thermal recovery capabilities of the finger. The method clearly highlighted the difference between PRP and SSc patients and provides useful information about the abnormalities of their thermal and thermoregulatory finger properties.

In consideration of the generally accepted theory that the different recovery curves of the patients are a reflection of the slow deterioration of the microcirculation, so that over time in the same patients it is possible to observe changes in the thermal recovery curves, the method described earlier could be used to monitor the clinical evolution of the disease. In addition, pharmacological treatment effects could be advantageously followed up.

32.4 Diagnosis of Varicocele and Follow-Up of the Treatment

Varicocele is a widely spread male disease consisting of a dilatation of the pampiniform venous plexus and of the internal spermatic vein. Consequences of such a dilatation are an increase of the scrotal temperature and a possible impairment of the potential fertility [29,30]. In normal men, testicular temperature is

3 to 4°C lower than core body temperature [29]. Two thermoregulatory processes maintain this lower temperature: heat exchange with the environment through the scrotal skin and heat clearance by blood flow through the pampiniform plexus. Venous stasis due to the varicocele may increase the temperature of the affected testicle or pampiniform plexus. Thus, an abnormal temperature difference between the two hemiscrota may suggest the presence of varicocele [6,29–31] (see Figure 32.8). Telethermography can reveal abnormal temperature differences between the two testicles and altered testicular thermal recovery after an induced cold stress. Affected testicles return to prestress equilibrium temperatures faster than do normal testicles [6]. The fIR imaging has been used to determine whether altered scrotal thermoregulation is related to subclinical varicocele [15]. In a study conducted in 2001, Merla and Romani enrolled 60 volunteers, 18 to 27 years of age (average age, 21 ± 2 years), with no symptoms or clinical history of varicocele. After clinical examination, echo color Doppler imaging (the gold standard) and fIR imaging were performed. The fIR imaging evaluation consisted of obtaining scrotal images, measuring the basal temperature at the level of the pampiniform plexus (T_p) and the testicles (T_t), and determining thermal recovery of the scrotum after cold thermal stress. The temperature curve of the hemiscrotum during rewarming showed an exponential pattern and was, therefore, fitted to an exponential curve. The time constant τ of the best exponential fit depends on the thermal properties of the scrotum and its blood perfusion [15]. Therefore τ provides a quantitative parameter assessing how much the scrotal thermoregulation is affected by varicocele. Cooling was achieved by applying a dry patch to the scrotum that was 10°C colder than the basal scrotal temperature. The fIR measurements were performed according to usual standardization procedures [15]. The basal prestress temperature and the recovery time constant τ_p at the level of the pampiniform plexus and of the testicles (τ_t) were evaluated on each hemiscrotum. A basal testicular temperature greater than 32°C and basal pampiniform plexus temperature greater than 34°C were considered warning thresholds. Temperature differences among testicles (ΔT_t) or pampiniform plexus ΔT_p and temperature greater than 1.0°C were also considered warning values, as were τ_p and $\Delta\tau_t$ values longer than 1.5 min. The fIR imaging evaluation classified properly the stages of disease, as confirmed by the echo color Doppler imaging and clinical examination in a blinded manner. In 38 subjects, no warning basal temperatures or differences in rewarming temperatures were observed. These subjects were considered to be normal according to fIR imaging. Clinical examination and echo color Doppler imaging confirmed the absence of varicocele ($p < .01$, one-way ANOVA test).

In 22 subjects, one or more values were greater than the warning threshold for basal temperatures or differences in rewarming temperatures. Values for ΔT_p and the $\Delta\tau_p$ were higher than the warning thresholds in 8 of the 22 subjects, who were classified as having grade 1 varicocele. Five subjects had ΔT_t and $\Delta\tau_t$ values higher than the threshold. In 9 subjects, 3 or more infrared functional imaging values were greater than the warning threshold values. The fIR imaging classification was grade 3 varicocele. Clinical examination and echo color Doppler imaging closely confirmed the fIR imaging evaluation of the stage of the varicocele. The fIR imaging yielded no false-positive or false-negative results. All participants with positive results on fIR imaging also had positive results on clinical examination and echo color Doppler imaging. The sensitivity and specificity of fIR test were 100 and 93%, respectively. An abnormal change in the temperature of the testicles and pampiniform plexus may indicate varicocele, but the study demonstrated that impaired thermoregulation is associated with varicocele-induced alteration of blood flow. Time of recovery of prestress temperature in the testicles and pampiniform plexus appears to assist in classification of the disease. The fIR imaging accurately detected 22 nonsymptomatic varicocele.

The control of the scrotum temperature should improve after varicocelectomy as a complementary effect of the reduction of the blood reflux. Moreover, follow-up of the changes in scrotum thermoregulation after varicocelectomy may provide early indications on possible relapses of the disease.

To answer these questions, Merla et al. [9] used fIR imaging to study changes in the scrotum thermoregulation of 20 patients (average age, 27 ± 5 years) that were judged eligible for varicocelectomy on the basis of the combined results of the clinical examination, Echo color Doppler imaging, and spermiogram. No bilateral varicoceles were included in the study.

Patients underwent clinical examination, echo color Doppler imaging and instrument varicocele grading, and infrared functional evaluation before varicocelectomy and every 2 weeks thereafter, up to the

24th week. Out of 20, 14 patients suffered from grade 2 left varicocele. All of them were characterized by basal temperatures and recovery time after cold stress according to Reference 15. Varicoceles were surgically treated via interruption of the internal spermatic vein using modified Palomo's technique. The fIR imaging documented changes in the thermoregulatory control of the scrotum after the treatment as following: 13 out of the 14 grade 2 varicocele patients exhibited normal basal T_t, T_p on the varicocele side of the scrotum, and normal temperature differences ΔT_t and ΔT_p starting from the 4th week after varicocelectomy. Their $\Delta \tau_t$ and $\Delta \tau_p$ values returned to normal range from the 4th to the 6th week. Four out of the Six grade 3 varicocele patients exhibited normal basal T_t, T_p on the varicocele side of the scrotum, and normal temperature differences ΔT_t and ΔT_p starting from the 6th week after varicocelectomy. Their $\Delta \tau_t$ and $\Delta \tau_p$ values returned to normal range from the 6th to the 8th week. The other three patients did not return to normal values of the above-specified parameters. In particular, $\Delta \tau_t$ and $\Delta \tau_p$ remained much longer than the threshold warming values [6,15] up to the last control (Figure 32.8). Echo color Doppler imaging and clinical examination assessed relapses of the disease. The study proved that the surgical treatment of the varicocele induces modification in the thermoregulatory properties of the scrotum, reducing the basal temperature of the affected testicle and pampiniform plexus, and slowing down its recovery time after thermal stress. Among the 17 with no relapse, 4 exhibited return to normal T_t, T_p, ΔT_t, and ΔT_p for the latero-anterior side of the scrotum, while the posterior side of the scrotum remained hyperthermal or characterized by ΔT_t and ΔT_p higher than the threshold warning value. This fact suggested that the surgical treatment via interruption of the internal spermatic vein using Palomo's technique may not be the most suitable method for those varicoceles. The time requested by the scrotum to restore normal temperature distribution and control seems to be positively correlated to the volume and duration of the blood reflux lasting: the greater the blood reflux, the longer the time. The study

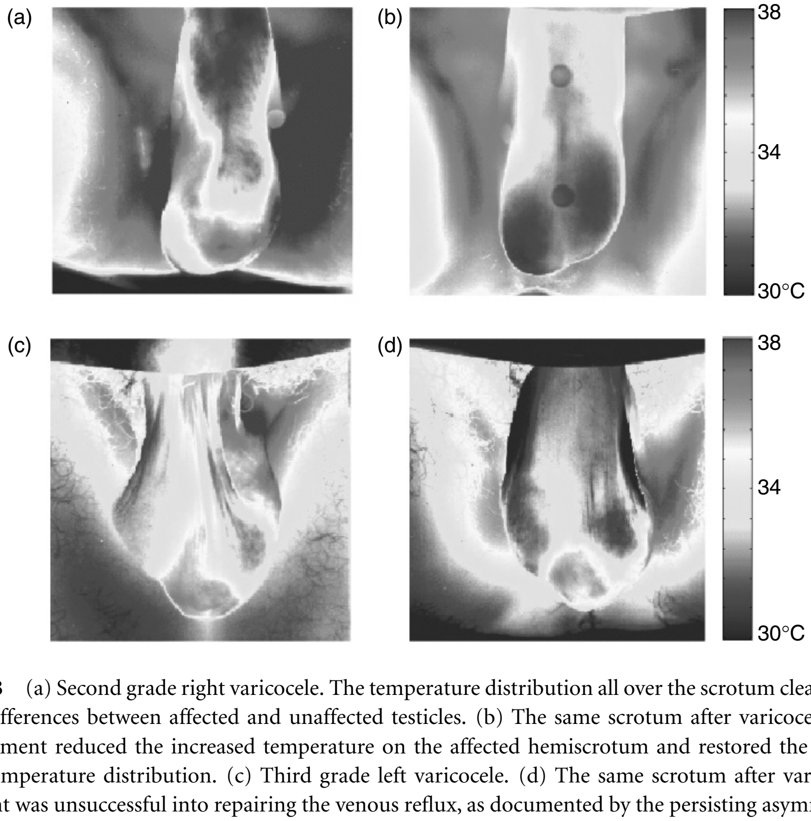


FIGURE 32.8 (a) Second grade right varicocele. The temperature distribution all over the scrotum clearly highlights significant differences between affected and unaffected testicles. (b) The same scrotum after varicocelectomy. The surgical treatment reduced the increased temperature on the affected hemiscrotum and restored the symmetry in the scrotal temperature distribution. (c) Third grade left varicocele. (d) The same scrotum after varicocelectomy. The treatment was unsuccessful into repairing the venous reflux, as documented by the persisting asymmetric scrotal distribution.

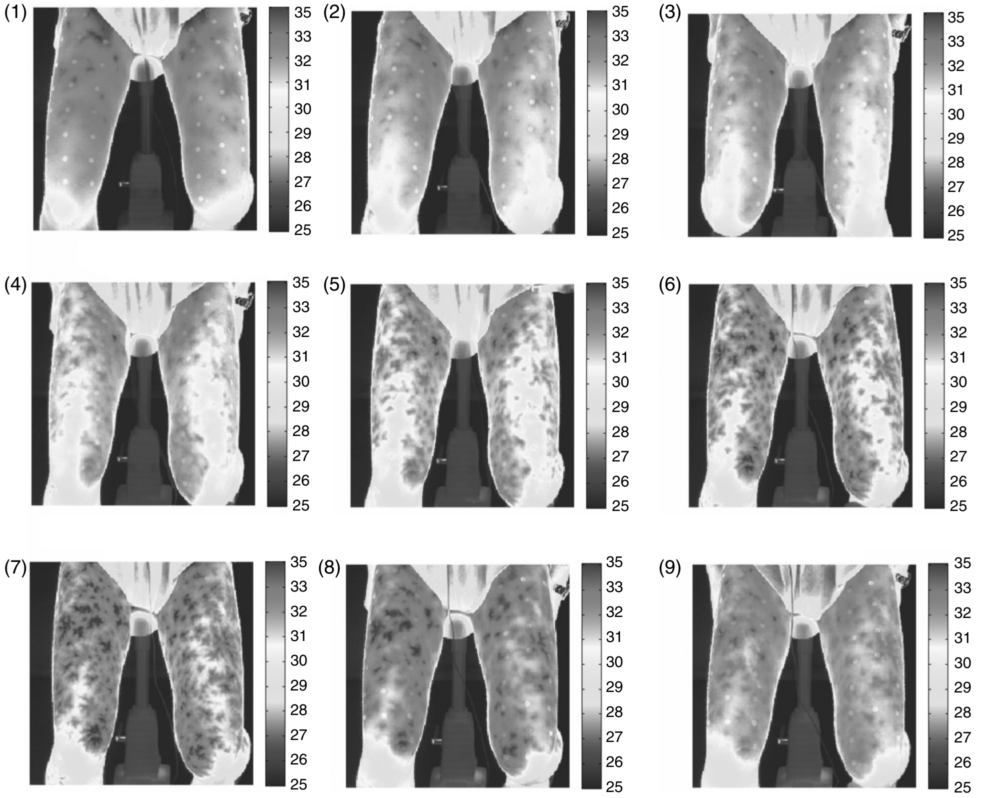


FIGURE 32.9 From top right to bottom left, fIR image sequence recorded along the exercise for a fit level subject. Images 1 and 2 were recorded during the warm up. Images 3 to 5 were recorded during the load phase of the exercise, with constant increasing working load. Image 6 corresponds to the maximum load and the beginning of the recovery phase. Note the arterovenous shunts opening to improve the core muscle layer cooling. (From Merla et al. in *Proceedings of the 24th IEEE Engineering in Medicine and Biology Society Conference*, Houston, October 23–25, 2002. With permission.)

demonstrated that IR imaging may provide early indication on the possible relapsing of the disease and may be used as a suitable complementary follow-up tool.

32.5 Analysis of Skin Temperature During Exercise

Skin temperature depends on complex thermal exchanges between skin tissue, skin blood, the environment, and the inner warmer tissue layers [31]. Johnson [32], Kenney and Johnson [33], and Zontak et al. [34] documented modifications in blood flow of the skin during: a reduction of the skin blood flow characterizes the initial phase of the exercise; while increase in skin blood flow accompanies increased working load. Zontak et al. [34] showed — by means of thermography — that dynamics of hand skin temperature response during leg exercise depends on the type of the exercise (graded vs. constant load): graded load exercise results in a constant decrease of finger temperature reflecting a constantly dominating vasoconstrictor response; steady state exercise causes a similar initial temperature decrease followed by rewarming of hands reflecting dominance of the thermoregulatory reflexes at a later stage of the exercise.

Merla et al. [18] studied directly the effects of graded load exercise on the thigh skin temperature, by means of fIR imaging. The study was aimed to assess possible relationship between the dynamics of the skin temperature along the exercise, the way the exercise is executed, and the fitness level of the subject.

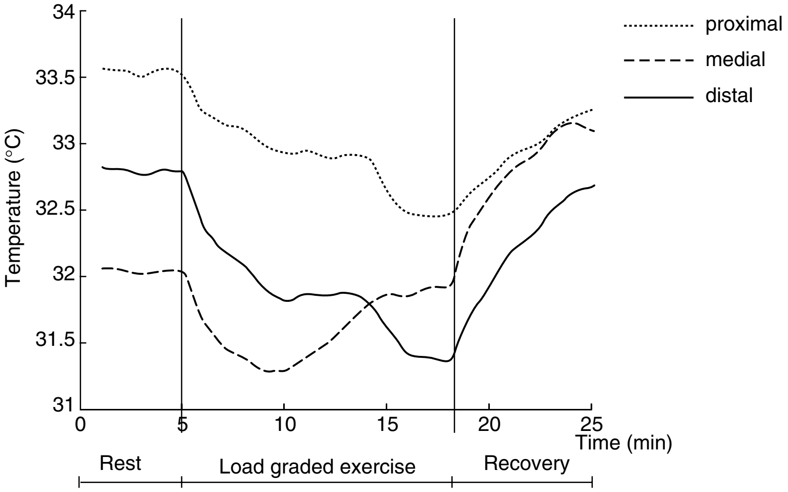


FIGURE 32.10 T_{ts} vs. time curves of each third of the left thigh for the subject in Figure 32.9. The distal third exhibits a different behavior with respect to the more proximal ones, probably because of a different involving in the execution of the exercise. (From Merla et al., in *Proceedings of the 24th IEEE Engineering in Medicine and Biology Society Conference.* With permission.)

In their study, the authors recruited 35 volunteers (24 M/11 F; age = 22 ± 1.5 yr; height = 173 ± 2.8 cm; body weight = 68 ± 4.2 kg), with normal response to cardiac effort test. The thigh skin temperature (T_{ts}) of the subject was continuously recorded under conditions of rest, increasing physical load, and recovery from exercise. Three different regions T_{ts} (proximal, medial, and distal third of thigh length, frontal view) were studied. After 2.5-min rest fIR recording, exercise testing was performed using a calibrated bicycle ergometer. The ergometer was controlled by a special software, that after measurement of Body Mass Index, total body fat percentage, arterial pressure, biceps strength, and back flexibility, extrapolates VO_2 (oxygen consumption) values during the several steps of the exercise test, and the fitness level monitoring the heartbeat rate (HR). The subjects underwent a load graded exercise test — 25 W every 3 min increasing workloads, 50 rpm — up to reaching the submaximal HR condition. The recovery phase was defined as the time necessary to restore the rest HR by means of recovery exercise and rest. T_{ts} distribution was continuously monitored, while fIR images were recorded every 1-min along the exercise and the recovery phases (Figure 32.9). For all of the subjects, the temperature featured an exponential pattern, decreasing in the warming up and increasing during the recovery. Time constants of the T_{ts} vs. time curves were calculated for the graded exercise (τ_{ex}) and the recovery (τ_{rec}) phases, respectively (Figure 32.10).

Artery–venous shunt openings were noticed during the restoring phase to improve the cooling of the core of the muscle. Different T_{ts} vs. time curves were observed for the three chosen recording regions according to the specific involvement of the underlying muscular structures (Figure 32.10). Each subject produced different T_{ts} vs. time curves, according to the fitness level and VO_2 consumption: the greater VO_2 consumption and better fitness, the faster τ_{ex} and τ_{rec} (see Table 32.1).

The results obtained by Merla et al. are consistent with previous observations [32–34]: the muscles demand increased blood flow at the beginning of the exercise that results in skin vasoconstriction. Prolonged work causes body temperature increase that invokes thermal regulatory processes with heat conduction to the skin. Therefore, skin thermoregulatory processes are also invoked by the exercise and can be recorded by fIR imaging. Executing of graded exercise determines a decreasing of the skin temperature throughout the exercise period, in association with increasing demand of blood from the working muscles. As the muscles stop working and the recovery and the restoration phase start, thermoregulatory control is invoked to limit body temperature increase and the skin vessels dilate to increase heat conduction to the skin and, then, to the environment [34]. One remarkable result is the fact that different T_{ls} vs. time curves for contralateral-homologous regions or different thirds have been found in some of the

TABLE 32.1 VO_2 Consumption and Average Constant Time on the Whole Thigh Length

Fitness level	Number of subjects	VO_2 (ml/kg min)	τ_{ex} (min)	τ_{rec} (min)
Excellent	12	56.7 ± 2.5	8.5 ± 2.2	2.4 ± 1.5
Fit	9	30.4 ± 3.8	10.1 ± 2.5	4.8 ± 2.0
Fair	8	25.9 ± 3.4	12.5 ± 2.6	7.3 ± 1.8
Needs work	6	18.3 ± 2.9	15.3 ± 2.3	10.3 ± 1.9

Source: Taken from Merla et al. in *Proceedings of the 24th IEEE Engineering in medicine and Biology Society Conference*, Houston, October 23–25, 2002.

subjects. This fact may be likely explained as a different involvement of the several parts of the muscles used or as expression of different work capabilities of the lower arms in the exercise. The fIR recording of skin temperature during exercise may be therefore regarded as a tool to get information about the quality of the execution of the exercise and the fitness level state of the subjects.

32.6 Discussion and Conclusion

The fIR imaging is a biomedical imaging technique that relies on high-resolution IR imaging and on the modeling of the heat exchange and control processes at the skin layer. The fIR imaging is aimed to provide quantitative diagnostic parameters through the functional investigation of the thermoregulatory processes. It is also aimed to provide further information about the studied disease to the physicians, like explanation of the possible physics reasons of some thermal behaviors and their relationships with the physiology of the involved processes. One of the great advantages of fIR imaging is the fact that it is not invasive and it is a touchless imaging technique. The fIR is not a static imaging investigation technique. Therefore, data for fIR imaging need to be processed adequately for movement. Adequate bio heat modeling is also required. The medical fields for possible applications of fIR imaging are numerous, ranging from those described in this chapter, to psychometrics, cutaneous blood flow modeling, peripheral nervous system activity, and some angiopathies. The applications described in this chapter show that fIR imaging provides highly effective diagnostic parameters. The method is highly sensitive, and also highly specific to discriminating different conditions of the same disease. For the studies reported hereby, fIR imaging is sensitive and specific as the corresponding golden standard techniques, at least. In some cases, fIR represents a useful follow-up tool (like in varicocelectomy to promptly assess possible relapses) or even an elective diagnostic tool, as in the Raynaud's phenomenon. The τ image technique represents an innovative technique that provides useful and additional information to the golden standard techniques. Thanks to it, the whole functional processes associated to a disease can be depicted and summarized into just a single image.

References

[1] Aweruch, M.S., Thermography: its current diagnostic status in muscular–skeletal medicine, *Med. J. Aust.*, 154, 441, 1991.

[2] Prescott et al., Sequential dermal microvascular and perivascular changes in the development of scleroderma, *J. Pathol.,* 166, 255, 1992.

[3] Herrick, A.L. and Clark, S., Quantifying digital vascular disease in patients with primary Raynaud's phenomenon and systemic sclerosis, *Ann. Rheum. Dis.,* 57, 70, 1998.

[4] Darton, K. and Black, C.M., Pyroelectric vidicon thermography and cold challenge quantify the severity of Raynaud's phenomenon, *Br. J. Rheumatol.,* 30, 190, 1991.

[5] Javanetti, S. et al., Thermography and nailfold capillaroscopy as noninvasive measures of circulation in children with Raynaud's phenomenon, *J. Rheumatol.,* 25, 997, 1998.

[6] Merla, A. et al., Dynamic digital telethermography: a novel approach to the diagnosis of varicocele, *Med. Biol. Eng. Comp.,* 37, 1080, 1999.

[7] Merla, A. et al., Correlation of telethermographic and ultrasonographic reports in the therapeutic monitoring of second-class muscular lesions treated by hyperthermia, *Med. Biol. Eng. Comp.,* 37, 942, 1999.

[8] Merla, A., Biomedical applications of functional infrared imaging, presented at the *21st Annual Meeting of Houston Society of Engineering in Medicine and Biology*, Houston, TX, February 12–13, 2004.

[9] Merla, A. et al., Assessment of the effects of the varicocelectomy on the thermoregulatory control of the scrotum, *Fertil. Steril.,* 81, 471, 2004.

[10] Merla, A. et al., Recording of the sympathetic thermal response by means of infrared functional imaging, in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Cancun, Mexico, September 17–21, 2003.

[11] Merla, A. et al., Infrared functional imaging applied to the study of emotional reactions: preliminary results, in *Proceedings of the 4th International Non-Invasive Functional Source Imaging*, Chieti, Italy, September 9–13, 2003.

[12] Merla, A. and Romani, G.L., Skin blood flow rate mapping through functional infrared imaging, in *Proceedings of the World Congress of Medical Physics WC2003*, Sidney, August 24–29, 2003.

[13] Merla, A. Cianflone, F., and Romani, G.L., Skin blood flow rate estimation through functional infrared imaging analysis, in *Proceedings of the 5th International Federation of Automatic Control Symposium on Modelling and Control in Biomedical Systems*, Melbourne, August 19–23, 2003.

[14] Merla, A. et al., Raynaud's phenomenon: infrared functional imaging applied to diagnosis and drugs effects, *Int. J. Immun. Pharm.* 15, 41, 2002.

[15] Merla, A. et al., Use of infrared functional imaging to detect impaired thermoregulatory control in men with asymptomatic varicocele, *Fertil. Steril.,* 78, 199, 2002.

[16] Merla, A. et al., Infrared functional imaging applied to Raynaud's phenomenon, *IEEE Eng. Med. Biol. Mag.,* 21, 73, 2002.

[17] Merla, A. et al., Quantifying the relevance and stage of disease with the tau image technique. *IEEE Eng. Med. Biol. Mag.,* 21, 86, 2002.

[18] Merla, A. et al., Infrared functional imaging: analysis of skin temperature during exercise, in *Proceedings of the 24th IEEE Engineering in Medicine and Biology Society Conference*, Houston, October 23–25, 2002.

[19] Merla, A. et al., Time recovery image: a diagnostic image technique based on the dynamic digital telethermography, *Thermol. Int.,* 10, 142, 2000.

[20] Merla, A. et al., Tau image: a diagnostic imaging technique based on the dynamic digital telethermography, *in Proceedings of the WC2000 Chicago World Congress on Medical Physics and Biomedical Engineering and 22nd International Conference of IEEE Engineering in Medicine and Biology Society*, Digest of Papers CD, track 1,TU-FXH, July 2000, Chicago.

[21] Allen, E.V. and Brown, G.E., Raynaud's disease: a critical review of minimal requisites for diagnosis, *Am. J. Med. Sci.,* 183, 187, 1932.

[22] Subcommittee for Scleroderma Criteria of the American Rheumatism Association Diagnostic and Therapeutic Criteria Committee, Preliminary criteria for the classification of systemic sclerosis (scleroderma), *Arthr. Rheum.*, 23, 581, 1980.

[23] O'Reilly, D. et al., Measurement of cold challenge response in primary Raynaud's phenomenon and Raynaud's phenomenon associated with systemic sclerosis, *Ann. Rheum. Dis.,* 51, 1193, 1992.

[24] Clarks, S. et al., The distal–dorsal difference as a possible predictor of secondary Raynaud's phenomenon, *J. Rheumatol.,* 26, 1125, 1999.

[25] Schuhfried, O. et al., Thermographic parameters in the diagnosis of secondary Raynaud's phenomenon, *Arch. Phys. Med. Rehab.,* 81, 495, 2000.

[26] Ring, E.F.J., Ed., Cold stress test for the hands, in *The thermal image in Medicine and Biology*, Uhlen Verlag, Wien, 1995.

[27] Merla, A. et al., Combined approach to the initial stage Raynaud's phenomenon diagnosis by means of dynamic digital telethermography, capilloroscopy and pletismography: preliminary findings, *Med. Biol. Eng. Comp.,* 37, 992, 1999.

[28] Shitzer, A. et al., Lumped parameter tissue temperature–blood perfusion model of a cold stressed finger, *J. Appl. Physiol.*, 80, 1829, 1996.

[29] Mieusset, R. and Bujan, L., Testicular heating and its possible contributions to male infertility: a review, *Int. J. Andr.,* 18, 169, 1995.

[30] Trum, J.W., The value of palpation, varicoscreen contact thermography and colour Doppler ultrasound in the diagnosis of varicocele, *Hum. Reprod.,* 11, 1232, 1996.

[31] Brengelmann, G.L. et al., Altered control of skin blood flow during exercise at high internal temperatures, *J. Appl. Physiol.,* 43, 790, 1977.

[32] Johnson, J.M., Exercise and the cutaneous circulation, *Exerc. Sport Sci. Rev.* 20, 59, 1992.

[33] Kenney, W.L. and Johnson, J.M., Control of skin blood flow during exercise, *Med. Sci. Sports Exerc.,* 24, 303, 1992.

[34] Zontak, A. et al., Dynamic thermography: analysis of hand temperature during exercise, *Ann. Biomed. Eng.,* 26, 988, 1998.

[35] ASHRAE: handbook fundamentals SI edition, ASHRAE, Atlanta, GA, 1985.

[36] Cooke, E.D. et al., Reflex sympathetic dystrophy and repetitive strain injury: temperature and microcirculatory changes following mild cold stress, *J.R. Soc. Med.,* 86, 690, 1993.

[37] Di Benedetto, M., Regional hypothermia in response to minor injury, *Am. J. Phys. Med. Rehab.,* 75, 270, 1996.

[38] Garagiola, U., Use of telethermography in the management of sports injuries, *Sports Med.,* 10, 267, 1990.

[39] Maricq, H.R. et al., Diagnostic potential of in vivo capillary microscopy in scleroderma and related disorders, *Arthr. Rheum.,* 23, 183, 1980.

[40] Rodnan, G.P., Myerowitz, R.I., and Justh, G.O., Morphological changes in the digital arteries of patients with progressive systemic sclerosis and Raynaud phenomenon, *Medicine*, 59, 393, 1980.

[41] Tucker, A., Infrared thermographic assessment of the human scrotum, *Fertil. Steril.,* 74, 802, 2000.

33

Thermal Imaging in Surgery

Paul Campbell
Ninewells Hospital

Roderick Thomas
Swansea Institute of Technology

33.1 Overview

Advances in miniaturization and microelectronics, coupled with enhanced computing technologies, have combined to see modern infrared imaging systems develop rapidly over the past decade. As a result, the instrumentation has become considerably improved, not only in terms of its inherent resolution (spatial *and* temporal) and detector sensitivity (values ca. 25 mK are typical) but also in terms of its portability: the considerable reduction in bulk has resulted in light, camcorder (or smaller) sized devices. Importantly, cost has also been reduced so that entry to the field is no longer prohibitive. This attractive combination of factors has led to an ever increasing range of applicability across the medical spectrum. Whereas the mainstay application for medical thermography over the past 40 years has been with rheumatological and associated conditions, usually for the detection and diagnosis of peripheral vascular diseases such as Raynaud's phenomenon, the latest generations of thermal imaging systems have seen active service within new surgical realms such as orthopaedics, coronary by-pass operations, and also in urology. The focus of this chapter relates not to a specific area of surgery per se, but rather to a generic and pervasive aspect of all modern surgical approaches: the use of *energized* instrumentation during surgery. In particular, we will concern ourselves with the use of thermal imaging to accurately monitor temperature within the tissue locale surrounding an energy-activated instrument. The rationale behind this is that it facilitates optimization of operation specific protocols that may either relate to thermally based therapies, or else to reduce the extent of collateral damage that may be introduced when inappropriate power levels, or excessive pulse durations, are implemented during surgical procedures.

33.2 Energized Systems

Energy-based instrumentation can considerably expedite fundamental procedures such as vessel sealing and dissection. The instrumentation is most often based around ultrasonic, laser, or radio-frequency (RF)-current based technologies. Heating tissue into distinct temperature regimes is required in order to achieve the desired effect (e.g., vessel sealing, cauterization, or cutting). In the context of electrical current heating, the resultant effect of the current on tissue is dominated by two factors: the temperature attained by the tissue; and the duration of the heating phase, as encapsulated in the following equation:

$$T - T_0 = \frac{1}{\sigma \rho c} J^2 \delta t \qquad (33.1)$$

where T and T_0 are the final and initial temperatures (in degrees Kelvin [K]) respectively, σ is the electrical conductivity (in S/m), ρ is the tissue density, c is the tissue specific heat capacity (J kg^{-1} K^{-1}), J is the current density (A/m^2), and δt is the duration of heat application. The resultant high temperatures are not limited solely to the tissue regions in which the electrical current flow is concentrated. Heat will flow away from hotter regions in a time dependence fashion given by the Fourier equation:

$$Q(r, t) = -k \nabla T(r, t) \qquad (33.2)$$

where Q is the heat flux vector, the proportionality constant k is a scalar quantity of the material known as the thermal conductivity, and $\nabla T(r, t)$ is the temperature gradient vector. The overall spatio-temporal evolution of the temperature field is embodied within the differential equation of heat flow (alternatively known as the diffusion equation)

$$\frac{1}{\alpha} \frac{\partial T(r, t)}{\partial t} = \nabla^2 T(r, t) \qquad (33.3)$$

where α is the thermal diffusivity of the medium defined in terms of the physical constants, k , ρ, and c thus:

$$\alpha = k/\rho c \qquad (33.4)$$

and temperature T is a function of both the three dimensions of space (r) and also of time t. In other words, high temperatures are not limited to the region specifically targeted by the surgeon, and this is often the source of an added surgical complication caused by collateral or proximity injury. Electrosurgical damage, for example, is the most common cause of iatrogenic bowel injury during laparoscopic surgery and 60% of mishaps are missed, that is, the injury is not recognized during surgery and declares itself with peritonitis several days after surgery or even after discharge from hospital. This level of morbidity can have serious consequences, in terms of both the expense incurred by re-admission to hospital, or even the death of the patient. By undertaking *in vivo* thermal imaging during energized dissection it becomes possible to determine, in real time, the optimal power conditions for the successful accomplishment of specific tasks, and with minimal collateral damage. As an adjunct imaging modality, thermal imaging may also improve surgical practice by facilitating easier identification and localization of tissues such as arteries, especially by less experienced surgeons. Further, as tumors are more highly vascularized than normal tissue, thermal imaging may facilitate their localization and staging, that is, the identification of the tumor's stage in its growth cycle. Figure 33.1 shows a typical set-up for implementation of thermography during surgery.

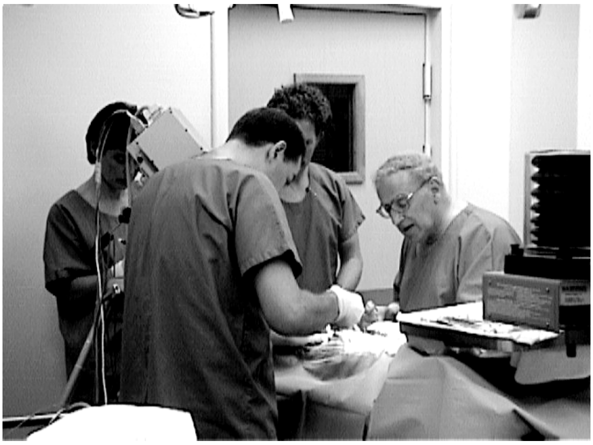


FIGURE 33.1 Typical set-up for a thermal imaging in surgery. The camera is tripod mounted toward the foot of the operating table and aimed at the surgical access site (camera visible over the left shoulder of the nearmost surgeon).

33.3 Thermal Imaging Systems

As skin is a close approximation to an ideal black body (the emissivity, ε, of skin is 0.98, whereas that of an ideal black body has $\varepsilon = 1$), then we can feel reasonably confident in applying the relevant physics directly to the situation of thermography in surgery. One important consideration must be the waveband of detector chosen for thermal observations of the human body. It is known from the thermal physics of black bodies, that the wavelength at which the maximum emissive power occurs, $\lambda_{\max}$ (i.e., the peak in the Planck curve), is related to the body's temperature T through Wien's law:

$$\lambda_{\max} T = 0.002898 \tag{33.5}$$

Thus for bodies at 310 K (normal human body temperature), the peak output is around 10 μm, and the majority of the emitted thermal radiation is limited to the range from 2 to 20 μm. The optimal detectors for passive thermal imaging of normal skin should thus have best sensitivity around the 10 μm range, and this is indeed the case with many of the leading thermal imagers manufactured today, which often rely on GaAs quantum well infrared photodetectors (QWIPs) with a typical waveband of 8–9 μm. A useful alternative to these longwave detectors involves the use of indium–antimonide (InSb) based detectors to detect radiation in the mid-wave infrared (3–5 μm). Both these materials have the benefit of enhanced temperature sensitivity (ca. 0.025 K), and are both wholly appropriate even for quantitative imaging of hotter surfaces, such as may occur in energized surgical instrumentation.

33.4 Calibration

Whilst the latest generation of thermal imaging systems are usually robust instruments exhibiting low drift over extended periods, it is sensible to recalibrate the systems at regular intervals in order to preserve the integrity of captured data. For some camera manufacturers, recalibration can be undertaken under a service agreement and this usually requires shipping of the instrument from the host laboratory. However for other systems, recalibration must be undertaken in-house, and on such occasions, a black body source (BBS) is required.

Most BBS are constructed in the form of a cavity at a known temperature, with an aperture to the cavity that acts as the black body, effectively absorbing all incident radiation upon it. The cavity temperature must be measured using a high accuracy thermometric device, such as a platinum resistance thermometer

(PRT), with performance characteristics traceable to a thermometry standard. Figure 33.2b shows one such system, as developed by the UK National Physical Laboratory at Teddington, and whose architecture relies on a heat-pipe design. The calibration procedure requires measurement of the aperture temperature at a range of temperature set-points that are simultaneously monitored by the PRT (e.g., at intervals of 5° between temperature range of 293 and 353 K). Direct comparison of the radiometric temperature measured by the thermal camera with the standard temperature monitored via the PRT allows a calibration table to be generated across the temperature range of interest. During each measurement, sufficient time must be allowed in order to let the programmed temperature set-point equilibrate, otherwise inaccuracies will result. Further, the calibration procedure should ideally be undertaken under similar ambient conditions to those under which usual imaging is undertaken. This may include aspects such as laminar, or even fan-assisted, flow around the camera body which will affect the heat transfer rate from the camera to the ambient and in turn may affect the performance of the detector (viz Figure 33.2b).

33.5 Thermal Imaging During Energized Surgery

Fully remote-controlled cameras may be ideally suited to overhead bracket mountings above the operating table so that a bird's eye view over the surgical site is afforded. However, without a robotized arm to fully control pitch and location, the view may be restrictive. Tripod mounting, as illustrated in Figure 33.1, and with a steep look-down angle from a distance of about 1 m to the target offers the most versatile viewing without compromising the surgeon's freedom of movement. However, this type of set-up demands that a camera operator be on hand continually in order to move the imaging system to those positions offering best viewing for the type of energized procedure being undertaken.

33.5.1 RF Electrosurgery

As mentioned earlier, the most common energized surgical instrumentation employ a physical system reliant on either (high frequency) electrical current, an ultrasonic mechanism, or else incident laser energy in order to induce tissue heating. Thermal imaging has been used to follow all three of these procedures. There are often similarities in approach between the alternative modalities. For example, vessel sealing often involves placement of elongated forcep-style electrodes across a target vessel followed by ratcheted compression, and then a pulse of either RF current, or alternatively ultrasonic activation of the forceps, is applied through the compressed tissue region. The latest generations of energized instrumentation may have active feedback control over the pulse to facilitate optimal sealing with minimal thermal spread (e.g., the Valleylab *Ligasure* instrument) however under certain circumstances, such as with calcified tissue or in excessively liquid environments the performance may be less predictable.

Figure 33.3 illustrates how thermal spread may be monitored during the instrument activation period of one such "intelligent" feedback device using RF current. The initial power level for each application is determined through a fast precursor voltage scan that determines the natural impedance of the compressed tissue. Then, by monitoring the temperature dependence of impedance (of the compressed tissue) during current activation, the microprocessor controlled feedback loop automatically maintains an appropriate power level until a target impedance is reached indicating that the seal is complete. This process typically takes between 1 and 6 sec, depending on the nature of the target tissue. Termination of the pulse is indicated by an audible tone burst from the power supply box. The performance of the system has been evaluated in preliminary studies involving gastric, colonic and small bowel resection [1]; hemorraoidectomy [2]; prostatectomy [3]; and cholecystectomy [4].

Perhaps most strikingly, the facility for real time thermographic monitoring, as illustrated in Figure 33.3, affords the surgeon immediate appreciation of the instrument temperature, providing a visual cue that automatically alerts to the potential for iatrogenic injury should a hot instrument come into close contact with vital structures. By the same token, the *in situ* thermal image also indicates when the tip of the instrument has cooled to ambient temperature. It should be noted that the amount by which the activated

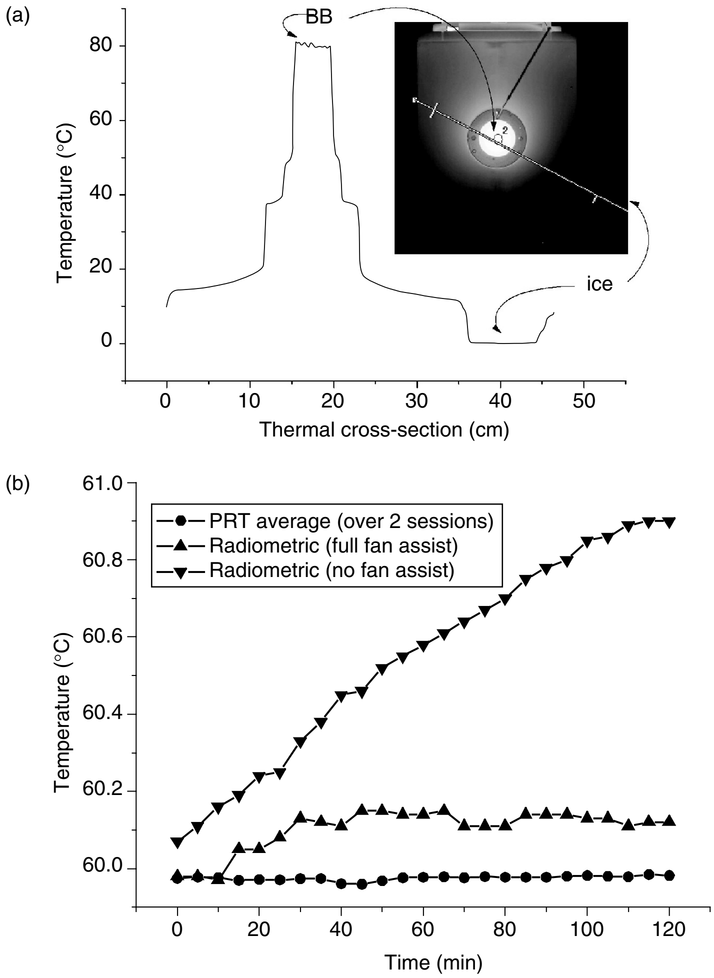


FIGURE 33.2 Thermal cross-section (profile) through the black body calibration source together with equilibrated crushed ice, which acts as a convenient secondary temperature gauge *in situ*. (Insert [left] thermal view with linear region of interest highlighted, and [right] optical view of the black body cavity and beaker of [equilibrated] crushed ice to the lower right.) (b) Radiometric detector drift during start up under two different ambient conditions. The detector readout is centered on the black body cavity source shown in (a), which was itself maintained at a target temperature of 59.97°C throughout the measurements (solid circles). Without fan-assisted cooling of the camera exterior, the measured temperature drifted by 0.8°C over 2 h, hence the importance of calibration under typical operating conditions. With fan-assisted cooling, the camera "settles" within around 30 min of switching on. (Camera: Raytheon Galileo [Raytheon Systems].)

head's temperature rises is largely a function of device dimensions, materials, and the power levels applied together with the pulse duration.

33.5.2 Analysis of Collateral Damage

Whilst thermograms typical of Figure 33.3 offer a visually instructive account of the thermal scene and its temporal evolution, a quantitative analysis of the sequence is more readily achieved through the identification of a linear region of interest (LROI), as illustrated by the line bisecting the device head in Figure 33.4a. The data constituted by the LROI is effectively a snapshot thermal profile across those pixels lying on this designated line (Figure 33.4b). A graph can then be constructed to encapsulate the

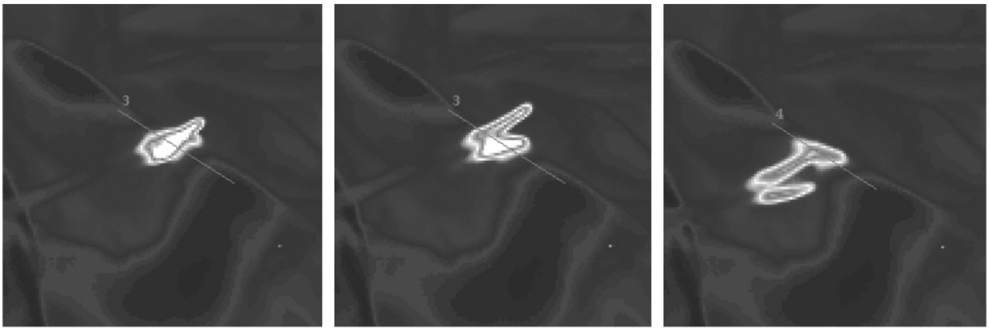


FIGURE 33.3 Thermographic sequence taken with the Dundee thermal imaging system and showing (33.1) energized forceps attached to bowel (white correlates with temperature), (33.2) detachment of the forceps revealing hot tissue beneath, (33.3) remnant hot-spot extending across the tissue and displaying collateral thermal damage covering 4.5 mm either side of the instrument jaws.

time dependent evolution of the LROI. This is displayed as a 3D surface (a function of spatial co-ordinate along the LROI, time, and temperature) upon which color-mapped contours are evoked to represent the different temperature domains across the LROI (Figure 33.4c). In order to facilitate measurement of the thermal spread, the 3D surface, as represented in matrix form, can then be interrogated with a mathematical programming package, or alternatively inspected manually, a process that is most easily undertaken after projecting the data to the 2D coordinate-time plane, as illustrated in Figure 33.4d. The critical temperature beyond which tangible heat damage can occur to tissue is assumed to be 45°C [5]. Thermal spread is then calculated by measuring the maximum distance between the 45°C contours on the planar projection, then subtracting the electrode "footprint" diameter from this to get the total spread. Simply dividing this result by two gives the thermal spread either side of the device electrodes.

The advanced technology used in some of the latest generations of vessel sealing instrumentation can lead to a much reduced thermal spread, compared with the earlier technologies. For example with the Ligasure LS1100 instrument, the heated peripheral region is spatially confined to less than 2 mm, even when used on thicker vessels/structures. A more advanced version of the device (LS1200 [*Precise*]) consistently produces even lower thermal spreads, typically around 1 mm (viz Figure 33.4). This performance is far superior to other commercially available energized devices.

For example, Kinoshita and co-workers [6] have observed (using infrared imaging) that the typical lateral spread of heat into adjacent tissue is sufficient to cause a temperature of over 60°C at radial distances of up to 10 mm from the active electrode when an ultrasonic scalpel is used. Further, when standard bipolar electro-coagulation instrumentation is used, the spread can be as large as 22 mm. Clearly, the potential for severe collateral and iatrogenic injury is high with such systems unless power levels are tailored to the specific procedure in hand and real time thermal imaging evidently represents a powerful adjunct technology to aid this undertaking.

Whilst the applications mentioned thusfar relate to "open" surgical procedures requiring a surgical incision to access the site of interest, thermal imaging can also be applied as a route to protocol optimization for other less invasive procedures also. Perhaps the most important surgical application in this regime involves laser therapy for various skin diseases/conditions. Application of the technique in this area is discussed below.

33.6 Laser Applications in Dermatology

33.6.1 Overview

Infrared thermographic monitoring (ITM) has been successfully used in medicine for a number of years and much of this has been documented by Prof. Francis Ring [http://www.medimaging.org/], who has

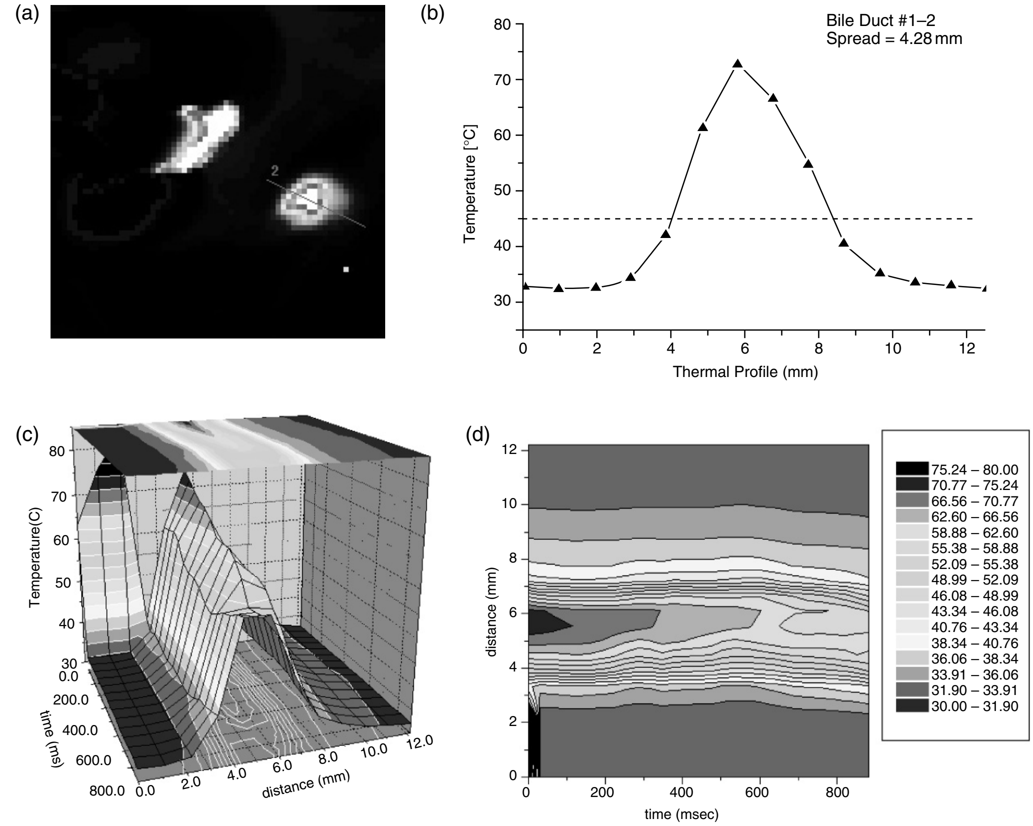


FIGURE 33.4 (a) Mid-infrared thermogram taken at the instant an energized forceps (Ligasure LS1200 "*Precise*") is removed from the surgical scene after having conducted a seal on the bile duct. The hot tips of the forceps are clearly evident in the infrared view (just left of center), as is the remnant hot-spot where the seal has occurred on the vessel. By generating a linear region of interest (LROI) through the hot-spot, as indicated by the highlighted line in the figure, it is possible to monitor the evolution of the hot-spot's temperature in a quantitative fashion. (b) Thermal profile corresponding to the LROI shown in (a). (c) By tracking the temporal evolution of the LROI, it is possible to generate a 3D plot of the thermal profile by simply stacking the individual profiles at each acquisition frame. In this instance the cooling behavior of the hot-spot is clearly identified. Manual estimation of the thermal spread is most easily achieved by resorting to the 2D contour plot of the thermal profile's temporal evolution, as shown in (d). In this instance, the maximal spread of the 45°C contours is measured as 4.28 mm. By subtracting the forcep "footprint" (2.5 mm for the device shown) and dividing the result by 2, we arrive at the thermal spread for the device. The average thermal spread (for 6 bile-duct sealing events) was 0.89 ± 0.35 mm.

established a database and archive within the Department of Computing at the University of Glamorgan, UK, spanning over 30 years of ITM applications. Examples include monitoring abnormalities such as malignancies, inflammation, and infection that cause localized increases in skin temperature, which show as hot spots or as asymmetrical patterns in an infrared thermogram.

A recent medical example that has benefited by the intervention of ITM is the treatment by laser of certain dermatological disorders. Advancements in laser technology have resulted in new portable laser therapies, examples of which include the removal of vascular lesions (in particular Port Wine Stains [PWS]), and also cosmetic enhancement approaches such as hair-(depilation) and wrinkle removal.

In these laser applications it is a common requirement to deliver laser energy uniformly without overlapping of the beam spot to a sub-dermal target region, such as a blood vessel, but with the minimum of collateral damage to the tissue locale. Temperature rise at the skin surface, and with this the threshold to burning/scarring is of critical importance for obvious reasons. Until recently, this type of therapy had not yet benefited significantly from thermographic evaluation. However, with the introduction of the latest generation thermal imaging systems, exhibiting the essential qualities of portability, high resolution, and high sensitivity, significant inroads to laser therapy are beginning to be made.

Historically, lasers have been used in dermatology for some 40 years [25]. In recent years there have been a number of significant developments particularly regarding the improved treatment of various skin disorders most notably the removal of vascular lesions using dye lasers [8,12,15,17,19] and depilation using ruby lasers [9,14,16]. Some of the general indicators as to why lasers are the preferred treatment of choice are summarized in Table 33.1.

33.7 Laser-Tissue Interactions

The mechanisms involved in the interaction between light and tissue depend on the characteristics of the impinging light and the targeted human tissue [24]. To appreciate these mechanisms the optical properties of tissue must be known. It is necessary to determine the tissue reflectance, absorption, and scattering properties as a function of wavelength. A simplified model of laser light interaction with the skin is illustrated in Figure 33.5.

Recent work has shown that laser radiation can penetrate through the epidermis and basal structure to be preferentially absorbed within the blood layers located in the lower dermis and subcutis. The process is termed selective photothermolysis, and is the specific absorption of laser light by a target tissue in order to eliminate that target without damaging surrounding tissue. For example, in the treatment of Port Wine

TABLE 33.1 Characteristics of Laser Therapy during and after Treatment

General indicators	Dye laser vascular lesions	Ruby laser depilation
During treatment	Varying output parameters Portable Manual and scanned Selective destruction of target chromophore (Haemoglobin)	Varying output parameters Portable Manual and scanned Selective destruction of target chromophore (melanin)
After treatment (desired effect)	Slight bruising (purpura)	Skin returns to normal coloring (no bruising)
	Skin retains its elasticity	Skin retains surface markings
	Skin initially needs to be protected from UV and scratching	Skin retains its ability to tan after exposure to ultraviolet light
	Hair follicles are removed	Hair removed

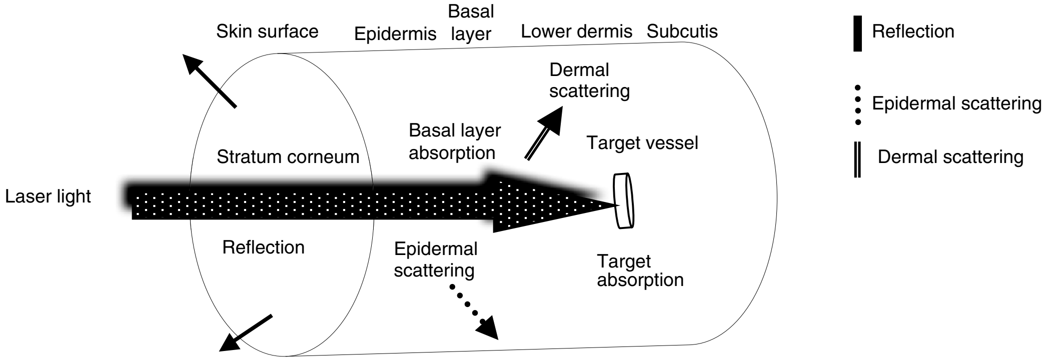


FIGURE 33.5 Passage of laser light within skin layers.

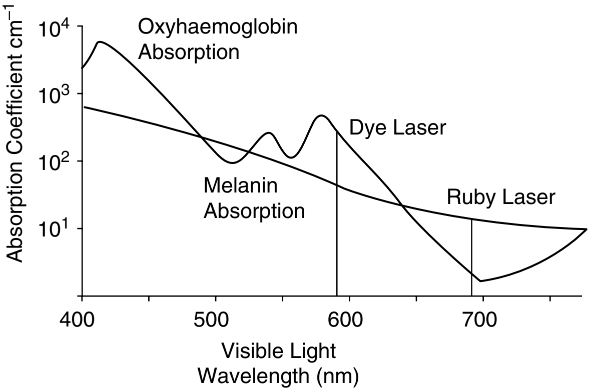


FIGURE 33.6 Spectral absorption curves for human blood and melanin.

TABLE 33.2 Interaction Effects of Laser Light and Tissue

Effect	Interaction
Photothermal	
Photohyperthermia	Reversible damage of normal tissue (37–42°C)
Photothermolysis	Loosening of membranes (odema), tissue welding (45–60°C)
Photocoagulation	Thermal-dynamic effects, micro-scale overheating
Photocarbonization	Coagulation, necrosis (60–100°C)
Photovaporization	Drying out, vaporization of water, carbonization (100–300°C)
	Pyrolysis, vaporization of solid tissue matrix (>300°C)
Photochemical	
Photochemotherapy	Photodynamic therapy, black light therapy
Photoinduction	Biostimulation
Photoionization	
Photoablation	Fast thermal explosion, optical breakdown, mechanical shockwave

Stains (PWS), a dye laser of wavelength 585 nm has been widely used [10] where the profusion of small blood vessels that comprise the PWS are preferentially targeted at this wavelength. The spectral absorption characteristics of light through human skin have been well established [7] and are replicated in Figure 33.6 for the two dominant factors: melanin and oxyhaemoglobin.

There are three types of laser/tissue interaction, namely: photothermal, photochemical, and protoionization (Table 33.2), and the use of lasers on tissue results in a number of differing interactions including photodisruption, photoablation, vaporization, and coagulation, as summarized in Figure 33.7.

The application of appropriate laser technology to medical problems depends on a number of laser operating parameters including matching the optimum laser wavelength for the desired treatment. Some typical applications and the desired wavelengths for usage are highlighted in Table 33.3.

33.8 Optimizing Laser Therapies

There are a number of challenges in optimizing laser therapy, mainly related to the laser parameters of wavelength, energy density, and spot size. Combined with these are difficulties associated with poor positioning of hand-held laser application that may result in uneven treatment [overlapping spots and/or uneven coverage (stippling) of spots], excessive treatment times, and pain. Therefore, for enhanced efficacy an improved understanding of the thermal effects of laser–tissue interaction benefits therapeutic

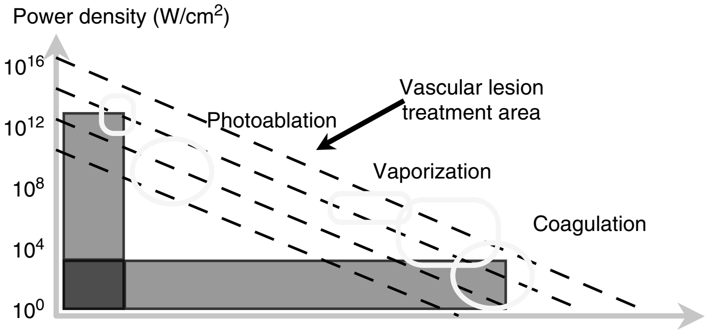


FIGURE 33.7 Physiological characteristics of laser therapy. (From Thomas et al., 2002, *Proceedings of SPIE*, 1–4 April, Orlando, USA. With permission.)

TABLE 33.3 Laser Application in Dermatology

Laser	Wavelength (nm)	Treatment
Flashlamp short-pulsed dye	510	Pigmented lesions, for example, freckles, tattoos
Flashlamp long-pulsed dye	585	PWS in children, warts, hypertrophic scars
Ruby single-pulse or Q-switched	694	Depilation of hair
Alexandrite Q-switched	755	Multicolored tattoos, viral warts, depilation
Diode variable	805	Multicolored tattoos, viral warts
Neodymium yitrium aluminum (Nd-YAG) Q-switched	1064	Pigmented lesions; adult port-wine stains, black/blue tattoos
Carbon dioxide continuous pulsed	10600	Tissue destruction, warts, tumors

approaches. Here, variables for consideration include:

1. Thermal effects of varying spot size.
2. Improved control of hand-held laser minimising overlapping and stippling.
3. Establishment of minimum gaps.
4. Validation of laser computer scanning.

Evaluation (Figure 33.8) was designed to elucidate whether or not measurements of the surface temperature of the skin are reproducible when illuminated by nominally identical laser pulses. In this case a 585 nm dye laser and a 694 nm ruby laser were used to place a number of pulses manually on tissue. The energy emitted by the laser is highly repeatable. Care must be taken to ensure that both the laser and radiometer position are kept constant and that the anatomical location used for the test had uniform tissue pigmentation.

Figure 33.8 shows the maximum temperature for each of twenty shots fired on the forearm of a representative Caucasian male with type 2 skin*. Maximum temperature varies between 48.90 and 48.10°C representing a variance of 1°C (±0.45°C). This level of reproducibility is pleasing since it shows that, despite the complex scenario, the radiometer is capable of repeatedly and accurately measuring surface tissue temperatures. In practice the radiometer may be used to inform the operator when any accumulated temperature has subsided allowing further treatment without exceeding some damage threshold.

Energy density is also an important laser parameter and can be varied to match the demands of the application. It is normal in the discipline to measure energy density (fluence) in J/cm^2. In treating vascular lesions most utilize an energy density for therapy of 5 to 10 J/cm^2 [13]. The laser operator needs to be sure that the energy density is uniform and does not contain hot-spots that may take the temperature above the damage threshold inadvertently. Preliminary characterization of the spot with thermal imaging can

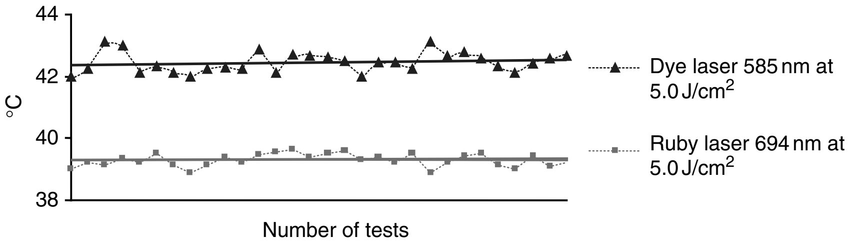


FIGURE 33.8 Repeatability of initial maximum skin temperatures (°C) of two lasers with similar energy density but different wavelengths.

then aid with fine tuning of the laser and reduce the possibility of excessive energy density and with that the possibility of collateral damage.

33.9 Thermographic Results of Laser Positioning

During laser therapy the skin is treated with a number of spots, applied manually depending on the anatomical location and required treatment. It has been found that spot size directly affects efficacy of treatment. The wider the spot size the higher the surface temperature [22]. The type and severity of lesion also determines the treatment required. Its color severity (dark to light) and its position on skin (raised to level). Therefore the necessary treatment may require a number of passes of the laser over the skin. It is therefore essential as part of the treatment that there is a physical separation between individual spots so that:

1. The area is not over treated with overlapping spots that could otherwise result in local heating effects from adjacent spots resulting in skin damage
2. The area is not under treated leaving stippled skin
3. The skin has cooled sufficiently before second or subsequent passes of the laser

Figure 33.9 shows two laser shots placed next to each other some 4 mm apart. The time between the shots is 1 sec. There are no excessive temperatures evident and no apparent temperature build-up in the gap. This result, which concurs with Lanigan [18], suggests a minimum physical separation of 5 mm between all individual spot sizes.

The intention is to optimize the situation leading to a uniform therapeutic and aesthetic result without either striping or thermal build-up. This is achieved by initially determining the skin color (Chromotest) for optimum energy settings, followed by a patch test and subsequent treatment. Increasing the number of spots to 3 with the 4 mm separation reveals a continuing trend, as shown in Figure 33.10. The gap between the first two shots is now beginning to merge in the 2 sec period that has lapsed. The gap between shots 2 and 3 remains clear and distinct and there are clearly visible thermal bands across the skin surface of between 38–39 and 39–40°C. These experimental results supply valuable information to support the development of both free-hand treatment and computer-controlled techniques.

33.10 Computerized Laser Scanning

Having established the parameters relating to laser spot positioning, the possibility of achieving reproducible laser coverage of a lesion by automatic scanning becomes a reality. This has potential advantages, which include:

1. Accurate positioning of the spot with the correct spacing from the adjacent spots
2. Accurate timing allowing the placement at a certain location at the appropriate lapsed time

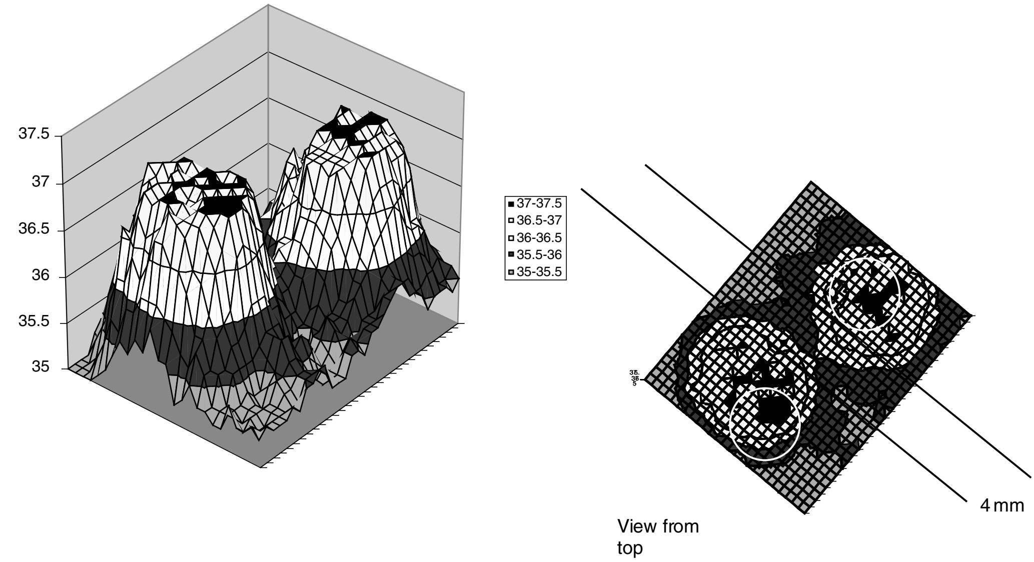


FIGURE 33.9 Two-dye laser spots with a minimum of 4 mm separation (585 nm at 4.5 J/cm^2, 5 mm spot).

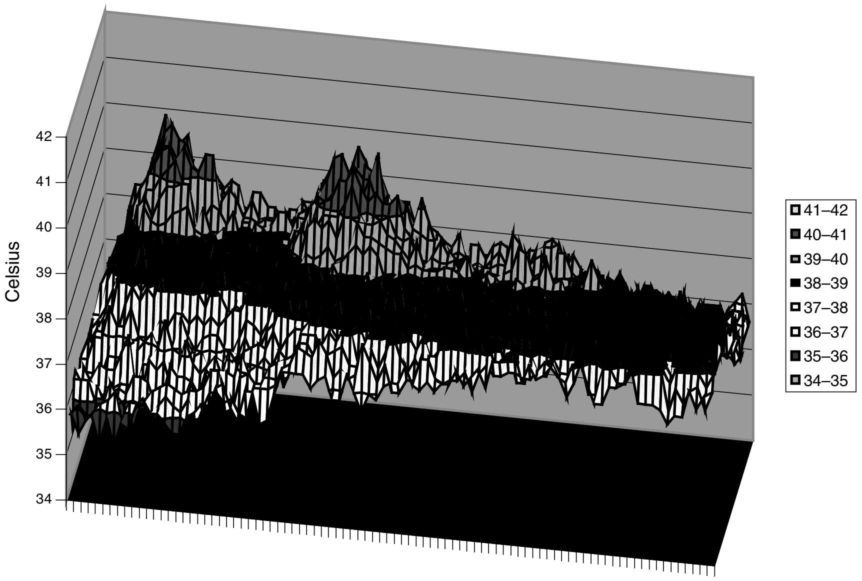
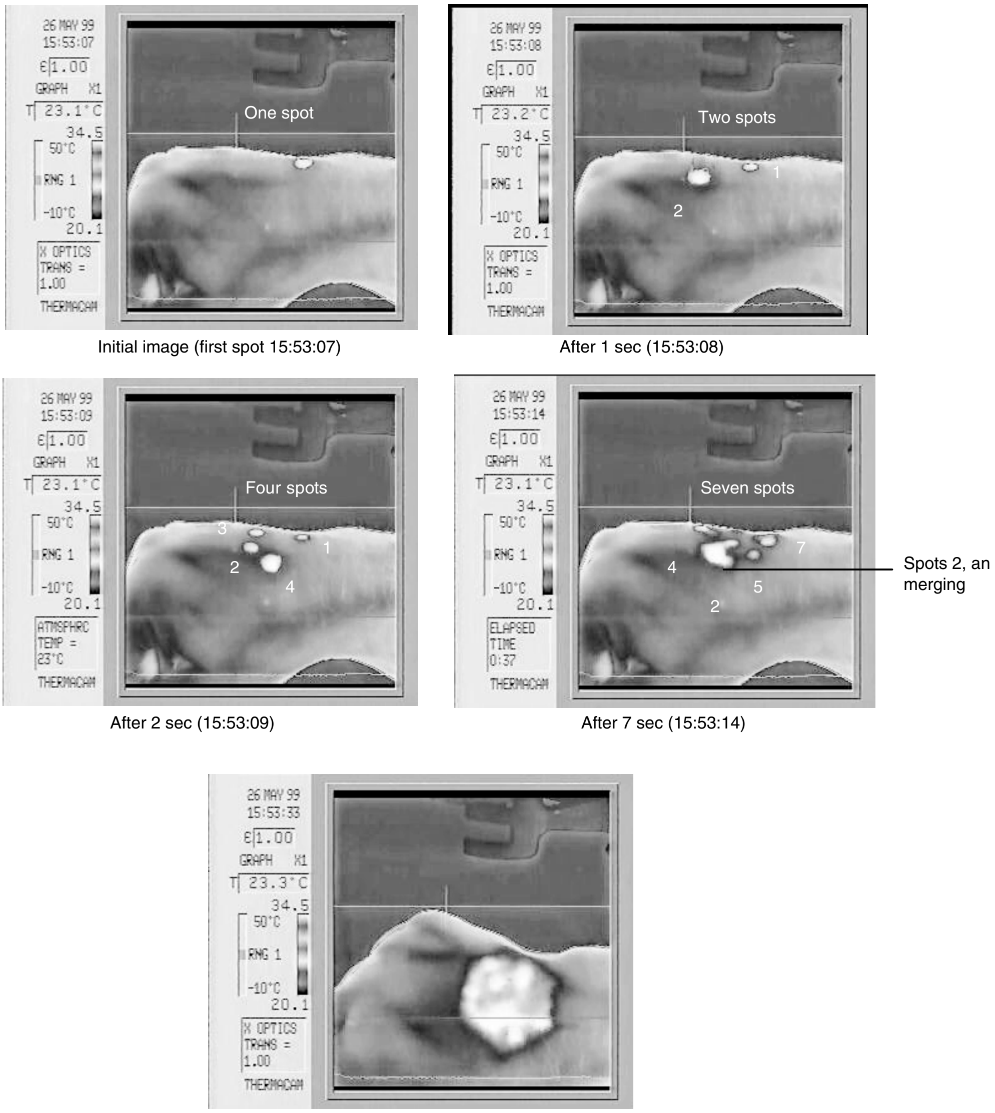


FIGURE 33.10 Three-dye laser spots, 2 sec apart with a 5 mm separation (585 nm at 5 J/cm^2, 5 mm spot).

There are some disadvantages that include the need for additional equipment and regulatory approvals for certain market sectors

A computerized scanning system has been developed [9] that illuminates the tissue in a pre-defined pattern. Sequential pulses are not placed adjacent to an immediately preceding pulse thereby ensuring the minimum of thermal build-up. Clement et al. [9] carried out a trial, illustrating treatment coverage using a hand-held system compared to a controlled computer scanning system. Two adjacent areas (lower arm) were selected and shaved. A marked hexagonal area was subjected to 19 shots using a hand-held system, and an adjacent area of skin was treated with a scanner whose computer control is designed to uniformly fill the area with exactly 19 shots. Such tests were repeated and the analyzed statistics showed that, on



Initial image (first spot 15:53:07)

After 1 sec (15:53:08)

After 2 sec (15:53:09)

After 7 sec (15:53:14)

After 26 sec

FIGURE 33.11 Sample sequences during computer laser scanning.

average, only 60% of area is covered by laser spots. The use of thermography allowed the validation and optimization of this automated system in a way that was impossible without thermal imaging technology. The following sequence of thermal images, Figure 33.11, captures the various stages of laser scanning of the hand using a dye laser at 5.7 J/cm^2. Thermography confirms that the spot temperature from individual laser beams will merge and that both the positioning of spots and the time duration between spots dictate the efficacy of treatment.

33.10.1 Case Study 1: Port Wine Stain

Vascular naevi are common and are present at birth or develop soon after. Superficial lesions are due to capillary networks in the upper or mid dermis, but larger angiomas can be located in the lower dermis and subcutis. An example of vascular naevi is the Port-Wine Stain (PWS) often present at birth, is an irregular

TABLE 33.4 Vasculature Treatment Types

Treatment type	Process	Possible concerns
Camouflage	Applying skin colored pigments to the surface of the skin. Enhancement to this technique is to tattoo skin colored inks into the upper layer of the lesion	Only a temporary measure and is very time consuming. Efficacy dependant on flatter lesions
Cryosurgery	Involves applying super-cooled liquid nitrogen to the lesion to destroy abnormal vasculature	May require several treatments
Excision	Common place where the lesion is endangering vital body functions	Not considered appropriate for purely cosmetic reasons. Complex operation resulting in a scar. Therefore, only applicable to the proliferating haemangioma lesion.
Radiation therapy	Bombarding the lesion with radiation to destroy vasculature	Induced number of skin cancer in a small number of cases
Drug therapy	Widely used administering steroids	Risk of secondary complications affecting bodily organs

red or purple macule which often affects one side of the face. Problems can arise if the naevus is located close to the eye and some cases where a PWS involves the trigeminal nerve's ophthalmic division may have an associated intracranial vascular malformation known as Sturge Weber Syndrome. The treatment of vascular naevi can be carried out a number of ways often dependent on the nature, type, anatomical and severity of lesion location, as highlighted in Table 33.4.

A laser wavelength of 585 nm is preferentially absorbed by haemoglobin within the blood, but there is partial absorption in the melanin rich basal layer in the epidermis. The objective is to thermally damage the blood vessel, by elevating its temperature, while ensuring that the skin surface temperature is kept low. For a typical blood vessel, the temperature–time graph appears similar to Figure 33.12. This suggests that it is possible to selectively destroy the PWS blood vessels, by elevating them to a temperature in excess of 100°C, causing disruption to the small blood vessels, whilst maintaining a safe skin surface temperature. This has been proven empirically via thermographic imaging with a laser pulsing protocol that was devised and optimized on the strength of Monte-Carlo based models [26] of the heat dissipation processes [11]. The two-dimensional Cartesian thermal transport equation is:

$$\nabla T^2 + \frac{Q(x,y)}{k} = \frac{1}{\alpha}\frac{\partial T}{\partial t} \tag{33.6}$$

where temperature T has both an implied spatial and temporal dependence and the volumetric source term, $Q(x,y)$, is obtained from the solution of the Monte-Carlo radiation transport problem [27].

33.10.2 Case Study 2: Laser Depilation

The 694 nm wavelength laser radiation is preferentially absorbed by melanin, which occurs in the basal layer and particularly in the hair follicle base, which is the intended target using an oblique angle of laser beam (see Figure 33.13). A Monte-Carlo analysis was performed in a similar manner to *Case Study 1* above, where the target region in the dermis is the melanin rich base of the hair follicle. Figures 33.14a,b show the temperature–time profiles for 10 and 20 J cm^2 laser fluence [23]. These calculations suggest that it is possible to thermally damage the melanin-rich follicle base whilst restricting the skin surface temperature to values that cause no superficial damage. Preliminary clinical trials indicated that there is indeed a beneficial effect, but the choice of laser parameters still required optimizing.

Thermographic analysis has proved indispensable in this work. Detailed thermometric analysis is shown in Figure 33.15a. Analysis of this data shows that in this case, the surface temperature is raised to

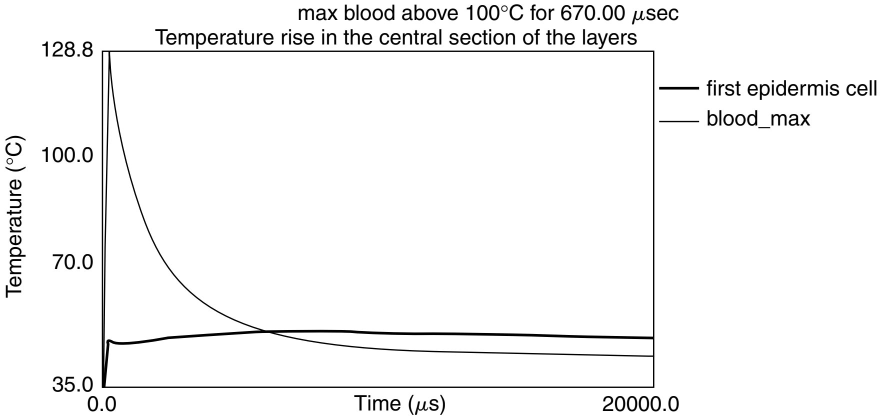


FIGURE 33.12 Typical temperatures for PWS problem, indicating thermal disruption of blood vessel, while skin surface temperature remains low.

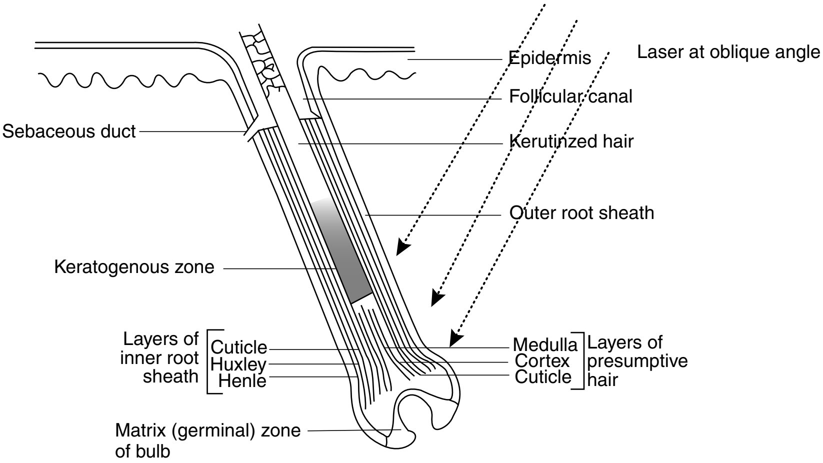


FIGURE 33.13 Oblique laser illumination of hair follicle.

about 50°C. The thermogram also clearly shows the selective absorption in the melanin-dense hair. The temperature of the hair is raised to over 207°C. This thermogram illustrates direct evidence for selective wavelength absorption leading to cell necrosis. Further clinical trials have indicated a maximum fluence of 15 J cm^2 for type III caucasian skin. Figure 33.15b illustrates a typical thermographic image obtained during the real-time monitoring.

33.11 Conclusions

The establishment, development, and consequential success of medical infrared thermographic (MIT) intervention with laser therapy is primarily based on the understanding of the following, that are described in more detail below:

1. Problem/condition to be monitored
2. Set-up and correct operation of infrared system (appropriate and validated training)
3. Appropriate conditions during the monitoring process
4. Evaluation of activity and development of standards and protocol

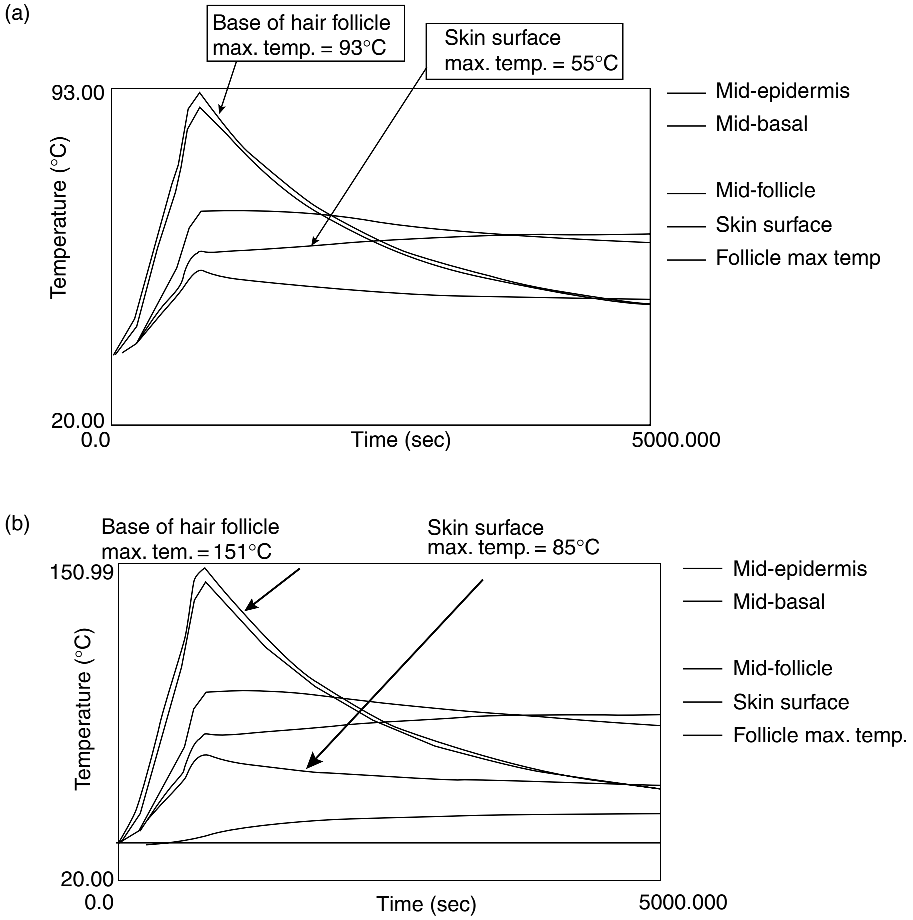


FIGURE 33.14 (a) Temperature–time profiles at 10 J cm^2 ruby (694 nm), 800 μsec laser pulse on caucasian skin type III. (b) Temperature–time profiles for 20 J cm^2 ruby (694 nm), 800 μsec laser pulse on Caucasian skin type III.

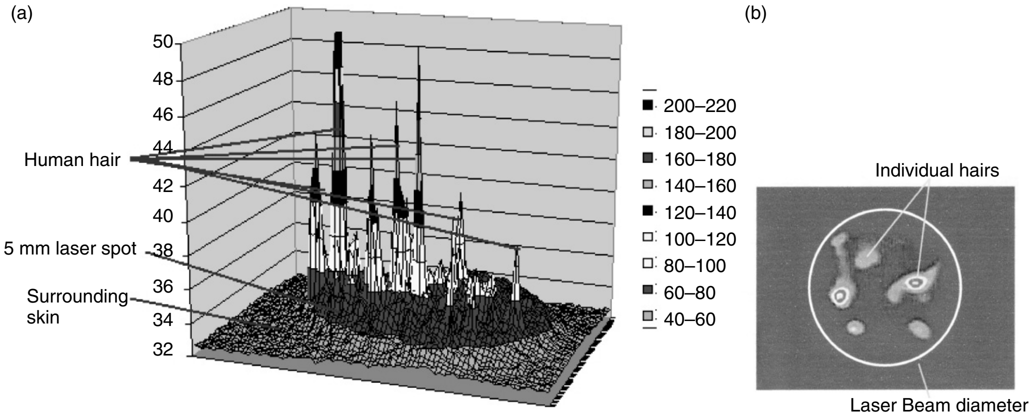


FIGURE 33.15 (a) Post-processed results of 5 mm. (b) Simplified thermogram diameter 694 nm 20 J cm^2 800 μsec ruby pulse of ruby laser pulse, with 5 mm spot at 20 J cm^2.

With reference to (1) above in the conclusions, the condition to be monitored, there needs to be a good knowledge as to the physiological aspects of the desired medical process; in laser therapy an understanding as to the mechanisms involved in laser–tissue interaction. A good reference source of current practice can be found in the Handbook of Optical Biomedical Diagnostics, published by The International Society for Optical Engineering (SPIE).

In this application fast data-capture (>50 Hz), good image quality (256 × 256 pixels), temperature sensitivity, and repeatability were considered important and an Inframetrics SC1000 Focal Plane Array Radiometer (3.4 to 5μm, CMOS PtSi Cooled Detector) with a real-time data acquisition system (Dynamite) was used. There are currently very fast systems available with data acquisition speeds in terms of hundreds of Hertz with detectors that provide excellent image quality. In (2) the critical aspect is training [21]. Currently, infrared equipment manufacturers design systems with multiple applications in mind. This has resulted in many aspects of good practice and quality standards. This is one of the reasons why industrial infrared thermography is so successful. This has not necessarily been the case in medicine. However, it is worth noting that there are a number of good infrared training organizations throughout the world, particularly in the United States. The advantages of adopting training organizations such as these is that they have experience of training with reference to a very wide range and type of infrared thermographic systems, in a number of different applications. This will help in the identification of the optimum infrared technology. In (3) consideration as to the conditions surrounding the patient and the room environment are important for optimum results. In the United Kingdom for example, Prof. Francis Ring, University of Glamorgan has led the way in the development and standardizations of clinical infrared practice [20]. Finally, (4) the evaluation of such practice is crucial if lessons are to be learnt and protocol and standards are to emerge.

Infrared thermal imaging provides an important tool for optimizing energized surgical interventions and facilitates validation of theoretical models of evolving temperature fields.

References

[1] Heniford, B.T., Matthews, B.D., Sing, R.F., Backus, C., Pratt, P., and Greene, F.L. (2001) Initial results with an electrothermal bipolar vessel sealer. *Surg. Endosc.* 15: 799–801.

[2] Palazzo, F.F., Francis, D.L., and Clifton, M.A. (2002) Randomised clinical trial of ligasure versus open haemorrhoidectomy. *Br. J. Surg.* 89, 154–157.

[3] Sengupta, S. and Webb, D.R. (2001) Use of a computer controlled bipolar diathermy system in radical prostatectomies and other open urological surgery. *ANZ J. Surg.* 71: 538–540.

[4] Schulze, S., Krztiansen, V.B., Fischer-Hansen, B., and Rosenberg, J. (2002) Sealing of the cystic duct with bipolar electrocoagulation. *Surg. Endosc.* 16: 342–344.

[5] Reidenbach, H.D. and Buess, G. (1992). Anciliary technology: electrocautery, thermoregulation and laser. In Cuschieri, A., Buess, G., and Perrisat, L. Eds., *Operative Manual of Endoscopic Surgery.* Springer-Verlag, Berlin-Heidelberg-New York, pp. 44–60.

[6] Kinoshita, T., Kanehira, E., Omura, K., Kawakami, K., and Watanabe, Y. (1999) Experimental study on heat production by a 23.5 kHz ultrasonically activated device for endoscopic surgery. *Surg. Endosc.* 13: 621–625.

[7] Andersen, R.R. and Parrish, J.A. (1981) Microvasculature can be selectively damaged using dye lasers. *Lasers Surg. Med.* 1: 263–270.

[8] Barlow, R.J., Walker, N.P.J., and Markey, A.C. (1996) Treatment of proliferative haemangiomas with 585nm pulsed dye laser. *Br. J. Dermatol.* 134: 700–704.

[9] Clement, R.M., Kiernan, M.N., Thomas, R.A., Donne, K.E., and Bjerring, P.J. (1999) The use of thermal imaging to optimise automated laser irradiation of tissue, *Skin Research and Technology.* Vol. 5, No. 2, *6th Congress of the International Society for Skin Imaging,* July 4–6, 1999, Royal Society London.

[10] Clement, R.M., Donne, K.D., Thomas, R.A., and Kiernan, M.N. (2000) Thermographic condition monitoring of human skin during laser therapy, *Quality Reliability Maintenance, 3rd International Conference*, St Edmund Hall, University of Oxford, 30–31 March 2000.

[11] Daniel, G. (2002) An investigation of thermal radiation and thermal transport in laser–tissue interaction, PhD Thesis, Swansea Institute.

[12] Garden, J.M., Polla, L.L., and Tan, O.T. (1988) Treatment of port wine stains by pulsed dye laser — analysis of pulse duration and long term therapy. *Arch. Dermatol.* 124: 889–896.

[13] Garden, J.M. and Bakus, W. (1996) Clinical efficacy of the pulsed dye laser in the treatment of vascular lesions. *J. Dermatol. Surg. Oncol.* 19: 321–326.

[14] Gault, D., Clement, R.M., Trow, R.B., and Kiernan, M.N. (1998) Removing unwanted hairs by laser. *Face* 6: 129–130.

[15] Glassberg, E., Lask, G., Rabinowitz, L.G., and Tunnessen, W.W. (1989) Capillary haemangiomas: case study of a novel laser treatment and a review of therapeutic options. *J. Dermatol. Surg. Oncol.* 15: 1214–1223.

[16] Grossman et al. (1997) Damage to hair follicle by normal mode ruby laser pulse. *J. Amer. Acad. Dermatol.* 889–894.

[17] Kiernan, M.N. (1997) An analysis of the optimal laser parameters necessary for the treatment of vascular lesions, PhD Thesis, The University of West of England.

[18] Lanigan, S.W. (1996) Port wine stains on the lower limb: response to pulsed dye laser therapy. *Clin. Exp. Dermatol.* 21: 88–92.

[19] Motley, R.J., Katugampola, G., and Lanigan, S.W. (1996) Microvascular abnormalities in port wine stains and response to 585 nm pulsed dye laser treatment. *Br. J. Dermatol.* 135: Suppl. 47: 13–14.

[20] Ring, E.F.J. (1995) History of thermography. In Ammer, K. and Ring, E.F.J., Eds., *The Thermal Image in Medicine and Biology.* Uhlen Verlag, Vienna, pp. 13–20.

[21] Thomas, R.A. (1999) *Thermography.* Coxmoor Publishers, Oxford, pp. 79–103.

[22] Thomas, R.A., Donne, K.E., Clement, R.M., and Kiernan, M. (2002) Optimised laser application in dermatology using infrared thermography, *Thermosense XXIV*, *Proceedings of SPIE*, April 1–4, Orlando, USA.

[23] Trow, R. (2001) The design and construction of a ruby laser for laser depilation, PhD Thesis, Swansea Institute.

[24] Welsh, A.J. and van Gemert, M.V.C. (1995) *Optical–Thermal Response of Laser-Irradiated Tissue.* Plenum Press, ISBN 0306449269.

[25] Wheeland, R.G. (1995) Clinical uses of lasers in dermatology. *Lasers Surg. Med.* 16: 2–23.

[26] Wilson, B.C. and Adam, G. (1983) A Monte Carlo model for the absorption and flux distributions of light in tissue. *Med. Phys. Biol.* 1.

[27] Donne, K.E. (1999) Two dimensional computer model of laser tissue interaction. Private communication.

34

Infrared Imaging Applied to Dentistry

Barton M. Gratt
University of Washington

34.1 The Importance of Temperature

Temperature is very important in all biological systems. Temperature influences the movement of atoms and molecules and their rates of biochemical activity. Active biological life is, in general, restricted to a temperature range of 0 to 45°C [1]. Cold-blooded organisms are generally restricted to habitats in which the ambient temperature remains between 0 and 40°C. However, a variety of temperatures well outside of this occurs on earth, and by developing the ability to maintain a constant body temperature, warm-blooded animals; for example, birds, mammals, including humans have gained access to a greater variety of habitats and environments [1].

With the application of common thermometers, elevation in the core temperature of the body became the primary indicator for the diagnosis of fever. Wunderlich introduced fever measurements as a routine procedure in Germany, in 1872. In 1930, Knaus inaugurated a method of basal temperature measurement, achieving full medical acceptance in 1952. Today, it is customary in hospitals throughout the world to take body temperature measurements on all patients [2].

The scientists of the first part of the 20th century used simple thermometers to study body temperatures. Many of their findings have not been superseded, and are repeatedly confirmed by new investigators using new more advanced thermal measuring devices. In the last part of the 20th century, a new discipline termed "thermology" emerged as the study of surface body temperature in both health and in disease [2].

34.2 The Skin and Skin-Surface Temperature Measurement

The skin is the outer covering of the body and contributes 10% of the body's weight. Over 30% of the body's temperature-receptive elements are located within the skin. Most of the heat produced within the body is dissipated by way of the skin, through radiation, evaporation, and conduction. The range of ambient temperature for thermal comfort is relatively broad (20 to 25°C). Thermal comfort is dependent upon humidity, wind velocity, clothing, and radiant temperature. Under normal conditions there is a steady flow of heat from the inside of a human body to the outside environment. Skin temperature distribution within specific anatomic regions; for example, the head vs. the foot, are diverse, varying by as much as ±15°C. Heat transport by convection to the skin surface depends on the rate of blood flow through the skin, which is also variable. In the trunk region of the body, blood flow varies by a factor of 7; at the foot, blood flow varies by a factor of 30; while at the fingers, it can vary by a factor of 600 [3].

It appears that measurements of body (core) temperatures and skin (surface) temperature may well be strong physiologic markers indicating health or disease. In addition, skin (surface) temperature values appear to be unique for specific anatomic regions of the body.

34.3 Two Common Types of Body Temperature Measurements

There are two common types of body temperature measurements that are made and utilized as diagnostic indicators.

1. *The Measurement of Body Core Temperature.* The normal core temperature of the human body remains within a range of 36.0 to 37.5°C [1]. The constancy of human core temperature is maintained by a large number of complex regulatory mechanisms [3]. Body core temperatures are easily measured orally (or anally) with contacting temperature devices including: manual or digital thermometers, thermistors, thermocouples, and even layers of liquid temperature sensitive crystals, etc. [4–6].
2. *The Measurement of Body Surface Temperature.* While body core temperature is very easy to measure, the body's skin surface temperature is very difficult to measure. Any device that is required to make contact with the skin cannot measure the body's skin surface temperature reliably. Since skin has a relatively low heat capacity and poor lateral heat conductance, skin temperature is likely to change on contact with a cooler or warmer object [2]. Therefore, an indirect method of obtaining skin surface temperature is required, a common thermometer on the skin, for example, will not work.

Probably the first research efforts that pointed out the diagnostic importance of the infrared emission of human skin and thus initiated the modern era of thermometry were the studies of Hardy in 1934 [7,8]. However, it took 30 years for modern thermometry to be applied in laboratories around the world. To conduct noncontact thermography of the human skin in a clinical setting, an advanced computerized infrared imaging system is required. Consequently, clinical thermography required the advent of microcomputers developed in the late 1960s and early 1970s. These sophisticated electronic systems employed advanced microtechnology, requiring large research and development costs.

Current clinical thermography units use single detector infrared cameras. These work as follows: infrared radiation emitted by the skin surface enters the lens of the camera, passes through a number of rapidly spinning prisms (or mirrors), which reflect the infrared radiation emitted from different parts of the field of view onto the infrared sensor. The sensor converts the reflected infrared radiation into electrical signals. An amplifier receives the electric signals from the sensor and boosts them to electric potential signals of a few volts that can be converted into digital values. These values are then fed into a computer. The computer uses this input, together with the timing information from the rotating mirrors, to reconstruct a digitized thermal image from the temperature values of each small area within the field of observation. These digitized images are easily viewed and can be analyzed using computer software and stored on a computer disk for later reference.

34.4 Diagnostic Applications of Thermography

In 1987, the *International Bibliography of Medical Thermology* was published and included more than 3000 cited publications on the medical use of thermography, including applications for anesthesiology, breast disease, cancer, dermatology, gastrointestinal disorders, gynecology, urology, headache, immunology, musculoskeletal disorders, neurology, neurosurgery, ophthalmology, otolaryngology, pediatrics, pharmacology, physiology, pulmonary disorders, rheumatology, sports medicine, general surgery, plastic and reconstructive surgery, thyroid, cardiovascular and cerebrovascular, vascular problems, and veterinary medicine [9]. In addition, changes in human skin temperature has been reported in conditions involving the orofacial complex, as related to dentistry, such as the temporomandibular joint [10–25], and nerve damage and repair following common oral surgery [25–27]. Thermography has been shown not to be useful in the assessment of periapical granuloma [28]. Reports of dedicated controlled facial skin temperature studies of the orofacial complex are limited, but follow findings consistent with other areas of the body [29,30].

34.5 The Normal Infrared Facial Thermography

The pattern of heat dissipation over the skin of the human body is normally symmetrical and this includes the human face. It has been shown that in normal subjects, the difference in skin temperature from side-to-side on the human body is small, about 0.2°C [31]. Heat emission is directly related to cutaneous vascular activity, yielding enhanced heat output on vasodilatation and reduced heat output on vasoconstriction. Infrared thermography of the face has promise, therefore, as a harmless, noninvasive, diagnostic technique that may help to differentiate selected diagnostic problems. The literature reports that during clinical studies of facial skin temperature a significant difference between the absolute facial skin temperatures of men vs. women was observed [32]. Men were found to have higher temperatures over all 25 anatomic areas measured on the face (e.g., the orbit, the upper lip, the lower lip, the chin, the cheek, the TMJ, etc.) than women. The basal metabolic rate for a normal 30-year-old male, 1.7 m tall (5 ft, 7 in.), weighing 64 kg (141 lbs), who has a surface area of approximately 1.6 m^2, is approximately 80 W; therefore, he dissipates about 50 W/m^2 of heat [33]. On the other hand, the basal metabolic rate of a 30-year-old female, 1.6 m tall (5 ft, 3 in.), weighing 54 kg (119 lbs), with a surface area of 1.4 m^2, is about 63 W, so that she dissipates about 41 W/m^2 of heat [33,34]. Assuming that there are no other relevant differences between males and females, women's skin is expected to be cooler, since less heat is lost per unit (per area of body surface). Body heat dissipation through the face follows this prediction. In addition to the effect of gender on facial temperature, there are indications that age and ethnicity may also affect facial temperature [32].

When observing patients undergoing facial thermography, there seems to be a direct correlation between vasoactivity and pain, which might be expected since both are neurogenic processes. Differences in facial skin temperature, for example, asymptomatic adult subjects (low temperatures differences) and adult patients with various facial pain syndromes (high temperature differences) may prove to be a useful criterion for the diagnosis of many conditions [35]. Right- vs. left-side temperature differences (termed: delta T or ΔT) between many specific facial regions in normal subjects were shown to be low (<0.3°C) [40], while similar ΔT values were found to be high (>0.5°C) in a variety of disorders related to dentistry [35].

34.6 Abnormal Facial Conditions Demonstrated with Infrared Facial Thermography

34.6.1 Assessing Temporomandibular Joint (TMJ) Disorders with Infrared Thermography

It has been shown that normal subjects have symmetrical thermal patterns over the TMJ regions of their face. Normal subjects had ΔT values of 0.1°C (±0.1°C) [32,36]. On the other hand, TMJ pain patients were found to have asymmetrical thermal patterns, with increased temperatures over the affected TMJ region, with ΔT values of +0.4°C (±0.2°C) [37]. Specifically, painful TMJ patients with internal derangement and painful TMJ osteoarthritis were both found to have asymmetrical thermal patterns and increased temperatures over the affected TMJ, with mean area TMJ ΔT of +0.4°C (±0.2°C) [22,24]. In other words, the correlation between TMJ pain and hyper perfusion of the region seems to be independent of the etiology of the TMJ disorder (osteoarthritis vs. internal derangement). In addition, a study of mild-to-moderate TMD (temporomandibular joint dysfunction) patients indicated that area ΔT values correlated with the level of the patient's pain symptoms [38]. And a more recent double-blinded clinical study compared active orthodontic patients vs. TMD patients vs. asymptomatic TMJ controls, and showed average ΔT values of +0.2, +0.4, and +0.1°C; for these three groups respectively. This study showed that thermography could distinguish between patients undergoing active orthodontic treatment and patients with TMD [39].

34.6.2 Assessing Inferior Alveolar Nerve (IAN) Deficit with Infrared Thermography

The thermal imaging of the chin has been shown to be an effective method for assessing inferior alveolar nerve deficit [40]. Whereas normal subjects (those without inferior alveolar nerve deficit) show a symmetrical thermal pattern, (ΔT of +0.1°C [±0.1°C]); patients with inferior alveolar nerve deficit had elevated temperature in the mental region of their chin (ΔT of +0.5°C [±0.2°C]) on the affected side [41]. The observed vasodilatation seems to be due to blockage of the vascular neuronal vasoconstrictive messages, since the same effect on the thermological pattern could be invoked in normal subjects by temporary blockage of the inferior alveolar nerve, using a 2% lidocaine nerve block injection [42].

34.6.3 Assessing Carotid Occlusal Disease with Infrared Thermography

The thermal imaging of the face, especially around the orbits, has been shown to be an effective method for assessing carotid occlusal disease. Cerebrovascular accident (CVA), also called stroke, is well known as a major cause of death. The most common cause of stroke is atherosclerosotic plaques forming emboli, which travel within vascular blood channels, lodging in the brain, obstructing the brain's blood supply, resulting in a cerebral vascular accident (or stroke). The most common origin for emboli is located in the lateral region of the neck where the common carotid artery bifurcates into the internal and the external carotid arteries [43,44]. It has been well documented that intraluminal carotid plaques, which both restrict and reduce blood flow, result in decreased facial skin temperature [43–54]. Thermography has demonstrated the ability to detect a reduction of 30% (or more) of blood flow within the carotid arteries [55]. Thermography shows promise as an inexpensive painless screening test of asymptomatic elderly adults at risk for the possibility of stroke. However, more clinical studies are required before thermography may be accepted for routine application in screening toward preventing stroke [55,56].

34.6.4 Additional Applications of Infrared Thermography

Recent clinical studies assessed the application of thermography on patients with chronic facial pain (orofacial pain of greater than 4 month's duration). Thermography classified patients as being "normal" when selected anatomic ΔT values ranged from 0.0 to ±0.25°C, and "hot" when ΔT values were >+0.35°C, and "cold" when area ΔT values were $<-0.35°C$. The study population consisted of 164 dental pain patients and 164 matched (control) subjects. This prospective, matched study determined that subjects classified with "hot" thermographs had the clinical diagnosis of (1) sympathetically maintained pain, (2) peripheral nerve-mediated pain, (3) TMJ arthropathy, or (4) acute maxillary sinusitis. Subjects classified with "cold" areas on their thermographs were found to have the clinical diagnosis of (1) peripheral nerve-mediated pain, or (2) sympathetically independent pain. Subjects classified with "normal" thermographs included patients with the clinical diagnosis of (1) cracked tooth syndrome, (2) trigeminal neuralgia, (3) pretrigeminal neuralgia, or (4) psychogenic facial pain. This new system of thermal classification resulted in 92% (301 or 328) agreement in classifying pain patients vs. their matched controls. In brief, ΔT has been shown to be within $\pm 0.4°C$ in normal subjects, while showing values greater than $+0.7°C$ and less than $-0.6°$ C in abnormal facial pain patients [10], making "ΔT" an important diagnostic parameter in the assessment of orofacial pain [35].

34.6.5 Future Advances in Infrared Imaging

Over the last 20 years there have been additional reports in the dental literature giving promise to new and varied applications of infrared thermography [57–63]. While, infrared thermography is promising, the future holds even greater potential for temperature measurement as a diagnostic tool, the most promising being termed dynamic area telethermometry (DAT) [64,65]. Newly developed DAT promises to become a new more advanced tool providing quantitative information on the thermoregulatory frequencies (TRFs) manifested in the modulation of skin temperature [66]. Whereas the static thermographic studies discussed above demonstrate local vasodilatation or vasoconstriction, DAT can identify the mechanism of thermoregulatory frequencies and thus it is expected, in the future, to significantly improve differential diagnosis [66].

In summary, the science of thermology, including static thermography, and soon to be followed by DAT, appears to have great promise as an important diagnostic tool in the assessment of orofacial health and disease.

Acknowledgments

This chapter is dedicated to Professor Michael Anbar, of Buffalo, New York: A brilliant scientist, my thermal science mentor, and "The Father of Dynamic Area Telethermography."

References

[1] Grobklaus, R. and Bergmann, K.E. Physiology and regulation of body temperature. In *Applied Thermology: Thermologic Methods.* J.-M. Engel, U. Fleresch, and G. Stuttgen, Eds., Federal Republic of Germany: VCH (1985), pp. 11–20.

[2] *Applied Thermology: Thermologic Methods.* J.-M. Engel, U. Fleresch, and G. Stuttgen, Eds., Federal Republic of Germany: VCH (1985), pp. 11–20.

[3] Kirsch, K.A. Physiology of skin-surface temperature. In *Applied Thermology: Thermologic Methods.* J.-M. Engel, U. Fleresch, and G. Stuttgen, Eds., Federal Republic of Germany: VCH (1985), pp. 1–9.

[4] Anbar, M., Gratt, B.M., and Hong, D. Thermology and facial telethermography: part I. History and technical review. *Dentomaxillofac. Radiol.* (1998) 27: 61–67.

[5] Anbar, M. and Gratt, B.M. Role of nitric oxide in the physiopathology of pain. *J. Musc. Skeletal Joint Pain* (1997) 14: 225–254.

[6] Rost, A. Comparative measurements with an infrared and contact thermometer for thermal stress reaction. In *Thermological Methods.* J.-M. Engel, U. Flesch, and G. Stuttgen, Eds., Weinheim: VCH Verlag (1985), pp. 169–170.

[7] Hardy, J.D. The radiation of heat from the human body: I–IV. *J. Clin. Invest.* (1934) 13: 593–620.

[8] Hardy, J.D. The radiation of heat from the human body: I–IV. *J. Clin. Invest.* (1934) 13: 817–883.

[9] Abernathy, M. and Abernathy, T.B. International bibliography of thermology. *Thermology.* (1987) 2: 1–533.

[10] Berry, D.C. and Yemm, R. Variations in skin temperature of the face in normal subjects and in patients with mandibular dysfunction. *Br. J. Oral Maxillofac. Surg.* (1971) 8: 242–247.

[11] Berry, D.C. and Yemm, R. A further study of facial skin temperature in patients with mandibular dysfunction. *J. Oral Rehabil.* (1974) 1: 255–264.

[12] Kopp, S. and Haraldson, T. Normal variations in skin temperature of the face in normal subjects and in patients with mandibular dysfunction. *Br. J. Oral Maxillofac. Surg.* (1983) 8: 242–247.

[13] Johansson, A., Kopp, S., and Haraldson, T. Reproducibility and variation of skin surface temperature over the temporomandibular joint and masseter muscle in normal individuals. *Acta Odontol. Scand.* (1985) 43: 309–313.

[14] Tegelberg, A. and Kopp, S. Skin surface temperature over the temporo-mandibular and metacarpophalangeal joints in individuals with rheumatoid arthritis. *Odontol. Klin.*, Box 33070, 400 33 Goteborg, Sweden (1986) Report No. 31, pp. 1–31.

[15] Akerman, S. et al. Relationship between clinical, radiologic and thermometric findings of the temporomandibular joint in rheumatoid arthritis. *Odontol. Klin.*, Box 33070, 400 33 Goteborg, Sweden (1987) Report No. 41, pp. 1–30.

[16] Finney, J.W., Holt, C.R., and Pearce, K.B. Thermographic diagnosis of TMJ disease and associated neuromuscular disorders. *Special Report: Postgraduate Medicine* (March 1986), pp. 93–95.

[17] Weinstein, S.A. Temporomandibular joint pain syndrome — the whiplash of the 1980s, *Thermography and Personal Injury Litigation, Ch. 7.* S.D. Hodge, Jr., Ed., New York, USA: John Wiley & Sons (1987), pp. 157–164.

[18] Weinstein, S.A., Gelb, M., and Weinstein, E.L. Thermophysiologic anthropometry of the face in home sapiens. *J. Craniomand. Pract.* (1990) 8: 252–257.

[19] Pogrel, M.A., McNeill, C., and Kim, J.M. The assessment of trapezius muscle symptoms of patients with temporomandibular disorders by the use of liquid crystal thermography. *Oral Surg. Oral Med. Oral. Pathol. Oral Radiol. Endod.* (1996) 82: 145–151.

[20] Steed, P.A. The utilization of liquid crystal thermography in the evaluation of temporomandibular dysfunction. *J. Craniomand. Pract.* (1991) 9: 120–128.

[21] Gratt, B.M., Sickles, E.A., Graff-Radford, S.B., and Solberg, W.K. Electronic thermography in the diagnosis of atypical odontalgia: a pilot study. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod.* (1989) 68: 472–481.

[22] Gratt, B.M. et al. Electronic thermography in the assessment of internal derangement of the TMJ. *J. Orofacial Pain* (1994) 8: 197–206.

[23] Gratt, B.M., Sickles, E.A., Ross, J.B., Wexler, C.E., and Gornbein, J.A. Thermographic assessment of craniomandibular disorders: diagnostic interpretation versus temperature measurement analysis. *J. Orofacial Pain* (1994) 8: 278–288.

[24] Gratt, B.M., Sickles, E.A., and Wexler, C.E. Thermographic characterization of osteoarthrosis of the temporomandibular joint. *J. Orofacial Pain* (1994) 7: 345–353.

[25] Progrell, M. A., Erbez, G., Taylor, R.C., and Dodson, T.B. Liquid crystal thermography as a diagnostic aid and objective monitor for TMJ dysfunction and myogenic facial pain. *J. Craniomand. Disord. Facial Oral Pain* (1989) 3: 65–70.

[26] Dmutpueva, B.C., and Alekceeva, A.H. Applications of thermography in the evaluation of the postoperative patient. *Stomatologiia* (1986) 12: 29–30 (Russian).

[27] Cambell, R.L., Shamaskin, R.G., and Harkins, S.W. Assessment of recovery from injury to inferior alveolar and mental nerves. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod.* (1987) 64: 519–526.

[28] Crandall, C.E. and Hill, R.P. Thermography in dentistry: a pilot study. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod.* (1966) 21: 316–320.

[29] Gratt, B.M., Pullinger, A., and Sickles, E.A. Electronic thermography of normal facial structures: A pilot study. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod.* (1989) 68: 346–351.

[30] Weinstein, S.A., Gelb, M., and Weinstein, E.L. Thermophysiologic anthropometry of the face in homo sapiens. *J. Craniomand. Pract.* (1990) 8: 252–257.

[31] Uematsu, S. Symmetry of skin temperature comparing one side of the body to the other. *Thermology* (1985) 1: 4–7.

[32] Gratt, B.M. and Sickles, E.A. Electronic facial thermography: an analysis of asymptomatic adult subjects. *J. Orofacial Pain* (1995) 9: 222–265.

[33] Blaxter, K. Energy exchange by radiation, convection, conduction and evaporation. In *Energy Metabolism in Animals and Man.* New York: Cambridge University Press (1989), pp. 86–99.

[34] Blaxter, K. The minimal metabolism. In *Energy Metabolism in Animals and Man.* New York: Cambridge University Press (1989) 120–146.

[35] Gratt, B.M, Graff-Radford, S.B., Shetty, V., Solberg, W.K., and Sickles, E.A. A six-year clinical assessment of electronic facial thermography. *Dentomaxillofac. Radiol.* (1996) 25: 247–255.

[36] Gratt, B.M., and Sickles, E.A. Thermographic characterization of the asymptomatic TMJ. *J. Orofacial Pain* (1993) 7: 7–14.

[37] Gratt, B.M., Sickles, E.M., and Ross, J.B. Thermographic assessment of craniomandibular disorders: diagnostic interpretation versus temperature measurement analysis. *J. Orofacial Pain* (1994) 8: 278–288.

[38] Canavan, D. and Gratt, B.M. Electronic thermography for the assessment of mild and moderate TMJ dysfunction. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod.* (1995) 79: 778–786.

[39] McBeth, S.A., and Gratt, B.M. A cross-sectional thermographic assessment of TMJ problems in orthodontic patients. *Am. J. Orthod. Dentofac. Orthop.* (1996) 109: 481–488.

[40] Gratt, B.M., Shetty, V., Saiar, M., and Sickles, E.A. Electronic thermography for the assessment of inferior alveolar nerve deficit. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod.* (1995) 80: 153–160.

[41] Gratt, B.M., Sickles, E.A., and Shetty, V. Thermography for the clinical assessment of inferior alveolar nerve deficit: a pilot study. *J. Orofacial Pain* (1994) 80: 153–160.

[42] Shetty, V., Gratt, B.M., and Flack, V. Thermographic assessment of reversible inferior alveolar nerve deficit. *J. Orofacial Pain* (1994) 8: 375–383.

[43] Wood, E.H. Thermography in the diagnosis of cerebrovascular disease: preliminary report. *Radiology* (1964) 83: 540–546.

[44] Wood, E.H. Thermography in the diagnosis of cerebrovascular disease. *Radiology* (1965) 85: 207–215.

[45] Steinke, W., Kloetzsch, C., and Hennerici, M. Carotid artery disease assessed by color Doppler sonography and angiography. *AJR* (1990) 154: 1061–1067.

[46] Hu, H.-H. et al. Color Doppler imaging of orbital arteries for detection of carotid occlusive disease. *Stroke* (1993) 24: 1196–1202.

[47] Carroll, B.A., Graif, M., and Orron, D.E. Vascular ultrasound. In *Peripheral Vascular Imaging and Intervention.* D. Kim and D.E. Orron, Eds., St. Louis, MO, Mosby/Year Book (1992), pp. 211–225.

[48] Mawdsley, C., Samuel, E., Sumerling, M.D., and Young, G.B. Thermography in occlusive cerebrovascular diseases. *Br. Med. J.* (1968) 3: 521–524.

[49] Capistrant, T.D. and Gumnit, R.J. Thermography and extracranial cerebrovascular disease: a new method to predict the stroke-prone individual. *Minn. Med.* (1971) 54: 689–692.

[50] Karpman, H.L., Kalb, I.M., and Sheppard, J.J. The use of thermography in a health care system for stroke. *Geriatrics* (1972) 27: 96–105.

[51] Soria, E. and Paroski, M.W. Thermography as a predictor of the more involved side in bilateral carotid disease: case history. *Angiology* (1987) 38: 151–158.

[52] Capistrat, T.D. and Gumnit, R.J. Detecting carotid occlusive disease by thermography. *Stroke* (1973) 4: 57–65.

[53] Abernathy, M., Brandt, M.M., and Robinson, C. Noninvasive testing of the carotid system. *Am. Fam. Physic.* (1984) 29: 157–164.

[54] Dereymaeker, A., Kams-Cauwe, V., and Fobelets, P. Frontal dynamic thermography: improvement in diagnosis of carotid stenosis. *Eur. Neurol.* (1978) 17: 226–234.

[55] Gratt, B.M., Halse, A., and Hollender, L. A pilot study of facial infrared thermal imaging used as a screening test for detecting elderly individuals at risk for stroke. *Thermol. Int.* (2002) 12: 7–15.

[56] Friedlander A.H. and Gratt B.M. Panoramic dental radiography and thermography as an aid in detecting patients at risk for stroke. *J. Oral Maxillofac. Surg.* (1994) 52: 1257–1262.

[57] Graff-Radford, S.B., Ketalaer, M.-C., Gratt, B.M., and Solberg, W.K. Thermographic assessment of neuropathic facial pain: a pilot study. *J. Orofacial Pain* (1995) 9: 138–146.

[58] Pogrel, M.A., Erbez, G., Taylor, R.C., and Dodson, T.B. Liquid crystal thermography as a diagnostic aid and objective monitor for TMJ dysfunction and myogenic facial pain. *J. Craniobandib. Disord. Facial Oral Pain* (1989) 3: 65–70.

[59] Pogrel, M.A., Yen, C.K., and Taylor, R.C. Infrared thermography in oral and maxillo-facial surgery. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod.* (1989) 67: 126–131.

[60] Graff-Radford, S.B., Ketlaer, M.C., Gratt, B.M., and Solberg, W.K. Thermographic assessment of neuropathic facial pain. *J. Orofacial Pain* (1995) 9: 138–146.

[61] Biagioni, P.A., Longmore, R.B., McGimpsey, J.G., and Lamey, P.J. Infrared thermography: its role in dental research with particular reference to craniomandibular disorders. *Dentomaxillofac. Radiol.* (1996) 25: 119–124.

[62] Biagioni, P.A., McGimpsey, J.G., and Lamey, P.J. Electronic infrared thermography as a dental research technique. *Br. Dent. J.* (1996) 180: 226–230.

[63] Benington, I.C., Biagioni, P.A., Crossey, P.J., Hussey, D.L., Sheridan, S., and Lamel, P.J. Temperature changes in bovine mandibular bone during implant site preparation: an assessment using infra-red thermography. *J. Dent.* (1996) 24: 263–267.

[64] Anbar, M. Clinical applications of dynamic area telethermography. In *Quantitative Dynamic Telthermography in Medical Diagnosis.* CRC Press: Boca Raton, FL (1994), pp. 147–180.

[65] Anbar, M. Dynamic area telethermography and its clinical applications. *SPIE Proc.* (1995) 2473: 3121–3323.

[66] Anbar M., Grenn, M.W., Marino, M.T., Milescu, L., and Zamani, K. Fast dynamic area telethermography (DAT) of the human forearm with a Ga/As quantum well infrared focal plane array camera. *Eur. J. Therol.* (1997) 7: 105–118.

35

Use of Infrared Imaging in Veterinary Medicine

Ram C. Purohit
Auburn University

Tracy A. Turner
Private Practice

David D. Pascoe
Auburn University

35.1 Historical Perspective

In the mid-1960s and early 1970s, several studies were published indicating the value of IR (infrared) thermography in veterinary medicine [1–3]. In the 1965 research of Delahanty and George [2], the thermographic images required at least 6 min to produce a thermogram, a lengthy period of time during which the veterinarian had to keep the horse still while the scan was completed. This disadvantage was overcome by the development of high speed scanners using rotating IR prisms which then could produce instantaneous thermograms.

Stromberg [4–6] and Stromberg and Norberg [7] used thermography to diagnose inflammatory changes of the superficial digital flexor tendons in race horses. With thermography, they were able to document and detect early inflammation of the tendon, 1 to 2 weeks prior to the detection of lameness using clinical examination. They suggested that thermography could be used for early signs of pending lameness and it could be used for preventive measures to rest and treat race horses before severe lameness became obvious on physical examination.

In 1970, the Horse Protection Act was passed by the United States Congress to ban the use of chemical or mechanical means of "soring" horses. It was difficult to enforce this act because of the difficulty in obtaining measurable and recordable proof of violations. In 1975, Nelson and Osheim [8] documented that soring caused by chemical or mechanical means on the horse's digit could be diagnosed as having a definite abnormal characteristic IR emission pattern in the affected areas of the limb. Even though thermography at that time became the technique of choice for the detection of soring, normal thermography patterns in horses were not known. This prompted the USDA to fund research for the uses of thermography in veterinary medicine.

Purohit et al. [9] established a protocol for obtaining normal thermographic patterns of the horses' limbs and other parts of the body. This protocol was regularly used for early detection of acute and chronic inflammatory conditions in horses and other animal species. Studies at Auburn University vet school used an AGA 680 liquid cooled thermography system that had a black and white and an accessory color display units that allows the operator to assign the array of ten isotherms to temperature increments from 0.2 to 10.0°C. Images were captured within seconds rather than the 6 min required for earlier machines. In veterinary studies at Auburn University, the thermographic isotherms were imaged with nine colors and white assigned to each isotherm that varied in temperature between either 0.5 or 1.0°C.

In a subsequent study, Purohit and McCoy [10] established normal thermal patterns (temperature and gradients) of the horse, with special attention directed towards thoracic and pelvic limbs. Thermograms of various parts of the body were obtained 30 min before and after the exercise for each horse. Thermographic examination was also repeated for each horse on six different days. Thermal patterns and gradients were similar in all horses studied with a high degree of right to left symmetry in IR emission.

At the same time, Turner et al. [11] investigated the influence of the hair coat and hair clipping. This study demonstrated that the clipped leg was always warmer. After exercise, both clipped and unclipped legs had similar increases in temperature. The thermal patterns and gradients were not altered by clipping and/or exercise [10,11]. This indicated that clipping hair in horses with even hair coats was not necessary for thermographic evaluation. However, in some areas where the hair is long hair and not uniform, clipping may be required. Recently, concerns related to hair coat, thermographic imaging, and temperature regulation were investigated in llamas exposed to the hot humid conditions of the southeast [12]. While much of the veterinary research has focused on the thermographic imaging as a diagnostic tool, this study expanded its use into the problems of thermoregulation in various non endemic species.

Current camera technology has improved scanning capabilities that are combined with computer-assisted software programs. This new technology provides the practitioner with numerous options for image analysis, several hundred isotherms capable of capturing temperature differences in the hundredths of a degree Celsius, and better image quality. Miniaturized electronics have reduced the size of the units, allowing some systems to be housed in portable hand-held units. With lower cost of equipment, more thermographic equipment are being utilized in human and animal veterinary medicine and basic physiology studies.

It was obvious from initial studies by several authors that standards needed to be established for obtaining reliable thermograms in different animal species. The variations in core temperature and differences in the thermoregulatory mechanism responses between species emphasizes the importance of individually established norms for thermographic imagery.

A further challenge in veterinary medicine may occur when animal patient care may necessitate outdoor imaging.

35.2 Standards for Reliable Thermograms

Thermography provides an accurate, quantifiable, noncontact, noninvasive measure and map of skin surface temperatures. Skin surface temperatures are variable and change according to blood flow regulation to the skin surface. As such, IR thermography practitioner must be aware of the internal and external influences that alter this dynamic process of skin blood flow and temperature regulation. While imaging equipment can vary widely in price, these differences are often reflective of the wave-length capturing capability of the detectors and adjunct software that can aid in image analysis. The thermographer needs to understand the limitations of their IR system in order to make appropriate interpretations of their data. There have been some published studies that have not adhered to reliable standards and equipment prerequisites, thereby detracting from the acceptance of thermography as a valuable research and clinical technique. In some cases a simple cause–effect relationship was assumed to demonstrate the diagnosis of a disease or syndrome based on thermal responses as captured by thermographic images.

Internal and external factors have a significant effect on the skin surface temperature. Therefore, the use of thermography to evaluate skin surface thermal patterns and gradient requires an understanding of the dynamic changes which occur in blood flow at systemic, peripheral, regional, and local levels [9,10]. Thus, to enhance the diagnostic value of thermography, we recommend the following standards for veterinary medical imaging:

1. The environmental factors which interfere with the quality of thermography should be minimized. The room temperature should be maintained between 21 and 26°C. Slight variations in some cases may be acceptable, but room temperature should always be cooler than the animal's body temperature and free from air drafts.
2. Thermograms obtained outdoors under conditions of direct air drafts, sunlight, and extreme variations in temperature may provide unreliable thermograms in which thermal patterns are altered. Such observations are meaningless as a diagnostic tool.
3. When an animal is brought into a temperature controlled room, it should be equilibrated at least 20 min or more, depending on the external temperature from which the animal was transported. Animals transported from extreme hot or cold environments may require up to 60 min of equilibration time. Equilibration time is adequate when the thermal temperatures and patterns are consistently maintained over several minutes.
4. Other factors affecting the quality of thermograms are exercise, sweating, body position and angle, body covering, systemic and topical medications, regional and local blocks, sedatives, tranquilizers, anesthetics, vasoactive drugs, skin lesions such as scars, surgically altered areas, etc. As stated prior, the hair coat may be an issue with uneven hair length or a thick coat.
5. It is recommended that the infrared imaging should be performed using an electronic non contact cooled system. The use of long wave detectors is preferable.

The value of thermography is demonstrated by the sensitivity to changes in heat on the skin surface and its ability to detect temporal and spatial changes in thermal skin responses that corresponds to temporal and spatial changes in blood flow. Therefore, it is important to have well documented normal thermal patterns and gradients in all species under controlled environments prior to making any claims or detecting pathological conditions.

35.3 Dermatome Patterns of Horses and Other Animal Species

Certain chronic and acute painful conditions associated with peripheral neurovascular and neuromuscular injuries are easy to confuse with spinal injuries associated with cervical, thoracic, and lumbar-sacral areas [13,14]. Similarly, inflammatory conditions such as osteoarthritis, tendonitis, and other associated conditions may also be confused with other neurovascular conditions. Thus, studies have been done over the last 25 years at Auburn University to map cutaneous and differentiate the sensory-sympathetic dermatome patterns of cervical, thoracic, and lumbosacral regions in horses [13,14]. Infrared thermography was used to map the sensory-sympathetic dermatome in horses. The dorsal or ventral spinal nerve(s) were blocked with 0.5% of mepevacine as a local anesthetic. The sensory sympathetic spinal nerve block produced two effects. First, blocking the sympathetic portion of the spinal nerve caused increased thermal patterns and produced sweating of the affected areas. Second, the areas of insensitivity produced by the sensory portion of the block were mapped and compared with the thermal patterns. The areas of insensitivity were found to correlate with the sympathetic innervations.

Thermography was used to provide thermal patterns of various dermatome areas from cervical areas to epidural areas in horses. Clinical cases of cervical area nerve compression provided cooler thermal patterns, away from the site of injuries. In cases of acute injuries, associated thermal patterns were warmer than normal cases at the site of the injury. Elucidation of dermatomal (thermatom) patterns provided location for spinal injuries for the diagnosis of back injuries in horses. Similarly, in a case of a dog where the neck injury (subluxation of atlanto-axis) the diagnosis was determined by abnormal thermal patterns and gradients.

35.4 Peripheral Neurovascular Thermography

When there are alterations in skin surface temperature, it may be difficult to distinguish and diagnose between nerve and vascular injuries. The cutaneous circulation is under sympathetic vasomotor control. Peripheral nerve injuries and nerve compression can result in skin surface vascular changes that can be detected thermographically. It is well known that inflammation and nerve irritation may result in vasoconstriction causing cooler thermograms in the afflicted areas. Transection of a nerve and/or nerve damage to the extent that there is a loss of nerve conduction results in a loss in sympathetic tone which causes vasodilation indicated by an increase in the thermogram temperature. Of course, this simple rationale is more complicated with different types of nerve injuries (neuropraxia, axonotomesis, and neurotmesis). Furthermore, lack of characterization of the extent and duration of injuries may make thermographic interpretation difficult.

Studies were done on horses and other animal species to show that if thermographic examination is performed properly under controlled conditions, it can provide an accurate diagnosis of neurovascular injuries. The rationale for a neurovascular clinical diagnosis is provided in the following Horner's Syndrome case.

35.4.1 Horner's Syndrome

In four horses, Horner's Syndrome was also induced by transaction of vagosympathetic trunk on either left or right side of the neck [15]. Facial thermograms of a case of Horner's Syndrome were done 15 min before and after the exercise. Sympathetic may cause the affected side to be warm by 2–3°C more than the non-transected side. This increased temperature after denervation is reflective of an increase in blood flow due to vasodilation in the denervated areas [15, 16]. The increased thermal patterns on the affected side were present up to 6–12 weeks. In about 2–4 months, neurotraumatized side blood flow readjusted to the local demand of circulation. Thermography of both non-neuroectomized and neuroectomized sides looked similar and normal [16]. In some cases, this readjustment took place as early as five days and it was difficult to distinguish the affected side. The intravenous injection of 1 mg of epinephrine in a 1000 lb horse caused an increase in thermal patterns on the denervated side, the same as indicating the presence of Horner's Syndrome. Administration of I V acetyl promazine (30 mg/1000 lb horse) showed increased heat (thermal pattern) on the normal non-neuroectomized side, whereas acetylpromazine had no effect on the neurectomized side. Alpha-blocking drug acetylpromazine caused vasodilation and increased blood flow to normal non-neurectomized side, whereas no effect was seen in the affected neurectomized side due to the lack of sympathetic innervation [16–18].

35.4.2 Neurectomies

Thermographic evaluation of the thoracic (front) and pelvic (back) limbs were done before and after performing digital neurectomies in several horses. After posterior digital neurectomy there were significant increases in heat in the areas supplied by the nerves [17]. Within 3–6 weeks, readjustment of local blood flow occurred in the neurectomized areas, and it was difficult to differentiate between the non-neurectomized and the neurectomized areas. Ten minutes after administration of 0.06 mg/kg I V injection of acetylpromazine, a 2–3°C increase in heat was noted in normal non-neurectomized areas, whereas the neurectomized areas of the opposite limb were not affected.

35.4.3 Vascular Injuries

Thermography has been efficacious in the diagnosis of vascular diseases. It has been shown that the localized reduction of blood flow occurs in the horse with navicular disease [11]. This effect was more obvious on thermograms obtained after exercise than before exercise. Normally, 15–20 min of exercise will increase skin surface temperature by 2–2.5°C in horses [10,11]. In cases of arterial occlusion, the area distal to the occlusion in the horses' limb shows cooler thermograms. The effects of exercise or administration of alpha-blocking drugs like acetylpromazine causes increased blood flow to peripheral circulation in normal areas with intact vascular and sympathetic responses [17,18]. Thus, obtaining thermograms either after exercise or after administration of alpha-blocking drugs like acetylpromazine provides prognostic value for diagnosis of adequate collateral circulation. Therefore, the use of skin temperature as a measure of skin perfusion merits consideration for peripheral vascular flow, perfusion, despite some physical and physiological limitations, which are inherent in methodology [19].

Furthermore, interference with the peripheral vascular blood flow can result from neurogenic inhibition, vascular occlusion, and occlusion as a result of inflammatory vascular compression. Neurogenic inhibition can be diagnosed through the administration of alpha-blocking drugs which provide an increase in blood flow. Vascular impairment may also be associated with local injuries (inflammation, edema, swelling, etc.) which may provide localized cooler or hotter thermograms. Thus, evaluation using thermography should note the physical state and site of the injury.

35.5 Musculoskeletal Injuries

Thermography has been used in the clinical and subclinical cases of osteoarthritis, tendonitis, navicular disease, and other injuries such as sprains, stress fractures, and shin splints [10,11,20,21]. In some cases thermal abnormalities may be detected two weeks prior to the onset of clinical signs of lameness in horses, especially in the case of joint disease [21], tendonitis [10], and navicular problems [11,20].

Osteoarthritis is a severe joint disease in horses. Normally, diagnosis is made by clinical examination and radiographic evaluation. Radiography detects the problem after deterioration of the joint surface has taken place. Clinical evaluation is only done when horses show physical abnormalities in their gait due to pain. An early sign of osteoarthritis is inflammation, which can be detected by thermography prior to it becoming obvious on radiograms [21].

In studies of standard bred race horses, the effected tarsus joint can demonstrate abnormal thermal patterns indicating inflammation in the joint two to three weeks prior to radiographic diagnosis [21]. The abnormal thermograms obtained in this study were more distinct after exercise than before exercise. Thus, thermography provided a subclinical diagnosis of osteoarthritis in this study.

Thermography was used to evaluate the efficacy of corticosteroid therapy in amphotericine-B induced arthritis in ponies [22]. The intra-articular injection of 100 mg of methylprednisolone acetate was effective in alleviating the clinical signs of lameness and pain. It is important to note that when compared with clinical signs of non-treated arthritis, it was difficult to differentiate increased thermal patterns between corticosteroid treated vs. non-treated, arthritis-induced joints. However, corticosteroid therapy did not decrease the healing time of intercarpal arthritis, whereas corticosteroid therapy did decrease the time for return to normal thermographic patterns for tibiotarsal joints. In this study, thermography was useful in detecting inflammation in the absence of clinical signs of pain in corticosteroid treated joints and aiding the evaluation of the healing processes in amphotericin B-induced arthritis [22].

The chronic and acute pain associated with neuromuscular conditions can also be diagnosed by this technique. In cases where no definitive diagnosis can be made using physical examination and x-rays, thermography has been efficacious for early diagnosis of soft tissue injuries [10,23]. The conditions such as subsolar abscesses, laminitis, and other leg lameness can be easily differentiated using thermography [10,11]. We have used thermography for quantitative and qualitative evaluation of anti-inflammatory drugs such as phenylbutazone in the treatment of physical or chemically induced inflammation. The most useful application of thermography in veterinary medicine and surgery has been to aid early detection of an acute and chronic inflammatory process.

35.6 Thermography of the Testes and Scrotum in Mammalian Species

The testicular temperature of most mammalian species must be below body temperature for normal spermatogenesis. The testes of most domestic mammalian species migrates out of the abdomen and are retained in the scrotum, which provides the appropriate thermal environment for normal spermatogenesis [24,25]. The testicular arterial and venous structure is such that arterial coils are enmeshed in the pampiniform plexus of the testicular veins, which provides a counter current heating regulating mechanism by which arterial blood entering the testes is cooled by the venous blood leaving the testes [24,25]. In the ram, the temperature of the blood in the testicular artery decreases by 4°C from the external inguinal ring to the surface of the testes. Thus, to function effectively, the mammalian testes are maintained at a lower temperature.

Purohit [26,27] used thermography to establish normal thermal patterns and gradients of the scrotum in bulls, stallions, bucks, dogs, and llamas. The normal thermal patterns of the scrotum in all species studied is characterized by right to left symmetrical patterns, with a constant decrease in the thermal gradients from the base to the apex. In bulls, bucks, and stallions, a thermal gradient of 4–6°C from the base to apex with concentric hands signifies normal patterns. Inflammation of one testicle increased ipsilateral scrotal temperatures of 2.5–3°C [26,28] If both testes were inflamed, there was an overall increase of 2.5–3°C temperature and a reduction in temperature gradient was noted.

Testicular degeneration could be acute or chronic. In chronic testicular degeneration with fibrosis, there was a loss of temperature gradient, loss of concentric thermal patterns, and some areas were cooler than others with no consistent patterns [26]. Reversibility of degenerative changes depends upon the severity and duration of the trauma. The infrared thermal gradients and patterns in dogs [27] and llamas [27,29] are unique to their own species and the patterns are different from that of the bull and buck.

Thermography has also been used in humans, indicating a normal thermal pattern which is characterized by symmetric and constant temperatures between 32.5 and 34.5°C [30–33]. Increased scrotal infrared emissions were associated with intrascrotal tumor, acute and chronic inflammation, and varicoceles [34,35]. Thermography has been efficacious for early diagnosis of acute and/or chronic testicular degeneration in humans and many animal species. The disruption of the normal thermal patterns of the scrotum is directly related to testicular degeneration. The testicular degeneration may cause transient or permanent infertility in the male. It is well established that increases in scrotal temperature above normal causes disruption of spermatogenesis, affects sperm maturation, and contributes toward subfertile or infertile semen quality. Early diagnosis of pending infertility has a significant impact on economy and reproduction in animals.

35.7 Conclusions

The value of thermography can only be realized if it is used properly. All species studied thus far have provided remarkable bilateral symmetrical patterns of infrared emission. The high degree of right-to-left symmetry provides a valuable asset in diagnosis of unilateral problems associated with various inflammatory disorders. On the other hand, bilateral problems can be diagnosed due to changes in thermal gradient and/or overall increase or decrease of temperature, away from the normal established thermal patterns in a given area of the body. Various areas of the body on the same side have normal patterns and gradients. This can be used to diagnose a change in gradient patterns. Alteration in normal thermal patterns and gradients indicates a thermal pathology. If thermal abnormalities are evaluated carefully, early diagnosis can be made, even prior to the appearance of clinical signs of joint disease, tendonitis, and various musculoskeletal problems in various animal species. Thermography can be used as a screening device for early detection of an impending problem, allowing veterinarian institute treatment before the problem becomes more serious. During the healing process post surgery, animals may appear physically sound. Thermography can be used as a diagnostic aid in assessing the healing processes. In equine sports medicine, thermography can be used on a regular basis for screening to prevent severe injuries to the horse. Early detection and treatment can prevent financial losses associated with delayed diagnosis and treatment.

The efficacy of non contact electronic infrared thermography has been demonstrated in numerous clinical settings and research studies as a diagnostic tool for veterinary medicine. It has had a strong impact on veterinary medical practice and thermal physiology where accurate skin temperatures need to be assessed under normal conditions, disease pathologies, injuries, and thermal stress. The importance of infrared thermography as a research tool cannot be understated for improving the medical care of animals and for the contributions made through animal research models that improve our understanding of human structures and functions.

References

[1] Smith W.M. Application of thermography in veterinary medicine. *Ann. NY Acad. Sci.*, 121, 248, 1964.

[2] Delahanty D.D. and George J.R. Thermography in equine medicine. *J. Am. Vet. Med. Assoc.*, 147, 235, 1965.

[3] Clark J.A. and Cena K. The potential of infrared thermography in veterinary diagnosis. *Vet. Rec.*, 100, 404, 1977.

[4] Stromberg B. The normal and diseased flexor tendon in racehorses. *Acta Radiol.* [Suppl.] 305, 1, 1971.

[5] Stromberg B. Thermography of the superficial flexor tendon in race horses. *Acta Radiol.* [Suppl.] 319, 295, 1972.

[6] Stromberg B. The use of thermograph in equine orthopedics. *J. Am. Vet. Radiol. Soc.*, 15, 94, 1974.

[7] Stromberg B. and Norbert I. Infrared emission and Xe-disappearance rate studies in the horse. *Equine Vet. J.*, 1, 1–94, 1971.

[8] Nelson H.A. and Osheim D.L. Soring in Tennessee walking horses: detection by thermography. *USDA-APHIS*, *Veterinary Services Laboratories*, Ames, Iowa, pp. 1–14, 1975.

[9] Purohit R.C., Bergfeld II W.A. McCaoy M.D., Thompson W.M., and Sharman R.S. Value of clinical thermography in veterinary medicine. *Auburn Vet.*, 33, 140, 1977.

[10] Purohit R.C. and McCoy M.D. Thermography in the diagnosis of inflammatory processes in the horse. *Am. J. Vet. Res.*, 41, 1167, 1980.

[11] Turner T.A. et al. Thermographic evaluation of podotrochlosis in horses. *Am. J. Vet. Res.*, 44, 535, 1983.

[12] Heath A.M., Navarre C.B., Simpkins A.S., Purohit R.C., and Pugh D.G. A comparison of heat tolerance between sheared and non sheared alpacas (llama pacos). *Small Ruminant Res.*, 39, 19, 2001.

[13] Purohit R.C. and Franco B.D. Infrared thermography for the determination of cervical dermatome patterns in the horse. *Biomed. Thermol.*, 15, 213, 1995.

[14] Purohit R.C., Schumacher J, Molloy J.M., Smith, and Pascoe D.D. Elucidation of thoracic and lumbosacral dermatomal patterns in the horse. *Thermol. Int.*, 13, 79, 2003.

[15] Purohit R.C., McCoy M.D., and Bergfeld W.A. Thermographic diagnosis of Horner's syndrome in the horse. *Am. J. Vet. Res.*, 41, 1180, 1980.

[16] Purohit R.C. The diagnostic value of thermography in equine medicine. *Proc. Am. Assoc. Equine Pract.*, 26, 316–326, 1980.

[17] Purohit R.C. and Pascoe D.D. Thermographic evaluation of peripheral neurovascular systems in animal species. *Thermology*, 7, 83, 1997.

[18] Purohit R.C., Pascoe D.D., Schumacher J, Williams A., and Humburg J.H. Effects of medication on the normal thermal patterns in horses. *Thermol. Osterr.*, 6, 108, 1996.

[19] Purohit R.C. and Pascoe D.D. Peripheral neurovascular thermography in equine medicine. *Thermol. Osterr.*, 5, 161, 1995.

[20] Turner T.A., Purohit R.C., and Fessler J.F. Thermography: a review in equine medicine. *Comp. Cont. Education Pract. Vet.*, 8, 854, 1986.

[21] Vaden M.F., Purohit R.C. Mcoy, and Vaughan J.T. Thermography: a technique for subclinical diagnosis of osteoarthritis. *Am. J. Vet. Res.*, 41, 1175–1179, 1980.

[22] Bowman K.F., Purohit R.C., Ganjan, V.K., Peachman R.D., and Vaughan J.T. Thermographic evaluation of corticosteroids efficacy in amphotericin-B induced arthritis in ponies. *Am. J. Vet. Res.* 44, 51–56, 1983.

[23] Purohit R.C. Use of thermography in the diagnosis of lameness. *Auburn Vet.*, 43, 4, 1987.

[24] Waites G.M.H. and Setchell B.P. Physiology of testes, epididymis, and scrotum. In *Advances in Reproductive Physiology*. McLaren A., Ed., London, Logos, Vol. 4, pp. 1–21, 1969.

[25] Waites G.M.H. Temperature regulation and the testes. In *The Testis*, Johnson A.D., Grones W.R., and Vanderwork N.L., Eds., New York, Academy Press, Inc., Vol. 1, pp. 241–237, 1970.

[26] Purohit R.C., Hudson R.S., Riddell M.G., Carson R.L., Wolfe D.F., and Walker D.F. Thermography of bovine scrotum. *Am. J. Vet. Res.*, 46, 2388–2392, 1985.

[27] Purohit R.C., Pascoe D.D., Heath A.M. Pugh D.G., Carson R.L., Riddell M.G., and Wolfe D.F. Thermography: its role in functional evaluation of mammalian testes and scrotum. *Thermol. Int.*, 12, 125–130, 2002.

[28] Wolfe D.F., Hudson R.S., Carson R.L., and Purohit, R.C. Effect of unilateral orchiectomy on semen quality in bulls. *J. Am. Vet. Med. Assoc.*, 186, 1291, 1985.

[29] Heath A.M., Pugh D.G., Sartin E.A., Navarre B., and Purohit R.C. Evaluation of the safety and efficacy of testicular biopsies in llamas. *Theriogenology*, 58, 1125, 2002.

[30] Amiel J.P., Vignalou L., Tricoire J. et al. Thermography of the testicle: preliminary study. *J. Gynecol. Obstet. Biol. Reprod.*, 5, 917, 1976.

[31] Lazarus B.A. and Zorgiotti A.W. Thermo-regulation of the human testes. *Fertil. Steril.*, 26, 757, 1978.

[32] Lee J.T. and Gold R.H. Localization of occult testicular tumor with scrotal thermography. *J. Am. Med. Assoc.*, 1976, 236, 1976.

[33] Wegner G. and Weissbach Z. Application of palte thermography in the diagnosis of scrotal disease. *MMW*, 120, 61, 1978.

[34] Gold R.H., Ehrilich R.M., Samuels B. et al. Scrotal thermography. *Radiology*, 1221, 129, 1979.

[35] Coznhaire F., Monteyne R., and Hunnen M. The value of scrotal thermography as compared with selective retrograde venography of the internal spermatic vein for the diagnosis of subclinical varicoceles. *Fertil. Steril.*, 27, 694, 1976.

36

Standard Procedures for Infrared Imaging in Medicine

Kurt Ammer
Ludwig Boltzmann Research Institute for Physical Diagnostics and University of Glamorgan

E. Francis Ring
University of Glamorgan

36.1 Introduction

Infra red thermal imaging has been used in medicine since the early 1960s. Working groups within the European Thermographic Association (now European Association of Thermology) produced the first publications on standardization of thermal imaging in 1978 [1] and 1979 [2]. However, Collins and Ring established already in 1974 a quantitative thermal index [3], which was modified in Germany by J.-M. Engel in 1978 [4]. Both indices opened the field of quantitative evaluation of medical thermography.

Further recommendations for standardization appeared in 1983 [5] and 1984, the later related to essential techniques for the use of thermography in clinical drug trials [6]. J.-M. Engel published a booklet entitled "Standardized thermographic investigations in rheumatology and guideline for evaluation" in 1984 [7]. The author presented his ideas for standardization of image recording and assessment including some normal values for wrist, knee, and ankle joints. Engel's measurements of knee temperatures were first published in 1978 [4]. Normal temperature values of the lateral elbow, dorsal hands, anterior knee, lateral and medial malleolus and the 1st metatarsal joint were published by Collins in 1976 [8].

The American Academy of Thermology published technical guidelines in 1986 including some recommendations for thermographic examinations [9]. However, the American authors concentrated on determining the symmetry of temperature distribution rather than the normal temperature values of particular body regions. Uematsu in 1985 [10] and Goodman, 1986 [11] published the side to side variations of surface temperatures of the human body. These symmetry data were confirmed by E.F. Ring for the lower leg in 1986 [12].

In Japan, medical thermal imaging has been an accepted diagnostic procedure since 1981 [13]. Recommendations for the analysis of neuromuscular thermograms were published by Fujimasa et al. in 1986 [14]. Five years later more detailed proposals for the thermal image based analysis of physiological functions were published in *Biomedical Thermology* [15], the official journal of the Japanese Society of thermology. This paper was the result of a workshop on clinical thermography criteria.

Recently, the thermography societies in Korea have published a book, which summarizes in 270 pages general standards for imaging recording and interpretation of thermal images in various diseases [16].

As the relationship between skin blood flow and body surface temperature has been obvious from the initial use of thermal imaging in medicine, quantitative assessments were developed at an early stage. E.F. Ring developed a thermographic index for the assessment of ischemia in 1980, that was originally used for patients suffering from Raynauds' disease [17]. The European Association of Thermology published a statement in 1988 on the subject of Raynaud's Phenomenon [18]. Normal values for recovering after a cold challenge have been published since 1976 [19,20]. A range of temperatures were applied in this thermal challenge test, the technique was reviewed by E.F. Ring in 1997 [21].

An overview of recommendations gathered from, The Japanese Society of Biomedical Thermology and the European Association of Thermology was collated and published by Clark and Goff in 1997 [22]. This paper is based on the practical implications of the foregoing papers taken from the perspective of the modern thermal imaging systems available to medicine.

Finally, a project at the University of Glamorgan, aims to create an atlas of normal thermal images of healthy subjects [23]. This study, started in 2001, has generated a number of questions related to the influence of body positions on accuracy and precision of measurements from thermal images [24,25].

36.2 Definition of Thermal Imaging

Thermal imaging is regarded as a technique for temperature measurements based on the infrared radiation from objects. Unlike images created by x-rays or proton activation through magnetic resonance, thermal imaging is not related to morphology. The technique provides only a map of the distribution of temperatures on the surface of the object imaged.

Whenever infrared thermal imaging is considered as a method for measurement, the technique must meet all criteria of a measurement. The most basic features of measurement are accuracy (in the medical field also named validity) and precision (in medicine reliability). Anbar [26] has listed five other terms related to the precision of infrared based temperature measurements. When used as an outcome measure, responsiveness or sensitivity to change is an important characteristic.

36.2.1 Accuracy

Measurements are basic procedures of comparison namely to compare a standardized meter with an object to be measured. Any measurement is prone to error, thus a perfect measurement is impossible. However, the smaller the variation of a particular measurement from the standardized meter, the higher is the accuracy of the measurement or in other words, an accurate measurement is as close as possible to the true value of measurement. In medicine, accuracy is often named validity, mainly caused by the fact, that medical measurements are not often performed by the simple comparison of meter and object. For example, assessments from various features of a human being may be combined into a new construct, resulting in a innovative measurement of health.

36.2.2 Precision

A series of measurements can not achieve totally identical results. The smaller the variation between single results, the higher is the precision or repeatability (reliability) of the measurement. However, reliability without accuracy, is useless. For example, a sports archer who always hits the same peripheral sector of the

TABLE 36.1 Conditions Affecting, Accuracy, Precision and Responsiveness of Temperature Measures

Condition affecting	Accuracy	Precision	Responsiveness
Object or subject	X	X	X
Camera systems, standards, and calibration	X	X	X
Patient position and image capture		X	X
Information protocols and resources		X	X
Image analysis	X	X	X
Image exchange	X	X	X
Image presentation	X	X	X

target, has very high reliability, but no validity, because such an athlete must find the centre of the target to be regarded as accurate.

36.2.3 Responsiveness

Both, accuracy and precision, have an impact on the sensitivity to change of outcome measures. Validity is needed to define correctly the symptom to be measured. Precision will affect the responsiveness also, because a change of the symptom can only be detected if this change is bigger than the variation of repeated measurements.

36.3 Sources of Variability of Thermal Images

Table 36.1 shows conditions in thermal imaging that may affect accuracy, precision, and responsiveness.

36.3.1 Object or Subject

As the emittance of infrared radiation is the source of remote temperature measurements, knowledge of the emissivity of the object is essential for the calculation of temperature related to the radiant heat. In non living objects emissivity is mainly a function of the texture of the surface.

Seventy years ago, Hardy [27] showed that the human skin acts like an almost perfect black body radiator with an emissivity of 0.98. Studies from Togawa in Japan have demonstrated that the emissivity of the skin is unevenly distributed [28]. In addition, infrared reflection from the environment and substances applied on the skin may also alter the emissivity [29–31]. Water is an efficient filter for infrared rays and can be bound to the superficial corneal layer of the skin during immersion for at least 15 min [32,33] or in the case of severe edema [34]. This can affect the emissivity of the skin.

The hair coat of animal may show a different emissivity than the skin after clipping the hair [35]. Variation in the distribution of the hairy coat will influence the emissivity of the animal's surface [36]. Variation in emissivity will influence the accuracy of temperature measurements.

Homeothermic beings, maintain their deep body (core) temperature through variation of the surface (shell) temperature, and show a circadian rhythm of both the core and shell temperature [37–40]. Repeated temperature registrations not performed at the same time of the day will therefore affect the precision of these measurements.

36.3.2 Camera Systems, Standards, and Calibration

36.3.2.1 The Imaging System

A new generation of infra red cameras have become available for medical imaging. The older systems normally single element detectors using an optical mechanical scanning process, were mostly cooled by the addition of liquid nitrogen [41–43]. However, adding nitrogen to the system, affects the stability of temperature measurements for a period up to 60 min [44]. Nitrogen cooled scanners had the effect of limiting the angle at which the camera could be used which restricted operation.

Electronic cooling systems were then introduced, which provided the use of image capturing without restrictions of the angle between the object and the camera. The latest generation of focal plane array cameras can be used without cooling, providing almost maintenance free technology [45]. However, repeated calibration procedures built inside the camera can affect the stability of temperature measurements [46].

The infrared wavelength, recorded by the camera, will not affect the temperature readings as long as the algoritm of calculation temperature from emitted radiation is correct. However, systems equipped with sensors sensitive in different bands of the infrared spectrum are capable to determine the emissivity of objects [47].

36.3.2.2 Temperature Reference

Earlier reports stipulate the requirement for a separate thermal reference source for calibration checks on the camera [9,48,49]. Many systems now include an internal reference temperature, with manufacturers claiming that external checks are not required. Unless frequent servicing is obtained, it is still advisable to use an external source, if only to check for drift in the temperature sensitivity of the camera. An external reference, which may be purchased or constructed, can be left switched on throughout the day. This allows the operator to make checks on the camera, and in particular provides a check on the hardware and software employed for processing. These constant temperature source checks may be the only satisfactory way of proving the reliability of temperature measurements made from the thermogram [48]. Linearity of temperature measurements which may be questionable in focal plane array equipment, can be checked with two ore more external temperature references. New low cost reference sources, based on the triple point of particular chemicals, are currently under construction in the United Kingdom [44].

36.3.2.3 Mounting the Imager

A camera stand which provides vertical height adjustment is very important for medical thermography. Photographic tripod stands are inconvenient for frequent adjustment and often result in tilting the camera at an undefined angle to the patient. This is difficult to reproduce, and unless the patient is positioned so that the surface scanned is aligned at 90° to the camera lens, distortion of the image is unavoidable. Undefined angles of the camera view affects the precision of measurements.

In the case of temperature measurements from a curved surface, the angle between the radiating object and the capturing device may be the critical source of false measurements [50–52]. At an angle of view beyond 30° small losses of capturing the full band of radiation start to occur, at an angel of 60° the loss of information becomes critically and is followed by false temperature readings. The determination of the temperature of the same forefoot in different views shows clearly, that consideration of the angel of the viewing is a significant task [53]. Unless corrected, thermal images of of evenly curved objects lack accuracy of temperature measurements [54].

Studio camera stands are ideal, they provide vertical height adjustment with counterbalance weight compensation. It should be noted that the type of lens used on the camera will affect the working distance and the field of view, a wide angle lens reduces distance between the camera and the subject in many cases, but may also increase peripheral distortion of the image [55].

36.3.2.4 Camera Initialization

Start up time with modern cameras are claimed to be very short, minutes or seconds. However, the speed with which the image becomes visible is not an indication of image stability. Checks on calibration will usually show that a much longer period from 10 min to several hours with an uncooled system are needed to achieve stable conditions for temperature readings from infrared images [5,46].

36.3.3 Patient Position and Image Capture

Standardized positions of the body for image capture and clearly defined fields of view can reduce systematic errors and increases both accuracy and precision of temperature readings from thermal images recorded in such a manner. In radiography, standardized positions of the body for image capture have been included in the protocol for quality assurance for a long time. Although thermal imaging does not provide much anatomical information compared to other imaging techniques, variation of body positions and the related fields of view affects the precision of temperature readings from thermograms. However, the intra- and inter-rater repeatability of temperature values from the same thermal image was found to be excellent [56].

36.3.3.1 Location for Thermal Imaging

The size of investigation room does not influence the quality of temperature measurements from thermal images, unless the least distance in one direction is not shorter than the distance between the camera and an object of 1.2 m height [57]. Such a condition will result in thermal images out of focus. Other important features of the examination room are thermal insulation and prevention of any direct or reflected infrared radiation sources. Following this proposal will result in an increase of accuracy and precisison of measurements.

36.3.3.2 Ambient Temperature Control

This is a primary requirement for most clinical applications of thermal imaging. A range of temperatures from 18 to 25°C should be attainable and held for at least 1 h to better than 1°C. Due to the nature of human thermoregulation, stability of the room temperature is a critical feature. It have been shown, that subjects acclimatized for 40–60 min to a room temperature of 22°C showed differences in surface temperature at various measuring sites of the face after lowering the ambient temperature by 2°C [58]. While the nose cooled on average by 4°C, the forehead and the meatus decreased the surface temperature by only by 0.4–0.45%. Similar changes may occur at other acral sites such as tips of fingers or toes, as both regions are highly involved in heat exchange for temperature regulation.

At lower temperatures, the subject is likely to shiver, and over 25°C room temperature will cause sweating, at least in most European countries. Variations may be expected in colder or warmer climates, in the latter case, room temperatures may need to be 1 to 2°C higher [59].

Additonal techniques for cooling particular regions of the body have been developed [60,61]. Immersion of the hands in water at various tempeatures is a common challenge for the assessment of vasospastic disease [21].

Heat generated in the investigation room affects the room temperature. Possible heat sources are electronic equipment such as the scanner and its computer, but also human bodies. For this reason the air-conditioning unit should be capable of compensating for the maximum number of patients and staff likely to be in the room at any one time. These effects will be greater in a small room of 2 × 3 m or less.

Air convection is a very effective method of skin cooling and related to the wind speed. Therefore, air conditioning equipment should be located so that direct draughts are not directed at the patient, and that overall air speed is kept as low as possible. A suspended perforated ceiling with ducts diffusing the air distribution evenly over the room is ideal [62].

A cubicle or cubicles within the temperature controlled area is essential. These should provide privacy for disrobing and a suitable area for resting through the acclimatization period.

36.3.3.3 Pre-Imaging Equilibration

On arrival at the department, the patient should be informed of the examination procedure, instructed to remove appropriate clothing and jewellery, and asked to sit or rest in the preparation cubicle for a fixed time. The time required to achieve adequate stability in blood pressure and skin temperature is generally considered to be 15 min, with 10 min as a minimum [63–65]. After 30 min cooling, oszillations of the skin temperature can be detected, in different regions of the body with different amplitudes resulting in a temperature asymmetry between left and right sides [64].

Contact of body parts with the environment or with other body parts alters the surface temperature due to heat transfer by conduction. Therefore, during the preparation time the patient must avoid folding or crossing arms and legs, or placing bare feet on a cold surface. If the lower extremities are to be examined, a stool or leg rest should be provided to avoid direct contact with the floor [66]. If these requirements are not met, poor precision of measurements may result.

36.3.3.4 Positions for Imaging

As in anatomical imaging studies, it is preferable to standardize on a series of standard views for each body region. The EAT Locomotor Diseases Group recommendations include a triangular marker system to indicate anterior, posterior, lateral, and angled views [2,67]. However, reproduction of positions for angled views may be difficult, even when aids such as rotating platforms are used [68].

Modern image processing software provide comment boxes which can be used to encode the angle of view which will be stored with the image [69]. It should be noted that the position of the patient for scanning and in preparation must be constant. Standing, sitting, or lying down affect the surface area of the body exposed to the ambient, therefore an image recorded with the patient in a sitting position may not be comparable with one recorded on a separate occasion in a standing position. In addition, blood flow against the influence of gravity contributes to the skin temperature of fingers in various limb positions [70].

36.3.3.5 Field of View

Image size is dependent on the distance between the camera and the patient and the focal length of the infrared camera lens. The lens is generally fixed on most medical systems, so it is good practice to maintain a constant distance from the patient for each view, in order to acquire a reproducible field of view for the image. If in different thermograms different fields of the same subject are compared, the variable resolution can lead to false temperature readings [71]. However, maintaining the same distance between object and camera, cannot compensate for individual body dimensions, for example, big subjects will have big knees and therefore maintaining the same distance as for a tiny subjects knee is not applicable.

To overcome this problem, the field of view has been defined in the standard protocol at the University of Glamorgan in a two fold way, that is, body position and alignment of anatomical landmarks to the edge of the image [23]. These definitions enabled us to investigate the reproducibilty of body views using the distance in pixels between anatomical landmarks and the outline of the infrared images [24–72].

Figure 36.1 gives examples of the views, that have been investigated for the reproduciblity of body positions. Table 36.2 shows the mean value, standard deviation, and 95% confidence interval of the variation of body views of the upper and the lower part of the human body. Variations in views of the lower part of the body were bigger than in views of the upper part. The highest degree of variation was found in the view "Both Ankles Anterior," but the smallest variation in the view "Face."

36.3.4 Information Protocols and Resources

Human skin temperature is the product of heat dissipated from the vessels and organs within the body, and the effect of the environmental factors on heat loss or gain. There are a number of further influences which are controllable, such as cosmetics [29], alcohol intake [73–75], and smoking [76–78]. In general terms the patient attending for examination should be advised to avoid all topical applications such as ointments and cosmetics on the day of examination to all the relevant areas of the body [31,47,79,80]. Large meals and above average intake of tea or coffee should also be excluded, although studies supporting this recommendation are hard to find and the results are not conclusive [81,82].

Patients should be asked to avoid tight fitting clothing, and to keep physical exertion to a minimum. This particularly applies to methods of physiotherapy such as electrotherapy [83,85], ultrasound [86,87], heat treatment [88,90], cryotherapy [91–94], massage [95–97], and hydrotherapy [31,32,98,99], because thermal effects from such treatment can last for 4 to 6 h under certain conditions. Heat production by muscular exercise is a well documented phenomenon [65,100–103].

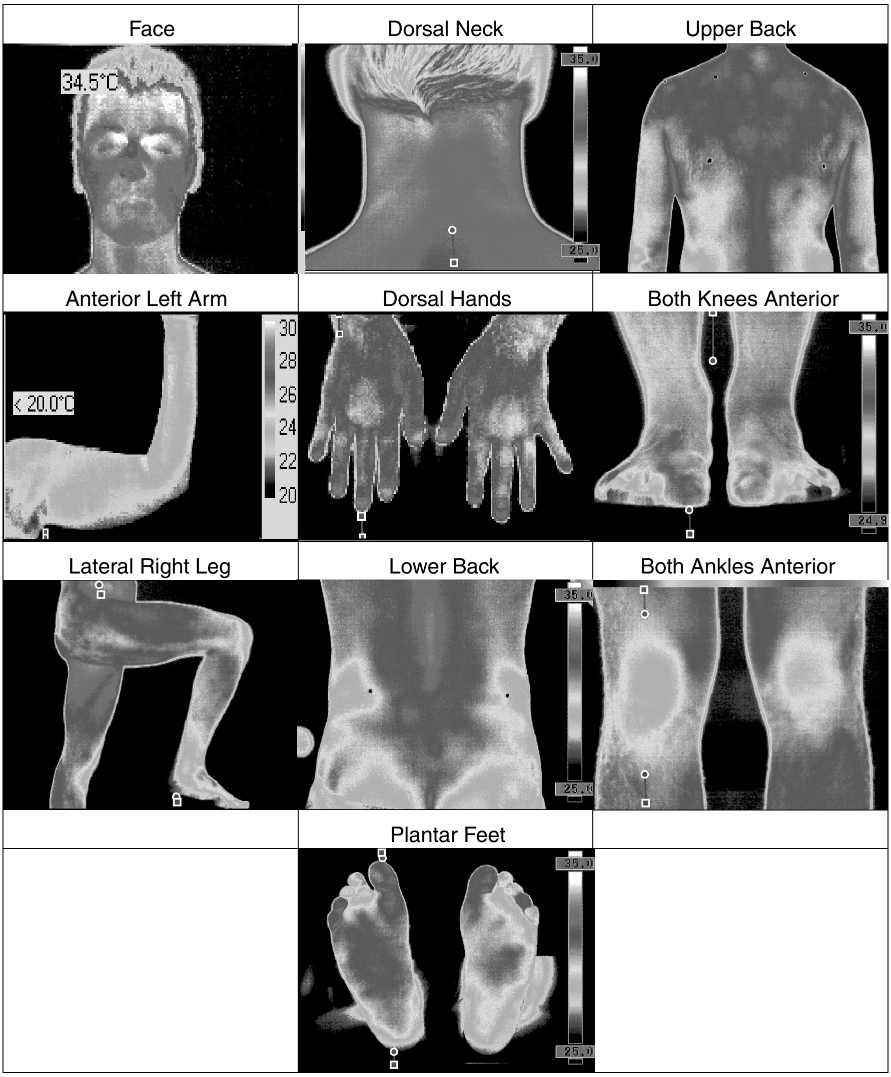


FIGURE 36.1 (See color insert following page 29-16.) Body views investigated.

TABLE 36.2 Variation of Positions of All the Investigated Views

View	Upper edge (pixel) mean ± SD (95% CI)	Lower edge (pixel) mean ± SD (95% CI)	Left side edge (pixel) mean ± SD (95% CI)
Face	0.5 ± 5.3 (−2.2 to 1.9)	4.0 ± 10.9 (−0.03 to 8.2)	
Dorsal neck	−8.4 ± 36.4 (−18.3 to 1.6)	122.6 ± 146.6 (82.6 to 162.6)	
Upper back	4.5 ± 9.9 (0.8 to 8.2)	28.1 ± 22.0 (19.9 to 36.4)	
Anterior left arm	22.4 ± 33.0 (8.7 to 36.0)	15.8 ± 15.4 (9.5 to 22.2)	12.5 ± 16.0 (5.9 to 19.1)
Dorsal hands	41.8 ± 17.8 (35.5 to 48.2)	33.2 ± 22.3 (25.3 to 41.5)	
Both knees anterior	80.7 ± 47.3 (60.7 to 100.7)	84.3 ± 37.0 (68.6 to 99.9)	
Lateral right leg	16.7 ± 21.0 (5.9 to 27.5)	17.2 ± 15.8 (9.0 to 25.3)	
Lower back	17.1 ± 4.2 (8.6 to 25.6)	16.3 ± 4.6 (16.3 ro 34.9)	
Both ankles anterior	158.8 ± 12.2 (133.6 to 184.1)	54.9 ± 9.1 (36.1 to 37.8)	
Plantar feet	31.0 ± 24.1 (23.2 to 38.7)	25.7 ± 23.1 (18.3 to 33.1)	

Drug treatment can also affect the skin temperature. This phenomenon was used to evaluate the therapeutic effects of medicaments [6]. Drugs affecting the cardiovascular system must be reported to the thermographer, in order that the correct interpretation of thermal images will be given [104–107].

Omiting just one of the above mentioned conditions will result in reduced precision of temperature measurements.

36.3.5 Image Processing

Every image or block of images must carry the indication of temperature range, with color code/temperature scale. The color scale itself should be standardized. Industrial software frequently provides a grey-scale picture and one or more color scales. However, modern image processing software permits to squeeze the color scale in already recorded images in order to increase the image contrast. Such a procedure will affect the temperature readings from thermal images as temperatures outside of the compressed temperature scale will not be included in the statistics of selected regions of interest. This will result in erroneous temperature readings, affecting both accuracy and precision of measurements.

36.3.6 Image Analysis

Almost all systems now use image processing techniques and provide basic quantitation of the image [108–110]. In some cases this may be operated from a chip within the camera, of may be carried out through an on- or off- line computer. For older equipment like the AGA 680 series several hardware adaptations have been reported to achieve quantitation of the thermograms [111–113].

It has to be emphasized that false color coding of infrared images does not provide means for temperature measurement. If colors are separated by a temperature distance of 1°C, the temperature difference between two points situated in adjacent colors may be between 0.1 and 1.9°C. It is obvious, that false colored images provide at its best an estimation of temperature, but not a measurement. The same is true for liquid crystal thermograms.

Nowadays, temperature measurements in thermal images are based on the definition of regions of interest (ROI). However, standards for shape, size and placement of these regions are not available or incomplete. Although a close correlation exists for ROI of different size in the same region [114], the precision of measurement is affected when ROIS of different size and location are used for repeated measurements.

The Glamorgan protocol [23] is the very first attempt to create a complete standard for the definition of regions of interest in thermal images based on anatomical limits. Furthermore, in the view "both knee anterior" the shape with the highest reproducibility was investigated. During one of the Medical Infrared Training-Courses at the University of Glamorgan, three newly trained investigators defined on the same thermal image of both anterior knees twice the region of interest in the shape of a box, an ellipsoid or as an hour-glass shape. Similar to the result of a pilot study that compared these shapes for repeateabilty, the highest reliability was found for temperature readings from the hour-glass shape, followed by readings from ellipsoids and boxes [53]. The repeatabilty of the regions on the view "Left Anterior Arm," "Both Ankles Anterior," "Dorsal Feet," and "Plantar Feet" were also investigated and resulted in reliabilty coefficients between 0.7 (right ankle) and 0.93 (forearm). The intraclass correlation coefficients ranged between 0.48 (upper arm) and 0.87 (forearm). Applying the Glamorgan protocol consequently, will result in precise temperature measurements from thermal images.

36.3.7 Image Exchange

Most of the modern infrared systems store the recorded thermal images in an own image format, which may not compatible with formats of thermal images from other manufacturers. However, most of this images can be transformed into established image formats such as TIF, JPEG, GIF, and others. As a thermal image is the pictographic representation of a temperature map, the sole image is not enough unless the related temperature information is not provided. Consequently, temperature measurements from standard computer images derived from thermograms is not possible.

Providing both temperature scale and a scale of grey shades, allows the exchange of thermal images over long distance and between different, but compatible image processing software [115]. The grey scale must be derived from the original grey shade thermal image. If it has been transformed from a false color image, the resulted black-and-white thermogram may not be representative for the original grey scale gradient as the grey scale of individual colors may deviate from the particular grey shade of the image. This can then result in false temperature readings.

36.3.8 Image Presentation

Image presentation does not influence the result of measurements from thermal images. However, if thermograms are read by eyes, their appearance will affect the credibility of the information in thermal images. This is for instance the case, when thermal images are use as evidence in legal trials [116].

It was stated, that for forensic acceptability of thermography standardization and repeatability of the technique are very important features [117]. This supports the necessity of quantitative evaluation of thermal images and standards strictly applied to the technique of infrared imaging will finally result in high accuracy and precision of this method of temperature measurement. At that stage it can be recommended as responsive outcome measure for clinical trials in rheumatology [6,8], angiopathies [107,118], neuromuscular disorders [119], surgery [120], and paedriatrics [121].

References

[1] Aarts, N.J.M. et al. Thermograp. terminology. *Acta Thermograp.*, 1978, Suppl. 2.

[2] Engel, J.M. et al. Thermography in locomotor diseases — recommended procedure. *Eur. J. Rheum. Inflamm.*, 2, 299, 1979.

[3] Collins, A.J. et al. Quantitation of thermography in arthritis using multi-isothermal analysis. *Ann. Rheum. Dis.*, 33, 113, 1974.

[4] Engel, J.-M. Quantitative Thermographie des Kniegelenks. *Z. Rheumatol.*, 37, 242, 1978.

[5] Ring, E.F.J. Standardisation of thermal imaging in medicine: physical and environmental factors, in *Thermal Assessment of Breast Health*, Gautherie, M., Albert, E., and Keith, L., Eds., MTP Press Ltd, Lancaster/Boston/The Hague, 1983, p. 29.

[6] Ring, E.F.J., Engel, J.M., and Page-Thomas, D.P. Thermologic methods in clinical pharmacology — skin temperature measurement in drug trials. *Int. J. Clin. Pharm. Ther. Tox.*, 22, 20, 1984.

[7] Engel, J.-M. and Saier, U. *Thermographische Standarduntersuchungen in der Rheumatologie und Richtlinien zu deren Befundung.* Luitpold, München, 1984.

[8] Collins, A.J. Anti-inflammatory drug assessment by the thermographic index. *Acta Thermograp.*, 1, 73, 1976.

[9] Pochaczevsky, R. et al. *Technical Guidelines*, 2nd ed. *Thermology*, 2, 108, 1986.

[10] Uematsu, S. Symmetry of skin temperatures comparing one side of the body to the other. *Thermology*, 1, 4, 1985.

[11] Goodman, P.H. et al. Normal temperature asymmetry of the back and extremities by computer-assisted infrared imaging. *Thermology*, 1, 195, 1986.

[12] Bliss, P. et al. Investigation of nerve root irritation by infrared thermography, in *Back Pain — Methods for Clinical Investigation and Assessment*, Hukins, D.W.L. and Mulholland, R.C., Eds., University Press, Manchester, 1986, p. 63.

[13] Atsumi, K. High technology applications of medical thermography in Japan. *Thermology*, 1, 79–80, 1985.

[14] Fujimasa, I. et al. A new computer image processing system for the analysis of neuromuscular thermograms: a feasibility study. *Thermology*, 1, 221, 1986.

[15] Fujimasa, I. A proposal for thermographic imaging diagnostic procedures for temperature related physiologic function analysis. *Biomed. Thermol.*, 11, 269, 1991.

[16] Lee, D.-I. (Ed.) *Practical Manual of Clinical Thermology*, ISBN 89-954013-04.

[17] Ring, E.F.J. A thermographic index for the assessment of ischemia. *Acta thermograp.*, 5, 35, 1980.

[18] Aarts, N. P. et al. Raynaud's phenomenon: assessment by thermography. *Thermology*, 3, 69, 1988.

[19] Acciarri, L., Carnevale, F., and Della Selva, A. Thermography in the hand angiopathy from vibrating tools. *Acta thermograp.*, 1, 18, 1976.

[20] Ring, E.F. and Bacon, P.A. Quantitative thermographic assessment of inositol nicotinate therapy in Raynaud's phenomena. *J. Int. Med. Res.*, 5, 217, 1977.

[21] Ring, E.F.J. Cold stress test for the hands, in *The Thermal Image in Medicine and Biology*, Ammer, K. and Ring, E.F.J., Eds., Uhlen-Verlag, Wien, 1995, p. 237.

[22] Clark, R.P. and de Calcina-Goff, M. Guidelines for Standardisation in Medical Thermography Draft International Standard Proposals. *Thermol. Osterr.*, 7, 47, 1997.

[23] Website address, Atlas of Normals, www.medimaging.org.

[24] Ammer, K. et al. Rationale for standardised capture and analysis of infrared thermal images, in *Proceedings Part II, EMBEC'02 2.European Medical & Biological Engineering Conference.* Hutten, H. and Krösel, P., Eds. IFMBE, Graz, 2002, p. 1608.

[25] Ring, E.F.J. et al. Errors and artefacts in thermal imaging, in *Proceedings Part II, EMBEC'02 2.European Medical & Biological Engineering Conference.* Hutten, H. and Krösel, P., Eds., IFMBE, Graz, 2002, p. 1620.

[26] Anbar, M. Recent technological developments in thermology and their impact on clinical applications. *Biomed. Thermol.*, 10, 270, 1990.

[27] Hardy, J.D. The radiation of heat from the human body. III. The human skin as a black body radiator. *J. Clin. Invest.*, 13, 615, 1934.

[28] Togawa, T. and Saito, H. Non-contact imaging of thermal properties of the skin. *Physiol. Meas.*, 15, 291, 1994.

[29] Engel, J.-M. Physical and physiological influence of medical ointments of infrared thermography, in *Recent Advances in Medical Thermology*, Ring, E.F.J. and Phillips, B., Eds., Plenum Press, New York, 1984, p. 177.

[30] Hejazi, S. and Anbar, M. Effects of topical skin treatment and of ambient light in infrared thermal images. *Biomed. Thermol.*, 12, 300, 1992.

[31] Ammer, K. The influence of antirheumatic creams and ointments on the infrared emission of the skin, in *Abstracts of the 10th International Conference on Thermogrammetry and Thermal Engineering in Budapest 18–20, June 1997*, Benkö, I. et al., Eds., MATE, Budapest, 1997, p. 177.

[32] Ammer, K. Einfluss von Badezusätzen auf die Wärmeabstrahlung der Haut. *ThermoMed*, 10, 71, 1994.

[33] Ammer, K. The influence of bathing on the infrared emission of the skin, in A*bstracts of the 9th International Conference on Thermogrammetry and Thermal Engineering in Budapest 14–16, June 1995*, Benkö, I., Lovak., and Kovacsics, I., Eds., MATE, Budapest, 1995, p. 115.

[34] Ammer, K. Thermographie in lymphedema, in *Advanced Techniques and Clinical Application in Biomedical Thermologie*, Mabuchi, K., Mizushina, S., and Harrison, B., Eds., Harwood Academic Publishers, Chur/Schweiz, 1994, p. 213.

[35] Heath, A.M. et al. A comparison of surface and rectal temperatures between sheared and non-sheared alpacas (*Lama pacos*). *Small Rumin. Res.*, 39, 19, 2001.

[36] Purohit, R.C. et al. Thermographic evaluation of animal skin surface temperature with and without haircoat. *Thermol. Int.*, 11, 83, 2001.

[37] Damm, F., Döring, G., and Hildebrandt, G. Untersuchungen über den Tagesgang von Hautdurchblutung und Hauttemperatur unter besonderer Berücksichtigung der physikalischen Temperaturregulation. *Z. Physik. Med. Rehabil.*, 15, 1, 1974.

[38] Reinberg, A. Circadian changes in the temperature of human beings. *Bibl. Radiol.*, 6, 128, 1975.

[39] Schmidt, K.-L., Mäurer, R., and Rusch, D. Zur Wirkung örtlicher Wärme und Kälteanwendungen auf die Hauttemperatur am Kniegelenk. *Z. Rheumatol.*, 38, 213, 1979.

[40] Kanamori, T. et al. Circadian rhythm of body temperature. *Biomed. Thermol.*, 11, 292, 1991.

[41] Friedrich, K.H. Assessment criteria for infrared thermography systems. *Acta thermograp*, 5, 68, 1980.

[42] Alderson, J.K.A. and Ring, E.F.J. "Sprite" high resolution thermal imaging system. *Thermology*, 1, 110, 1985.

[43] Dibley, D.A.G. Opto-mechanical systems for thermal imaging, in *The Thermal Image in Medicine and Biology*, Ammer, K., and Ring, E.F.J., Eds., Uhlen-Verlag, Wien, 1995, p. 33.

[44] Plassmann, P. Advances in image processing for thermology, *Presented at Int. Cong. of Thermology*, Seoul, June 5–6, 2004, p. 3.

[45] Kutas, M. Staring focal plane array for medical thermal imaging, in *The Thermal Image in Medicine and Biology*, Ammer, K., and Ring, E.F.J., Eds., Uhlen-Verlag, Wien, 1995, p. 40.

[46] Ring, E.F.J., Minchinton, M., and Elvins, D.M. A focal plane array system for clinical infrared imaging. *IEEE/EMBS Proceedings*, Atlanta 1999, p. 1120.

[47] Hejazi, S. and Spangler, R.A. A multi-wavelength thermal imaging system, in *Proc. 11th Annual Int. Conf. IEEE Engineering in Medicine and Biology Society*, II, 1989, p. 1153.

[48] Ring, E.F.J. Quality control in infrared thermography, in *Recent Advances in Medical Thermology*, Ring, E.F.J. and Phillips, B., Eds., Plenum Press, New York, 1984, p. 185.

[49] Clark, R.P. et al. Thermography and pedobarography in the assessment of tissue damage in neuropathicand atherosclerotic feet. *Thermology*, 3, 15, 1988.

[50] Clark, J.A. Effects of surface emissivity and viewing angle errors in thermography. *Acta thermograp*, 1, 138, 1976.

[51] Steketee, J. Physical aspects of infrared thermography, in *Recent Advances in Medical Thermology*, Ring, E.F.J. and Phillips, B., Eds., Plenum Press, New York, 1984, p. 167.

[52] Wiecek, B., Jung, A., and Zuber, J. Emissivity-Bottleneck and Challenge for thermography. *Thermol. Int.*, 10, 15, 2000.

[53] Ammer K. Need for standardisation of measurements, in *Thermal Imaging in Thermography and Lasers in Medicine*, Wiecek, B., Ed., Akademickie Centrum Graficzno-Marketigowe Lodart S.A, Lodz, 2003, p. 13.

[54] Anbar, M. Potential artifacts in infrared thermographic measurements. *Thermology*, 3, 273, 1991.

[55] Ring, E.F.J. and Dicks, J.M. Spatial resolution of new thermal imaging systems, *Thermol. Int.*, 9, 7, 1999.

[56] Melnizky, P., Schartelmüller, T., and Ammer, K. Prüfung der intra-und interindividuellen Verlässlichkeit der Auswertung von Infrarot-Thermogrammen. *Eur. J. Thermol.*, 7, 224, 1997.

[57] Ring, E.F.J. and Ammer, K. The technique of thermal imaging in medicine. *Thermol. Int.*, 10, 7, 2000.

[58] Khallaf, A. et al. Thermographic study of heat loss from the face. *Thermol. Österr.*, 4, 49, 1994.

[59] Ishigaki, T. et al. Forehead–back thermal ratio for the interpretation of infrared imaging of spinal cord lesions and other neurological disorders. *Thermology*, 3, 101, 1989.

[60] Schuber, T.R. et al. Directed dynamic cooling,a methodic contribution in telethermography. *Acta thermograp*, 1, 94, 1977.

[61] Di Carlo, A. Thermography in patients with systemic sclerosis. *Thermol. Österr.*, 4, 18, 1994.

[62] Love, T.J. Heat transfer considerations in the design of a thermology clinic. *Thermology*, 1, 88, 1985.

[63] Ring, E.F.J. Computerized thermography for osteo-articular diseases. *Acta thermograp.*, 1, 166, 1976.

[64] Roberts, D.L. and Goodman, P.H. Dynamic thermoregulation of back and upper extremity by computer-aided infrared imaging. *Thermology*, 2, 573, 1987.

[65] Mabuchi, K. et al. Development of a data processing system for a high-speed thermographic camera and its use in analyses of dynamic thermal phenomena of the living body, in *The Thermal Image in Medicine and Biology*, Ammer, K., and Ring, E.F.J., Eds., Uhlen-Verlag, Wien, 1995, p. 56.

[66] Cena, K. Environmental heat loss, in *Recent Advances in Medical Thermology*, Ring, E.F.J. and Phillips, B., Eds., Plenum Press, New York, 1984, p. 81.

[67] Engel, J.-M. Kennzeichnung von Thermogrammen, in *Thermologische Messmethodik*, Engel, J.-M., Flesch, U., and Stüttgen, G., Eds., Notamed, Baden–Baden, 1983, p. 176.

[68] Park, J.-Y. Current development of medical infrared imaging technology, *Presented at Int. Congr. of Thermology*, Seoul, June 5–6, 2004, p. 9.

[69] Plassmann, P. and Ring, E.F.J. An open system for the acquisition and evaluation of medical thermological images. *Eur. J. Thermol.* 7, 216, 1997.

[70] Abramson, D.I. et al. Effect of altering limb position on blood flow, O_2 uptake and skin temperature. *J. Appl. Physiol.*, 17, 191, 1962.

[71] Schartelmüller, T. and Ammer, K. Räumliche Auflösung von Infrarotkameras. *Thermol. Österr.*, 5, 28, 1995.

[72] Ammer, K. Update in standardization and temperature measurement from thermal images, *Presented at Int. Cong. of Thermology*, Seoul, June 5–6, 2004, p. 7.

[73] Mannara, G., Salvatori, G.C., and Pizzuti, G.P. Ethyl alcohol induced skin temperature changes evaluated by thermography. Preliminary results. *Boll. Soc. Ital. Biol. Sper.*, 69, 587, 1993.

[74] Melnizky, P. and Ammer, K. Einfluss von Alkohol und Rauchen auf die Hauttemperatur des Gesichts, der Hände und der Kniegelenke. *Thermol. Int.*, 10, 191, 2000.

[75] Ammer, K., Melnizky, P., and Rathkolb, O. Skin temperature after intake of sparkling wine, still wine or sparkling water. *Thermol. Int.*, 13, 99, 2003.

[76] Gershon-Cohen, J., Borden, A.G., and Hermel, M.B. Thermography of extremities after smoking. *Br. J. Radiol.*, 42, 189, 1969.

[77] Usuki, K. et al. Effects of nicotine on peripheral cutaneous blood flow and skin temperature. *J. Dermatol. Sci.*, 16, 173, 1998.

[78] Di Carlo, A. and Ippolito, F. Early effects of cigarette smoking in hypertensive and normotensive subjects. An ambulatory blood pressure and thermographic study. *Minerva Cardioangiol.*, 51, 387, 2003.

[79] Collins, A.J. et al. Some observations on the pharmacology of "deep-heat," a topical rubifacient. *Ann. Rheum. Dis.*, 43, 411, 1984.

[80] Ring, E.F. Cooling effects of Deep Freeze Cold gel applied to the skin, with and without rubbing, to the lumbar region of the back. *Thermol. Int.*, 14, 64, 2004.

[81] Federspil, G. et al. Study of diet-induced thermogenesis using telethermography in normal and obese subjects. *Recent Prog. Med.*, 80, 455, 1989.

[82] Shlygin, G.K. et al. Radiothermometric research of tissues during the initial reflex period of the specific dynamic action of food. *Med. Radiol. (Mosk)*, 36, 10, 1991.

[83] Danz, J. and Callies, R. Infrarothermometrie bei differenzierten Methoden der Niederfrequenztherapie. *Z. Physiother.*, 31, 35, 1979.

[84] Rusch, F., Neeck, G., and Schmidt, K.L. Über die Hemmung von Erythemen durch Capsaicin. 3.Objektivierung des Capsaicin-Erythems mittels statischer und dynamischer Thermographie, *Z. Phys. Med. Baln. Med. Klim.*, 17, 18, 1988.

[85] Mayr, H., Thür, H., and Ammer, K. Electrical stimulation of the stellate ganglia, in *The Thermal Image in Medicine and Biology*, Ammer, K., and Ring, E.F.J., Eds., Uhlen-Verlag, Wien, 1995, p. 206.

[86] Danz, J. and Callies R. Thermometrische Untersuchungen bei unterschiedlichen Ultraschallintensitäten. *Z. Physiother.*, 30, 235, 1978.

[87] Demmink, J.H., Helders, P.J., Hobaek, H., and Enwemeka, C. The variation of heating depth with therapeutic ultrasound frequency in physiotherapy. *Ultrasound Med. Biol.*, 29, 113–118, 2003.

[88] Rathkolb, O. and Ammer, K. Skin temperature of the fingers after different methods of heating using a wax bath. *Thermol Österr.*, 6, 125, 1996.

[89] Ammer, K. and Schartelmüller, T. Hauttemperatur nach der Anwendung von Wärmepackungen und nach Infrarot-A-Bestrahlung. *Thermol. Österr.*, 3, 51, 1993.

[90] Goodman, P.H., Foote, J.E., and Smith, R.P. Detection of intentionally produced thermal artifacts by repeated thermographic imaging. *Thermology*, 3, 253, 1991.

[91] Dachs, E., Schartelmüller, T., and Ammer, K. Temperatur zur Kryotherapie und Veränderungen der Hauttemperatur am Kniegelenk nach Kaltluftbehandlung. *Thermol. Österr.*, 1, 9, 1991.

[92] Rathkolb, O. et al. Hauttemperatur der Lendenregion nach Anwendung von Kältepackungen unterschiedlicher Größe und Applikationsdauer. *Thermol. Österr.*, 1, 15, 1991.

[93] Ammer, K. Occurrence of hyperthermia after ice massage. *Thermol. Österr.*, 6, 17, 1996.

[94] Cholewka, A. et al. Temperature effects of whole body cryotherapy determined by thermography. *Thermol. Int.*, 14, 57, 2004.

[95] Danz, J., Callies, R., and Hrdina, A. Einfluss einer abgestuften Vakuumsaugmassage auf die Hauttemperatur. *Z. Physiother.*, 33, 85, 1981.

[96] Eisenschenk, A. and Stoboy, H. Thermographische Kontrolle physikalisch-therapeutischer Methoden. *Krankengymnastik*, 37, 294, 1985.

[97] Kainz, A. Quantitative Überprüfung der Massagewirkung mit Hilfe der IR-Thermographie. *Thermol. Österr.*, 3, 79, 1993.

[98] Rusch, D. and Kisselbach, G. Comparative thermographic assessment of lower leg baths in medicinal mineral waters (Nauheim Springs), in *Recent Advances in Medical Thermology*, Ring, E.F.J. and Phillips, B., Eds., Plenum Press, New York, 1984, p. 535.

[99] Ring, E,F.J., Barker, J.R., and Harrison, R.A. Thermal effects of pool therapy on the lower limbs. *Thermology*, 3, 127, 1989.

[100] Konermann, H. and Koob, E. Infrarotthermographische Kontrolle der Effektivität krankengymnastischer Behandlungsmaßnahmen. *Krankengymnastik*, 27, 39, 1975.

[101] Smith, B.L., Bandler, M.K., and Goodman, P.H. Dominant forearm hyperthermia: a study of fifteen athletes. *Thermology*, 2, 25, 1986.

[102] Melnizky, P., Ammer, K., and Schartelmüller, T. Thermographische Überprüfung der Heilgymnastik bei Patienten mit Peroneusparese. *Thermol. Österr.*, 5, 97, 1995.

[103] Ammer, K. Low muscular acitivity of the lower leg in patients with a painful ankle. *Thermol. Österr.*, 5, 103, 1995.

[104] Ring, E.F., Porto, L.O., and Bacon, P.A. Quantitative thermal imaging to assess inositol nicotinate treatment for Raynaud's syndrome. *J. Int. Med. Res.*, 9, 393, 1981.

[105] Lecerof, et al. Acute effects of doxazosin and atenolol on smoking-induced peripheral vasoconstriction in hypertensive habitual smokers. *J. Hypertens.*, 8, S29, 1990.

[106] Tham, T.C., Silke, B., and Taylor, S.H. Comparison of central and peripheral haemodynamic effects of dilevalol and atenolol in essential hypertension. *J. Hum. Hypertens.*, 4, S77, 1990.

[107] Natsuda, H. et al. Nitroglycerin tape for Raynaud's phenomenon of rheumatic disease patients — an evaluation of skin temperature by thermography. *Ryumachi*, 34. 849, 1994.

[108] Engel, J.M. Thermotom- ein Softwarepaket für die thermographische Bildanalyse in der Rheumatologie, in *Thermologische Messmethodik*, Engel, J.-M., Flesch, U., and Stüttgen, G., Eds., Notamed, Baden–Baden, 1983, p. 110.

[109] Bösiger, P. and Scaroni, F. Mikroprozessor-unterstütztes Thermographie-System zur quantitativewn on-line Analyse von statischen und dynamischen Thermogrammen, in *Thermologische Messmethodik*, Engel, J.-M., Flesch, U., and Stüttgen, G., Eds., Notamed, Baden–Baden, 1983, p. 125.

[110] Brandes, P. PIC-Win-Iris Bildverarbeitungssoftware. *Thermol. Österr.*, 4, 33, 1994.

[111] Ring, E.F.J. Quantitative thermography in arthritis using the AGA integrator. *Acta thermograp.*, 2, 172, 1977.

[112] Parr, G. et al. Microcomputer standardization of the AGA 680 M system, in *Recent Advances in Medical Thermology*, Ring, E.F.J. and Phillips, B., Eds., Plenum Press, New York, 1984, pp. 211–214.

[113] Van Hamme, H., De Geest, G., and Cornelis, J. An acquisition and scan conversion unit for the AGA THV680 medical infrared camera. *Thermology*, 3, 205, 1990.

[114] Mayr, H. Korrelation durchschnittlicher und maximaler Temperatur am Kniegelenk bei Auswertung unterschiedlicher Messareale. *Thermol. Österr.*, 5, 89, 1995.

[115] Plassmann, P. On-line Communication for Thermography in Europe, *Presented at Int. Cong. of Thermology*, Seoul, June 5–6, 2004, p. 50.

[116] Ring, E.F.J. Thermal imaging in medico-legal claims. *Thermol. Int.*, 10, 97, 2000.

[117] Sella, G.E. Forensic criteria of acceptability of thermography. *Eur. J. Thermol.*, 7, 205, 1997.

[118] Hirschl, M. et al. Double-blind, randomised, placebo controlled low level laser therapy study in patients with primary Raynaud's phenomenon. *Vasa*, 31, 91, 2002.

[119] Schartelmüller, T., Melnizky, P., and Engelbert, B. Infrarotthermographie zur Evaluierung des Erfolges physikalischer Therapie bei Patienten mit klinischem Verdacht auf Thoracic Outlet Syndrome. *Thermol. Int.*, 9, 20, 1999.

[120] Kim, Y.S. and Cho, Y.E. Pre- and postoperative thermographic imaging in lumbar disc herniations, in *The Thermal Image in Medicine and Biology*, Ammer, K., and Ring, E.F.J., Eds., Uhlen-Verlag, Wien, 1995, p. 168.

[121] Siniewicz, K, et al. Thermal imaging before and after physial exercises in children with orthostatic disorders of the cardiovascular system. *Thermol. Int.*, 12, 139, 2002.

37

Infrared Detectors and Detector Arrays

Paul Norton
Stuart Horn
Joseph G. Pellegrino
Philip Perconti
U.S. Army Communications and Electronics Research, Development and Engineering Center (CERDEC) Night Vision and Electronic Sensors Directorate

There are two general classes of detectors: *photon* (or quantum) and *thermal* detectors [1,2]. Photon detectors convert absorbed photon energy into released electrons (from their bound states to conduction states). The material band gap describes the energy necessary to transition a charge carrier from the valence band to the conduction band. The change in charge carrier state changes the electrical properties of the material. These electrical property variations are measured to determine the amount of incident optical power. Thermal detectors absorb energy over a broad band of wavelengths. The energy absorbed by a detector causes the temperature of the material to increase. Thermal detectors have at least one inherent electrical property that changes with temperature. This temperature-related property is measured electrically to determine the power on the detector. Commercial infrared imaging systems suitable for medical applications use both types of detectors. We begin by describing the physical mechanism employed by these two detector types.

37.1 Photon Detectors

Infrared radiation consists of a flux of photons, the quantum-mechanical elements of all electromagnetic radiation. The energy of the photon is given by:

$$E_{\mathrm{ph}} = h\nu = hc/\lambda = 1.986 \times 10^{-19}/\lambda \ \mathrm{J}/\mu\mathrm{m} \qquad (37.1)$$

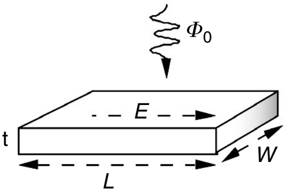


FIGURE 37.1 Photoconductive detector geometery.

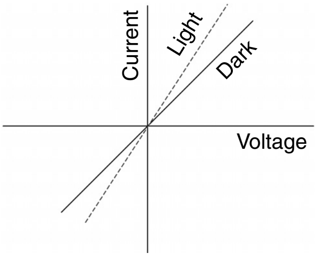


FIGURE 37.2 Current–voltage characteristics of a photoconductive detector.

where h is the Planck's constant, c is the speed of light, and λ is the wavelength of the infrared photon in micrometers (μm).

Photon detectors respond by elevating an bound electron in a material to a free or conductive state. Two types are photon detectors are produced for the commercial market:

- Photoconductive
- Photovoltaic

37.1.1 Photoconductive Detectors

The mechanism of photoconductive detectors is based upon the excitation of bound electrons to a mobile state where they can move freely through the material. The increase in the number of conductive electrons, n, created by the photon flux, Φ_0 allows more current to flow when the detective element is used in a bias circuit having an electric field E. The photoconductive detector element having dimensions of length L, width W, and thickness t is represented in Figure 37.1.

Figure 37.2 illustrates how the current–voltage characteristics of a photoconductor change with incident photon flux (Chapter 4).

The response of a photoconductive detector can be written as:

$$R = \frac{nqRE\tau(\mu_n + \mu_p)}{E_{\mathrm{ph}}L}(V/W) \qquad (37.2)$$

where R is the response in volts per Watt, η is the quantum efficiency in electrons per photon, q is the charge of an electron, R is the resistance of the detector element, τ is the lifetime of a photoexcited electron, and μ_{n} and μ_{p} are the mobilities of the electrons and holes in the material in volts per square centimeter per second.

Noise in photoconductors is the square root averaged sum of terms from three sources:

- Johnson noise
- Thermal generation-recombination
- Photon generation-recombination

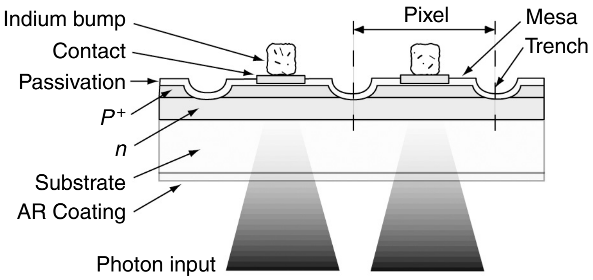


FIGURE 37.3 Photovoltaic detector structure example for mesa diodes.

Expressions for the total noise and each of the noise terms are given in Equation 37.3 to Equation 37.6

$$V_{\text{noise}} = \sqrt{V^2_{\text{Johnson}} + V^2_{\text{ph g-r}} + V^2_{\text{th g-r}}} \tag{37.3}$$

$$V_{\text{Johnson}} = \sqrt{4kTR} \tag{37.4}$$

$$V_{\text{ph g}-\text{r}} = \frac{\sqrt{\eta\phi(WL)}2qRE\tau(\mu_{\text{n}} + \mu_{\text{p}})}{L} \tag{37.5}$$

$$V_{\text{th g}-\text{r}} = \sqrt{\frac{np}{n+p}\tau\left(\frac{Wt}{L}\right)2qRE(\mu_{\text{n}} + \mu_{\text{p}})} \tag{37.6}$$

The figure of merit for infrared detectors is called D^*. The units of D^* are cm $(\text{Hz})^{1/2}$/W, but are most commonly referred to as Jones. D^* is the detector's signal-to-noise (SNR) ratio, normalized to an area of 1 cm^2, to a noise bandwidth of 1 Hz, and to a signal level of 1 W at the peak of the detectors response. The equation for D^* is:

$$D^*_{\text{peak}} = \frac{R}{V_{\text{noise}}}\sqrt{WL}(\text{Jones}) \tag{37.7}$$

where W and L are defined in Figure 37.1.

A special condition of D^* for a photoconductor is noted when the noise is dominated by the photon noise term. This is a condition in which the D^* is maximum.

$$D^*_{\text{blip}} = \frac{\lambda}{2hc}\sqrt{\frac{\eta}{E_{\text{ph}}}} \tag{37.8}$$

where "blip" notes background-limited photodetector.

37.1.2 Photovoltaic Detectors

The mechanism of photovoltaic detectors is based on the collection of photoexcited carriers by a diode junction. Photovoltaic detectors are the most commonly used photon detectors for imaging arrays in current production. An example of the structure of detectors in such an array is illustrated in Figure 37.3 for a mesa photodiode. Photons are incident from the optically transparent detector substrate side and

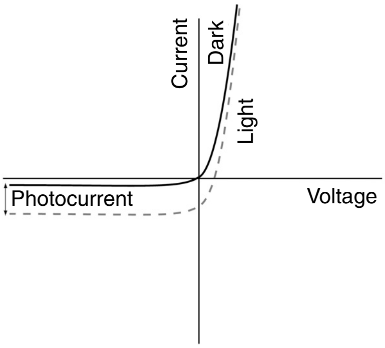


FIGURE 37.4 Current–voltage characteristics of a photovoltaic detector.

are absorbed in the *n*-type material layer. Absorbed photons create a pair of carriers, an electron and a hole. The hole diffuses to the *p*-type side of the junction creating a photocurrent. A contact on the *p*-type side of the junction is connected to an indium bump that mates to an amplifier in a readout circuit where the signal is stored and conveyed to a display during each display frame. A common contact is made to the *n*-type layer at the edge of the detector array. Adjacent diodes are isolated electrically from each other by a mesa etch cutting the *p*-type layer into islands.

Figure 37.4 illustrates how the current–voltage characteristics of a photodiode change with incident photon flux (Chapter 4).

The current of the photodiode can be expressed as:

$$I = I_0(\mathrm{e}^{qV/kT} - 1) - I_{\mathrm{photo}} \tag{37.9}$$

where I_0 is reverse-bias leakage current and I_{photo} is the photoinduced current. The photocurrent is given by:

$$I = I_0(\mathrm{e}^{qV/kT} - 1) - I_{\mathrm{photo}} \tag{37.10}$$

where Φ_0 is the photon flux in photons/cm^2/sec and A is the detector area.

Detector noise in a photodiode includes three terms: Johnson noise, thermal diffusion generation and recombination noise, and photon generation and recombination. The Johnson noise term, written in terms of the detector resistance $\mathrm{d}I/\mathrm{d}V = R_0$ at zero bias as:

$$i_{\mathrm{Johnson}} = \sqrt{4kT/R_0} \tag{37.11}$$

where k is Boltzmann's constant and T is the detector temperature. The thermal diffusion current is given by:

$$i_{\mathrm{diffusion\ noise}} = q\sqrt{2I_{\mathrm{s}}\left[\exp\left(\frac{eV}{kT}\right) - 1\right]} \tag{37.12}$$

where the saturation current, I_{s}, is given by:

$$I_{\mathrm{s}} = qn_i^2\left[\frac{1}{N_{\mathrm{a}}}\sqrt{\frac{D_{\mathrm{n}}}{\tau_{n_0}}} + \frac{1}{N_{\mathrm{d}}}\sqrt{\frac{D_{\mathrm{p}}}{\tau_{p_0}}}\right] \tag{37.13}$$

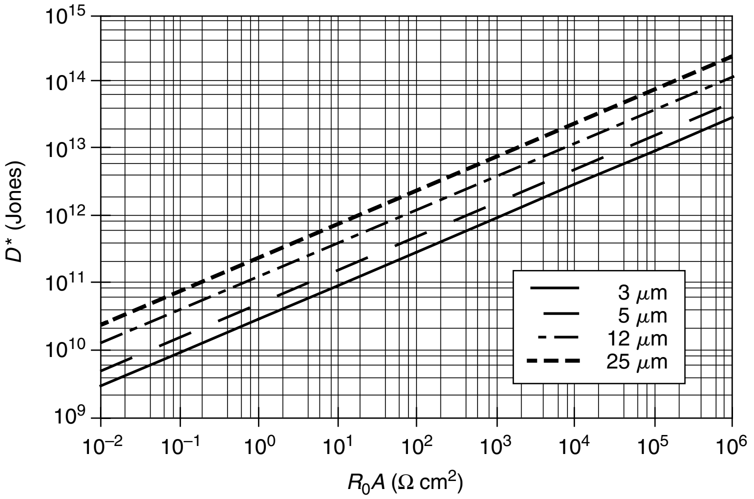


FIGURE 37.5 D^* as a function of the detector resistance-area product, R_0A. This condition applies when detector performance is limited by dark current.

where N_a and N_d are the concentration of p- and n-type dopants on either side of the diode junction, τ_{n0} and τ_{p0} are the carrier lifetimes, and D_n and D_p are the diffusion constants on either side of the junction, respectively.

The photon generation-recombination current noise is given by:

$$i_{\text{photon noise}} = q\sqrt{2\eta\Phi_0} \tag{37.14}$$

When the junction is at zero bias, the photodiode D^* is given by:

$$D_\lambda^* = \frac{\lambda}{hc}\eta e \frac{1}{\left[(4kT/R_0A) + 2e^2\eta\right]} \tag{37.15}$$

In the special case of a photodiode that is operated without sufficient cooling, the maximum D^* may be limited by the dark current or leakage current of the junction. The expression for D^* in this case, written in terms of the junction-resistance area product, R_0A, is given by:

$$D_\lambda^* = \frac{\lambda}{hc}\eta e\sqrt{\frac{R_0A}{4kT}} \tag{37.16}$$

Figure 37.5 illustrates how D^* is limited by the R_0A product for the case of dark-current limited detector conditions.

For the ideal case where the noise is dominated by the photon flux in the background scene, the peak D^* is given by:

$$D_\lambda^* = \frac{\lambda}{hc}\sqrt{\frac{\eta}{2E_{\mathrm{ph}}}} \tag{37.17}$$

Comparing this limit with that for a photoconductive detector in Equation 37.8, we see that the background-limited D^* for a photodiode is higher by a factor of square root of 2 ($\sqrt{2}$).

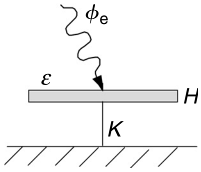


FIGURE 37.6 Abstract bolometer detector structure, where C is the thermal capacitance, G is the thermal conductance, and ε is the emissivity of the surface. Φ_e represents the energy flux in W/cm^2.

37.2 Thermal Detectors

Thermal detectors operate by converting the incoming photon flux to heat [3]. The heat input causes the thermal detector's temperature to rise and this change in temperature is sensed by a bolometer. A bolometer element operates by changing its resistance as its temperature is changed. A bias circuit across the bolometer can be used to convert the changing current to a signal output.

The coefficient α is used to compare the sensitivity of different bolometer materials and is given by:

$$\alpha = \frac{1}{R_d}\frac{dR}{dT} \tag{37.18}$$

where R_d is the resistance of the bolometer element, and dR/dT is the change in resistance per unit change in temperature. Typical values of α are 2 to 3%.

Theoretically, the bolometer structure can be represented as illustrated in Figure 37.6. The rise in temperature due to a heat flux ϕ_e is given by:

$$\Delta T = \frac{\eta P_0}{G(1+\omega^2\tau^2)^{1/2}} \tag{37.19}$$

where P_0 is the radiant power of the signal in watts, G is the thermal conductance (K/W), h is the percentage of flux absorbed, and ω is the angular frequency of the signal. The bolometer time constant, τ, is determined by:

$$\tau = \frac{C}{G} \tag{37.20}$$

where C is the heat capacity of detector element.

The sensitivity or D^* of a thermal detector is limited by variations in the detector temperature caused by fluctuations in the absorption and radiation of heat between the detector element and the background. Sensitive thermal detectors must minimize competing mechanisms for heat loss by the element, namely, convection and conduction.

Convection by air is eliminated by isolating the detector in a vacuum. If the conductive heat losses were less than those due to radiation, then the limiting D^* would be given by:

$$D^*(T,f) = 2.8\times 10^{16}\sqrt{\frac{\varepsilon}{T_2^5+T_1^5}}\,\text{Jone} \tag{37.21}$$

where T_1 is the detector temperature, T_2 the background temperature, and ε the value of the detector's emissivity and equally it's absorption. For the usual case of both the detector and background temperature at normal ambient, 300 K, the limiting D^* is 1.8×10^{10} Jones.

Bolometer operation is constrained by the requirement that the response time of the detector be compatible with the frame rate of the imaging system. Most bolometer cameras operate at a 30 Hz frame

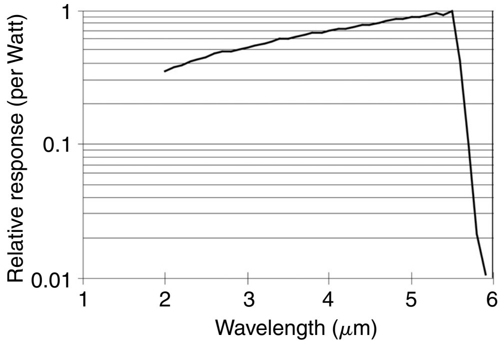


FIGURE 37.7 Spectral response per watt of an InSb detector at 80 K.

rate — 33 msec frame. Response times of the bolometer are usually designed to be on the order of 10 msec. This gives the element a fast enough response to follow scenes with rapidly varying temperatures without objectionable image smearing.

37.3 Detector Materials

The most popular commercial cameras for thermal imaging today use the following detector materials [4]:

- InSb for 5 μm medium wavelength infrared (MWIR) imaging
- $Hg_{1-x}Cd_xTe$ alloys for 5 and 10 μm long wavelength infrared (LWIR) imaging
- Quantum well detectors for 5 and 10 μm imaging
- Uncooled bolometers for 10 μm imaging

We will now review a few of the basic properties of these detector types.

Photovoltaic InSb remains a popular detector for the MWIR spectral band operating at a temperature of 80 K [5,6]. The detector's spectral response at 80 K is shown in Figure 37.7. The spectral response cutoff is about 5.5 μm at 80 K, a good match to the MWIR spectral transmission of the atmosphere. As the operating temperature of InSb is raised, the spectral response extends to longer wavelengths and the dark current increases accordingly. It is thus not normally used above about 100 K. At 80 K the R_0A product of InSb detectors is typically in the range of 10^5 to 10^6 Ω cm^2 — see Equation 37.16 and Figure 37.5 for reference.

Crystals of InSb are grown in bulk boules up to 3 in. in diameter. InSb materials is highly uniform and combined with a planar-implanted process in which the device geometry is precisely controlled, the resulting detector array responsivity is good to excellent. Devices are usually made with a p/n diode polarity using diffusion or ion implantation. Staring arrays of backside illuminated, direct hybrid InSb detectors in 256 × 256, 240 × 320, 480 × 640, 512 × 640, and 1024 × 1024 formats are available from a number of vendors.

HgCdTe detectors are commercially available to cover the spectral range from 1 to 12 μm [7–13]. Figure 37.8 illustrates representative spectral response from photovoltaic devices, the most commonly used type. Crystals of HgCdTe today are mostly grown in thin epitaxial layers on infrared-transparent CdZnTe crystals. SWIR and MWIR material can also be grown on Si substrates with CdZnTe buffer layers. Growth of the epitaxial layers is by liquid phase melts, molecular beams, or by chemical vapor deposition. Substrate dimensions of CdZnTe crystals are in the 25 to 50 cm^2 range and Si wafers up to 5 to 6 in. (12.5 to 15 cm) in diameter have been used for this purpose. The device structure for a typical HgCdTe photodiode is shown in Figure 37.3.

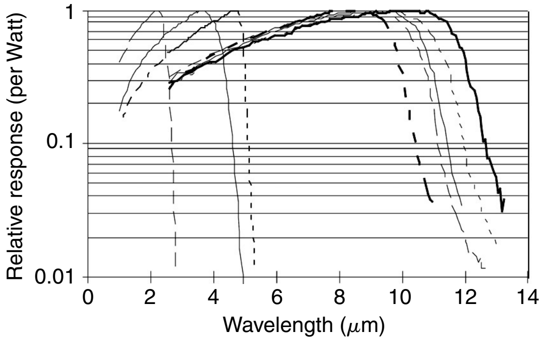


FIGURE 37.8 Representative spectral response curves for a variety of HgCdTe alloy detectors. Spectral cutoff can be varied over the SWIR, MWIR, and LWIR regions.

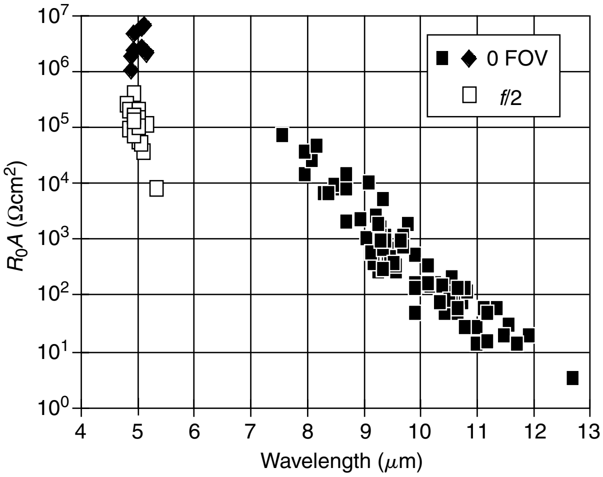


FIGURE 37.9 Values of R_0A product as a function of wavelength for HgCdTe photodiodes. Note that the R_0A product varies slightly with illumination — 0° field-of-view compared with $f/2$ — especially for shorter-wavelength devices.

At 80 K the leakage current of HgCdTe is small enough to provide both MWIR and LWIR detectors that can be photon-noise dominated. Figure 37.9 shows the R_0A product of representative diodes for wavelengths ranging from 4 to 12 μm.

The versatility of HgCdTe detector material is directly related to being able to grow a broad range of alloy compositions in order to optimize the response at a particular wavelength. Alloys are usually adjusted to provide response in the 1 to 3 μm short wavelength infrared (SWIR), 3 to 5 μm MWIR, or the 8 to 12 μm LWIR spectral regions. Short wavelength detectors can operate uncooled, or with thermoelectric coolers that have no moving parts. Medium and long wavelength detectors are generally operated at 80 K using a cryogenic cooler engine. HgCdTe detectors in 256 × 256, 240 × 320, 480 × 640, and 512 × 640 formats are available from a number of vendors.

Quantum well infrared photodetectors (QWIPs) consist of alternating layers of semiconductor material with larger and narrower bandgaps [14–20]. This series of alternating semiconductor layers is deposited one layer upon another using an ultrahigh vacuum technique such as molecular beam epitaxy (MBE). Alternating large and narrow bandgap materials give rise to quantum wells that provide bound and quasi-bound states for electrons or holes [1–5].

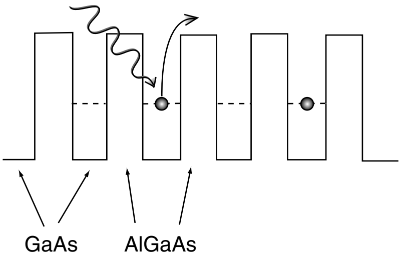


FIGURE 37.10 Quantum wells generate bound states for electrons in the conduction band. The conduction bands for a QWIP structure are shown consisting of $Al_xGa_{1-x}As$ barriers and GaAs wells. For a given pair of materials having a fixed conduction band offset, the binding energy of an electron in the well can be adjusted by varying the width of the well. With an applied bias, photoexcited electrons from the GaAs wells are transported and detected as photocurrent.

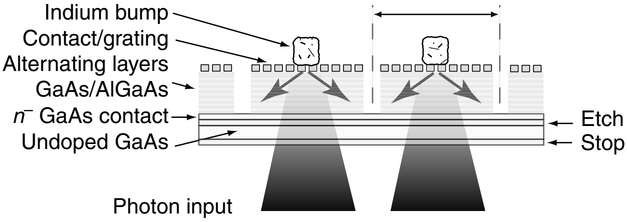


FIGURE 37.11 Backside illuminated QWIP structure with a top side diffraction grating/contact metal. Normally-incident light is coupled horizontally into the quantum wells by scattering off a diffraction grating located at the top of the focal plane array.

Many simple QWIP structures have used GaAs as the narrow bandgap quantum well material and $Al_xGa_{1-x}As$ as the wide bandgap barrier layers as shown in Figure 37.10. The properties of the QWIP are related to the structural design and can be specified by the well width, barrier height, and doping density. In turn, these parameters can be tuned by controlling the cell temperatures of the gallium, aluminum, and arsenic cells as well as the doping cell temperature. The quantum well width (thickness) is governed by the time interval for which the Ga and As cell shutters are left opened. The barrier height is regulated by the composition of the $Al_xGa_{1-x}As$ layers, which are determined by the relative temperature of the Al and Ga cells. QWIP detectors rely on the absorption of incident radiation within the quantum well and typically the well material is doped *n*-type at an approximate level of 5×10^{17}.

The QWIP detectors require that an electric field component of the incident radiation be perpendicular to the layer planes of the device. Imaging arrays use diffraction gratings as shown in Figure 37.11. In particular, the latter approach is of practical importance in order to realize two-dimensional detector arrays. The QWIP focal plane array is a reticulated structure formed by conventional photolithographic techniques. Part of the processing involves placing a two-dimensional metallic grating over the focal plane pixels. The grating metal is typically angled at 45° patterns to reflect incident light obliquely so as to couple the perpendicular component of the electric field into the quantum wells thus producing the photoexcitation. The substrate material (GaAs) is backside thinned and a chemical/mechanical polish is used to produce a mirrorlike finish on the backside. The front side of the pixels with indium bumps are flip-chip bonded to a readout IC. Light travels through the back side and is unabsorbed during its first pass through the epilayers; upon scattering with a horizontal propagation component from the grating some of it is then absorbed by the quantum wells, photoexciting carriers. An electric field is produced perpendicular to the layers by applying a bias voltage at doped contact layers. The structure then behaves as a photoconductor.

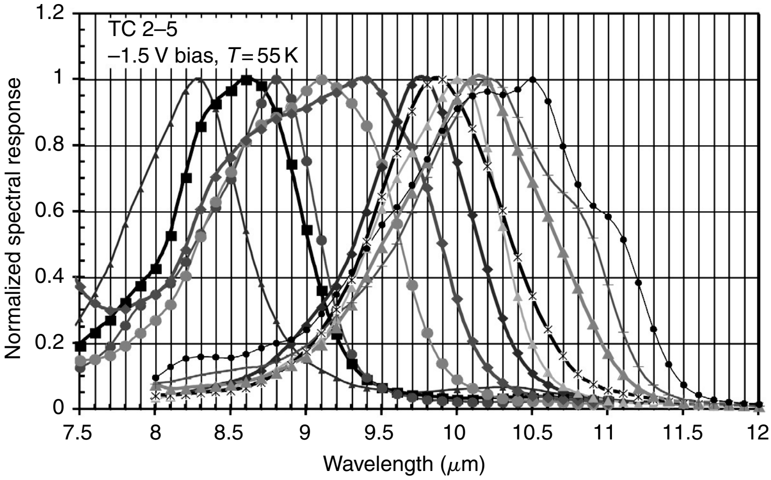


FIGURE 37.12 Representative spectral response of QWIP detectors.

The QWIP detectors require cooling to about 60 K for LWIR operation in order to adequately reduce the dark current. They also have comparatively low quantum efficiency, generally less than 10%. They thus require longer signal integration times than InSb or HgCdTe devices. However, the abundance of radiation in the LWIR band in particular allows QWIP detectors to still achieve excellent performance in infrared cameras.

The maturity of the GaAs-technology makes QWIPs particularly suited for large commercial focal plane arrays with high spatial resolution. Excellent lateral homogeneity is achieved, thus giving rise to a small fixed-pattern noise. QWIPs have an extremely small 1/f noise compared to interband detectors (like HgCdTe or InSb), which is particularly useful if long integration times or image accumulation are required. For these reasons, QWIP is the detector technology of choice for many applications where somewhat smaller quantum efficiencies and lower operation temperatures, compared to interband devices, are tolerable. QWIPs are finding useful applications in surveillance, night vision, quality control, inspection, environmental sciences, and medicine.

Quantum well infrared detectors are available in the 5- and 10-μm spectral region. The spectral response of QWIP detectors can be tuned to a wide range of values by adjusting the width and depth of quantum wells formed in alternating layers of GaAs and GaAlAs. An example of the spectral response from a variety of such structures is shown in Figure 37.12. QWIP spectral response is generally limited to fairly narrow spectral bandwidth — approximately 10 to 20% of the peak response wavelength. QWIP detectors have higher dark currents than InSb or HgCdTe devices and generally must be cooled to about 60 K for LWIR operation.

The quantum efficiencies of InSb, HgCdTe, and QWIP photon detectors are compared in Figure 37.13. With antireflection coating, InSb and HgCdTe are able to convert about 90% of the incoming photon flux to electrons. The QWIP quantum efficiencies are significantly lower, but work at improving them continues to occupy the attention of research teams.

We conclude this section with a description of Type-II superlattice detectors [21–26]. Although Type-II superlattice detectors are not yet used in arrays for in commercial camera system, the technology is briefly reviewed here because of its potential future importance. This material system mimics an intrinsic detector material such as HgCdTe, but is "bandgap engineered." Type-II superlattice structures are fabricated from multilayer stacks of alternating layers of two different semiconductor materials. Figure 37.14 illustrates the structure. The conduction band minimum is in one layer and the valence band minimum is in the adjacent layer (as opposed to both minima being in the same layer as in a Type-I superlattice).

The idea of using type-II superlattices for LWIR detectors was originally proposed in 1977. Recent work on the MBE growth of Type-II systems by [7] has led to the exploitation of these materials for IR

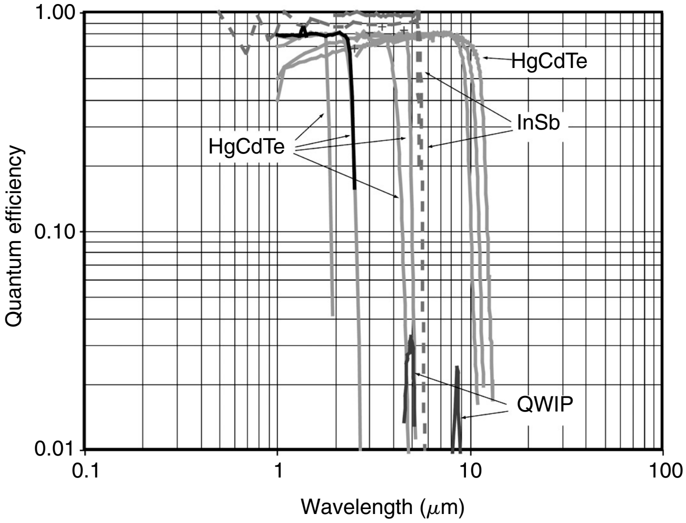


FIGURE 37.13 Comparison of the quantum efficiencies of commercial infrared photon detectors. This figure represents devices that have been antireflection coated.

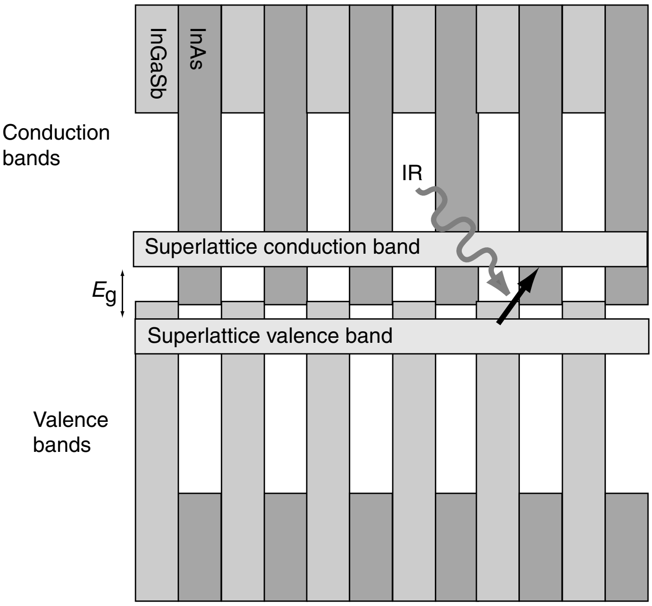


FIGURE 37.14 Band diagram of a short-period InAs/(In,Ga)Sb superlattice showing an infrared transition from the heavy hole (hh) miniband to the electron (e) miniband.

detectors. Short period superlattices of, for example, strain-balanced InAs/(Ga,In)Sb lead to the formation of conduction and valence minibands. In these band states heavy holes are largely confined to the (Ga,In)Sb layers and electrons are primarily confined to the InAs layers. However, because of the relatively low electron mass in InAs, the electron wave functions extend considerably beyond the interfaces and have significant overlap with heavy-hole wave functions. Hence, significant absorption is possible at the minigap energy (which is tunable by changing layer thickness and barrier height).

Cutoff wavelengths from 3 to 20 μm and beyond are potentially possible with this system. Unlike QWIP detectors, the absorption of normally incident flux is permitted by selection rules, obviating the need for grating structures or corrugations that are needed with QWIPs. Finally, Auger transition rates, which place intrinsic limits on the performance of these detectors and severely impact the lifetimes found in bulk, narrow-gap detectors, can be minimized by judicious choices of the structure's geometry and strain profile.

In the future, further advantages may be achievable by using the InAs/Ga(As,Sb) material system where both the InAs and Ga(As,Sb) layers may be lattice matched to InAs substrates. The intrinsic quality obtainable in these structures can be in principle superior to that obtained in InAs/(Ga,In)Sb structures. Since dislocations may be reduced to a minimum in the InAs/Ga(As,Sb) material system, it may be the most suitable Type-II material for making large arrays of photovoltaic detectors.

Development efforts for Type-II superlattice detectors are primarily focused on improving material quality and identifying sources of unwanted leakage currents. The most challenging problem currently is to passivate the exposed sidewalls of the superlattices layers where the pixels are etched in fabrication. Advances in these areas should result in a new class of IR detectors with the potential for high performance at high operating temperatures.

37.4 Detector Readouts

Detectors themselves are isolated arrays of photodiodes, photoconductors, or bolometers. Detectors need a readout to integrate or sample their output and convey the signal in an orderly sequence to a signal processor and display [27].

Almost all readouts are integrated circuits (ICs) made from silicon. They are commonly referred to as readout integrated circuits, or ROICs. Here we briefly describe the functions and features of these readouts, first for photon detectors and then for thermal detectors.

37.4.1 Readouts for Photon Detectors

Photon detectors are typically assembled as a hybrid structure, as illustrated in Figure 37.15. Each pixel of the detector array is connected to the unit cell of the readout through an indium bump. Indium bumps allow for a soft, low-temperature metal connection to convey the signal from the detector to the readout's input circuit.

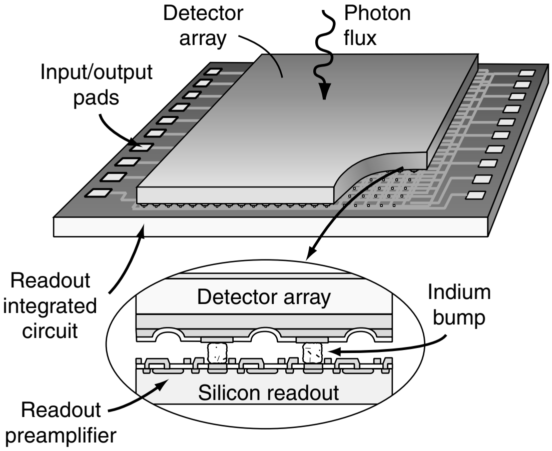


FIGURE 37.15 Hybrid detector array structure consists of a detector array connected to a readout array with indium metal bumps. Detector elements are usually photodiodes or photoconductors, although photocapacitors are sometimes used. Each pixel in the readout contains at least one addressable switch, and more often a preampflifier or buffer together with a charge storage capacitor for integrating the photosignal.

Commercial thermal imagers that operate in the MWIR and LWIR spectral regions generally employ a direct injection circuit to collect the detector signal. This is because this circuit is simple and works well with the relatively high photon currents in these spectral bands. The direct injection transistor feeds the signal onto an integrating capacitor where it stored for a time called the integration time. The integration time is typically around 200 μsec for the LWIR spectral band and 2 msec for the MWIR band, corresponding to the comparative difference in the photon flux available. The integration time is limited by the size of the integration capacitor. Typical capacitors can hold on the order of 3×10^7 electrons.

For cameras operating in the SWIR band, the lower flux levels typically require a more complicated input amplifier. The most common choice employs a capacitive feedback circuit, providing the ability to have significant gain at the pixel level before storage on an integrating capacitor.

Two readout modes are employed, depending upon the readout design:

- Snapshot
- Rolling frame

In the snapshot mode, all pixels integrate simultaneously, are stored, and then read out in sequence, followed by resetting the integration capacitors. In the rolling frame mode the capacitors of each row are reset after each pixel in that row is read. In this case each pixel integrates in different parts of the image frame. A variant of the rolling frame is an interlaced output. In this case the even rows are read out in the first frame and the odd rows in the next. This corresponds to how standard U.S. television displays function.

It is common for each column in the readout to have an amplifier to provide some gain to the signal coming from each row as it is read. The column amplifier outputs are then fed to the output amplifiers. Commercial readouts typically have one, two, or four outputs, depending upon the array size and frame rate. Most commercial cameras operate at 30 or 60 Hz.

Another common feature found on some readouts is the ability to operate at higher frame rates on a subset of the full array. This ability is called windowing. It allows data to be collected more quickly on a limited portion of the image.

37.4.2 Thermal Detector Readouts

Bolometer detectors have comparatively lower resistance than photon detectors and relatively slow inherent response times. This condition allows readouts that do not have to integrate the charge during the frame, but only need to sample it for a brief time. This mode is frequently referred to as pulse-biased.

The unit cell of the bolometer contains only a switch that is pulsed on once per frame to allow current to flow from each row in turn to the column amplifiers. Bias is supplied by the row multiplexer. Sample times for each detector are typically on the order of the frame time divided by the number of rows. Many designs employ differential input column amplifiers that are simultaneously fed an input from a dummy or blind bolometer element in order to subtract a large fraction of the current that flows when the element is biased.

The nature of bolometer operation means that the readout mode is rolling frame. Some designs also provide interlaced outputs for input to TV-like displays.

37.4.3 Readout Evolution

Early readouts required multiple bias supply inputs and multiple clock signals for operation. Today only two clocks and two bias supplies are typically required. The master clock sets the frame rate. The integration clock sets the time that the readout signal is integrated, or that the readout bias pulse is applied. On-chip clock and bias circuits generate the additional clocks and biases required to run the readout. Separate grounds for the analog and digital chip circuitry are usually employed to minimize noise.

Current development efforts are beginning to add on-chip analog-to-digital (A/D) converters to the readout. This feature provides a direct digital output, avoiding significant difficulties in controlling extraneous noise when the sensor is integrated with an imaging or camera system.

37.5 Technical Challenges for Infrared Detectors

Twenty-five years ago, infrared imagining was revolutionized by the introduction of the Probeye Infrared camera. At a modest 8 pounds, Probeye enabled handheld operation, a feature previously unheard of at that time when very large, very expensive IR imaging systems were the rule. Infrared components and technologies have advanced considerably since then. With the introduction of the Indigo Systems Omega camera, one can now acquire a complete infrared camera weighing less than 100 g and occupying 3.5 in.^{3}.

Many forces are at play enabling this dramatic reduction in camera size. Virtually all of these can be traced to improvements in the silicon IC processing industry. Largely enabled by advancements in photolithography, but additionally aided by improvements in vacuum deposition equipment, device feature sizes have been steadily reduced. It was not too long ago that the minimum device feature size was just pushing to break the 1-μm barrier. Today, foundries are focused on production implementation of 65 to 90 nm feature sizes.

The motivation behind such significant improvements has been the high-dollar/high-volume commercial electronics business. Silicon foundries have expended billions of dollars in capitalization and R&D aimed at increasing the density and speed of the transistors per unit chip area. Cellular telephones, personal data assistants (PDAs), and laptop computers are all applications demanding smaller size, lower power, and more features — performance — from electronic components. Infrared detector arrays and cameras have taken direct advantage of these advancements.

37.5.1 Uncooled Infrared Detector Challenges

The major challenge for all infrared markets is to reduce the pixel size while increasing the sensitivity. Reduction from a 50-μm pixel to a 25-μm pixel, while maintaining or even reducing noise equivalent temperature difference (NETD), is a major goal that is now being widely demonstrated (see Figure 37.16). The trends are illustrated by a simple examination of a highly idealized bolometer: the DC response of a detector in which we neglect all noise terms except temperature fluctuation noise, and the thermal conductance value is not detector area dependent (i.e., we are not at or near the radiation conductance limit). Using these assumptions, reducing the pixel area by a factor of four will reduce the SNR by a factor

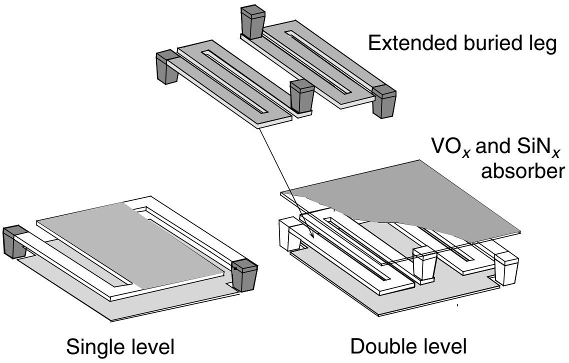


FIGURE 37.16 Uncooled microbolometer pixel structures having noise-equivalent temperature difference (NEΔT) values <50 mK: single level for 2 mil (50 μm) pixels in a 240 × 320 format and double level for 1 mil (25 μm) pixels in a 480 × 640 format (courtesy of Raytheon Vision Systems).

of eight as shown below:

$$\Delta T_{\text{signal}|\text{DCresponse}} = \frac{P_{\text{signalDC}}}{G_{\text{th}}} = \frac{\gamma A_{\text{D}} I_{\text{light}}}{G_{\text{th}}} \tag{37.22}$$

where P_{signalDC} is the DC signal from IR radiation (absorbed power) [W], A_{D} is the detector area [m^2], I_{light} is the light intensity [W/m^2], G_{th} is the thermal conductance [W/K], and γ is a constant that accounts for reflectivity and other factors not relevant to this analysis.

For a detector in the thermal fluctuation limit, the root mean square temperature fluctuation noise is a function of the incident radiation and the thermal conductance of the bolometer bridge.

$$\Delta T_{\text{noise}} \sqrt{\langle \Delta T^2 \rangle} = \sqrt{\frac{kT^2}{C_{\text{th}}}} \tag{37.23}$$

where T is the operating temperature in Kelvin, k is Boltzman's constant, and C_{th} is the total heat capacity of the detector in Joules per Kelvin [J/K].

The total heat capacity can be written as $C_{\text{th}} = c_{\text{p}} A_{\text{d}} Z_{\text{bridge}}$, where Z_{bridge} is the bolometer bridge thickness in meters and c_{p} is the specific heat of the detector in J/K-m^3 .

The signal to noise (SNR) is then

$$\frac{\Delta T_{\text{signal}}}{\Delta T_{\text{noise}}} = \frac{\gamma A_{\text{D}} I_{\text{light}}}{G_{\text{th}}} \sqrt{\frac{c_{\text{p}} A_{\text{D}} Z_{\text{bridge}}}{kT^2}} = \frac{\gamma A_{\text{D}} I_{\text{light}}}{G_{\text{th}}} A_{\text{D}}^{3/2} \sqrt{\frac{c_{\text{p}} Z_{\text{bridge}}}{kT^2}} \tag{37.24}$$

It can be seen that the SNR goes as the area to the three halves. Therefore, a 4× reduction in detector area reduces the SNR by a factor of eight for this ideal bolometer case. Thermal conductance is assumed constant, that is, the ratio of leg length to thickness remains constant as the detector area is reduced. In practical constructions, reducing the pixel linear dimensions by 2× also reduces the leg length by 2×, thus the thermal conductance increases and aggravates the problem. In order to improve the SNR caused by the 4× loss in area, one may be tempted to reduce the thermal conductance G_{th} by 8×. To accomplish this, the length of the legs must be increased and their thickness reduced. By folding the legs under the detector, as seen in Figure 37.10, one can achieve this result. However, an 8× reduction in thermal conductance would result in a detrimental increase in the thermal time constant.

The thermal time constant is given by $\tau_{\text{thermal}} = C_{\text{th}}/G_{\text{th}}$. The heat capacity is reduced by 4× because of the area loss. If G_{th} is reduced by a factor of 8×, then $\tau_{\text{thermal}} = 2C_{\text{th}}/G_{\text{th}}$ is increased by a factor of two. This image smear associated with this increased time constant would prove problematic for practical military applications.

In order to maintain the same time constant, the total heat capacity must be reduced accordingly. Making the detector thinner may achieve this result except that it also increases the temperature fluctuation noise. From this simple example one can readily see the inherent relationship between SNR and the thermal time constant.

We would like to maintain both an equivalent SNR and thermal time constant as the detector cell size is decreased. This can be achieved by maintaining the relationships between the thermal conductance, detector area, and bridge thickness as shown in the following.

The thermal time constant is given by the following:

$$\tau_{\text{thermal}} = \frac{C_{\text{th}}}{G_{\text{th}}} = \frac{c_{\text{p}} A_{\text{D}} Z_{\text{bridge}}}{G_{\text{th}}} \tag{37.25}$$

Equating the thermal time constant of the large and small pixels equal and doing the same with the SNR leads to the following relationships, where the primed variables are the parameters required for the new detector cell:

$$\tau_{\text{thermal}} = \frac{c_{\text{p}} A_{\text{D}} Z_{\text{bridge}}}{G_{\text{th}}} = \frac{c_{\text{p}} A'_{\text{D}} Z'_{\text{bridge}}}{G'_{\text{th}}} \tag{37.26}$$

$$\frac{\Delta T_{\text{signal}}}{\Delta T_{\text{noise}}} = \frac{\gamma A_{\text{D}} I_{\text{light}}}{G_{\text{th}}} \sqrt{\frac{c_{\text{p}} A_{\text{D}} Z_{\text{bridge}}}{kT^2}} = \frac{\gamma A'_{\text{D}} I_{\text{light}}}{G'_{\text{th}}} \sqrt{\frac{c_{\text{p}} A'_{\text{D}} Z'_{\text{bridge}}}{kT^2}} \tag{37.27}$$

Rearranging τ_{thermal} to find the ratio $G_{\text{th}}/G'_{\text{th}}$ and substituting into the SNR, we obtain:

$$\frac{Z'_{\text{bridge}}}{Z_{\text{bridge}}} = \frac{A'_D}{A_D}, \quad \text{and it follows that } \frac{G'_{\text{th}}}{G_{\text{th}}} = \left(\frac{Z'_{\text{bridge}}}{Z_{\text{bridge}}}\right)^2 \tag{37.28}$$

So, it becomes evident that a 4× reduction in pixel cell area requires a 16×, and not an 8×, reduction in thermal conductance to maintain equivalent SNR and thermal time constant. This gives some insight into the problems of designing small pixel bolometers for high sensitivity. It should be noted that in current implementations, the state-of-the-art sensitivity is about 10× from the thermal limits.

37.5.2 Electronics Challenges

Specific technology improvements spawned by the commercial electronics business that have enabled size reductions in IR camera signal processing electronics include:

- Faster digital signal processors (DSPs) with internal memory $\geq$1 MB)
- Higher-density field-programmable gate arrays (FPGAs) (>200 K gates and with an embedded processor core
- Higher-density static (synchronous?) random access memory >4 MB
- Low-power, 14-bit differential A/D converters

Another enabler, also attributable to the silicon industry, is reduction in the required core voltage of these devices (see Figure 37.17). Five years ago, the input voltage for virtually all-electronic components was 5 V. Today, one can buy a DSP with a core voltage as low as 1.2 V. Power consumption of the device is proportional to the square of the voltage. So a reduction from 5- to 1.2-V core represents more than an order of magnitude power reduction.

The input voltage ranges for most components (e.g., FPGAs, memories, etc.) are following the same trends. These reductions are not only a boon for reduced power consumption, but also these lower power devices typically come in much smaller footprints. IC packaging advancements have kept up with the

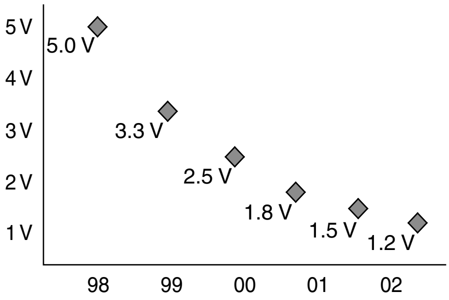


FIGURE 37.17 IC device core voltage vs. time

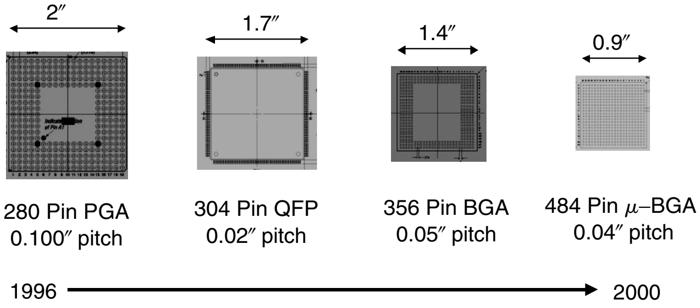


FIGURE 37.18 Advancements in component packaging miniaturization together with increasing pin count that enables reduced camera volume.

higher-density, lower-power devices. One can now obtain a device with almost twice the number of I/Os in 25% of the required area (see Figure 37.18).

All of these lower power, smaller footprint components exist by virtue of the significant demand created by the commercial electronics industry. These trends will continue. Moore's law (logic density in bits/in.^{2} will double every 18 months) nicely describes the degree by which we can expect further advancements.

37.5.3 Detector Readout Challenges

The realization of tighter design rules positively affects reduction in camera size in yet another way. Multiplexers, or ROICs, directly benefit from the increased density. Now, without enlarging the size of the ROIC die, more functions can be contained in the device. On-ROIC A/D conversion eliminates the need for a dedicated, discrete A/D converter. On-ROIC clock and bias generation reduces the number of vacuum Dewar feedthroughs to yield a smaller package as well as reducing the complexity and size of the camera power supply. Putting the nonuniformity correction circuitry on the ROIC reduces the magnitude of the detector output signal swing and minimizes the required input dynamic range of the A/D converter. All of these increases in ROIC functionality come with the increased density of the silicon fabrication process.

37.5.4 Optics Challenges

Another continuing advancement that has helped reduced the size of IR cameras is the progress made at increasing the performance of the uncooled detectors themselves. The gains made at increasing the sensitivity of the detectors has directly translated to reduction in the size of the optics. With a sensitivity goal of 100 mK, an $F/1$ optic has traditionally been required to collect enough energy. Given the recent sensitivity improvements in detectors, achievement of 100 mK can be attained with an $F/2$ optic. This reduction in required aperture size greatly reduces the camera size and weight. These improvements in detector sensitivity can also be directly traceable to improvements in the silicon industry. The same photolithography and vacuum deposition equipments used to fabricate commercial ICs are used to make bolometers. The finer geometry line widths translate directly to increased thermal isolation and increased fill factor, both of which are factors in increased responsivity.

Reduction in optics' size was based on a sequence of NEDT performance improvements in uncooled VO_x microbolometer detectors so that faster optics $F/1.4$ to $F/2$ could be utilized in the camera and still maintain a moderate performance level. As indicated by Equation 37.29 to Equation 37.33, the size of the optics is based on the required field-of-view (FOV), number of detectors (format of the detector array), area of the detector, and $F\#$ of the optics (see Figure 37.19). The volume of the optics is considered to be approximately a cylinder with a volume of $\pi r^2 L$. In Equation 37.29 to Equation 37.33, FL is the optics focal length equivalent to L, D_o is the optics diameter and $D_o/2$ is equivalent to r, A_{det} is the area of the

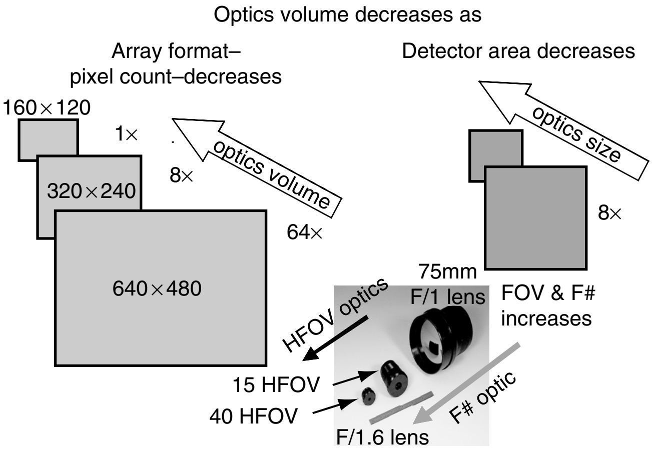


FIGURE 37.19 Trade-off between optics size and volume and $f/\#$, array format, and pixel size.

detector, *F#* is the *f*-number of the optics and HFOV is the horizontal field-of-view.

$$\mathrm{FL} = \frac{\#\text{ horizontal detectors}}{\mathrm{Tan(HFOV/2)}} = \frac{\sqrt{A_{\mathrm{det}}}}{2} \tag{37.29}$$

$$D_0 = \frac{\#\text{ horizontal detectors}}{\mathrm{Tan(HFOV/2)}} = \frac{\sqrt{A_{\mathrm{det}}}}{2\mathrm{F\#}} \tag{37.30}$$

$$\mathrm{F\#} = \frac{\mathrm{FL}}{D_0} \tag{37.31}$$

$$\begin{aligned}\mathrm{Volume_{optics}} &= \pi\left[\frac{\mathrm{D_0}}{2}\right]^2 = \mathrm{FL}\\ &= \pi\left[\frac{(\#\text{ horizontal detectors}/(\tan(\mathrm{HFOV}/2))) = (\sqrt{A_{\mathrm{det}}}/2\mathrm{F\#})}{2}\right]^2 = \mathrm{FL}\end{aligned} \tag{37.32}$$

$$\mathrm{Volume_{optics}} = \pi\left[\frac{(\#\text{ horizontal detectors}/(\tan(\mathrm{HFOV}/2))) = \sqrt{A_{\mathrm{det}}}}{32\mathrm{F\#}^2}\right]^3 \tag{37.33}$$

Uncooled cameras have utilized the above enhancements and are now only a few ounces in weight and require only about 1 W of input power.

37.5.5 Challenges for Third-Generation Cooled Imagers

Third-generation cooled imagers are being developed to greatly extend the range at which targets can be detected and identified [28–30]. U.S. Army rules of engagement now require identification prior to attack. Since deployment of first- and second-generation sensors there has been a gradual proliferation of thermal imaging technology worldwide. Third-generation sensors are intended to ensure that U.S. Army forces maintain a technological advantage in night operations over any opposing force.

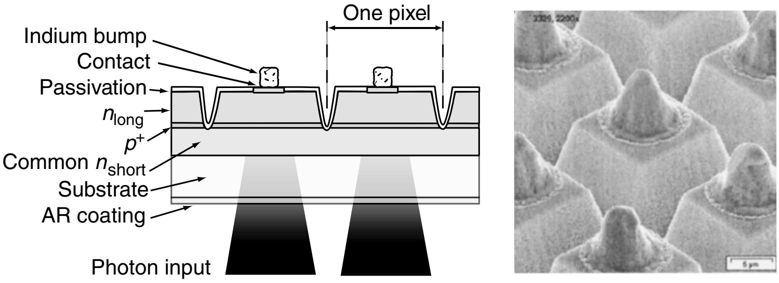


FIGURE 37.20 Illustration of a simultaneous two-color pixel structure — cross section and SEM. Simultaneous two-color FPAs have two indium bumps per pixel. A 50-μm simultaneous two-color pixel is shown.

Thermal imaging equipment is used to first detect an object, and then to identify it. In the detection mode, the optical system provides a wide field-of-view (WFOV — $f/2.5$) to maintain robust situational awareness [31]. For detection, LWIR provides superior range under most Army fighting conditions. Medium wavelength infrared offers higher spatial resolution sensing, and a significant advantage for long-range identification when used with telephoto optics (NFOV — $f/6$).

37.5.5.1 Cost Challenges — Chip Size

Cost is a direct function of the chip size since the number of detector and readout die per wafer is inversely proportion to the chip area. Chip size in turn is set by the array format and pixel size. Third-generation imager formats are anticipated to be in a high-definition 16 × 9 layout, compatible with future display standards, and reflecting the soldier's preference for a wide field-of-view. An example of such a format is 1280 × 720 pixels. For a 30 μm pixel this format yields a die size greater than 1.5 × 0.85 in. (22 × 38 mm). This will yield only a few die per wafer, and will also require the development of a new generation of dewar-cooler assemblies to accommodate these large dimensions. A pixel size of 20 μm results in a cost saving of more than 2×, and allows the use of existing dewar designs.

37.5.5.1.1 Two-Color Pixel Designs

Pixel size is the most important factor for achieving affordable third-generation systems. Two types of two-color pixels have been demonstrated. Simultaneous two-color pixels have two indium–bump connections per pixel to allow readout of both color bands at the same time. Figure 37.20 shows an example of a simultaneous two-color pixel structure. The sequential two-color approach requires only one indium bump per pixel, but requires the readout circuit to alternate bias polarities multiple times during each frame. An example of this structure is illustrated in Figure 37.21. Both approaches leave very little area available for the indium bump(s) as the pixel size is made smaller. Advanced etching technology is being developed in order to meet the challenge of shrinking the pixel size to 20 μm.

37.5.5.2 Sensor Format and Packaging Issues

The sensor format was selected to provide a wide field-of-view and high spatial resolution. Target detection in many Army battlefield situations is most favorable in LWIR. Searching for targets is more efficient in a wider field-of-view, in this case $F/2.5$. Target identification relies on having 12 or more pixels across the target to adequately distinguish its shape and features. Higher magnification, $F/6$ optics combined with MWIR optical resolution enhances this task.

Consideration was also given to compatibility with future standards for display formats. Army soldiers are generally more concerned with the width of the display than the height, so the emerging 16 : 9 width to height format that is planned for high-definition TV was chosen.

A major consideration in selecting a format was the packaging requirements. Infrared sensors must be packaged in a vacuum enclosure and mated with a mechanical cooler for operation. Overall array size was therefore limited to approximately 1 in. so that it would fit in an existing standard advanced dewar

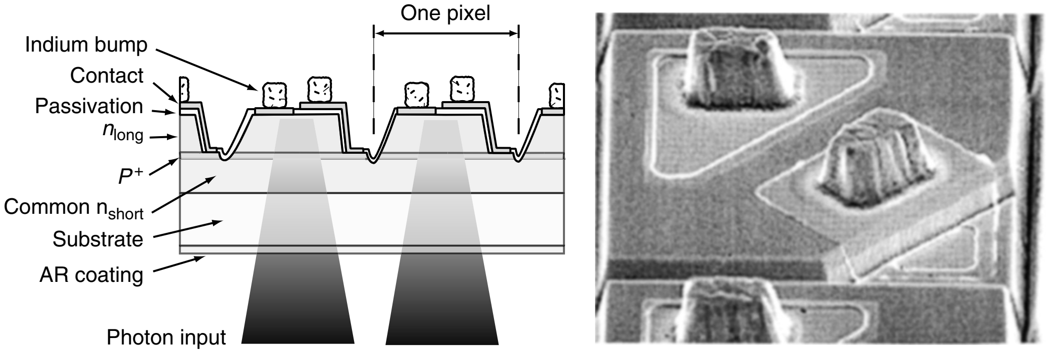


FIGURE 37.21 Illustration of a sequential two-color pixel structure — cross section and SEM. Sequential two-color FPAs have only one indium bump per pixel, helping to reduce pixel size. A 20 μm sequential two-color pixel is shown.

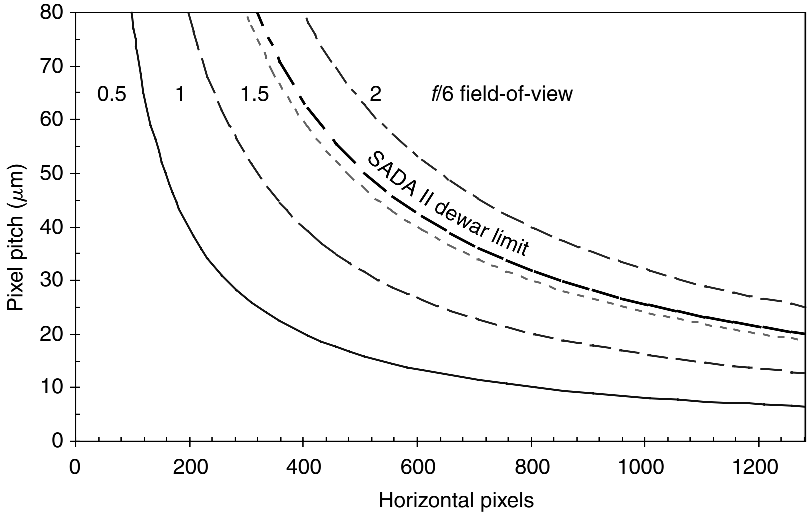


FIGURE 37.22 Maximum array horizontal format is determined by the pixel size and the chip size limit that will fit in an existing SADA dewar design for production commonality. For a 20-μm pixel and a 1.6° FOV, the horizontal pixel count limit is 1280. A costly development program would be necessary to develop a new, larger dewar.

assembly (SADA) dewar design. Figure 37.22 illustrates the pixel size/format/field-of-view trade within the design size constraints of the SADA dewar.

37.5.5.3 Temperature Cycling Fatigue

Modern cooled infrared focal plane arrays are hybrid structures comprising a detector array mated to a silicon readout array with indium bumps (see Figure 37.15).

Very large focal plane arrays may exceed the limits of hybrid reliability engineered into these structures. The problem stems from the differential rates of expansion between HgCdTe and Si, which results in large stress as a device is cooled from 300 K ambient to an operating temperature in the range of 77 to 200 K. Hybrids currently use mechanical constraints to force the contraction of the two components to closely match each other. This approach may have limits — when the stress reaches a point where the chip fractures.

Two new approaches exist that can extend the maximum array size considerably. One is the use of silicon as the substrate for growing the HgCdTe detector layer using MBE. This approach has shown excellent

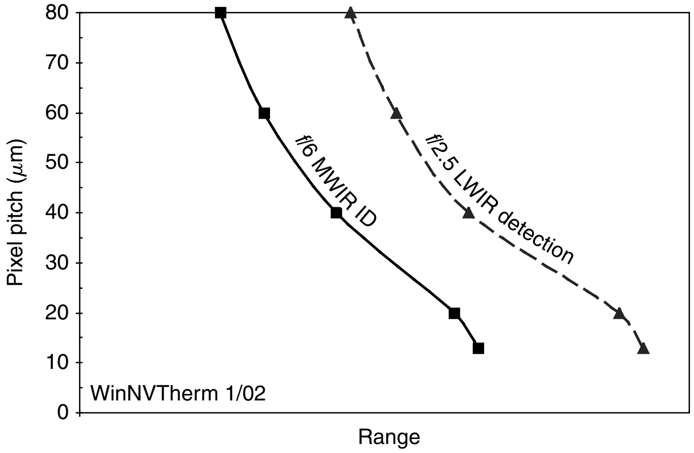


FIGURE 37.23 Range improves as the pixel size is reduced until a limit in optical blur is reached. In the examples above, the blur circle for the MWIR and LWIR cases are comparable since the *f*/number has been adjusted accordingly. D^* and integration time have been held constant in this example.

results for MWIR detectors, but not yet for LWIR devices. Further improvement in this approach would be needed to use it for Third-Generation MWIR/LWIR two-color arrays.

A second approach that has proven successful for InSb hybrids is thinning the detector structure. HgCdTe hybrids currently retain their thick, 500 μm, CdZnTe epitaxial substrate in the hybridized structure. InSb hybrids must remove the substrate because it is not transparent, leaving only a 10-μm thick detector layer. The thinness of this layer allows it to readily expand and contract with the readout. InSb hybrids with detector arrays over 2 in. (5 cm) on a side have been successfully demonstrated to be reliable.

Hybrid reliability issues will be monitored as a third-generation sensor manufacturing technology and is developed to determine whether new approaches are needed.

In addition to cost issues, significant performance issues must also be addressed for third-generation imagers. These are now discussed in the following section.

37.5.5.4 Performance Challenges

37.5.5.4.1 Dynamic Range and Sensitivity Constraints

A goal of third-generation imagers is to achieve a significant improvement in detection and ID range over Second-Generation systems. Range improvement comes from higher pixel count, and to a lesser extent from improved sensitivity. Figure 37.23 shows relative ID and detection range vs. pixel size in the MWIR and LWIR, respectively. Sensitivity (D^* and integration time) have been held constant, and the format was varied to keep the field-of-view constant.

Sensitivity has less effect than pixel size for clear atmospheric conditions, as illustrated by the clear atmosphere curve in Figure 37.24. Note that here the sensitivity is varied by an order of magnitude, corresponding to two orders of magnitude increase in integration time. Only a modest increase in range is seen for this dramatic change in SNR ratio. In degraded atmospheric conditions, however, improved sensitivity plays a larger role because the signal is weaker. This is illustrated in Figure 37.24 by the curve showing range under conditions of reduced atmospheric transmission.

Dynamic range of the imager output must be considered from the perspective of the quantum efficiency and the effective charge storage capacity in the pixel unit cell of the readout. Quantum efficiency and charge storage capacity determine the integration time for a particular flux rate. As increasing number of quanta are averaged, the SNR ratio improves as the square root of the count. Higher accuracy A/D converters are therefore required to cope with the increased dynamic range between the noise and signal levels. Figure 37.25 illustrates the interaction of these specifications.

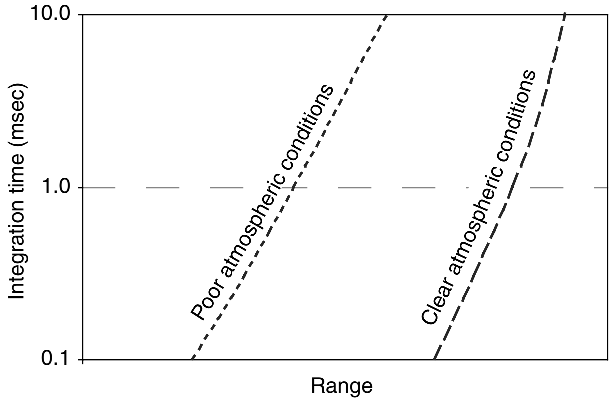


FIGURE 37.24 Range in a clear atmosphere improves only modestly with increased sensitivity. The case modeled here has a 20 μm pixel, a fixed D^*, and variable integration time. The 100× range of integration time corresponds to a 10× range in SNR. Improvement is more dramatic in the case of lower-atmospheric transmission that results in a reduced target signal.

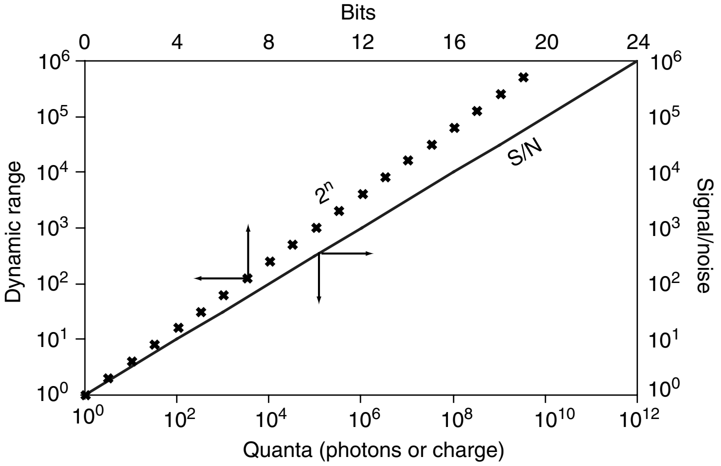


FIGURE 37.25 Dynamic range (2^n) corresponding to the number of digital bits (n) is plotted as a discrete point corresponding to each bit and referenced to the left and top scales. SNR ratio, corresponding to the number of quanta collected (either photons or charge) is illustrated by the solid line in reference to the bottom- and right-hand scales.

System interface considerations lead to some interesting challenges and dilemmas. Imaging systems typically specify a noise floor from the readout on the order of 300 μV. This is because system users do not want to encounter sensor signal levels below the system noise level. With readouts built at commercial silicon foundries now having submicrometer design rules, the maximum bias voltage applied to the readout is limited to a few volts — this trend has been downward from 5 V in the past decade as design rules have shrunk, as illustrated in Figure 37.26. Output swing voltages can only be a fraction of the maximum applied voltage, on the order of 3 V or less.

This means that the dynamic range limit of a readout is about 10,000 — 80 db in power — or less. Present readouts almost approach this constraining factor with 70 to 75 db achieved in good designs. In order to significantly improve sensitivity, the noise floor will have to be reduced.

If sufficiently low readout noise could be achieved, and the readout could digitize on chip to a level of 15 to 16 bits, the data could come off digitally and the system noise floor would not be an issue. Such developments may allow incremental improvement in third-generation imagers in the

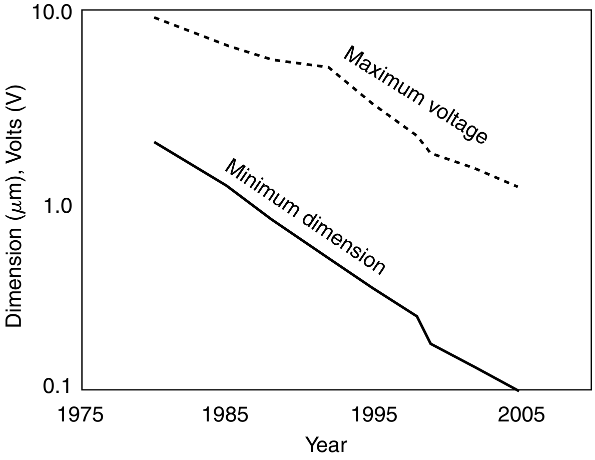


FIGURE 37.26 Trends for design rule minimum dimensions and maximum bias voltage of silicon foundry requirements.

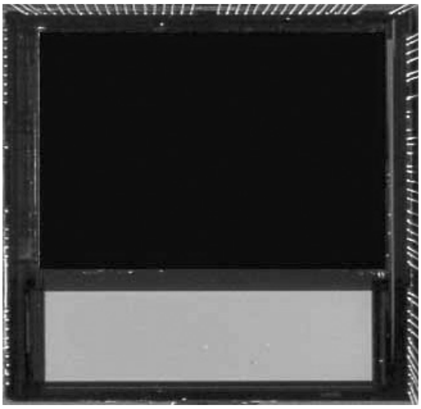


FIGURE 37.27 Focal planes with on-chip A/D converters have been demonstrated. This example shows a 900 × 120 TDI scanning format array. Photo supplied by Lester Kozlowski of Rockwell Scientific, Camarillo, CA.

future. Figure 37.27 illustrates an example of an on-chip A/D converter that has demonstrated 12 bits on chip.

A final issue here concerns the ability to provide high charge storage density within the small pixel dimensions envisioned for third-generation imagers. This may be difficult with standard CMOS capacitors. Reduced oxide thickness of submicrometer design rules does give larger capacitance per unit area, but the reduced bias voltage largely cancels any improvement in charge storage density. Promising technology in the form of ferroelectric capacitors may provide much greater charge storage densities than the oxide-on-silicon capacitors now used. Such technology is not yet incorporated into standard CMOS foundries. Stacked hybrid structures[1] [32] may be needed as at least an interim solution to incorporate the desired charge storage density in detector-readout-capacitor structures.

37.5.5.4.2 High Frame Rate Operation

Frame rates of 30 to 60 fps are adequate for visual display. In third-generation systems we plan to deploy high frame rate capabilities to provide more data throughput for advanced signal processing functions

[1] It should be noted that the third-generation imager will operate as an on-the-move wide area step-scanner wit automated ATR versus second-generation systems that rely on manual target searching. This allows the overall narrower field of view for the third-generation imager.

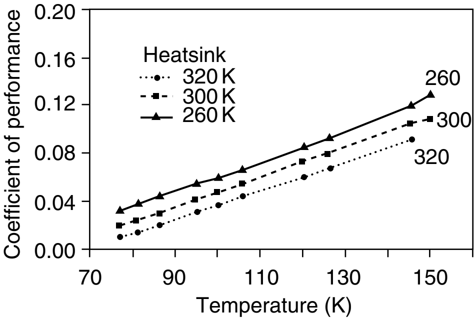


FIGURE 37.28 Javelin cooler coefficient of performance vs. temperature.

such as automatic target recognition (ATR), and missile and projectile tracking. An additional benefit is the opportunity to collect a higher percentage of available signal. Higher frame rates pose two significant issues. First, output drive power is proportional to the frame rate and at rates of 480 Hz or higher, this could be the most significant source of power dissipation on the readout. Increased power consumption on chip will also require more power consumption by the cryogenic cooler. These considerations lead us to conclude that high frame rate capabilities need to be limited to a small but arbitrarily positioned window of 64 × 64 pixels, for which a high frame rate of 480 Hz can be supported. This allows for ATR functions to be exercised on possible target locations within the full field-of-view.

37.5.5.4.3 Higher Operating Temperature

Current tactical infrared imagers operate at 77 K with few exceptions — notably MWIR HgCdTe, which can use solid-state thermoelectric (TE) cooling. Power can be saved, and cooler efficiency and cooler lifetime improved if focal planes operate at temperatures above 77 K.

Increasing the operating temperature results in a major reduction of input cryogenic cooler power. As can be seen from Figure 37.28 the coefficient of performance (COP) increases by a factor of 2.4 from 2.5 to 6% as the operating temperature is raised from 80 to 120 K with a 320 K heat sink. If the operating temperature can be increased to 150 K, the COP increases fourfold. This can have a major impact on input power, weight, and size.

Research is underway on an artificial narrow bandgap intrinsic-like material — strained-layer superlattices of InGaAsSb — which have the potential to increase operating temperatures to even higher levels [33]. Results from this research may be more than a decade away, but the potential benefits are significant in terms of reduced cooler operating power and maintenance.

The above discussion illustrates some of the challenges facing the development of third-generation cooled imagers. In addition to these are the required advances in signal processing and display technologies to translate the focal plane enhancements into outputs for the user. These advances can be anticipated to not only help to increase the range at which targets can be identified, but also to increase the rate of detection and identification through the use of two-color cues. Image fusion of the two colors in some cases is anticipated to help find camouflaged targets in clutter. Improved sensitivity and two-color response is further anticipated to minimize the loss of target contrast now encountered because of diurnal crossover. Future two-color imagers together with novel signal processing methods may further enhance the ability to detect land mines and find obscured targets.

37.6 Summary

Infrared sensors have made major performance strides in the last few years, especially in the uncooled sensors area. Cost, weight, and size of the uncooled have dramatically been reduced allowing a greater proliferation into the commercial market. Uncooled sensors will find greater use in the medical community as a result. High-performance cooled sensors have also been dramatically improved including the development of multicolor arrays. The high-performance sensors will find new medical applications because of the color discrimination and sensitivity attributes now available.

References

[1] D.G. Crowe, P.R. Norton, T. Limperis, and J. Mudar, Detectors, in *Electro-Optical Components*, W.D. Rogatto, Ed., Vol. 3, ERIM, Ann Arbor, MI; *Infrared & Electro-Optical Systems Handbook*, J.S. Accetta and D.L. Schumaker, Executive Eds., SPIE, Bellingham, WA, 1993, revised 1996, Chapter 4, pp. 175–283.

[2] P.R. Norton, Detector focal plane array technology, in *Encyclopedia of Optical Engineering*, Vol. 1, R.G. Driggers, Ed., Marcel Dekker, New York, 2003, pp. 320–348.

[3] P.W. Kruse and D.D. Skatrud, Eds., Uncooled infrared imaging arrays and systems in *Semiconductors and Semimetals*, R.K. Willardson and E.R. Weber, Eds., Academic Press, New York, 1997.

[4] P. Norton, Infrared image sensors, *Opt. Eng.*, 30, 1649–1663, 1991.

[5] T. Ashley, I.M. Baker, T.M. Burke, D.T. Dutton, J.A. Haigh, L.G. Hipwood, R. Jefferies, A.D. Johnson, P. Knowles, and J.C. Little, *Proc. SPIE*, 4028, 2000, pp. 398–403

[6] P.J. Love, K.J. Ando, R.E. Bornfreund, E. Corrales, R.E. Mills, J.R. Cripe, N.A. Lum, J.P. Rosbeck, and M.S. Smith, Large-format infrared arrays for future space and ground-based astronomy applications, *Proceedings of SPIE*; *Infrared Spaceborne Remote Sensing IX*, vol. 4486–38; pp. 373–384, 29 July–3 August, 2001; San Diego, USA.

[7] The photoconductive and photovoltaic detector technology of HgCdTe is summarized in the following references: D. Long and J.L. Schmidt, Mercury-cadmium telluride and closely related alloys, in *Semiconductors and Semimetals* 5, R.K. Willardson and A.C. Beer, Eds., Academic Press, New York, pp. 175–255, 1970; R.A. Reynolds, C.G. Roberts, R.A. Chapman, and H.B. Bebb, Photoconductivity processes in 0.09 eV bandgap HgCdTe, in *Proceedings of the 3rd International Conference on Photoconductivity*, E.M. Pell, Ed., Pergamon Press, New York, p. 217, 1971; P.W. Kruse, D. Long, and O.N. Tufte, Photoeffects and material parameters in HgCdTe alloys, in *Proceedings of the 3rd International Conference on Photoconductivity*, E.M. Pell, Ed., Pergamon Press, New York, p. 233, 1971; R.M. Broudy and V.J. Mazurczyk (HgCd) Te photoconductive detectors, in *Semiconductors and Semimetals,* 18, R.K. Willardson and A.C. Beer, Eds., chapter 5, Academic Press, New York, pp. 157–199, 1981; M.B. Reine, A.K. Sood, and T.J. Tredwell, Photovoltaic infrared detectors, in *Semiconductors and Semimetals*, 18, R.K. Willardson and A.C. Beer, Eds., chapter 6, pp. 201–311; D. Long, Photovoltaic and photoconductive infrared detectors, in *Topics in Applied Physics* 19, *Optical and Infrared Detectors*, R.J. Keyes, Ed., Springer-Verlag, Heidelberg, pp.101–147, 1970; C.T. Elliot, infrared detectors, in *Handbook on Semiconductors* 4, C. Hilsum, Ed., chapter 6B, North Holland, New York, pp. 727–798, 1981.

[8] P. Norton, Status of infrared detectors, *Proc. SPIE*, 2274, 82–92, 1994.

[9] I.M., Baker, Photovoltaic IR detectors in *Narrow-gap II–VI Compounds for Optoelectronic and Electromagnetic Applications*, P. Capper, Ed., Chapman and Hall, London, pp. 450–73, 1997.

[10] P. Norton, Status of infrared detectors, *Proc. SPIE,* 3379, 102–114, 1998.

[11] M. Kinch., HDVIP® FPA technology at DRS, *Proc. SPIE,* 4369, pp. 566–578, 1999.

[12] M.B. Reine., Semiconductor fundamentals — materials: fundamental properties of mercury cadmium telluride, in *Encyclopedia of Modern Optics,* Academic Press, London, 2004.

[13] A. Rogalski., HgCdTe infrared detector material: history, status and outlook, *Rep. Prog. Phys.* 68, 2267–2336, 2005.

[14] S.D. Guanapala, B.F. Levine, and N. Chand, *J. Appl. Phys.*, 70, 305, 1991.

[15] B.F. Levine, *J. Appl. Phys.*, 47, R1–R81, 1993.

[16] K.K. Choi., *The Physics of Quantum Well Infrared Photodetectors,* World Scientific, River Edge, New Jersey, 1997.

[17] S.D. Gunapala, J.K. Liu, J.S. Park, M. Sundaram, C.A. Shott, T. Hoelter, T.-L. Lin, S.T. Massie, P.D. Maker, R.E. Muller, and G. Sarusi, 9 μm Cutoff 256×256 GaAs/AlGaAs quantum well infrared photodetector hand-held camera, *IEEE Trans. Elect. Dev.*, 45, 1890, 1998.

[18] S.D. Gunapala, S.V. Bandara, J.K. Liu, W. Hong, M. Sundaram, P.D. Maker, R.E. Muller, C.A. Shott, and R. Carralejo, Long-wavelength 640×480 GaAs/AlGaAs quantum well infrared photodetector snap-shot camera, *IEEE Trans. Elect. Dev.,* 44, 51–57, 1997.

[19] M.Z. Tidrow et al., Device physics and focal plane applications of QWIP and MCT, *Opto-Elect. Rev.*, 7, 283–296, 1999.

[20] S.D. Gunapala and S.V. Bandara, Quantum well infrared photodetector (QWIP) focal plane arrays, in *Semiconductors and Semimetals*, R.K. Willardson and E.R. Weber, Eds., 62, Academic Press, New York, 1999.

[21] G.A. Sai-Halasz, R. Tsu, and L. Esaki, *Appl. Phys. Lett.* 30, 651, 1977.

[22] D.L. Smith and C. Mailhiot, Proposal for strained type II superlattice infrared detectors, *J. Appl. Phys.*, 62, 2545–2548, 1987.

[23] S.R. Kurtz, L.R. Dawson, T.E. Zipperian, and S.R. Lee, Demonstration of an InAsSb strained-layer superlattice photodiode, *Appl. Phys. Lett.*, 52, 1581–1583, 1988.

[24] R.H. Miles, D.H. Chow, J.N. Schulman, and T.C. McGill, Infrared optical characterization of InAs/GaInSb superlattices, *Appl. Phys. Lett.* 57, 801–803, 1990.

[25] F. Fuchs, U.Weimar, W. Pletschen, J. Schmitz, E. Ahlswede, M. Walther, J. Wagner, and P. Koidl, *J. Appl. Phys. Lett.*, 71, 3251, 1997.

[26] Gail J. Brown, Type-II InAs/GaInSb superlattices for infrared detection: an overview, *Proceedings of SPIE*, 5783, pp. 65–77, 2005.

[27] J.L. Vampola, Readout electronics for infrared sensors, in *electro-optical components*, chapter 5, vol. 3, W.D. Rogatto, Ed., *Infrared & Electro-Optical Systems Handbook,* J.S. Accetta and D.L. Schumaker, Executive Eds., *ERIM*, Ann Arbor, MI and *SPIE*, Bellingham, WA, pp. 285–342, 1993, revised 1996.

[28] D. Reago, S. Horn, J. Campbell, and R. Vollmerhausen, Third generation imaging sensor system concepts, *SPIE,* 3701, 108–117, 1999.

[29] P. Norton*, J. Campbell III, S. Horn, and D. Reago, Third-generation infrared imagers, *Proc. SPIE*, 4130, 226–236, 2000.

[30] S. Horn, P. Norton, T. Cincotta, A. Stoltz, D. Benson, P. Perconti, and J. Campbell, Challenges for third-generation cooled imagers, *Proc. SPIE*, 5074, 44–51, 2003.

[31] S. Horn, D. Lohrman, P. Norton, K. McCormack, and A. Hutchinson, Reaching for the sensitivity limits of uncooled and minimally cooled thermal and photon infrared detectors, *Proc. SPIE*, 5783, 401–411, 2005.

[32] W. Cabanskia, K. Eberhardta, W. Rodea, J. Wendlera, J. Zieglera, J.Fleißnerb, F. Fuchsb, R. Rehmb, J. Schmitzb, H. Schneiderb, and M. Walther, 3rd gen focal plane array IR detection modules and applications, *Proc. SPIE*, 5406, 184–192, 2004.

[33] S. Horn, P. Norton, K. Carson#, R. Eden, and R. Clement, Vertically-integrated sensor arrays — VISA, *Proc. SPIE*, 5406, 332–340, 2004.

[34] R. Balcerak and S. Horn, Progress in the development of vertically-integrated sensor arrays, *Proc. SPIE*, 5783, 384–391, 2005.

38

Infrared Camera Characterization

Joseph G. Pellegrino
Jason Zeibel
Ronald G. Driggers
Philip Perconti
U.S. Army Communications and Electronics Research, Development and Engineering Center (CERDEC) Night Vision and Electronic Sensors Directorate

Many different types of infrared (IR) detector technology are now commercially available and the physics of their operation has been described in an earlier chapter. IR imagers are classified by different characteristics such as scan type, detector material, cooling requirements, and detector physics. Thermal imaging cameras prior to the 1990s typically contained a relatively small number of IR photosensitive detectors. These imagers were known as *cooled scanning systems* because they required cooling to cryogenic temperatures and a mechanical scan mirror to construct a two-dimensional (2D) image of the scene. Large 2D arrays of IR detectors, or staring arrays, have enabled the development of *cooled staring systems* that maintain sensitivity over a wide range of scene flux conditions, spectral bandwidths, and frame rates. Staring arrays consisting of small bolometric detector elements, or microbolometers, have enabled the development of *uncooled staring systems* that are compact, lightweight, and low power (see Figure 38.1).

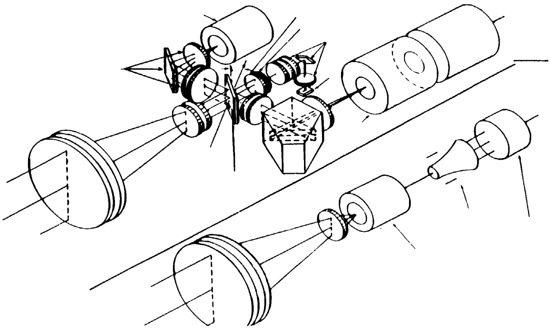


FIGURE 38.1 Scanning and staring system designs.

The sensitivity, or thermal resolution, of uncooled microbolometer focal plane arrays has improved dramatically over the past decade, resulting in IR video cameras that can resolve temperature differences under nominal imaging conditions as small as twenty millidegrees Kelvin using f/1.0 optics. Advancements in the manufacturing processes used by the commercial silicon industry have been instrumental in this progress. Uncooled microbolometer structures are typically fabricated on top of silicon integrated circuitry (IC) designed to readout the changes in resistance for each pixel in the array. The silicon-based IC serves as an electrical and mechanical interface for the IR microbolometer.

The primary measures of IR sensor performance are sensitivity and resolution. When measurements of end-to-end or human-in-the-loop (HITL) performance are required, the visual acuity of an observer through a sensor is included. The sensitivity and resolution are both related to the hardware and software that comprises the system, while the HITL includes both the sensor and the observer. Sensitivity is determined through radiometric analysis of the scene environment and the quantum electronic properties of the detectors. Resolution is determined by analysis of the physical optical properties, the detector array geometry, and other degrading components of the system in much the same manner as complex electronic circuit/signals analysis. The sensitivity of cooled and uncooled staring IR video cameras has improved by more than a factor of ten compared to scanning systems commercially available in the 1980s and early 1990s.[11,12]

Sensitivity describes how the sensor performs with respect to input signal level. It relates noise characteristics, responsivity of the detector, light gathering of the optics, and the dynamic range of the sensor. Radiometry describes how much light leaves the object and background and is collected by the detector. Optical design and detector characteristics are of considerable importance in sensor sensitivity analysis. In IR systems, noise equivalent temperature difference (NETD) is often a first order description of the system sensitivity. The three-dimensional (3D) noise model [1] describes more detailed representations of sensitivity parameters. The sensitivity of scanned long-wave infrared (LWIR) cameras operating at video frame rates is typically limited by very short detector integration times on the order of tens or hundreds of microseconds. The sensitivity of staring IR systems with high quantum efficiency detectors is often limited by the charge integration capacity, or well capacity, of the readout integrated circuit (ROIC). The detector integration time of staring IR cameras can be tailored to optimize sensitivity for a given application and may range from microseconds to tens of milliseconds.

The second type of measure is resolution. Resolution is the ability of the sensor to image small targets and to resolve fine detail in large targets. Modulation transfer function (MTF) is the most widely used resolution descriptor in IR systems. Alternatively, it may be specified by a number of descriptive metrics such as the optical Rayleigh Criterion or the instantaneous field-of-view of the detector. Where these metrics are component-level descriptions, the system MTF is an all-encompassing function that describes the system resolution. Sensitivity and resolution can be competing system characteristics and they are the most important issues in initial studies for a design. For example, given a fixed sensor aperture diameter, an increase in focal length can provide an increase in resolution, but may decrease sensitivity [2]. A more detailed consideration of the optical design parameters is included in the next chapter.

Quite often metrics, such as NETD and MTF, are considered separable. However, in an actual sensor, sensitivity and resolution performance are interrelated. As a result, minimum resolvable temperature difference (MRT or MRTD) has become a primary performance metric for IR systems.

This chapter addresses the parameters that characterize a camera's performance. A website advertising IR camera would in general contain a specification sheet that contains some variation of the terms that follow. A goal of this section is to give the reader working knowledge of these terms so as to better enable them to obtain the correct camera for their application:

- Three-dimensional noise
- NETD (Noise equivalent temperature difference)
- Dynamic range
- MTF

- MRT (minimum resolvable temperature) and MDT (minimum detectable temperature)
- Spatial resolution
- Pixel size

38.1 Dimensional Noise

The 3D noise model is essential for describing the sensitivity of an optical sensor system. Modern imaging sensors incorporate complex focal plane architectures and sophisticated postdetector processing and electronics. These advanced technical characteristics create the potential for the generation of complex noise patterns in the output imagery of the system. These noise patterns are deleterious and therefore need to be analyzed to better understand their effects upon performance. Unlike classical systems where "well behaved" detector noise predominates, current sensor systems have the ability to generate a wide variety of noise types, each with distinctive characteristics temporally, as well as along the vertical and horizontal image directions. Earlier methods for noise measurements at the detector preamplifier port that ignored other system noise sources are no longer satisfactory. System components following the stage that include processing may generate additional noise and even dominate total system noise.

Efforts at the Night Vision and Electronic Sensor Directorate to measure 2nd generation IR sensors uncovered the need for a more comprehensive method to characterize noise parameters. It was observed that the noise patterns produced by these systems exhibited a high degree of directionality. The data set is 3D with the temporal dimension representing the frame sequencing and the two spatial dimensions representing the vertical and horizontal directions within the image (see Figure 38.2).

To acquire this data cube, the field of view of a camera to be measured is flooded with a uniform temperature reference. A set number *n* (typically around 100) of successive frames of video data are then collected. Each frame of data consists of the measured response (in volts) to the uniform temperature source from each individual detector in the 2D focal plane array (FPA). When many successive frames of data are "stacked" together, a uniform source data cube is constructed. The measured response may be either analog (RS-170) or digital (RS-422, Camera Link, Hot Link, etc.) in nature depending on the camera interface being studied.

To recover the overall temporal noise, first the temporal noise is calculated for each detector in the array. A standard deviation of the *n* measured voltage responses for each detector is calculated. For an *h* by *v* array, there are *hv* separate values where each value is the standard deviation of *n* voltage measurements. The median temporal noise among these *hv* values is stated as the overall temporal noise in volts.

Following the calculation of temporal noise, the uniform source data cube is reduced along each axis according to the 3D noise procedure. There are seven noise terms as part of the 3D noise definition. Three components measure the average noise present along on each axis (horizontal, vertical, and temporal) of the data cube (σ_h, σ_v, and σ_t). Three terms measure the noise common to any given pair of axes in the data cube (σ_{tv}, σ_{th}, and σ_{vh}). The final term measures the uncorrelated random noise (σ_{tvh}). To calculate

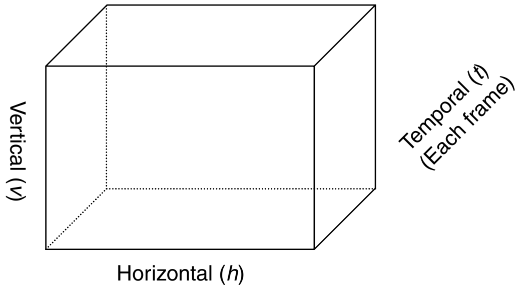


FIGURE 38.2 An example of a uniform source data cube for 3D noise measurements. The first step in the calculation of 3D noise parameters is the acquisition of a uniform source data cube.

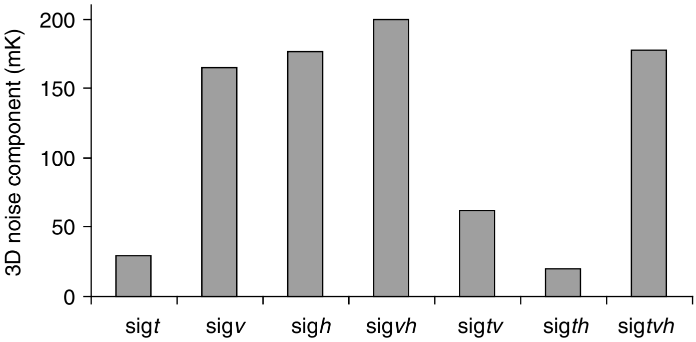


FIGURE 38.3 The 3D noise values for a typical data cube. The spatial nonuniformity can be seen in the elevated values of the spatial only 3D noise components σ_h, σ_v, and σ_{vh}. The white noise present in the system (σ_{tvh}) is roughly the same magnitude as the spatial 3D noise components.

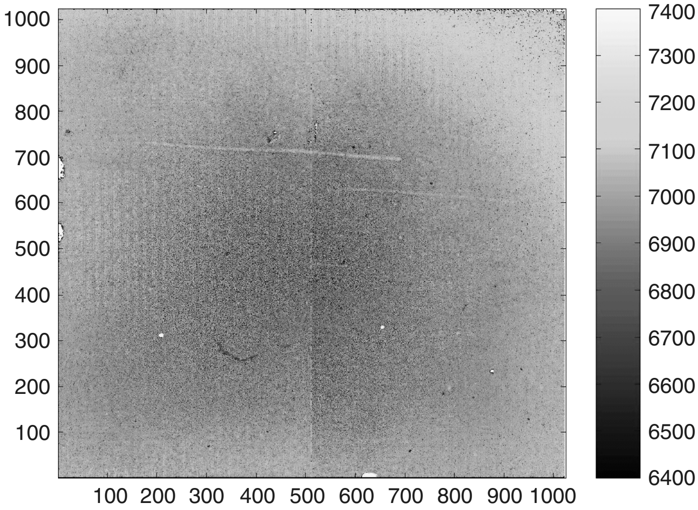


FIGURE 38.4 An example of a camera system with high spatial noise components and very low temporal noise components.

the spatial noise for the camera, each of the 3D noise components that are independent of time (σ_v, σ_h, and σ_{vh}) are added in quadrature. The result is quoted as the spatial noise of the camera in volts.

In order to represent a data cube in a 2D format, the cube is averaged along one of the axes. For example, if a data cube is averaged along the temporal axis, then a time averaged array is created. This format is useful for visualizing purely spatial noise effects as three of the components are calculated after temporal averaging (σ_h, σ_v, and σ_{vh}). These are the time independent components of 3D Noise. The data cube can also be averaged along both spatial dimensions. The full 3D Noise calculation for a typical data cube is shown in Figure 38.3.

Figure 38.4 shows an example of a data cube that has been temporally averaged. In this case, many spatial noise features are present. Column noise is clearly visible in Figure 38.4, however the dominant spatial noise component appears to be the "salt and pepper" fixed pattern noise. The seven 3D Noise components are shown in Figure 38.5. σ_{vh} is clearly the dominant noise term, as was expected due to the high fixed pattern noise. The column noise σ_h and the row noise σ_v are the next dominant. In this example, the overall bulls-eye variation in the average frame dominates the σ_v and σ_h terms. Vertical stripes present in the figure add to σ_v, but this effect is small in comparison, leading to similar values for

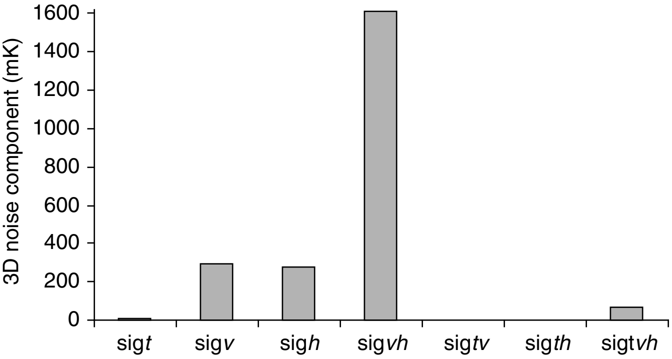


FIGURE 38.5 The 3D noise components for the data cube used to generate. Note that the amount of row and column noise is significantly smaller than the fixed pattern noise. All the 3D noise values with a temporal component are significantly smaller than the purely spatial values.

σ_v and σ_h. In this example, the temporal components of the 3D noise are two orders of magnitude lower than σ_vh. If this data cube were to be plotted as individual frames, we would see that successive frames would hardly change and the dominant spatial noise would be present (and constant) in each frame.

38.2 Noise Equivalent Temperature Difference

In general, imager sensitivity is a measure of the smallest signal that is detectable by a sensor. For IR imaging systems, noise equivalent temperature difference (NETD) is a measure of sensitivity. Sensitivity is determined using the principles of radiometry and the characteristics of the detector. The system intensity transfer function (SITF) can be used to estimate the noise equivalent temperature difference. NEDT is the system noise rms voltage over the noise differential output. It is the smallest measurable signal produced by a large target (extended source), in other words the minimum measurable signal.

Equation below describes NETD as a function of noise voltage and the system intensity transfer function. The measured NETD values are determined from a line of video stripped from the image of a test target, as depicted in Figure 38.10. A square test target is placed before a blackbody source. The delta T is the difference between the blackbody temperature and the mask. This target is then placed at the focal point of an off axis parabolic mirror. The mirror serves the purpose of a long optical path length to the target, yet relieves the tester from concerns over atmospheric losses to the temperature difference. The image of the target is shown in Figure 38.6. The SITF slope for the scan line in Figure 38.6 is the $\Delta\Sigma/\Delta T$, where $\Delta\Sigma$ is the signal measured for a given ΔT. The N_rms is the background signal on the same line.

$$\mathrm{NETD} = \frac{N_\mathrm{rms}\ [\mathrm{volts}]}{\mathrm{SITF\_Slope}\ [\mathrm{volts/K}]}$$

After calculating both the temporal and spatial noise, a signal transfer function (SiTF) is measured. The field of view of the camera is again flooded with a uniform temperature source. The temperature of the source is varied over the dynamic range of the camera's output while the mean array voltage response is recorded. The slope of the resulting curve yields the SiTF responsivity in volts per degree Kelvin change in the scene temperature. Once both the SiTF curve and the temporal and spatial noise in volts are known, the NETD can be calculated. This is accomplished by dividing the temporal and spatial noise in volts by the responsivity in volts per degree Kelvin. The resulting NETD values represent the minimum discernable change in scene temperature for both spatial and temporal observation.

The SiTF of an electro-optical (EO) or IR system is determined by the signal response once the dark offset signal has been subtracted off. After subtracting off the offset due to non flux effects, the SiTF

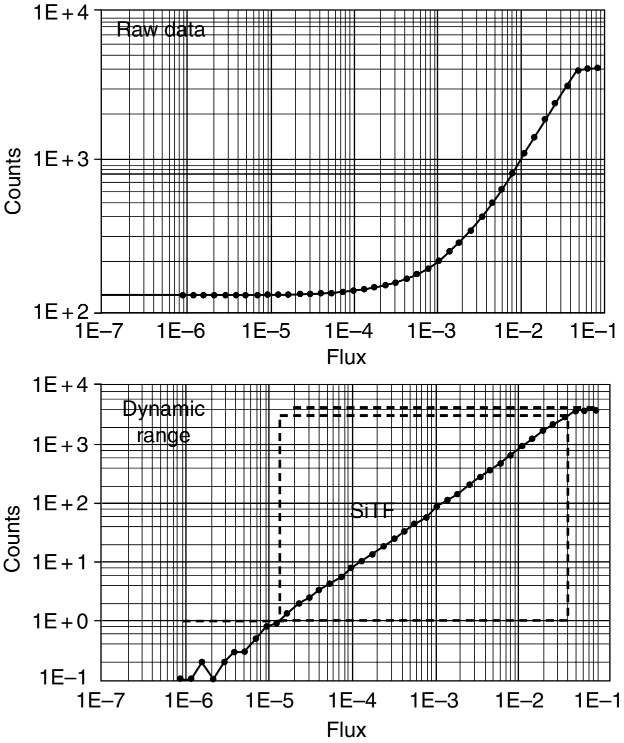


FIGURE 38.6 Dynamic range and system transfer function.

plots the counts output relative to the input photon flux. The SiTF is typically represented in response units of voltage, signal electrons, digital counts, etc. vs. units of the source: blackbody temperature, flux, photons, and so on. If the system behaves linearly within the dynamic range then the slope of the SiTF is constant. The dynamic range of the system, which may be defined by various criteria, is determined by the minimum (i.e., signal to noise ratio = 1) and maximum levels of operation.

38.3 Dynamic Range

The responsivity function also provides dynamic range and linearity information. The camera dynamic range is the maximum measurable input signal divided by the minimum measurable signal. The NEDT is assumed to be the minimum measurable signal. For AC systems, the maximum output depends on the target size and therefore the target size must be specified if dynamic range is a specification. Depending upon the application, the maximum input value may be defined by one of several methods. One method for specifying the dynamic range of a system involves having the ΔV_{sys} signal reach some specified level, say 90% of the saturation level as shown in Figure 38.7. Another method to assess the maximum input value is based on the signal's deviation from linearity. The range of data points that fall within a specified band is designated as the dynamic range. A third approach involves specifying the minimum SiTF of the system.

For most systems, the detector output signal is adjusted both in gain and offset so that the dynamic range of the A/D converter is maximized. Figure 38.8 shows a generic detector system that contains an 8-bit A/D converter. The converter can handle an input signal between 0 and 1 volt and an output between 0 and 255 counts. By selecting the gain and offset, any detector voltage range can be mapped into the digital output. Figure 38.9 shows 3 different system gains and offsets. When the source flux level is less

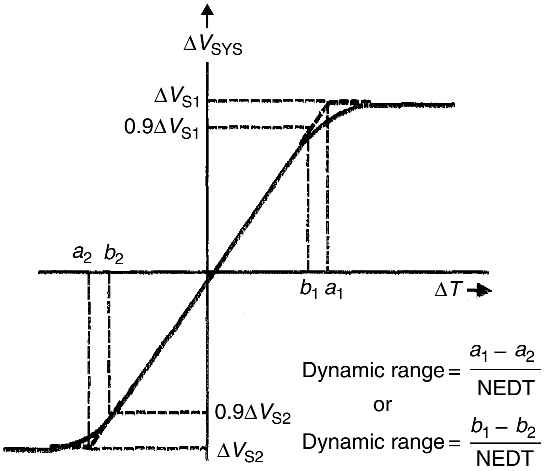


FIGURE 38.7 Dynamic range defined by linearity.

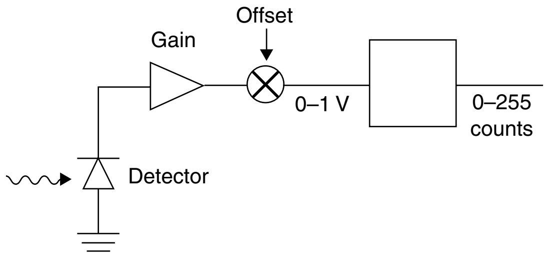


FIGURE 38.8 System with 8-bit A/D converter.

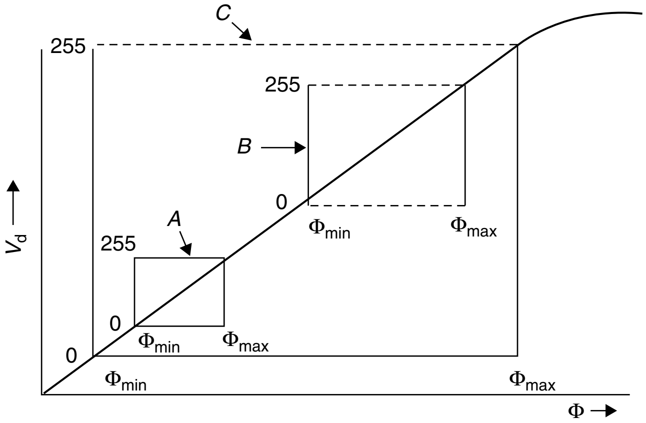


FIGURE 38.9 Different gains and voltage offsets affect the input-to-output transition.

than Φ_{min}, the source will not be seen (i.e. it will appear as 0 counts). When the flux level is greater than Φ_{max}, the source will appear as 255 counts, and the system is said to be saturated. The gain parameters, Φ_{min} and Φ_{max} are redefined for each gain and offset level setting.

Output A below occurs with maximum gain. Point B occurs with moderate gain and C with minimum gain. For the various gains, the detector output gets mapped into the full dynamic range of the A/D converter.

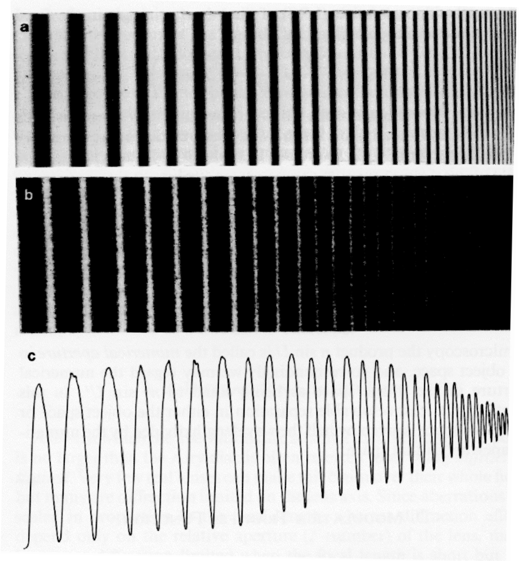


FIGURE 38.10 This figure shows the falloff in MTF as spatial frequency increases. Panel a is a sinusoidal test pattern, panel b is the optical system's (negative) response, and panel c shows the contrast as a function of spatial frequency.

38.4 Modulation Transfer Function

The modulation transfer function (MTF) of an optical system measures a system's ability to faithfully image a given object. Consider for example the bar pattern shown in Figure 38.10, with the cross section of each bar being a sine wave. Since the image of a sine wave light distribution is always a sine wave, no matter how bad the aberrations may be, the image is always a sine wave. The image will therefore have a sine wave distribution with intensity shown in Figure 38.10.

When the bars are coarsely spaced, the optical system has no difficulty faithfully reproducing them. However, when the bars are more tightly spaced, the contrast,

$$\text{Contrast} = \frac{\text{bright} - \text{dark}}{\text{bright} + \text{dark}}$$

begins to fall off as shown in panel c. If the dark lines have intensity = 0, the contrast = 1, and if the bright and dark lines are equally intense, contrast = 0. The contrast is equal to the MTF at a specified spatial frequency. Furthermore, it is evident that the MTF is a function of spatial frequency and position within the field.

38.5 Minimum Resolvable Temperature

Each time a camera is turned on, the observer subconsciously makes a judgement about image quality. The IR community uses the MRT and the MDT as standard measures of image quality. The MRT and MDT depend upon the IR imaging system's resolution and sensitivity. MRT is a measure of the ability to resolve detail and is inversely related to the MTF, whereas the MDT is a measure to detect something. The MRT and MDT deal with an observer's ability to perceive low contrast targets which are embeddedd in noise.

MRT and MDT are not absolute values rather they are temperature differentials relative to a given background. They are sometimes referred to as the minimum resolvable temperature difference (MRTD) and the minimum detectable temperature difference (MDTD).

The theoretical MRT is

$$\mathrm{MRT}(f_x) = \frac{k \cdot (\mathrm{NEDT})}{\mathrm{MTF}_{\mathrm{perceived}}(f_x)} \cdot \sqrt{}\{\beta_1 + \cdots + \beta_n\}$$

where $\mathrm{MTF}_{\mathrm{perceived}} = \mathrm{MTF}_{\mathrm{SYS}}\,\mathrm{MTF}_{\mathrm{MONITOR}}\,\mathrm{MTF}_{\mathrm{EYE}}$. The $\mathrm{MTF}_{\mathrm{system}}$ is defined by the product $\mathrm{MTF}_{\mathrm{sensor}}$ $\mathrm{MTF}_{\mathrm{optics}}$ $\mathrm{MTF}_{\mathrm{electronics}}$. Each β_i in the equation is an eye filter that is used to interpret the various components of noise. As certain noise sources increase, the MRT also increases. MRT has the same ambient temperature dependence as the NEDT; as the ambient temperature increases, MRT decreases. Because the MTF decreases as spatial frequency increases, the MRT increases with increasing spatial frequency. Overall system response depends on both sensitivity and resolution. The MRT parameter is bounded by sensitivity and resolution. Figure shows that different systems may have different MRTs. System A has a better sensitivity because it has a lower MRT at low spatial frequencies. At mid-range spatial frequencies, the systems are approximately equivalent and it can be said that they provide equivalent performance. At higher frequencies, System B has better resolution and can display finer detail than system A. In general, neither sensitivity, resolution, nor any other single parameter can be used to compare systems; many quantities must be specified for complete system-to-system comparison.

38.6 Spatial Resolution

The term resolution applies to two different concepts with regard to vision systems. Spatial resolution refers to the image size in pixels — for a given scene, more pixels means higher resolution. The spatial resolution is a fixed characteristic of the camera and cannot be increased by the frame grabber of post-processing techniques. Zooming techniques, for example, merely interpolate between pixels to expand an image without adding any new information to what the camera provided. It is easy to decrease the resolution, however, by simply ignoring part of the data. National Instruments frame grabbers provide for this with a "scaling" feature that instructs the frame grabber to sample the image to return a 1/2, 1/4, 1/8, and so on, scaled image. This is convenient when system bandwidth is limited and you don't require any precision measurements of the image.

The other use of the term "resolution" is commonly found in data acquisition applications and refers to the number of quantization levels used in A/D conversions. Higher resolution in this sense means that you would have improved capability of analyzing low-contrast images. This resolution is specified by the A/D converter; the frame grabber determines the resolution for analog signals, whereas the camera determines it for digital signals (the frame grabber must have the capability of supporting whatever resolution the camera provides, though).

38.6.1 Pixel Size

Camera pixel size consists of the tiny dots that make up a digital image. So let us say that a camera is capable of taking images at 640 × 480 pixels. A little math shows us that such an image would contain 307,200 pixels or 0.3 megapixels. Now let's say the camera takes 1024 × 768 images. That gives us 0.8 megapixels. So the larger the number of megapixels, the more image detail you get. Each pixel can be one of 16.7 million colors.

The detector pixel size refers to the size of the individual sensor elements that make up the detector part of the camera. If we had two charge-coupled devices (CCDs) detectors with equal Quantum Efficiency (QEs) but one has 9 μm pixels and the other has 18 μm pixels (i.e., the pixels on CCD#2 are twice the linear size of those on CCD #1) and we put both of these CCDs into cameras that operate identically, then the image taken with CCD#1 will require 4X the exposure of the image taken with CCD#2. This seeming discrepancy is due in its entirety to the area of the pixels in the two CCDs and could be compared to the effectiveness of rain gathering gauges with different rain collection areas: A rain gauge with a 2-in. diameter throat will collect 4X as much rain water as a rain gauge with a 1-in. diameter throat.

References

[1] J. D'Agostino and C. Webb, 3-D analysis framework and measurement methodology for imaging system noise. *Proc. SPIE*, 1488, 110–121 (1991).

[2] R.G. Driggers, P. Cox, and T. Edwards, *Introduction to Infrared and Electro-Optical Systems*, Artech House, Boston, MA, 1998, p. 8.

39

Infrared Camera and Optics for Medical Applications

Michael W. Grenn
Jay Vizgaitis
Joseph G. Pellegrino
Philip Perconti
U.S. Army Communications and Electronics Research, Development and Engineering Center (CERDEC) Night Vision and Electronic Sensors Directorate

The infrared radiation emitted by an object above 0 K is passively detected by infrared imaging cameras without any contact with the object and is nonionizing. The characteristics of the infrared radiation emitted by an object are described by Planck's blackbody law in terms of spectral radiant emittance.

$$M_\lambda = \varepsilon(\lambda)\frac{c_1}{\lambda^5(e^{c_2/\lambda T} - 1)}\ \mathrm{W/cm^2}\ \mu\mathrm{m}$$

where c_1 and c_2 are constants of 3.7418×10^4 W μm^4/cm^2 and 1.4388×10^4 μm K. The wavelength, λ, is provided in micrometers and $\varepsilon(\lambda)$ is the emissivity of the surface. A blackbody source is defined as an object with an emissivity of 1.0, so that it is a perfect emitter. Source emissions of blackbodies at nominal terrestrial temperatures are shown in Figure 39.1. The radiant exitance of a blackbody at a 310 K, corresponding to a nominal core body temperature of 98.6°F, peaks at approximately 9.5 μm in the LWIR. The selection of an infrared camera for a specific application requires consideration of many factors including sensitivity, resolution, uniformity, stability, calibratability, user controllability, reliability, object of interest phenomenology, video interface, packaging, and power consumption.

Planck's equation describes the spectral shape of the source as a function of wavelength. It is readily apparent that the peak shifts to shorter wavelengths as the temperature of the object of interest increases.

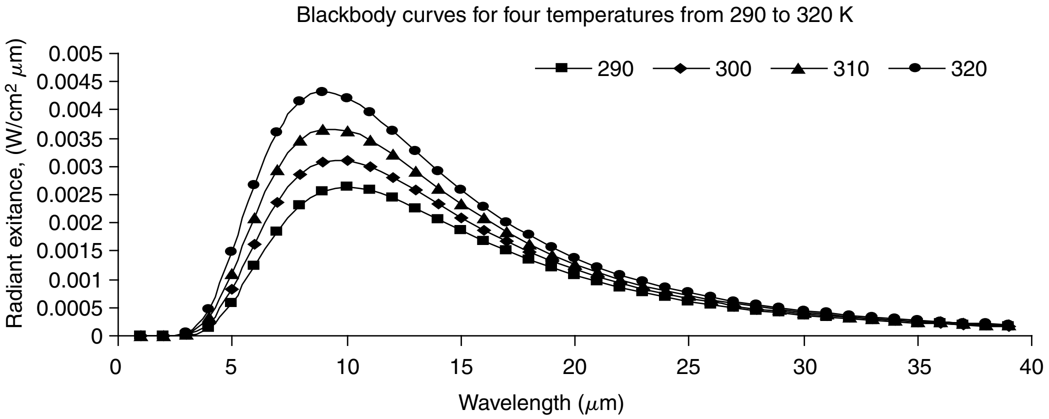


FIGURE 39.1 Planck's blackbody radiation curves.

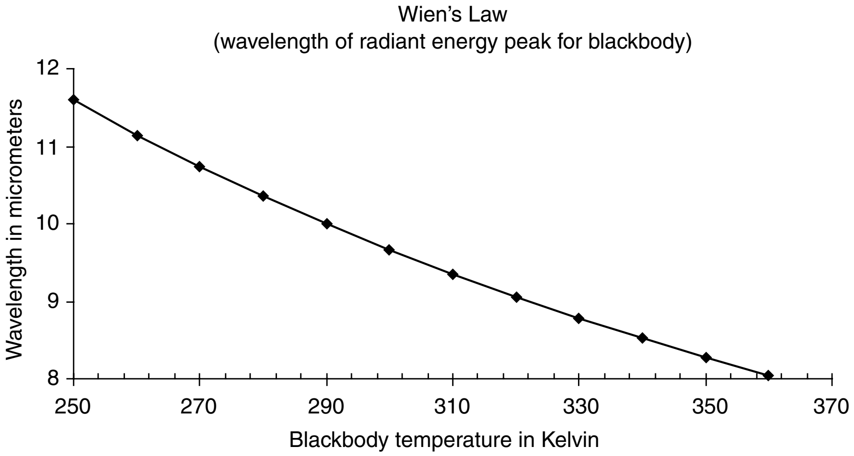


FIGURE 39.2 Location of peak of blackbody radiation, Wien's Law.

If the temperature of a blackbody approaches that of the sun, or 5900 K, the peak of the spectral shape would shift to 0.55 μm or green light. This peak wavelength is described by Wien's displacement law

$$\lambda_{\max} = 2898/T \ \mu\text{m}$$

Figure 39.2 shows the radiant energy peak as a function of temperature in the LWIR. It is important to note that the difference between the blackbody curves is the "signal" in the infrared bands. For an infrared sensor, if the background temperature is 300 K and the object of interest temperature is 302 K, the signal is the 2 K difference in flux between these curves. Signals in the infrared ride on very large amounts of background flux. This is not the case in the visible. For example, consider the case of a white object on a black background. The black background is generating no signal, while the white object is generating a maximum signal assuming the sensor gain is properly adjusted. The dynamic range may be fully utilized in a visible sensor. For the case of an IR sensor, a portion of the dynamic range is used by the large background flux radiated by everything in the scene. This flux is never a small value, hence sensitivity and dynamic range requirements are much more difficult to satisfy in IR sensors than in visible sensors.

A typical infrared imaging scenario consists of two major components, the object of interest and the background. In an IR scene, the majority of the energy is emitted from the constituents of the scene. This emitted energy is transmitted through the atmosphere to the sensor. As it propagates through the atmosphere it is degraded by absorption and scattering. Obscuration by intervening objects and additional energy emitted by the path also affect the target energy. This effect may be very small in short range imaging applications under controlled conditions. All these contributors, which are not the object of interest, essentially reduce one's ability to discriminate the object. The signal is further degraded by the optics of the sensor. The energy is then sampled by the detector array and converted to electrical signals. Various electronics amplify and condition this signal before it is presented to either a display for human interpretation or an algorithm like an automatic target recognizer for machine interpretation. A linear systems approach to modeling allows the components' transfer functions to be treated separately as contributors to the overall system performance. This approach allows for straightforward modifications to a performance model for changes in the sensor or environment when performing tradeoff analyses.

The photon flux levels (photons per square centimeter per second) on Earth is 1.5×10^{17} in the daytime and around 1×10^{10} at night in the visible. In the MWIR, the daytime and nighttime flux levels are 4×10^{15} and 2×10^{15}, respectively, where the flux is a combination of emitted and solar reflected flux. In the LWIR, the flux is primarily emitted where both day and night yield a 2×10^{17} level. At first look, it appears that the LWIR flux characteristics are as good as a daytime visible system, however, there are two other factors limiting performance. First, the energy bandgaps of infrared sensitive devices are much smaller than in the visible, resulting in significantly higher detector dark current. The detectors are typically cooled to reduce this effect. Second, the reflected light in the visible is modulated with target and background reflectivities that typically range from 7 to 20%.

In the infrared, where all terrestrial objects emit, a two-degree equivalent blackbody difference in photon flux between object and background is considered high contrast. The flux difference between two blackbodies of 302 K compared to 300 K can be calculated in a manner similar to that shown in Figure 39.1. The flux difference is the signal that provides an image, hence the difference in signal compared to the ambient background flux should be noted. In the LWIR, the signal is 3% of the mean flux and in the MWIR it is 6% of the mean flux. This means that there is a large flux pedestal associated with imaging in the infrared.

There are two major challenges accompanying the large background pedestal in the infrared. First, the performance of a typical infrared detector is limited by the background photon noise and this noise term is determined by the mean of the pedestal. This value may be relatively large compared to the small signal differences. Second, the charge storage capacity of the silicon input circuit mated to each infrared detector in a staring array limits the amount of integrated charge per frame, typically around 10^7 charge carriers. An LWIR system in a hot desert background would generate 10^{10} charge carriers in a 33 msec integration time. The optical f-number, spectral bandwidth, and integration time of the detector are typically tailored to reach half well for a given imaging scenario for dynamic range purposes. This well capacity limited condition results in a sensitivity, or noise equivalent temperature difference (NETD) of 10 to 30 times below the photon limited condition. Figure 39.3 shows calculations of NETD as a function of background temperature for MWIR and LWIR staring detectors dominated by the photon noise of the incident IR radiation. At 310 K, the NETD of high quantum efficiency MWIR and LWIR focal plane arrays (FPAs) is nearly the same, or about 3 millidegrees K, when the detectors are permitted to integrate charge up to the frame time, or in this case about 33 msec. The calculations show the sensitivity limits from the background photon shot noise only and does not include the contribution of detector and system temporal and spatial noise terms. The effects of residual spatial noise on NETD are described later in the chapter. The well capacity assumed here is 10^9 charge carriers to demonstrate sensitivity that could be achieved under large well conditions. The MWIR device is photon limited over the temperature range and begins to reach the well capacity limit near 340 K. The 24 μm pitch 9.5 μm cutoff LWIR device is well capacity limited over the entire temperature range. The 18 μm pitch 9.5 μm cutoff LWIR device becomes photon limited around 250 K. Various on-chip signal processing techniques, such as charge skimming and

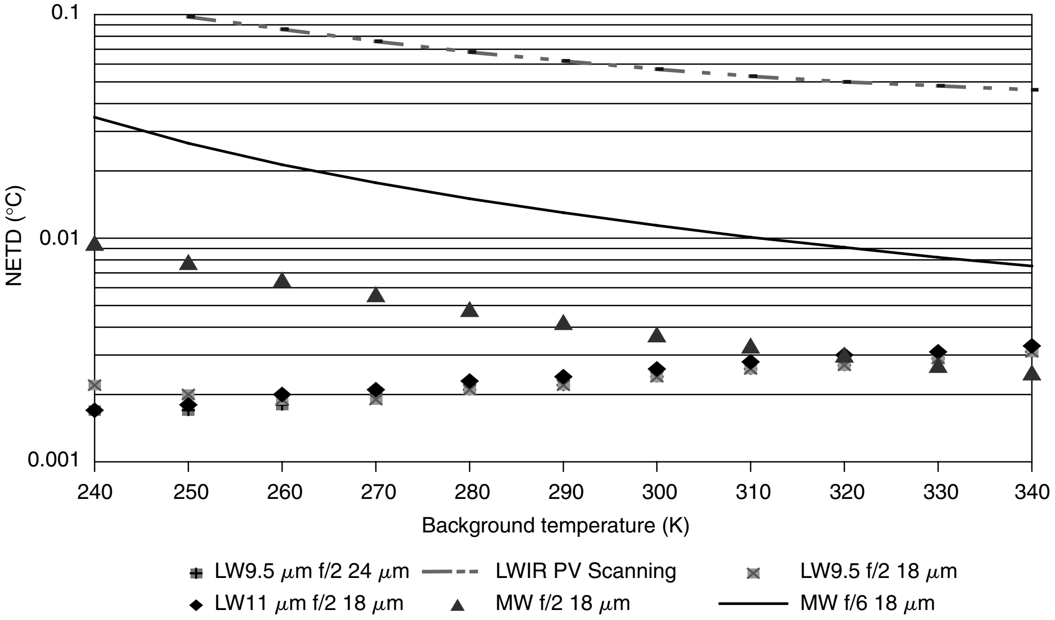


FIGURE 39.3 Background limited NETD for high quantum efficiency MWIR and LWIR detectors.

charge partitioning, have been investigated to increase the charge capacity of these devices. In addition, as the minimum feature sizes of the input circuitry decreases, more real estate in the unit cell can be allocated to charge storage.

Another major difference between infrared and visible systems is the size of the detector and diffraction blur. Typical sizes for MWIR and LWIR detectors, or pixels, range from 20 to 50 μm. Visible detectors less than 6 μm are commercially available today. The diffraction blur for the LWIR is more than ten times larger than the visible blur and MWIR blur is eight times larger than visible blur. Therefore, the image blur due to diffraction and detector size is much larger in an infrared system than a visible system. It is very common for infrared staring arrays to be sampling limited where the sample spacing is larger than the diffraction blur and the detector size. Dither and microscanning are frequently used to enhance performance. A more detailed discussion of the optical considerations of infrared sensors is provided later in the chapter.

Finally, infrared staring arrays consisting of cooled photon detectors or uncooled thermal detectors may have responsivities that vary dramatically from pixel to pixel. It is common practice to correct for the resulting nonuniformity using a combination of factory preset tables and user inputs. The nonuniformity can cause fixed pattern noise in the image that can limit the performance of the system even more than temporal noise and these effects are demonstrated in the next section.

39.1 Infrared Sensor Calibration

Significant advancement in the manufacturing of high-quality FPAs operating in the SWIR, MWIR, and LWIR has enabled industry to offer a wide range of affordable camera products to the consumer. Commercial applications of infrared camera technology are often driven by the value of the information it provides and price points set by the marketplace. The emergence of uncooled microbolometer FPA cameras with sensitivity less than 0.030°C at standard video rates has opened many new applications of the technology. In addition to the dramatic improvement in sensitivity over the past several years, uncooled microbolometer FPA cameras are characteristically compact, lightweight, and low power. Uncooled cameras are commercially available from a variety of domestic and foreign vendors including Agema, BAE Systems, CANTRONIC Systems, Inc., DRS and DRS Nytech, FLIR Systems, Inc., Indigo Systems, Inc.,

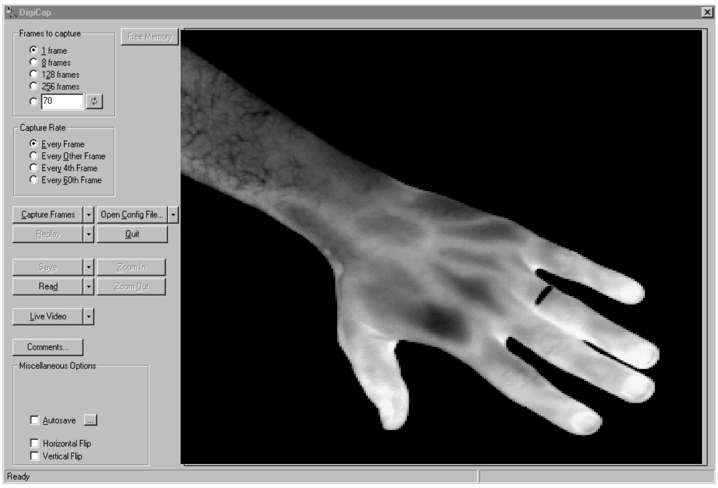


FIGURE 39.4 Windows-based GUI developed at NVESD for an uncooled medical imaging system.

Electrophysics Corp., Inc., Infrared Components Corp., IR Solutions, Inc., Raytheon, Thermoteknix Systems Ltd., ompact, low power The linearity, stability, and repeatability of the SiTF may be measured to determine the suitability of an infrared camera for accurate determination of the apparent temperature of an object of interest. LWIR cameras are typically preferred for imaging applications that require absolute or relative measurements of object irradiance or radiance because emitted energy dominates the total signal in the LWIR. In the MWIR, extreme care is required to ensure the radiometric accuracy of data. Thermal references may be used in the scene to provide a known temperature reference point or points to compensate for detector-to-detector variations in response and improve measurement accuracy. Thermal references may take many forms and often include temperature controlled extended area sources or uniformly coated metal plates with contact temperature sensors. Depending on the stability of the sensor, reference frames may be required in intervals from minutes to hours depending on the environmental conditions and the factory presets. Many sensors require an initial turn-on period to stabilize before accurate radiometric data can be collected. An example of a windows-based graphical user interface (GUI) developed at NVESD for an uncooled imaging system for medical studies is shown in Figure 39.4. The system allows the user to operate in a calibrated mode and display apparent temperature in regions of interest or at any specified pixel location including the pixel defined by the cursor. Single frames and multiple frames at specified time intervals may be selected for storage. Stability of commercially available uncooled cameras is provided earlier.

The LTC 500 thermal imager had been selected as a sensor to be used in a medical imaging application. Our primary goal was to obtain from the imagery calibrated temperature values within an accuracy of approximately a tenth of a degree Celsius. The main impediments to this goal consisted of several sources of spatial nonuniformity in the imagery produced by this sensor, primarily the spatial variation of radiance across the detector FPA due to self heating and radiation of the internal camera components, and to a lesser extent the variation of detector characteristics within the FPA. Fortunately, the sensor provides a calibration capability to mitigate the effects of the spatial nonuniformities.

We modeled the sensor FPA as a 2D array of detectors, each having a gain G and offset K, both of which are assumed to vary from detector to detector. In addition we assumed an internally generated radiance Y for each detector due to the self-heating of the internal sensor components (also varying from detector to detector, as well as slowly with time). Lastly, there is an internally programmable offset C for each detector which the sensor controls as part of its calibration function. Therefore, given a radiance X incident on some detector of the FPA from the external scene, the output Z for that detector is given by:

$$Z = GX + GY + K + C$$

39.2 Gain Detector

Individual detector gains were calculated by making two measurements. First, the sensor was allowed to run for several hours in order for the internal temperatures to stabilize. A uniform blackbody source at temperature T_1 (20°C) was used to fill the field of view (FOV) of the sensor and an output image Z_1 was collected. Next the blackbody temperature was set to T_2 (40°C) and a second output image Z_2 was collected. Since the measurement interval was small (<1 to 2 min) we assume the Y values remain constant, we have (for each detector):

$$Z_1 = GX_1 + GY + K + C$$

$$Z_2 = GX_2 + GY + K + C$$

where X_1 and X_2 refer to the external scene radiance corresponding to temperatures T_1 and T_2 incident on the detector and were calculated by integrating Planck's blackbody function over the 8 to 12 μm spectral band of the sensor. Taking the difference, we have (for each detector):

$$G = \frac{Z_2 - Z_1}{X_2 - X_1}$$

where the numbers here refer to measurements at different temperatures and, again, the value G (as well as Z and X) are assumed to vary from detector to detector (detector subscripts were omitted for clarity).

39.2.1 Nonuniformity Calibration

The LTC 500 provides the capability for nonuniformity calibration that allows the user to remove nonuniformities across the FPA assuming they do not change too rapidly with time. The procedure involves placing a uniform blackbody source across the FOV of the sensor and pressing the calibrate button. At this point, the sensor internally adjusts the value of a programmable offset for each detector so that the output Z of each detector is equal to a constant that we will denote Z_{CAL} (which the sensor sets to the midpoint of the digital pixel range, i.e., 16384).

Let D_1 and D_2 be two detectors selected from the 2D FPA, the outputs Z_1, Z_2 are then given by:

$$Z_1 = G_1X_1 + G_1Y_1 + K_1 + C_1$$

$$Z_2 = G_2X_2 + G_2Y_2 + K_2 + C_2$$

where now the numbers refer to different detectors, and as before G is gain, X is the incident radiance from the external scene, Y is the internal self heating radiance on the detectors, K is a possible offset variation from detector to detector, and C represents the programmable calibration offset for each detector. If we fill the FOV with a uniform blackbody at some temperature (T_{CAL}) producing a uniform radiance X_{CAL} on the FPA and activate the calibration function, we have:

$$Z_1 = Z_{\mathrm{CAL}} = G_1X_{\mathrm{CAL}} + G_1Y_1 + K_1 + C_1$$

$$Z_2 = Z_{\mathrm{CAL}} = G_2X_{\mathrm{CAL}} + G_2Y_2 + K_2 + C_2$$

so

$$C_1 = Z_{\mathrm{CAL}} - G_1X_{\mathrm{CAL}} - G_1Y_1 - K_1$$

$$C_2 = Z_{\mathrm{CAL}} - G_2X_{\mathrm{CAL}} - G_2Y_2 - K_2$$

where X_{CAL} is calculated by spectrally integrating Planck's function from 8 to 12μm at $T = T_{\mathrm{CAL}}$. C_1 and C_2 will now retain these values until the sensor is either recalibrated or powered down. Now, for some arbitrary externally supplied radiance X_1, X_2 on the FPA we have:

$$Z_1 = G_1X_1 + G_1Y_1 + K_1 + C_1$$

$$Z_1 = G_1X_1 + G_1Y_1 + K_1 + Z_{\mathrm{CAL}} - G_1X_{\mathrm{CAL}} - G_1Y_1 - K_1$$

$$Z_1 = G_1X_1 + Z_{\mathrm{CAL}} - G_1X_{\mathrm{CAL}}$$

$$Z_1 = G_1(X_1 - X_{\mathrm{CAL}}) + Z_{\mathrm{CAL}}$$

and, similarly

$$Z_2 = G_2(X_2 - X_{\mathrm{CAL}}) + Z_{\mathrm{CAL}}$$

therefore, the output of each detector depends only on the individual detector gain (which we know) and the external radiance incident on the detector. The spatially varying components (Y and K) have been removed.

Rearranging to solve for radiance input X as a function of the output intensity Z we have:

$$X = \frac{(Z - Z_{\mathrm{CAL}})}{G} + X_{\mathrm{CAL}}$$

Given a precomputed look up table RAD2TEMP of T, X pairs we can take the radiance value X and look up the corresponding temperature T for any pixel in the image. Hence we have computed temperature T_{C} as a function of radiance X on any detector:

$$T_{\mathrm{C}} = \mathrm{RAD2TEMP}[X]$$

39.3 Operational Considerations

Upon testing the system in a scenario that more accurately reflected the operational usage anticipated (i.e., with up to 10 ft between the sensor and the measured object), we encountered an unexpected discrepancy between the computed temperature values and the actual values as reported by the blackbody temperature display. We decided to assume that actual temperature values would vary linearly with the values computed by the above method. Therefore, we added a second step to the calibration procedure that requires the user to collect an image at a higher temperature than the calibration temperature T_{CAL}. Also, this second measurement would be made at a sensor to blackbody distance of approximately 10 ft. So now, we have two computed temperatures and two actual corresponding temperatures. Then we compute a slope and y-intercept describing the (assumed) linear relationship between the computed and actual temperatures.

$$T_{\mathrm{A}} = MT_{\mathrm{C}} + B$$

where

$$M = \frac{T_{\mathrm{A}_2} - T_{\mathrm{A}_1}}{T_{\mathrm{C}_2} - T_{\mathrm{C}_1}}$$

and

$$B = T_{\mathrm{A}_1} - MT_{\mathrm{C}_1}$$

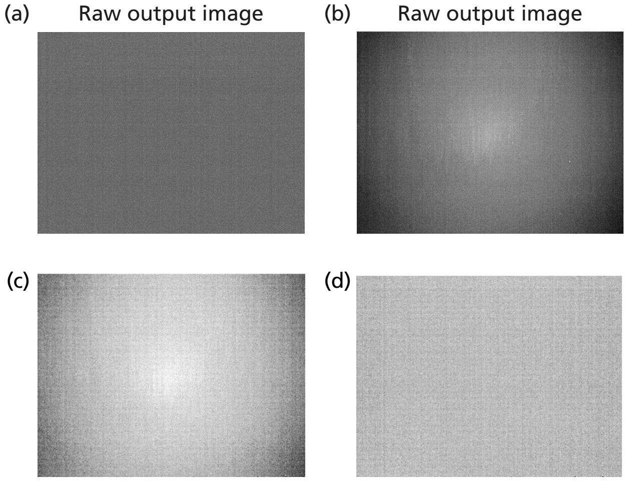


FIGURE 39.5 Gain calculation, (a) *low temperature*: uniform 30°C black body source, calibrated at 30°, $\mu = 16381$, $\sigma = 1.9$ (raw counts). (b) *High temperature*: uniform 40°C black body source, calibrated at 30°, $\mu = 16842, \sigma = 11.3$ (raw counts). (c) *Processed using uniform gain*: uniform 35°C black body source, calibrated at 30°, $\mu = 34.74, \sigma = 0.12$ (°C). (d) *Processed using computed gain*: uniform 35°C black body source, calibrated at 30°, $\mu = 34.74, \sigma = 0.04$ (°C).

where T_A is the adjusted temperature, T_C is the computed temperature from the previously described methodology. M and B are recomputed during each nonuniformity calibration.

From a camera perspective, the system intensity transfer function (SiTF) in digital counts for a DC $= A + BeT^4$

$$T = \left(\frac{DN - A}{B\varepsilon}\right)^{1/4}$$

By adding this second step to the calibration process, we were able to improve the accuracy of the computed temperature to within a tenth of a degree for the test data set (Figure 39.5).

39.4 Infrared Optical Considerations

This section focuses mainly on the MWIR and LWIR since optics in the NIR and SWIR and very similar to that of the visible. This area assumes a basic knowledge of optics, and applies that knowledge to the application of infrared systems.

39.4.1 Resolution

Designing an IR optical system is first initiated by developing a set of requirements that are needed. These requirements will be used to determine the focal plane parameters and desired spectral band. These parameters in turn drive the first order design, evolving into the focal length, entrance pupil diameter, FOV, and f/number.

If we start with the user inputs of target distance, size, cycle criteria (or pixel criteria), and spectral band, we can begin designing our sensor. First we calculate the minimum resolution angle (α, in radians). This parameter is also known as the instantaneous field of view (IFOV), and can be calculated by

$$\alpha = \frac{\mathrm{Size}_{\mathrm{tar}}}{(\mathrm{Range})(2 \times \mathrm{Cyles})}$$

or

$$\alpha = \frac{\text{Size}_{\text{tar}}}{(\text{Range})(\text{Pixels})}$$

Based on the wavelength, we can determine the minimum entrance pupil diameter that is necessary to distinguish between the two blur spots.

$$\text{EPD} = \frac{1.22\lambda}{\alpha}$$

Knowing the detector size and pixel pitch we can then determine the minimum focal length based on our IFOV that is required to meet our resolution requirements. Longer focal lengths will provide better spatial resolution.

$$\text{EFL} = \frac{\text{Pitch}}{\alpha}$$

Once we know our focal length, we can determine our FOV based on the height of the detector and the focal length. The vertical and horizontal fields of view are calculated separately based on their respective dimensions.

$$\theta = 2\tan\left[\frac{0.5h}{\text{EFL}}\right]$$

where h is the full detector height (or width). Depending on the system requirements, the size of the FPA may want to be scaled to match the desired FOV. Arrays with more pixels provide for greater resolution for a given FOV. However, smaller arrays cost less. Scaling an optical system to match the FOV for a different array format results in the scaling of the focal length, and thus the resolution.

The f/number is then calculated as the ratio of the focal length to the entrance pupil diameter.

$$f/\text{number} = \frac{\text{EFL}}{\text{EPD}}$$

The f/number of the system can be further optimized based on two parameters: the sensitivity and the blur circle. The minimum f/number is already set based on the calculated minimum entrance pupil diameter and focal length. The f/number can be adjusted to improve sensitivity by trying to optimize the blur circle to match the diagonal dimension of the detector pixel. This method provides a way to maximize the amount of energy on the pixel while minimizing aliasing.

$$f/\text{number} = \frac{\text{Pixel}_{\text{diagonal}}}{2.44\lambda}$$

A faster f/number is good in many ways as it can improve the resolution by reducing the blur spot and increasing the optics cutoff frequency. It also allows more signal to the detector and gives a boost in the signal to noise ratio. A fast f/number is absolutely necessary for uncooled systems because they have to overcome the noise introduced from operation at warmer temperatures. Faster f/numbers also help in environments with poor thermal contrast. However, a faster f/number also means that the optics will be larger, and the optical designs will be more difficult. Faster f/numbers introduce more aberrations into each lens making all aberrations more difficult to correct, and a diffraction limited system harder to achieve. A cost increase may also occur due to larger optics and tighter tolerances. A tradeoff has to occur to find the optimal f/number for the system. The table below shows the tradeoffs between optics diameter, focal length, resolution and FOV.

39.5 Spectral Requirement

The spatial resolution is heavily dependent on the wavelength of light and the f/number. Diffraction limits the minimum blur size based on these two parameters.

$$d_{\text{spot}} = 2.44\lambda(f/\text{number})$$

The table compares blur sizes for various wavelengths and f/numbers.

	Spot size (μm)			
f/number	$\lambda = 0.6$	$\lambda = 2$	$\lambda = 4$	$\lambda = 10$
1	1.5	4.9	9.8	24.4
2.5	3.7	12.2	24.4	61.0
4	5.9	19.5	39.0	97.6
5.5	8.1	26.8	53.7	134.2
7	10.2	34.2	68.3	170.8

First Order Parameters Resolution

	f/number	Focal length	Field of view	Entrance pupil diameter
Impact resolution	A faster f/number results in a smaller optics blur due to diffraction. The result is an improved diffraction limit, and thus better spatial resolution. However, two things can adversely impact this improvement. Aberrations increase with faster f/number, potentially moving a system out of being diffraction limited, in which case the faster f/number can potentially hurt you. Also, a fixed front aperture system will have to reduce its focal length to accommodate the faster f/number, thus reducing spatial resolution through a change in focal length	Increasing the focal length will increase spatial resolution. However, the amount of improvement may be limited by the size of the allowed aperture, as having to go to a slower f/number can reduce some of the gains. Longer focal lengths also result in narrower FOVs, which can be limited by stabilization issues	Narrower FOVs are the direct result of longer focal lengths. Longer focal lengths result in improved spatial resolution. The Field of view can also vary by changing the size of the FPA. If all other parameters are maintained, and the FPA size is increased merely through the addition of more pixels, then the FOV increases without impacting resolution. If the number of pixels stay the same, but the pixel size is increased. then resolution is decreased. If pixel size is constant, number of pixels is increased, and focal length is scaled to maintain a constant FOV, then the resolution scales with the focal length	The entrance pupil diameter (EPD) can impact the resolution in three ways. Increasing the EPD while maintaining focal length improves resolution by utilizing a faster f/number. Increasing the EPD while maintaining a constant f/number results in a longer focal length, and thus increase spatial resolution. Maintaining a constant EPD while increasing focal length results in an improved spatial resolution due to the focal length, but a reduced resolution due to diffraction. The point where there is no longer significant improvement is dependent on the pixel size

39.5.1 Depth of Field

It is often desired to image targets that are located at different distances in object space. Two targets that are separated by a distance will both appear to be equally in focus if they fall within the depth of field

(DOF) of the optics. A target that is closer to the optics than the DOF will appear defocused. In order to bring the out-of-focus target back in focus, it is required to refocus the optics so that the image plane shifts back to the location of the detector. Far targets focus shorter than near targets. If it is assumed that the optics are focused for an infinite target distance, the near DOF, known as the hyperfocal distance (HFD), can be found by

$$\text{HFD} = \frac{D^2}{2\lambda}$$

where D is the entrance pupil diameter, and λ is the wavelength in the same units as the diameter. This approximation can be made in the infrared because we can assume that we are utilizing a diffraction limited system.

This formula is dependent only on the aperture diameter and wavelength. It is easily seen that shorter wavelengths and larger apertures have larger HFDs. This relationship is based on the Rayleigh limit which states that as long as the wavefront error is within a $\frac{1}{4}$ wavelength of a true spherical surface, it is essentially diffraction limited. The depth of field can then be improved by focusing the optics to be optimally focused at the HFD. The optics are then in focus from that point to infinity based on the $\frac{1}{4}$ wave criteria, but also for a $\frac{1}{4}$ wave near that target distance. The full DOF of the system then becomes half the HFD to infinity. This is approximated by

$$\text{DOF} = \frac{D^2}{4\lambda}$$

If the region of interest is not within these bounds, it is possible to approximate the near and far focus points for a given object distance with

$$Z_{\text{near}} = \frac{\text{HFD} \times Z_o}{\text{HFD} + (Z_o - f)}$$

$$Z_{\text{far}} = \frac{\text{HFD} \times Z_o}{\text{HFD} - (Z_o - f)}$$

where X is the object distance that the objects are focused for and f is the focal length of the optics.

Work has been done with digital image processing techniques to improve the DOF by applying a method known as Wavefront Coding, developed by Cathey and Dowksi at the University of Colorado. This effectiveness of this technique has been well documented and demonstrated for the visible spectral band. Efforts are underway to demonstrate the effectiveness with LWIR uncooled cameras.

39.6 Selecting Optical Materials

The list of optical materials that transmit in the MWIR and LWIR spectrum is very short compared to that found in the visible spectrum. There are 21 crystalline materials and a handful chalcogenide glasses that transmit radiation at these wavelengths. Of the 21 crystalline materials, only six are practical to use in the LWIR, and nine are usable in the MWIR. The remaining possess poor characteristics such as being hygroscopic, toxic to the touch, etc., making them impractical for use in a real system. The list grows shorter for multi-spectral applications as only four transmit in the visible, MWIR, and LWIR. The chalcogenide glasses are an amorphous conglomerate of two or three infrared transmitting materials. A table of the practical infrared materials is listed along with a chart of their spectral bands. Transmission losses due to absorption can be calculated from the absorption coefficient of the material at the specified wavelength.

$$T_{\text{abs}} = \mathrm{e}^{-at}$$

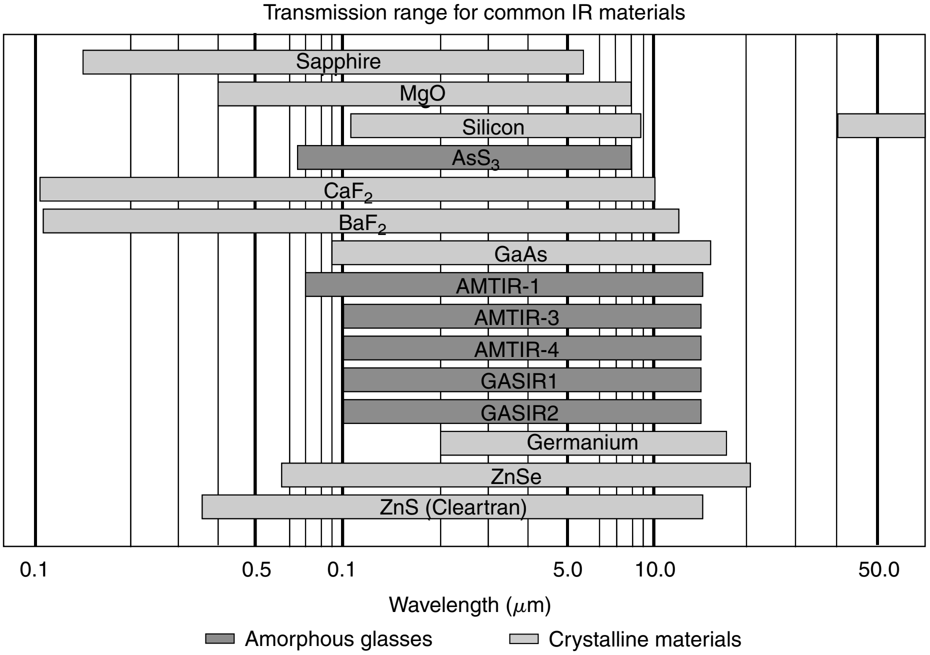


Table of Infrared Optical Materials

	Refractive Index			
Material	4 μm	=10 μ	dn/dT (K^{-1})	Spectral range (μm)
Germanium	4.0243	4.0032	0.000396	2.0–17.0
Gallium arsenide	3.3069	3.2778	0.000148	0.9–16.0
ZnSe	2.4331	2.4065	0.000060	0.55–20.0
ZnS (cleartran)	2.2523	2.2008	0.000054	0.37–14.0
AMTIR-1	2.2514	2.4976	0.000072	0.7–14.0
AMTIR-3	2.6200	2.6002	0.000091	1.0–14.0
AMTIR-4	2.6487	2.6353	−0.000030	1.0–14.0
GASIR 1	2.5100	2.4944	0.000055	1.0–14.0
GASIR 2	2.6038	2.5841	0.000058	1.0–14.0
Silicon	3.4255	N/A	0.000160	1.2–9.0
Sapphire	1.6753	N/A	0.000013	0.17–5.5
BaF_2	1.4580	1.4014	−0.000015	0.15–12.5
CaF_2	1.4097	1.3002	−0.000011	0.13–10.0
As_2S_3	2.4112	2.3816	−0.0000086	0.65–8.0
MgO	1.6679	N/A	0.000011	0.4–8.0

39.6.1 Special Considerations

In the LWIR, germanium is by far the best material to use for color correction and simplicity of design due to its high index of refraction and low dispersion. It is possible to design entire systems with germanium, but there are some caveats to this choice. Temperature plays havoc on germanium in two ways. It has a very high dn/dT (0.000396 K^{-1}), defocusing a lens that changes temperature of only a few degrees. It also has an absorption property in the LWIR for temperatures greater than 57°C. As the optic temperature rises above this point, the absorption coeffcient increases, reducing the transmission. The high cost of germanium can also be a factor. It is not always good choice for a low-cost sensor, as the lens may end up costing more than the FPA. A good thermal match to compensate for the $\mathrm{d}n/\mathrm{d}T$ of germanium is AMTIR-4. Its negative $\mathrm{d}n/\mathrm{d}T$ provides for an excellent compensator for the germanium. It is possible to design a two lens germanium/AMTIR-4 lens system that does not require refocusing for over a 60°C temperature range.

The low-cost optics are silicon, ZnS, and the chalcogenide glasses. Silicon is only usable in the MWIR, and although the material is very inexpensive, its hardness can make it very difficult to diamond turn, and thus expensive. Although it does not diamond turn well, it does grind and polish easily, providing a very inexpensive solution when complex surfaces such as aspheres and diffractives are not used. ZnS is relatively inexpensive to germanium and ZnSe, but is relatively expensive to silicon and the chalcogenide glasses. The chalcogenide glasses are by far the least expensive to make and manufacture making them an excellent solution for low-cost system design. There are three types of the chalcogenide glasses that are moldable. AMTIR-4, a product of Amorphous Materials, Inc., has a lower melting point that the other chalcogenides making it the easiest to mold. Another material GASIR1 and GASIR2, products of Umicore, have also been demonstrated as being moldable. They have similar optical properties to that of AMTIR-1 and AMTIR-3.

39.6.2 Coatings

The high indices of refraction of most infrared materials lead to large fresnel losses, and thus require AR coatings. The transmission for a plane uncoated surface is shown below. In air, $n_1 = 1$.

$$T = 1 - \left(\frac{n_1 - n_2}{n_1 + n_2}\right)^2$$

The total reflectance off both sides of an uncoated plate is the multiplication of the two surfaces. This is in turn multiplied by the transmission of the material due to absorption. An example is given below:

Example
Uncoated Zinc Selenide flat, $n = 2.4$ in air, $t = 1.5$ cm thickness, absorption $= 0.0005$.

$$\text{Total transmission through both sides} = (0.83)(0.999)(0.83) = 0.688$$

Standard AR coatings are readily available for all of the materials previously listed. Generally, better than 99% can be expected for an AR-coated lens for either the MWIR or LWIR. If dual band operation is required, expect this performance to drop to 96%, and for the price to go up. The multispectral coatings are more difficult to design, and result in having many more layers. Infrared beamsplitter coatings can be difficult and expensive to manufacture. Care should be taken in specifying both the transmission and reflection properties of the beamsplitter. Specifications that are too stringent often lead to huge costs and schedule delays. Also, it is very important to note that the choice of which wavelength passes through and which wavelength is reflected can make a significant impact on the performance of a beamsplitter. In general, transmitting the longer wavelength and reflecting the shorter wavelength will boost performance and reduce cost.

39.6.3 Reflective Optics

Reflective optics can be a very useful and effective design too in the infrared. Reflective optics have no chromatic aberrations, and allow for diffraction limited solutions for very wide spectral bands. However, the use of reflective optics is somewhat limited to narrow FOVs and have difficulty with fast f/numbers. In addition, the type of reflective system can impact the performance fairly significantly for longer wavelengths. The most common type of design, the Cassegrain, two mirrors that are aligned on the same optical axis. The secondary mirror acts as an obscuration to the primary, which results in a degraded MTF due to diffraction around the obscuration. This effect is not apparent in most wavebands because the MTF loss occurs after the Nyquist frequency of the detector. However, this is not the case for the LWIR where the MTF drop occurs before Nyquist. To overcome this effect, most reflective systems used in the LWIR are off-axis reflective optics. These optics will provide diffraction limited MTF as long as the f/numbers do not get too fast, and the FOVs do not get too large. The off-axis nature makes these reflective systems hard to align, and expensive to manufacture.

Acknowledgments

Thanks to John O'Neill, Jason Zeibel, Tim Mikulski, Kent McCormack for data. EO-IR Measurements, Inc. for developing the graphical user interface for the medical infrared imaging camera. Leonard Bonnell, Vipera Systems, Inc., for development of the infrared endoscope.

References

[1] J. D'Agostino and C. Webb, 3-D analysis framework and measurement methodology for imaging system noise. *Proc. SPIE*, 1488, 110–121 (1991).

[2] R.G. Driggers, P. Cox, and T. Edwards, *Introduction to Infrared and Electro-Optical Systems.* Artech House, Boston, MA, 1998, p. 8.

[3] G.C. Holst, *Electro-Optical Imaging System Performance.* JCD Publishing, Winter Park, FL, 1995, p. 347.

[4] M.W. Grenn, Recent advances in portable infrared imaging cameras. *Proc. IEEE-EMBS*, 1996, Amsterdam.

[5] M.W. Grenn, Performance of portable staring infrared cameras. *Proc. IEEE/EMBS* Oct.30–Nov. 2, 1997 Chicago, IL, USA.

Further Information

D'Agostino, J. and Webb, C. "3-D Analysis Framework and Measurement Methodology for Imaging System Noise," *Proc. SPIE*, 1488, 110–121 (1991).

Holst, G.C. *Electro-Optical Imaging System Performance.* JCD Publishing, Winter Park, FL, 1995, p. 432.

Johnson, J. "Analysis of Image Forming Systems," *Proceedings of IIS*, 249–273 (1958).

O.H. Schade, "Electro-optical Characteristics of Television Systems," *RCA Review*, IX(1–4) (1948).

Ratches, J.A. "NVL Static Performance Model for Thermal Viewing Systems," USA Electronics Command Report ECOM 7043, AD-A011212 (1973).

Sendall, R. and Lloyd, J.M. "Improved Specifications for Infrared Imaging Systems," *Proceedings of IRIS*, 14, 109–129 (1970).

Vollmerhausen, R.H. and Driggers, R.G. "NVTHERM: Next Generation Night Vision Thermal Model," *Proceedings of IRIS Passive Sensors*, 1 (1999).

IV

Medical Informatics

Luis G. Kun
IRMC/National Defense University

WHAT IS IT? In the summer of 2004 I was invited to lecture at Dartmouth. Dr. Rosen, my host, asked me if I could address my interpretation of Medical Informatics during my lecture. This presentation made me review and reflect on some old and new concepts that have appeared in the literature, in some cases. I have observed, for example, that the definition varies greatly depending on the profession of the person who answers this question. A few years earlier, while at the CDC, I had the opportunity to participate in a lecture by Ted Shortliffe, in which he explained a model of Medical Informatics that appears in his book (*Handbook of Medical Informatics*). I also like portions of the content that appears at the Vanderbilt University website in its program of medical informatics (MI) and finally the work of Musen/Von Bemmel reflected on their *Handbook of Medical Informatics.*

Healthcare informatics has been defined by these authors, respectively as:

- "A field of study concerned with the broad range of issues in the management and use of biomedical information, including medical computing and the study of the nature of medical information itself." Shortliffe E.H., Perreault L.E., Eds. *Medical Informatics: Computer Applications in Health Care and Biomedicine.* New York: Springer, 2001.
- "The science that studies the use and processing of data, information, and knowledge applied to medicine, health care and public health." Von Bemmel J.H., Musen M.A., Eds. *Handbook of Medical Informatics.* AW Houten, Netherlands: Bohn Stafleu Van Loghum; Heidelberg, Germany: Springer Verlag, 1997.

Vanderbilt's MI program uses a *"simplistic definition": Computer applications in medical care* and a *"better definition": Biomedical Informatics is an emerging discipline that has been defined as the study, invention, and implementation of structures and algorithms to improve communication, understanding, and management of medical information. The end objective of biomedical informatics is the coalescing of data, knowledge, and the tools necessary to apply that data and knowledge in the decision making process, at the time and place that a decision needs to be made. The focus on the structures and algorithms necessary to manipulate the information separates Biomedical Informatics from other medical disciplines where information content is the focus.*

The Vanderbilt model shows Medical Informatics as the intersection of three different domains: Biological Science, Information Analysis and Presentation (i.e., informatics, computation, statistics) and Clinical Health Services Research (i.e., Policy, Outcomes). The first two at the intersection create Bioinformatics, while the second and third create health informatics [through the translation from bench to bedside]. Some more definitions:

The noun informatics has one meaning; the sciences concerned with gathering and manipulating and storing and retrieving, and classifying recorded information [1]. in.for.mat.ics n. *Chiefly British.* Information science, and bi.o.in.for.mat.ics n. Information technology as applied to the life sciences, especially the technology used for the collection, storage, and retrieval of genomic data [2].

In the *Handbook of Medical Informatics* (Musen/VonBemmel et al.) medical informatics is located at the intersection of information technology and the different disciplines of medicine and health care. These authors decided not to enter into a fundamental discussion of the possible differences between medical informatics and health informatics, however several definitions of medical informatics (medical information science, health informatics) were given. Some of these take into account both the scientific and the applied sides of the field. They cited two definitions:

1. *Medical information science is the science of using system-analytic tools. . .to develop procedures (algorithms) for management, process control, decision making, and scientific analysis of medical knowledge* [3].
2. *Medical informatics comprises the theoretical and practical aspects of information processing and communication, based on knowledge and experience derived from processes in medicine and health care* [4].

According to Wikipedia: Medical informatics is the name given to the application of information technology to healthcare. It is the: "understanding, skills, and tools that enable the sharing and use of information to deliver healthcare and promote health" *(British Medical Informatics Society).*

Medical informatics is often called *healthcare informatics* or *biomedical informatics*, and forms part of the wider domain of *eHealth.* These later-generation terms reflect the substantive contribution of the citizen

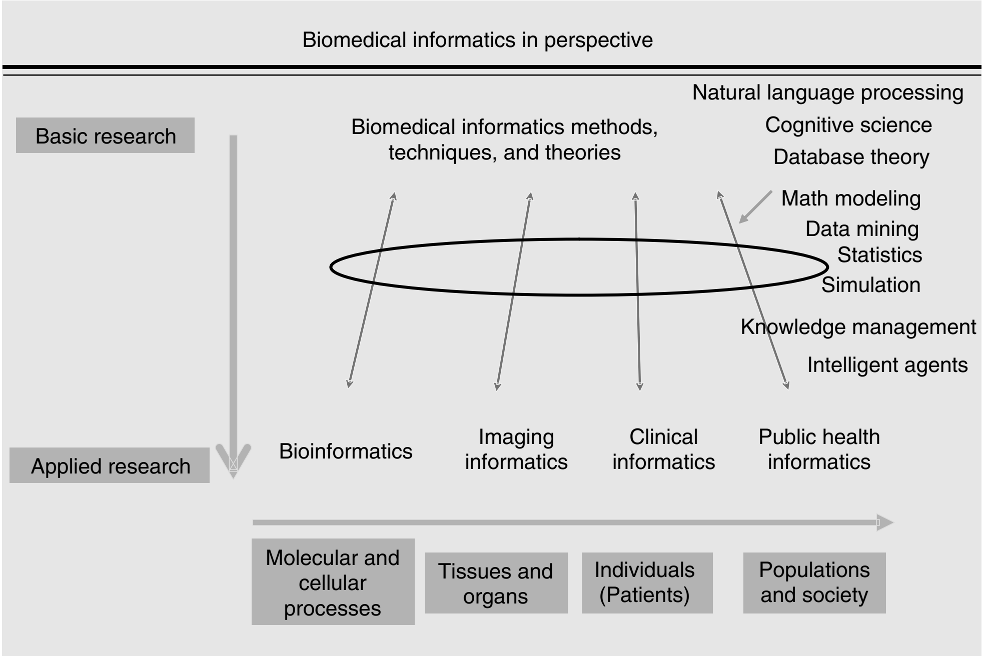


FIGURE IV.1 This figure is a modification of Ted Shortliffe's: "biomedical informatics in perspective." (Adapted from T. Shortliffe: Editorial, *Journal of Biomedical Informatics* 35 (2002) 279–280; republished with permission of the author).

and non-medical professions to the generation and usage of healthcare data and related information. Additionally, medical informaticians are active in bioinformatics and other fields not strictly defined as health care.

Biomedical Informatics: An Evolving Perspective

Some view medical informatics as a basic medical science with a wide variety of potential areas of applications. At the same time, the development and evaluation of new methods and theories are a primary focus of activities in this field. Using the results of experiences, allows, for example, the understanding, structuring, and encoding of knowledge, thus allowing its use in information processing by others, in the same field of specialty (i.e., within the field of clinical informatics) or in other areas (i.e., nursing informatics, dental informatics, and veterinary informatics). In Shortliffe's "biomedical informatics in perspective," his fundamental diagram shows how a clinician or researcher could "move" from basic research to applied research looking at molecular and cellular processes (bioinformatics), tissues and organs (imaging informatics), individual patients (clinical informatics) and populations and society (public health informatics). The important concept is, that a core series of informatics methods, tools and techniques, are the same regardless of the applied research area chosen.

This core includes: natural language processing, cognitive science, mathematical modeling and simulation, database theory, statistics, data mining, knowledge management, and intelligent agents. Many of these information technologies were developed in other fields and later applied by the "medical/health" community. In other cases the reverse has occurred. For example, the skeleton of an expert system developed for a medical diagnosis and treatment application, could be used by others to diagnose and correct problems in a computer system.

The basic core mentioned in the previous paragraph, then requires expertise in many different fields that include: biology, biomathematics, medicine, nursing, dentistry, veterinary, computer science–electrical

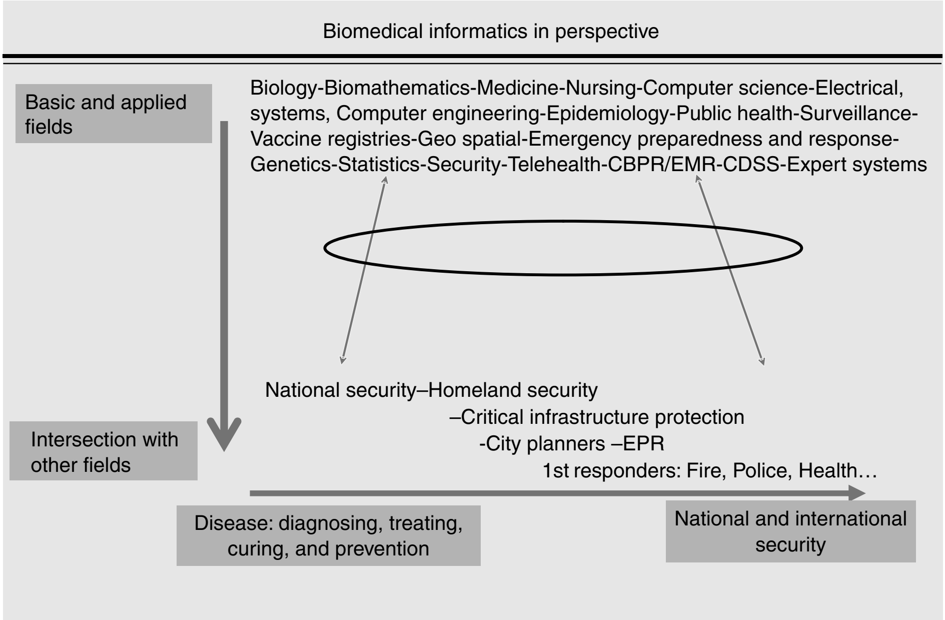

Biomedical informatics in perspective
Basic and applied fields
Biology-Biomathematics-Medicine-Nursing-Computer science-Electrical, systems, Computer engineering-Epidemiology-Public health-Surveillance-Vaccine registries-Geo spatial-Emergency preparedness and response-Genetics-Statistics-Security-Telehealth-CBPR/EMR-CDSS-Expert systems
National security–Homeland security
–Critical infrastructure protection
-City planners –EPR
1st responders: Fire, Police, Health…
Intersection with other fields
Disease: diagnosing, treating, curing, and prevention
National and international security

FIGURE IV.2 This is an example of applying the model shown in Figure IV.1, to other disciplines. It depicts biomedical informatics: tools, methods, techniques, and theories applied to individuals interested in the use of medical and public health informatics from a national security perspective. It also shows at the top, some of the different areas of expertise that may be touched by this field.

systems, computer engineering, epidemiology, public health, surveillance, geo-spatial information systems, emergency preparedness and response, genetics, statistics, security, telehealth, computer-based patient records, clinical decision support systems, expert systems, data mining/warehousing, etc. Figure IV.2, shows how if we start with the basic sciences and applied fields, we can develop a model that allows, through the use of medical and public health informatics, deal with issues such as national and international security and the critical infrastructure protection (CIP) of the public health infrastructure. This figure also shows many of the different professions or areas of expertise involved on the field of medical informatics.

Societal changes have occurred, in particular during the past four decades to information processing. In the last decade, for those that have observed the Internet and World Wide Web (WWW) evolution, many of these changes come as part of a new resulting culture. The convergence of computers and communications have played a key role in this evolving field of medical informatics. The way physicians, nurses, biomedical engineers, veterinarians, dentists, laboratory technicians, public health specialists, and other healthcare professionals do their "business" regardless if it is clinical, administrative, research and development or academic, requires and demands not only a clear understanding of these terms, but what their "business process" is all about.

For example, around the 1986 timeframe as technical manager for the requirements definition of the nursing point of care system for IBM Federal Systems in Gaithersburg, I had the opportunity to better understand the nursing process of information. Let us assume that a patient fall to the ground as he was having a heart attack. He cut his forehead, his left hand, arm, knee, and twisted his left ankle. When brought into an emergency room in a hospital the medical diagnosis given is: myocardial infarction. Yet from a nursing point of view, the patient diagnosis is different. He is assessed and evaluated, classified and care plans are developed for each of the diagnosis (for each of the injuries) according to a list that would include not only the heart attack (and the follow-up ordered by the physician), but the forehead, left hand, left arm, left knee, and left ankle. Following the NANDA diagnosis there will be a process to be followed for each wound or injury and that would facilitate the patient's discharge.

Each of the professionals mentioned above, that is, healthcare providers, etc. may have a piece of the patient's health puzzle which may need to be included in their record. In my professional life I have been involved in different activities that involved the utilization of different pieces of computer hardware, systems and application software and other devices that acquire, transmit, process, and store information, and in different forms (i.e., signals, voice, images from different modalities: CT, MRI, ultrasound, nuclear, x-ray, scanned/document images, etc.). In my early years at IBM in the late 70s, the concepts of "patient-centered" information was one of the leading concepts helping us drive towards the implementation of electronic health records. But in this day and age, we have a significant number of new devices, which were not made for that particular purpose and yet they are being used to collect, transmit, process, or store patient-related information. Biomedical engineers are and will be faced with these new challenges, where patient information may not be flowing from conventional data collection devices. Information now could come from sensing devices, RFID tags and through a myriad of wireless devices and networks. An RFID tag was approved in the 4Q 2004, by the FDA, that could be implanted into a patient. Aside from the issues of privacy and security of that information, the repercussions of such actions can be magnificent. It may improve the life of some or even save lives. Imagine, for example, patients suffering from either Alzheimer's or Parkinson's disease being admitted into an emergency room after an accident and not being able to convey any personal or medical-related information. Having available some basic but critical medical information can identify who the patient is, what medications they may be allergic, etc.

In the last 30 years, the field of medical informatics has grown tremendously both in its complexity and in content. As a result, two sections will be written in this handbook. The first one, represented in the chapters, will be devoted to areas that form a key "core" of computer technologies. These include: hospital information systems (HIS), computer-based patient records (CBPR or CPR), imaging, communications, standards, and other related areas. The second section includes the following topics: artificial intelligence, expert systems, knowledge-based systems neural networks, and robotics. Most of the techniques describe in the second section will require the implementation of systems explained in this first section. We could call most of these chapters the information infrastructure required to apply medical informatics techniques to medical data. These topics are crucial because they not only lay the foundation required to treat a patient within the walls of an institution but they also provide the roadmap required to deal with the patient's lifetime record while allowing selected groups of researchers and clinicians to analyze the information and generate outcomes research and practice guidelines information.

As an example, a network of associated hospitals in the East Coast (a health care provider network) may want to utilize an expert system that was created and maintained at Stanford University. This group of hospitals, HMOs, clinics, physician's offices, and the like would need a "standard" computer-based patient record (CPR) that can be used by the clinicians from any of the physical locations. In addition in order to access the information, all these institutions require telecommunications and networks that will allow for the electronic "dialogue." The different forms of the data, particularly clinical images, will require devices for displaying purposes, and the information stored in the different HIS, Clinical Information Systems (CIS), and departmental systems needs to be integrated. The multimedia type of record would become the data input for the expert system which could be accessed remotely (or locally) from any of the enterprise's locations. On the application side, the expert system could provide these institutions which techniques that can help in areas such as diagnosis and patient treatment. However, several new trends such as: total quality management (TQM), outcome research, and practice guidelines could be followed. It should be obvious to the reader that to have the ability to compare information obtained in different parts of the world by dissimilar and heterogeneous systems, certain standards need to be followed (or created) so that when analyzing the data, the information obtained will make sense.

Many information systems issues described in this introduction will be addressed in this section. The artificial intelligence chapters which follow should be synergistic with these concepts. A good understanding of the issues in this section is required prior to the utilization of the actual expert system. These issues are part of this section of medical informatics, other ones, however, for example, systems integration and process reengineering, will not be addressed here in detail but will be mentioned by the different authors. I encourage the reader to follow up on the referenced material at the end of each chapter, since the citations contain very valuable information.

Several perspectives in information technologies need to be taken in consideration when reading this section. One of them is described very accurately in the book entitled Globalization, Technology and Competition [5]. The first chapter of this book talks about new services being demanded by end users which include the integration of computers and telecommunications. From their stages theory point of view, the authors described very appropriately, that "we are currently" nearing the end of the micro era and the beginning of the network era. From an economy point of view, the industrial economy (1960s and 1970s) and the transitional economy (1970s and 1980s) moved into an information economy (1990s and beyond).

In the prior editions I mentioned that: "*Many other questions and answers reflected some of the current technological barriers and users needs. Because of these trends it was essential to include in this Handbook technologies that today may be considered state of the art but when read about 10 years from now will appear to be transitional only. Information technologies are moving into a multimedia environment which will require special techniques for acquiring, displaying, storing, retrieving, and communicating the information. We are in the process of defining some of these mechanisms.*" This is precisely what has occurred. In some instances such as imaging, this Handbook contains a full section dedicated to the subject. That section contains the principles, the associated math algorithms, and the physics related to all medical imaging modalities. The intention in this section was to address issues related to imagining as a form of medical information. These concepts include issues related then to acquisition, storage and retrieval, display and communications of document and clinical images, for example, picture archival and communications systems (PACS). From a CPR point of view, clinical and document images will become part of this electronic chart, therefore many of the associated issues will be discussed in this section more extensively. The state of the telecommunications has been described as a revolution; data and voice communications as well s full-motion video have come together as a new dynamic field. Much of what is happening today is a result of technology evolution and need. The connecting thread between evolutionary needs and revolutionary ideas is an integrated perspective of both sides of multiple industries. This topic will also be described in more detail in this section.

A Personal (Historical) Perspective on Information Technology in Healthcare: During my first 14 "professional years" I worked with IBM (1978–1992) across the country: Los Angeles, Dallas, Gaithersburg, Dallas and Houston. As I moved from city to city I noticed that while financial transactions would simply follow me (from place-to-place), that is, with the use of the same credit cards; this was not the case with my medical records. In fact, there wasn't an "electronic" version available. In today's terms my "paper-trail record" was a "cut and paste" document that I had to create and take with me wherever I went.

At work, I was engaged since 1983, on the development of concepts of the "All Digital Medical Record" (ADMR), bedside terminals, PACS, Integrated Diagnostics Systems, and other related topics such as Telemedicine. In the 1986 timeframe and while at IBM, Al Potvin (former IEEE-EMBS President) asked me to become part of the IEEE-USA Health Care Engineering Policy Committee, (HCEPC), and I did. I formed and chaired then, the Electronic Medical Record/High Performance Computers and Communications (EMR/HPCC) working group. Many of these concepts were presented at EMBS [6] and AAMI [7, 8] annual meetings, in Biomedical Engineering classes [9, 10], and even at National [11–13] and International Conferences/Meetings [14–19]. Nine and ten years ago respectively, we (the HCEPC) had 2 meetings where the Role of Technology in the Cost of Health Care were explored [20, 21]. At that time, (in the early 1990s.), and given the rate at which the American Health Care system was growing, the HCEPC asked me to organize a special Technology Policy session during the 1993 EMBS meeting in San Diego [22]. (After the formal presentations, the participants, who represented the United States, Canada and Europe had a terrific discussion that lasted well over 3 h from the allotted time for the session.) Costs were the fundamental and center piece of these discussions. The Clinton-Gore administration encouraged the use of information technology for health care reform and as a consequence we organized a meeting in 1995 [23] to address these needs.

Near seven years later the book entitled: To Err is Human[24] changed the focal point from "costs" to "human lives taken by medical errors". Experts estimated that as many as 98,000 people die (in the United States) in any given year from medical errors that occur in hospitals. That's more than die from motor vehicle accidents, breast cancer, or AIDS — three causes that receive far more public attention. Indeed, more people die annually from medication errors than from workplace injuries. Add the financial cost to the human tragedy, and medical error easily rose to the top ranks of urgent, widespread public problems. This book broke the silence that has surrounded medical errors and their consequence — but not by pointing fingers at caring health care professionals who make honest mistakes. After all, to err is human. Instead, it set forth a national agenda — with state and local implications — for reducing medical errors and improving patient safety through the design of a safer health system. It revealed the often startling statistics of medical error and the disparity between the incidence of error and public perception of it, given many patients' expectations that the medical profession always performs perfectly. A careful examination was made of how the surrounding forces of legislation, regulation, and market activity influenced the quality of care provided by health care organizations and then looked at their handling of medical mistakes. (Using a detailed case study, the book reviews the current understanding of why these mistakes happen. A key theme is that legitimate liability concerns discourage reporting of errors — which begs the question, "How can we learn from our mistakes?") Balancing regulatory versus market-based initiatives and public versus private efforts, the Institute of Medicine presented wide-ranging recommendations for improving patient safety, in the areas of leadership, improved data collection and analysis, and development of effective systems at the level of direct patient care. The bottom line was that it asserted that the problem is not bad people in health care — it is that good people are working in bad systems that need to be made safer.

A series of efforts, including both private and public sectors that is, e-Health Initiative [25], occurred which prompted a series of actions, documents and even proposed legislation in 2003, that is, HR 2915 The National Health Information Infrastructure (NHII); and also a National meeting where the requirements for such an infrastructure were agreed upon. A follow-up meeting took place on July 20–23, 2004 here in Washington, D.C.

The Secretarial Summit on Health Information Technology launching the National Health Information Infrastructure 2004: Cornerstones for Electronic Healthcare was well attended by over 1500 people representing the private and public healthcare industry. In challenging both sectors of the healthcare industry, Secretary Tommy G. Thompson stated, "Health information technology can improve quality of care and reduce medical errors, even as it lowers administrative costs. It has the potential to produce savings of 10% of our total annual spending on health care, even as it improves care for patients and provides new support for health care professionals." A report, titled "*The Decade of Health Information Technology: Delivering Consumer-centric and Information-rich Health Care*," ordered by President George W. Bush in April, was presented on July 21st by David Brailer, the National Coordinator for Health Information Technology, whom the president appointed to the new position in May. The report lays out the broad steps needed to achieve always-current, always-available electronic health records (EHR) for Americans. This responds to the call by President Bush to achieve EHRs for most Americans within a decade. The report identifies goals and action areas, as well as a broad sequence needed to achieve the goals, with joint private or public cooperation and leadership. The heads of every agency within DHHS (i.e., AHRQ, NIH, CDC, FDA, CMS, HRSA, etc.) were present and each made a presentation on how the NHII would affect their own area of service (e.g., research and development, education, reimbursement, etc.)

For more details about the news coverage on the NHII conference, please link to the following sites:

1. HHS Fact Sheet-HIT Report at-a-glance 7/21/04:
http://www.hhs.gov/news/press/2004pres/20040721.html
2. NY Times 7/21/04:
http://www.nytimes.com/2004/07/21/technology/21record.html

3. GovExec.com 7/21/04:
   http://www.governmentexecutive.com/dailyfed/0704/072104dk1.htm
4. Center for Health Transformation 7/21/04:
   http://www.healthtransformation.net/news/chtnews.asp
5. iHealthbeat 7/26/04:
   http://www.ihealthbeat.org/index.cfm?Action=dspItem&itemid=104473
6. USNews.com 8/2/04:
   http://www.usnews.com/usnews/tech/articles/040802/2wired.htm

New issues of the Information Age

As time passes by, we ("the educated world population") are becoming more accustomed to see how the Internet for example is being used to manage the health of the elderly and for homecare purposes. People living in urban, suburban, and rural areas are using telehealth as one of several mechanisms to deal with their personal health. This is one of the many ways that in my opinion healthcare costs have the potential to be reduced [26]. The Balanced Budget Act of 1997 was in fact the first time that the U.S. government decided to measure cost and medical effectiveness through telemedicine, involving elderly patients with diabetes. (A grant of about $28 million was given to Columbia University Medical Center in March 2000 and was renewed in 2004.)

In the United States the constant rising prices of medications however, are also making the elderly (in particular) look for alternative mechanisms to purchase them, that is, the Internet. The new questions that should be raised are: How can we ensure the quality of those drugs ordered through this mechanism? (i.e., Who is accountable? Who is the producer of such medications? etc.) How many people may be dying (or affected negatively) by those that are self-prescribing medications?

Several bills in the U.S. Congress are trying to deal, with some of these issues. The language has the potential to affect prescribing medications during interactive video consultations. The intent of this legislation, to ensure patients have access to safe and appropriate medicine, is a good one; however it may limit what we will be able to do through this type of technology. Some Internet Rx Legislation Under Consideration. (Bills can be accessed online at http://Thomas.loc.gov.):

- S 2464: Internet Pharmacy Consumer Protection Act aka Ryan Haight Act. *Sen. Coleman (MN). Co-sponsor: Feinstein.* Requires an in-person medical evaluation in order for a practitioner to dispense a prescription to a patient.
- HR 3880: Internet Pharmacy Consumer Protection Act. *Rep. Davis (VA) Cosponsors: Waxman.* Requires an in-person medical evaluation in order for a practitioner to dispense a prescription to a patient.
- S 2493: Safe Importation of Medical Products and Other Rx Therapies Act of 2004. *Sen. Gregg (NH). Co-sponsors: Smith.* Requires a treating provider to perform a documented patient evaluation (including a patient history and physical examination) of an individual to establish the diagnosis for which a prescription drug is prescribed.
- HR 3870: Prescription Drug Abuse Elimination Act of 2004. *Rep. Norwood (GA). Co-sponsor: Strickland.* Defines treating provider as a health care provider who has performed a documented patient evaluation of the individual involved (including a patient history and physical examination) to establish the diagnosis for which a prescription drug involved is prescribed.
- HR 2652: Internet Pharmacy Consumer Protection Act. *Rep. Stupak (MI).*States that a person may not introduce a prescription drug into interstate commerce or deliver the prescription drug for introduction into such commerce pursuant to a sale state requires an in-person medical evaluation in order for a practitioner to dispense a prescription to a patient.

In some cases, the defined requirements or the language used (i.e., in person patient evaluation) may act precisely against the use of that technology (i.e., tele-consultation). It is in the best interest of society, that biomedical engineers, among others get involved in "educating" the law-makers regarding the intrinsic value that the technology may bring into the system.

Summary of the Medical Informatics Section Covered in the Prior Edition(s)

In the first chapter Allan Pryor provided us with a tutorial on hospital information systems (HIS). He described not only the evolution of HIS and departmental systems and clinical information systems (CIS), but also their differences. Within the evolution he followed these concepts with the need for the longitudinal patient record and the integration of patient data. This chapter included patient database strategies for the HIS, data acquisition, patient admission, transfer and discharge functions. Also discussed were patient evaluation and patient management issues. From an end-user point of view, a terrific description on the evolution of data-driven and time-driven systems was included, culminating with some critical concepts on HIS requirements for decision support and knowledge base functionality. His conclusions were good indication of his vision.

Michael Fitzmaurice followed with "Computer-Based Patient Records" (CBPR or CPR). In the introduction, it was explained what is the CPR and why it is a necessary tool for supporting clinical decision making and how it is enhanced when it interacts with medial knowledge sources. This was followed by clinical decision support systems (CDSS): knowledge server, knowledge sources, medical logic modules (MLM), and nomenclature. This last issue in particular was one which needed to be well understood. The nomenclature used by physicians and by the CPRs differ among institutions. Applying logic to the wrong concepts can produce misinterpretations. The scientific evidence in this chapter included: patient care process, CDSS hurdles, CPR implementation, research data bases, telemedicine, hospital, and ambulatory care systems. A table of hospital and ambulatory care computer-based patient records systems concluded this chapter.

Because of the fast convergence of computers, telecommunications, and healthcare applications; already in the last edition it was impossible to separate these elements (i.e., communications and networks). Both are part of information systems. Soumitra Sengupta provided us in this chapter with a tutorial-like presentation which included an introduction and history, impact of clinical data, information types, and platforms. The importance of this section was reflected both in the contents reviewed under current technologies — LANs, WANs, middleware, medical domain middleware; integrated patient data base, and medical vocabulary — as well as in the directions and challenges section which included improved bandwidth, telemedicine, and security management. In the conclusions the clear vision is that networks will become the de facto fourth utility after electricity, water, and heat. Since the publishing of the last edition, both the Internet and the World Wide Web (WWW) had taken off in multiple directions, creating a societal-vision, which is much different than the one that most people expected then. For example, the use of telemedicine was seen as a tool for rural medicine and for emergency medicine and not an accepted one for Home care for the elderly and for persons suffering from a large number Chronic Diseases, both here in the United States and in the rest of the world.

"Non-AI Decision Making" was covered by Ron Summers and Ewart Carson. This chapter included an introduction which explained the techniques of procedural or declarative knowledge. The topics covered in this section included: analytical models, and decision theoretic models, including clinical algorithms, and decision trees. The section that followed covered a number of key topics which appear while querying large clinical databases to yield evidence of either diagnostic or treatment or research value; statistical models, database search, regression analysis, statistical pattern analysis, bayesian analysis, Depster-Shafer theory, syntactic pattern analysis, causal modeling, artificial neural networks. In the summary the authors clearly advised the reader to read this section in conjunction with the expert systems chapters that followed.

The standards section was closely associated with the CPR chapter of this section. Jeff Blair did a terrific job with his overview of standards related to the emerging health care information infrastructure. This chapter gave the reader not only an overview of the major existing and emerging health care information standards but an understanding of all (the "then," current) efforts, national and international, to coordinate, harmonize, and accelerate these activities. The introduction summarized how this section was organized. It included identifier standards (patient's, site of care, product, and supply labeling), communications (message format) standards, content, and structure standards. This section was followed by a summary of clinical data representations, guidelines for confidentiality, data, security, and authentication. After that quality indicators and data sets were described along with international standards. Coordinating and promotion organizations were listed at the end of this chapter including points of contact which proved to be very beneficial for those who needed to follow up.

Design issues in developing clinical decision support and monitoring systems by John Goethe and Joseph Bronzino provided insight for the development of clinical decision support systems. In their introduction and throughout this chapter, the authors provided a step-by-step tutorial with practical advice and make recommendations on design of the systems to achieve end-user acceptance. After that a description of a clinical monitoring system, developed and implemented by them for a psychiatric practice, was presented in detail. In their conclusions, the human engineering issue was discussed.

What is New in This Edition

There are two new chapters. One is an Introduction to Nursing Informatics by Kathleen McCormick et al. and one is on Medical Informatics and Biomedical Emergencies: New Training and Simulation Technologies for First Responders, by Joseph Rosen et al. Since nurses work side by side with physicians, biomedical engineers, information technologists, and others, the understanding becomes crucial of this profession and what they do in their daily hospital activities becomes crucial.

Delivering detection, diagnostic, and treatment information to first responders remains a central challenge in disaster management. This is particularly true in biomedical emergencies involving highly infectious agents. Adding inexpensive, established information technologies to existing response system will produce beneficial outcomes. In some instances, however, emerging technologies will be necessary to enable an immediate, continuous response. This article identifies and describes new training, education and simulation technologies that will help first responders cope with bioterrorist events.

Two revisions from the prior edition, were made by Summers and Carson and Michael Fitzmaurice respectively.

The authors of this section represented industry, academia, and government. Their expertise in many instances is multiple from developing to actual implementing these technical ideas. I am very grateful for all our discussions and their contributions.

References

[1] WordNet 1.7.1 Copyright© 2001 by Princeton University. All rights reserved.

[2] The American Heritage® Dictionary of the English Language, 4th ed. Copyright© 2004, 2000 by Houghton Mifflin Company. Published by Houghton Mifflin Company. All rights reserved.

[3] Shortliffe E.H. The science of biomedical computing. *Med. Inform.* 1984; 9: 185–93.

[4] Von Bemmel J.H. The structure of medical informatics. *Med. Inform.* 1984; 9: 175–80.

[5] (The Fusion of Computers and Telecommunications in the 1990s) by Bradley, Hausman and Nolan (1993).

[6] Invited Lecturer for the IEEE-EMBS, Dallas/Fort Worth Section. Session Topic — "Paperless Hospital" University of Texas, Arlington, TX; October 1989.

[7] Kun, L., Chairman during the *24th Annual AAMI-Association for the Advancement of Medical Instrumentation-Conference*, Session Topic — "The paperless hospital." Presentation: "Trends for the creation of an All Digital Medical Record," St. Louis, Missouri; May 1989.

[8] Kun, L., Guest Speaker to the *Association for the Advancement of Medical Instrumentation.* Session: Technology Management in 21st Century Health Care. Topic: "Electronic Community Medical Records." Anaheim, CA; May 1995.

[9] Kun, L., Invited lecturer Trinity College/*The Hartford Graduate Center Biomedical Engineering Seminar Series Fall 1990.* Session Topic — "The Digital Medical Record" Hartford, CT.

[10] Kun, L., UTA Biomedical Engineering Dept. Graduate Seminar. Topic —"All Digital Medical Record". Arlington, TX; October 1988.

[11] Kun, L., Guest lecturer for the *Emergency Physicians Training Conference.* Session Topic — "All Digital Medical Record at the ER", Infomart, Dallas, TX; June 1990.

[12] Kun, L., Guest Speaker to the *1st National Symposium on Coupling Technology to National Need.* Topic: "Impact of the Electronic Medical Record and the High Performance Computing and Communications in Health Care. Albuquerque, New Mexico; August 1993.

[13] Kun, L., Distinguished Lecturer and part of a debate panel on the topic: "The Vision of the Future of Health Care Technology and Health Care Reform." Presentation topic: "The Role of the EMR and the HPCC in Controlling the Cost of Health Care." *Sigma-Xi Distinguished Lectureship Program at the University of Connecticut,* April 1994.

[14] Kun, L., Applications of Microcomputers in Government, Industry and University. Centro Internacional de Fisica y la Universidad Nacional Quito, Ecuador , Course 4 Topic: *"The All Digital Medical Record. Image, Voice, Signals, Text and Graphics used for Patient Care";* July 1987.

[15] Kun, L., Universidad Nacional, Santo Domingo, Dominican Republic. Course 4 Topic: "The All Digital Medical Record. Image, Voice, Signals, Text and Graphics used for Patient Care"; July 1987.

[16] Kun, L., *Seminar Manager for the 1 week IBM Europe Institute: "Medical Computing: The Next Decade."* Presentation — "All Digital Medical Record", Garmisch-Partenkirchen, Germany. August 1989.

[17] Kun, L., *Seminar Technical Chairman IBM Research/IBM France.* Topic - "Medical Imaging Computing: The Next Decade." Presentation — "The Paperless Hospital", Lyon, France; June 1990.

[18] Kun, L., Guest Speaker and Session Chairman to the 1994 *2nd International Symposium Biomedical Engineering in the 21st Century.* Topics: "Telemedicine" and "The Computer Based Patient Record." Center for Biomedical Engineering, College of Medicine, National Taiwan University, Taipei Convention Center, Taipei, Taiwan, R.O.C. September 1994.

[19] Kun, L., Invited lecturer to the Director of Health Industry Marketing Asia/Pacific. Session Topic - "The All Digital Medical Record", Tokyo, Japan. Sponsor IBM Japan/Asia Pacific HQ; April 1991.

[20] Kun, L., Guest Speaker to the Role of Technology in the Cost of Health Care Conference. Session: "The Role of Information Systems in Controlling Costs". Topic: "The Electronic Medical Record and the High Performance Computing and Communications Controlling Costs of Health Care." Washington, DC; April 1994.

[21] Kun, L., Guest Speaker to the Role of Technology in the Cost of Health Care: Providing the Solutions Conference. Health Care Technology Policy II Topic: " Transfer & Utilization of Government Technology Assets to the Private Sector in the Fields of Health Care and Information Technologies". Moderator Session: "Use of Information Systems as a Management Tool". Washington, DC; May 1995.

[22] Kun, L., Session Chairman of a Special Mini-Symposia on: "Engineering Solutions to Healthcare Problems: Policy Issues", during the *15th annual International Conference IEEE-EMBS.* Also presenter of the lecture: "Health Care Reform Addressed by Technology Transfer: The Electronic Medical Record & the High Performance Computers & Communications Efforts. San Diego, California; October 1993.

[23] Kun, L., Conference Chairman — Health Care Information Infrastructure, Part of *SPIE's 1995 Symposium on Information, Communications and Computer Technology, Applications and Systems.* Philadelphia, Pennsylvania; October 1995

[24] Kohn, L., Janet M. Corrigan, and Molla S. Donaldson, Editors; Committee on Quality of Health Care in America, Institute of Medicine, "To Err is Human: Building a Safer Health System", (2000) Institute of Medicine.

[25] Foundation for eHealth Initiative: http://www.ncpdp.org/pdf/eHIOverview.pdf

[26] Kun, L., Guest lecture: "Healthcare of the elderly in the 21st century. Can we afford not to use telemedicine?" CIMIC, Rutgers University, NJ; December 12, 1996: http://cimic.rutgers.edu/seminars/kun.html

40

Hospital Information Systems: Their Function and State

T. Allan Pryor
University of Utah

The definition of a hospital information system (HIS) is unfortunately not unique. The literature of both the informatics community and health care data processing world is filled with descriptions of many differing computer systems defined as an HIS. In this literature, the systems are sometimes characterized into varying level of HISs according to the functionally present within the system. With this confusion from the literature, it is necessary to begin this chapter with a definition of an HIS. To begin this definition, I must first describe what it is not. The HIS will incorporate information from the several departments within the hospital, but an HIS is not a departmental system. Departmental systems such as a pharmacy or a radiology system are limited in their scope. They are designed to manage only the department that they serve and rarely contain patient data captured from other departments. Their function should be to interface with the HIS and provide portions of the patient medical/administrative record that the HIS uses to manage the global needs of the hospital and patient.

A clinical information system is likewise not an HIS. Again, although the HIS needs clinical information to meets its complete functionality, it is not exclusively restricted to the clinical information supported by the clinical information systems. Examples of clinical information systems are ICU systems, respiratory care systems, nursing systems. Similar to the departmental systems, these clinical systems tend to be one-dimensional with a total focus on one aspect of the clinical needs of the patient. They provide little support for the administrative requirements of the hospital.

If we look at the functional capabilities of both the clinical and departmental systems, we see many common features of the HIS. They all require a database for recording patient information. Both types of systems must be able to support data acquisition and reporting of patient data. Communication of information to other clinical or administrative departments is required. Some form of management support can be found in all the systems. Thus, again looking at the basic functions of the system one cannot differentiate the clinical/departmental systems from the HIS. It is this confusion that makes defining the HIS difficult and explains why the literature is ambiguous in this matter.

The concept of the HIS appears to be, therefore, one of integration and breadth across the patient or hospital information needs. That is, to be called an HIS the system must meet the global needs of those it is to serve. In the context, if we look at the hospital as the customer of the HIS, then the HIS must be able to provide global and departmental information on the state of the hospital. For example, if we consider the capturing of charges within the hospital to be an HIS function, then the system must capture all patient charges no matter which departmental originated those charges. Likewise all clinical information about the patient must reside within the database of the HIS and make possible the reporting and management of patient data across all clinical departments and data sources. It is totality of function that differentiates the HIS from the departmental or restricted clinical system, not the functions provided to a department or clinical support incorporated within the system.

The development of an HIS can take many architectural forms. It can be accomplished through interfacing of a central system to multiple departmental or clinical information systems. A second approach which has been developed is to have, in addition to a set of global applications, departmental or clinical system applications. Because of the limitation of all existing systems, any existing comprehensive HIS will in fact be a combination of interfaces to departmental/clinical systems and the applications/database of the HIS purchased by the hospital.

The remainder of this chapter will describe key features that must be included in today's HIS. The features discussed below are patient databases, patient data acquisition, patient admission/bed control, patient management and evaluation applications, and computer-assisted decision support. This chapter will not discuss the financial/administrative applications of an HIS, since those applications for the purposes of this chapter are seen as applications existing on a financial system that may not be integral application of the HIS.

40.1 Patient Database Strategies for the HIS

The first HISs were considered only an extension of the financial and administrative systems in place in the hospital. With this simplistic view many early systems developed database strategies that were limited in their growth potential. Their databases mimicked closely the design of the financial systems that presented a structure that was basically a "flat file" with well-defined fields. Although those fields were adequate for capturing the financial information used by administration to track the patient's charges, they were unable to adapt easily to the requirement to capture the clinical information being requested by health care providers. Today's HIS database should be designed to support a longitudinal patient record (the entire clinical record of the patient spanning multiple inpatient, outpatient encounters), integration of all clinical and financial data, and support of decision support functions.

The creation of a longitudinal patient record is now a requirement of the HIS. Traditionally the databases of the HISs were encounter-based. That is, they were designed to manage a single patient visit to the hospital to create a financial record of the visit and make available to the care provider data recorded during the visit. Unfortunately, with those systems the care providers were unable to view the progress of the patient across encounters, even to the point that in some HISs critical information such as patient allergies needed to be entered with each new encounter. From the clinical perspective, the management of a patient must at least be considered in the context of a single episode of care. This episode might include one or more visits to the hospital's outpatient clinics, the emergency department, and multiple inpatient stays. The care provider to manage properly the patient, must have access to all the information recorded from those multiple encounters. The need for a longitudinal view dictates that the HIS database structure must both allow for access to the patient's data independent of an encounter and still provide for encounter-based access to adapt to the financial and billing requirements of the hospital.

The need for integration of the patient data is as important as the longitudinal requirement. Traditionally the clinical information tended to be stored in separate departmental files. With this structure it was easy to report from each department, but the creation of reports combining data from the different proved difficult if not impossible. In particular in those systems where access to the departmental data was provided only though interfaces with no central database, it was impossible to create an integrated patient evaluation report. Using those systems the care providers would view data from different screens at their terminal and extract with pencil onto paper the results from each departmental (clinical laboratory, radiology, pharmacy, and so on) the information they needed to properly evaluate the patient. With the integrated clinical database the care provider can view directly on a single screen the information from all departments formatted in ways that facilitate the evaluation of the patient.

Today's HIS is no longer merely a database and communication system but is an assistant in the management of the patient. That is, clinical knowledge bases are an integral part of the HIS. These knowledge bases contain rules and/or statistics with which the system can provide alerts or reminders or implement clinical protocols. The execution of the knowledge is highly dependent on the structure of the clinical database. For example, a rule might be present in the knowledge base to evaluate the use of narcotics by the patient. Depending on the structure of the database, this may require a complex set of rules looking at every possible narcotic available in the hospital's formulary or a single rule that checks the presence of the class narcotics in the patient's medical record. If the search requires multiple rules, it is probably because the medical vocabulary has been coded without any structure. With this lack of structure there needs to be a specific rule to evaluate every possible narcotic code in the hospital's formulary against the patient's computer medication record. With a more structured data model a single rule could suffice. With this model the drug codes have been assigned to include a hierarchical structure where all narcotics would fall into the same hierarchical class. Thus, a single rule specific only to the class "narcotics" is all that is needed to compare against the patient's record.

These enhanced features of the HIS database are necessary if the HIS is going to serve the needs of today's modern hospital. Beyond these inpatient needs, the database of the HIS will become part of an enterprise clinical database that will include not only the clinical information for the inpatient encounters but also the clinical information recorded in the physician's office or the patient's home during outpatient encounters. Subsets of these records will become part of state and national health care databases. In selecting, therefore, and HIS, the most critical factor is understanding the structure and functionality of its database.

40.2 Data Acquisition

The acquisition of clinical data is key to the other functions of the HIS. If the HIS is to support an integrated patient record, then its ability to acquire clinical data from a variety of sources directly affect its ability to support the patient evaluation and management functions described below. All HIS systems provide for direct terminal entry of data. Depending on the system this entry may use only the keyboard or other "point and click" devices together with the keyboard.

Interfaces to other systems will be necessary to compute a complete patient record. The physical interface to those systems is straightforward with today's technology. The difficulty comes in understanding the data that are being transmitted between systems. It is easy to communicate and understand ASCII textual information, but coded information from different systems is generally difficult for sharing between systems. This difficulty results because there are no medical standards for either medical vocabulary or the coding systems. Thus, each system may have chose an entirely different terminology or coding system to describe similar medical concepts. In building the interface, therefore, it may be necessary to build unique translation tables to store the information from one system into the databases of the HIS. This requirement has limited the building of truly integrated patient records.

Acquisition of data from patient monitors used in the hospital can either be directly interfaced to the HIS or captured through an interface to an ICU system. Without these interfaces the acquisition of the monitoring data must be entered manually by the nursing personnel. It should be noted that whenever possible automated acquisition of data is preferable to manual entry. The automated acquisition is more accurate and reliable and less resource intensive. With those HISs which do not have interfaces to patient monitors, the frequency of data entry into the system is much less. The frequency of data acquisition affects the ability of the HIS to implement real-time medical decision logic to monitor the status of the patient. That is, in the ICU where decisions need to be made on a very timely manner, the information on which the decision is based must be entered as the critical event is taking place. If there is no automatic entry of the data, then the critical data needed for decision making may not be present, thus preventing the computer from assisting in the management of the patient.

40.3 Patient Admission, Transfer, and Discharge Functions

The admission application has three primary functions. The first is to capture for the patient's computer record pertinent demographic and financial/insurance information. A second function is to communicate that information to all systems existing on the hospital network. The third is to link the patient to previous encounters to ensure that the patient's longitudinal record is not compromised. This linkage also assists in capturing the demographic and financial data needed for the current encounter, since that information captured during a previous encounter may need not to be reentered as part of this admission. Unfortunately in many HISs the linkage process is not as accurate as needed. Several reasons explain this inaccuracy. The first is the motivation of the admitting personnel. In some hospitals they perceive their task as a business function responsible only for ensuring that the patient will be properly billed for his or her hospital stay. Therefore, since the admission program always allows them to create a new record and enter the necessary insurance/billing information, their effort to link the patient to his previous record may not be as exhaustive as needed.

Although the admitting program may interact with many financial and insurance files, there normally exists two key patient files that allow the HIS to meet its critical clinical functions. One is a master patient index (MPI) and the second is the longitudinal clinical file. The MPI contains the unique identifier for the patient. The other fields of this file are those necessary for the admitting clerk to identify the patient. During the admitting process the admitting clerk will enter identifying information such as name, sex, birth date, social security number. This information will be used by the program to select potential patient matches in the MPI from which the admitting clerk can link to the current admission. If no matches are detected by the program, the clerk creates a new record in the MPI. It is this process that all too frequently fails. That is, the clerk either enters erroneous data and finds no match or for some reason does not select as a match one of the records displayed. Occasionally the clerk selects the wrong match causing the data from this admission to be posted to the wrong patient. In the earlier HISs where no longitudinal record existed, this problem was not critical, but in today's system, errors in matching can have serious clinical consequences. Many techniques are being implemented to eliminate this problem including probabilistic matching, auditing processes, postadmission consolidation.

The longitudinal record may contain either a complete clinical record of the patient or only those variables that are most critical in subsequent admissions. Among the data that have been determined as most critical are key demographic data, allergies, surgical procedures, discharge diagnoses, and radiology reports. Beyond these key data elements more systems are beginning to store the complete clinical record. In those systems the structure of the records of the longitudinal file contain information regarding the encounter, admitting physician, and any other information that may be necessary to view the record from an encounter view or as a complete clinical history of the patient.

40.4 Patient Evaluation

The second major focus of application development for the HIS is creation of patient evaluation applications. The purpose of these evaluation programs is to provide to the care giver information about the patient which assists in evaluating the medical status of the patient. Depending on the level of data integration in the HIS, the evaluation applications will be either quite rudimentary or highly complex. In the simplest form these applications are departmentally oriented. With this departmental orientation the care giver can access through terminals in the hospital departmental reports. Thus, laboratory reports, radiology reports, pharmacy reports, nursing records, and the like can be displayed or printed at the hospital terminals. This form of evaluation functionality is commonly called results review, since it only allows the results of tests from the departments to be displayed with no attempt to integrate the data from those departments into an integrated patient evaluation report.

The more clinical HISs as mentioned above include a central integrated patient database. With those systems patient reports can be much more sophisticated. A simple example of an integrated patient evaluation report is a diabetic flowsheet. In this flowsheet the caregiver can view the time and amount of insulin given, which may have been recorded by the pharmacy or nursing application, the patient's blood glucose level recorded in the clinical laboratory or again by the nursing application. In this form the caregiver has within single report, correlated by the computer, the clinical information necessary to evaluate the patient's diabetic status rather than looking for data on reports from the laboratory system, the pharmacy system, and the nursing application. As the amount and type of data captured by the HIS increases, the system can produce ever-more-useful patient evaluation reports. There exist HISs which provide complete rounds reports the summarize on one to two screens all the patient's clinical record captured by the system. These reports not only shorten the time need by the caregiver to locate the information, but because of the format of the report, can present the data in a more intuitive and clinically useful form.

40.5 Patient Management

Once the caregiver has properly evaluated the state of the patient, the next task is to initiate therapy that ensures an optimal outcome for the patient. The sophistication of the management applications is again a key differentiation of HISs. At the simplest level management applications consist of order-entry applications. The order-entry application is normally executed by a paramedical personnel. That is, the physician writes the order in the patient's chart, and another person reviews from the chart the written order and enters it into the computer. For example, if the order is for a medication, then it will probably be a pharmacist who actually enters the order into the computer. For most of the other orders a nurse or ward clerk is normally assigned this task. The HIS records the order in the patient's computerized medical record and transmits the order to the appropriate department for execution. In those hospitals where the departmental systems are interfaced to the HIS, the electronic transmission of the order to the departmental system is a natural part of the order entry system. In many systems the transmission of the order is merely a printout of the order in the appropriate department.

The goal of most HISs is to have the physician responsible for management of the patient enter the orders into the computer. The problem that has troubled most of the HISs in achieving this goal has been the inefficiency of the current order-entry programs. For these programs to be successful they have to complete favorably with the traditional manner in which the physician writes the order. Unfortunately, most of the current order-entry applications are too cumbersome to be readily accepted by the physician. Generally they have been written to assist the paramedic in entering the order resulting with far too many screens or fields that need to be reviewed by the physician to complete the order. One approach that has been tried with limited success is the use of order sets. The order sets have been designed to allow the physician to easily from a single screen enter multiple orders. The use of order sets has improved the acceptability of the order-entry application to the physician, but several problems remain preventing universal acceptance by the physicians. One problem is that the order set will never be sufficiently complete to contain all orders that the physician would want to order. Therefore, there is some subset of patients orders that will have to be entered using the general ordering mechanisms of the program. Depending on the frequency of those orders, the acceptability of the program changes. Maintenance issues also arise with order sets, since it may be necessary to formulate order sets for each of the active physicians. Maintaining of the physician-specific order sets soon becomes a major problem for the data processing department.

It becomes more problematic if the HIS to increase the frequency of a given order being present on an order set allows the order sets to be not only physician-defined but problem-oriented as well. Here it is necessary to again increase the number of order sets or have the physicians all agree on those orders to be included in an order set for a given problem.

Another problem, which makes use of order entry by the physician difficult, is the lack of integration of the application into the intellectual tasks of the physician. That is, in most of the systems the physicians are asked to do all the intellectual work in evaluating and managing the care of the patient in the traditional manner and then, as an added task, enter the results of that intellectual effort into the computer. It is at this last step that is perceived by the physician as a clerical task at which the physician rebels. Newer systems are beginning to incorporate more efficiently the ordering task into other applications. These applications assist the physical throughout the entire intellectual effort of patient evaluation and management of the patient. An example of such integration would be the building of evaluation and order sets in the problem list management application. Here when the care provider looks at the patient problem list he or she accesses problem-specific evaluation and ordering screens built into the application, perhaps shortening the time necessary for the physician to make rounds on the patient.

Beyond simple test ordering, many newer HISs are implementing decision support packages. With these packages the system can incorporate medical knowledge usually as rule sets to assist the care provider in the management of patients. Execution of the rule sets can be performed in the foreground through direct calls from an executing application or in the background with the storing of clinical data in the patient's computerized medical record. This latter mode is called data-driven execution and provides an extremely powerful method of knowledge execution and alerting. that is, after execution of the rule sets, the HIS will "alert" the care provider of any outstanding information that may be important regarding the status of the patient or suggestions on the management of the patient. Several mechanisms have been implemented to direct the alerts to the care provider. In the simplest form notification is merely a process of storing the alert in the patient's medical record to be reviewed the next time the care provider accesses that patient's record. More sophisticated notification methods have included directed printouts to individuals whose job it is to monitor the alerts, electronic messages sent directly to terminals notifying the users that there are alerts which need to be viewed, and interfacing to the paging system of the hospital to direct alert pages to the appropriate personnel.

Execution of the rule sets are sometimes, time-driven. This mode results in sets of rules being executed at a particular point in time. The typical scenario for time-driven execution is to set a time of day for selected rule set execution. At that time each day the system executes the given set of rules for a selected population in the hospital. Time drive has proven to be a particularly useful mechanism of decision support for those applications that require hospitalwide patient monitoring.

The use of decision support has ranged from simple laboratory alerts to complex patient protocols. The responsibility of the HIS is to provide the tools for creation and execution of the knowledge base. The hospitals and their designated "experts" are responsible for the actual logic that is entered into the rule sets. Many studies are appearing in the literature suggesting that the addition of knowledge base execution to the HIS is the next major advancement to be delivered with the HIS. This addition will become a tool to better manage the hospital in the world of managed care.

The inclusion of decision support functionality in the HIS requires that the HIS be designed to support a set of knowledge tools. In general a knowledge bases system will consist of a knowledge base and an inference engine. The knowledge base will contain the rules, frames, and statistics that are used by the inference applications to substantiate a decision. We have found that in the health care area the knowledge base should be sufficiently flexible to support multiple forms of knowledge. That is, no single knowledge representation sufficiently powerful to provide a method to cover all decisions necessary in the hospital setting. For example, some diagnostic decisions may well be best suited for bayesian methods, whereas other management decisions may follow simple rules. In the context of the HIS, I prefer the term application manager to inference engine. The former is intended to imply that different applications may require different knowledge representations as well as different inferencing strategies to traverse the knowledge base. Thus, when the user selects the application, he or she is selecting a particular inference engine that may be unique to that application. The tasks, therefore, of the application manager are to provide the "look and feel" of the application, control the functional capabilities of the application, and invoke the appropriate inference engine for support of any "artificial intelligence" functionality.

40.6 Conclusion

Today's HIS is no longer the financial/administrative system that first appeared in the hospital. It has extended beyond that role to become an adjunct to the care of the patient. With this extension into clinical care the HIS has not only added new functionality to its design but has enhanced its ability to serve the traditional administrative and financial needs of the hospital as well. The creation of these global applications which go well beyond those of the departmental/clinical systems is now making the HIS the patient-focused system. With this global information the administrators and clinical staff together can accurately access where there are inefficiencies in the operation of the hospital from the delivery of both the administrative and medical care. This knowledge allows changes in the operation of the hospital that will ensure that optimal care continues to be provided to the patient at the least cost to the hospital. These studies and operation changes will continue to grow as the use of an integrated database and implementation of medical knowledge bases become increasingly routine in the functionality of the HIS.

References

[1] Pryor T.A., Gardner R.M., Clayton P.D. et al. (1983) The HELP system. *J. Med. Syst.* 7: 213.

[2] Pryor T.A., Clayton P.D., Haug P.J. et al. (1987) Design of a knowledge driven HIS. *Proc. 11th SCAMC*, 60.

[3] Bakker A.R. (1984) The development of an integrated and co-operative hospital information system. *Med. Inf.* 9: 135.

[4] Barnett G.O. (1984) The application of computer-based medical record systems in ambulatory practice. *N. Engl. J. Med.* 310: 1643.

[5] Bleich H.L., Beckley R.F., Horowitz G.L. et al. (1985) Clinical computing in a teaching hospital. *N. Engl. J. Med* 312: 756.

[6] Whiting-O'Keefe Q.E., Whiting A., and Henke J. (1988) The STOR clinical information system. *MD Comput.* 5: 8.

[7] Hendrickson G., Anderson R.K., Clayton P.D. et al. (1992). The integrated academic information system at Columbia-Presbyterian Medical Center. *MD Comput.* 9: 35.

[8] Safran C., Slack W.V., and Bleich H.L. (1989) Role of computing in patient care in two hospitals. *MD Comput.* 6: 141.

[9] Bleich H.L., Safran C., and Slack W.V. (1989) Departmental and laboratory computing in two hospitals. *MD Comput.* 6: 149.

[10] ASTM E1238-91 (1992) Specifications for transferring clinical observations between independent computer systems. Philadelphia, American Society for Testing and Materials.

[11] Tierney W.M., Miller M.E., and Donald C.J. (1990) The effect on test ordering of informing physicians of the charges for outpatient diagnostic tests. *N. Engl. J. Med.* 322: 1499.

[12] Stead W.W. and Hammond W.E. (1983) Functions required to allow TMR to support the information requirements of a hospital. *Proc. 7th SCAMC*, 106.

[13] Safran C., Herrmann F., Rind D. et al. (1990) Computer-based support for clinical decision making. *MD Comput.* 7: 319.

[14] Tate K.E., Gardner R.M., and Pryor T.A. (1989) Development of a computerized laboratory alerting system. *Comp. Biomed. Res.* 22: 575.

[15] Orthner H.F. and Blum B.I. (Eds.) (1989) *Implementing Health Care Information Systems*, Springer-Verlag.

[16] Dick R.S. and Steen E.B. (Eds.) (1991) *The Computer-Based Patient Record*, National Academy Press.

41

Computer-Based Patient Records

J. Michael Fitzmaurice
Agency for Healthcare Research and Quality

The objective of this section is to present the computer-based patient record (CPR) as a powerful tool for organizing patient care data to improve patient care and strengthen communication of patient care data among healthcare providers. The CPR is even more powerful when used in a system that retrieves applicable medical knowledge to support clinical decision making, improving patient safety, and promoting quality improvement. Evidence exists that the use of CPR systems (CPRS) can change both physician behavior and patient outcomes of care. As the speed and cost efficiency of computers rise, the cost of information storage and retrieval falls, and the breadth of ubiquitous networks becomes broader, it is essential that CPRs and systems that use them be evaluated for the improvements in health care that they can bring, and for their protection of the confidentiality of individually identifiable patient information.

Any opinions expressed in this chapter are those solely of the author and not of the U.S. Department of Health and Human Services or of the Agency for Healthcare Research and Quality.

The primary role of the CPR is to support the delivery of medical care to a particular patient. Serving this purpose, ideally the CPR brings past and current information about a particular patient to the physician, promotes communication among healthcare givers about that patient's care, and documents the process of care and the reasoning behind the choices that are made. Thus, the data in a CPR should be acquired as part of the normal process of healthcare delivery, by the providers of care and their institutions to improve data accuracy and timeliness of decision support. And these data should be shared for the benefit of the patient's care, perhaps with the permission or direction of the patient to safeguard confidentiality.

The CPR can also be an instrument for building a clinical data repository that is useful for collecting information about which medical treatments are effective in the practice of medicine in the community and for improving population-based health care. A clinical data repository may be provider-based or patient-based; it may be disease specific and geographically specific. Additional applications of CPR data beyond direct patient care can improve population-based care. These applications bring personal and public benefits, but also raise issues that must be addressed by healthcare policy makers.

Because patient information is likely to be located in the medical records of several of the patient's providers, providing high quality of care often requires exchanges of this information among providers. The vision of health information technology applications to improving the quality of health care contains a role for a national health information infrastructure. This infrastructure could take many forms but the most likely form is a combination of local or regional networks through which the required exchanges of patient information could take place. Currently most of these exchanges are done using faxed messages or phone calls. Sometimes the patient is just given the information to carry to the next provider. The CPR can be an even more powerful tool when it is connected to an electronic network and interoperable with other CPRs. Like the use of CPR applications, the use of health information networks also brings issues to be addressed by healthcare policy makers.

Clinical data standards, personal health identification, and communication networks, all critical factors for using CPRs effectively, are also addressed separately in other sections of this book.

41.1 Computer-Based Patient Record

A CPR is a collection of data about a patient's health care in electronic form. The CPR, also called an electronic health record (EHR) is part of a system (a CPRS, usually maintained in a hospital, physician's office, or an Internet or application service provider if it is web-based) that encompasses data entry and presentation, storage, and access to the clinical decision maker — usually a physician or nurse. The data are entered by keyboard, dictation and transcription, voice recognition and interpretation, light pen, touch screen, hand-held computerized notepad or a hand-held personal digital assistant (perhaps wireless) with gesture, and character recognition and grouping capabilities. Entry may also be by other means, for example, by direct instrumentation from electronic patient monitors and bedside terminals, nursing stations, bar code readers, radio-frequency identification (RFID), analyses by other linked computer systems such as laboratory autoanalyzers, and intensive care unit monitors, or even another provider's CPRS via a secure network. While the CPR could include patient-entered data, some medical providers may question the validity of such information for making their decisions; others may rely on such data for diagnosis and treatment.

Patient care data collected by a CPRS may be stored centrally or they may be stored in many places (e.g., distributed among the patient's providers) for retrieval at the request of an authorized user (most likely with the patient's authorization) through a database management system. The CPR may present data to the physician as text, tables, graphs, sound, images, full-motion video, and signals on an electronic screen, cell phones, pagers, or even paper. The CPR may also point to the location of additional patient data that cannot be easily incorporated into the CPR.

In too many current clinical settings (hospitals, physicians' offices, and ambulatory care centers), data pertaining to a patient's medical care are recorded and stored in a paper medical record. If the paper record is out of its normal location, or accompanying the patient during a procedure or an off-site study, it is

TABLE 41.1 Core Functions of an Electronic Health Record System

Health information and data
Results management
Order entry/management
Decision support
Electronic communication and connectivity
Patient support
Administrative processes
Reporting and population health management

Source: Institute of Medicine [2003]. *Patient Safety: Achieving a New Standard for Care.* Aspden P., Corrigan J.M., Wolcott J., and Erickson S.M. (eds). National Academic Press, Washington, D.C.

not available to the nurse, the attending physician, or the expert consultant. In paper form, data entries are often illegible and not easily retrieved and read by multiple users one at a time. On the other hand, an electronic form provides legible, clinical information which can be available to all users simultaneously, thus improving timely access to patient care data and communication among care providers.

Individual hospital departments (e.g., laboratory or pharmacy) often lose the advantages of automated data when their own computer systems print the computerized results onto paper. The pages are then sent to the patient's hospital floor and assembled into a paper record. The lack of standards for the electronic exchange of this data and the lack of implementation of existing standards, such as using the Logical Observation Identifiers, Names and Codes (LOINC) standard for reporting laboratory results, hinders the integration of computerized departmental systems. Searching electronic files is often more efficient than searching through paper. Weaknesses of paper medical record systems for supporting patient care and health care providers have long been known [Korpman, 1990, 1991].

Many of the functions of a CPR and how it operates within a healthcare information system to satisfy user demands are explained in the Institute of Medicine's (IOM) report, The Computer-Based Patient Record: An Essential Technology for Health Care [1991, 1997]. In response to a request by the Agency for Healthcare Research and Quality, IOM provided guidance to DHHS in 2003 on a set of "basic functionalities" that an electronic health record system should possess to promote patient safety [IOM, 2003], shown in Table 41.1.

This guidance is the basis for a new Health Level Seven (HL7, a standard developing organization) standard that specifies the functions of an EHR that will be useful for EHR purchasers to specify what functions they want and for vendors of EHR systems to describe the functions they offer [HL7, 2004].

41.2 Clinical Decision Support Systems

One of the roles of the CPR is to enable a clinical decision support system (CDSS) — computer software designed to aid clinical decision making — to provide the physician with medical knowledge that is pertinent to the care of the patient. Diagnostic suggestions, testing prompts, therapeutic protocols, practice guidelines, alerts of potential drug–drug and drug–food reactions, evidence-based treatment suggestions, and other decision support services can be obtained through the interaction of the CPR with a CDSS.

41.2.1 Knowledge Server

Existing knowledge about potential diagnoses and treatments, practice guidelines, and complicating factors pertinent to the patient's diagnosis and care is needed at the time treatment decisions are made. The go-between that makes this link is a "knowledge server," which acquires the necessary information for the decision maker from the knowledge server's information sources. The knowledge server can assist the clinical decision maker to put this information, that is, specific data and information about the patient's identification and condition(s) and medical knowledge, into the proper context for treating the patient [Tuttle et al., 1994].

41.2.2 Knowledge Sources

Knowledge sources include a range of options, from internal development and approval by a hospital's staff, for example, to sources outside the hospital, such as the National Guidelines Clearinghouse, see www.guideline.gov, initiated by the Agency for Healthcare Research and Quality (AHRQ), the American Medical Association, and the Association of American Health Plans; the Physicians Data Query program at the National Cancer Institute at http://www.nci.nih.gov/cancertopics/pdq; other consensus panel guidelines sponsored by the National Institutes of Health; guidelines developed by medical and other specialty societies and others; and specialized information from private-sector knowledge vendors. Additional sources of knowledge include the medical literature, which can be searched for high quality, comprehensive review articles, and for particular subjects using the "PubMed" program to explore the MEDLINE literature database available through the National Library of Medicine at http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed&itool=toolbar. AHRQ-supported Evidence-based Practice Center reports summarizing scientific evidence on specific topics of medical interest are available to support guideline development, see http://www.guideline.gov/resources/epc_reports.aspx.

41.2.3 Medical Logic Modules

If medical knowledge needs are anticipated, acquired beforehand, and put into a medical logic module (MLM), software can provide rule-based alerts, reminders and suggestions for the care provider at the point (time and place) of health service delivery. One format for MLMs is the Arden Syntax, which standardizes the logical statements [ASTM, 1992]. For example, an MLM might be interpreted as, "If sex is female, and age is greater than 50 years, and no Pap smear test result appears in the CPR, then recommend a Pap smear test to the patient." If MLMs are to have a positive impact on physician behavior and the patient-care process, then physicians using MLMs must agree on the rules in the logical statements or conditions and recommended actions that are based on interactions with patient-care data in the CPR. Another format is GLIF (the Guideline Interchange Format), a computer-interpretable language framework for modeling and executing clinical practice guidelines. GLIF uses GELLO, a guideline expression language that is better suited for GLIF's object-oriented data model, is extensible, and allows implementation of expressions that are not supported by the Arden Syntax [Wang 2004; also see http://www.openclinical.org/gmm_glif.html].

Because MLMs are usually independent, the presence or absence of one MLM does not affect the operation of other MLMs in the system. If done carefully and well, MLMs developed in one healthcare organization can be incorporated in the CPRSs of other healthcare organizations. However, this requires much more than using accepted medical content and logical structure. If the medical concept terminology (the nomenclature and code sets used by physicians and by the CPR) differs among organizations, the knowledge server may misinterpret what is in the CPR, apply logic to the wrong concept, or select the wrong MLM. Further, the physician receiving its message may misinterpret the MLM [Pryor and Hripcsak, 1994].

41.2.4 Nomenclature

For widespread use of CDSSs, a uniform medical nomenclature, consistent with the scientific literature is necessary. Medical knowledge is information that has been evaluated by experts and converted into useful medical concepts, options, and rules for decisions. For CDSSs to search through a patient's CPR, identify the medical concepts, retrieve appropriate patient data and information, and provide a link to the relevant knowledge, the CDSS has to recognize the names used in the CPR for the concepts [Cimino, 1993]. Providing direction for coupling terms and codes found in patient records to medical knowledge is the goal of the Unified Medical Language System (UMLS) project of the National Library of Medicine [NLM, 2005a]. “The UMLS Metathesaurus supplies information that computer programs can use to create standard data, interpret user inquiries, interact with users to refine their questions, and convert the users’ terms into the vocabulary used in relevant information sources” [IOM, 2005b].

The developers of medical informatics applications need a controlled medical terminology so that their applications will work across various sites of care and medical decision making. The desiderata, or requirements, of a controlled medical terminology as described by Cimino [1998] include: “vocabulary content, concept orientation, concept permanence, nonsemantic concept identifiers, polyhierarchy, formal definitions, rejection of ‘not elsewhere classified’ terms, multiple granularities, multiple consistent views, context representation, graceful evolution, and recognized redundancy.”

The National Committee on Vital and Health Statistics (NCVHS) is an 18-private-sector member, federal advisory committee with a 50-year history of advising the Secretary of Health and Human Services (HHS) on issues relating to health data, statistics, privacy, and national health information policy [http://aspe.dhhs.gov/ncvhs/]. Recognizing that “[w]ithout national standard vocabularies, precise clinical data collection and accurate interpretation of such data is difficult to achieve,” NCVHS recommended in 2000 that the Secretary of HHS “should provide immediate funding to accelerate the development and promote early adoption of PMRI standards.” This recommendation included clinical terminology activities of the National Library of Medicine, the Agency for Healthcare Research and Quality, and the Food and Drug administration to augment, develop and test clinical vocabularies, and to make them available publicly at low cost [NCVHS, PMRI, 2000]. The Consolidated Health Informatics work group, a White House, Office of Management and Budget, interagency, eGovernment Initiative dominated by HHS, Department of Veterans Affairs (VA), and Department of Defense (DoD) staff, made similar recommendations over the next 3 years. This encouragement led HHS to negotiate a national license for free use of SNOMED-CT (Systematized Nomenclature for Medicine — Clinical Terms), adopt 20 clinical data standards, and begin developing drug terminology and structured drug information, mapping vocabularies to SNOMED-CT, and investigating the standardization of data elements used for reporting patient safety adverse events.

41.3 Scientific Evidence

41.3.1 Patient Care Processes

Controlled trials have shown the effectiveness of CDSS for modifying physician behavior using preventive care reminders. In an early review of the scientific literature up, Johnston et al. [1994] reported that controlled trials of CDSSs have shown significant, favorable effects on care processes from (1) providing preventive care information to physicians and patients [McDonald et al., 1984; Tierney et al., 1986], (2) supporting diagnosis of high-risk patients [Chase et al., 1983], (3) determining the toxic drug dose for obtaining the desired therapeutic levels [White et al., 1987], and (4) aiding active medical care decisions [Tierney, 1988]. Johnston found clinician performance was generally improved when a CDSS was used and, in a small number of cases (3 of 10 trials), significant improvements in patient outcomes.

In a randomized, controlled clinical trial, one that randomly assigned some teams of physicians to computer workstations with screens designed to promote cost-effective ordering (e.g., of drugs and laboratory tests), Tierney et al. [1993] reported patient lengths of stay were 0.89 days shorter, and charges generated by the intervention teams were $887 lower, than for the control teams of physicians. These gains were not without an offset. Time and motion studies showed that intervention physician teams spent 5.5 min longer per patient during 10-h observation periods. This study is a rare controlled trial that sheds light on the resource impact and the effectiveness of using a CDSS.

In this setting, physician behavior was changed and resources were reduced by the application of logical algorithms to computer-based patient record information. Nevertheless, a different hospital striving to attain the same results would have to factor in the cost of the equipment, installation, maintenance, and software development plus the need to provide staff training in the use of a CDSS.

Additional evidence, rigorously obtained, shows beneficial effects of the use of EHRs within a CDSS on medical practices. As reported in the response of the Department of Health and Human Services to the GAO report, *Health and Human Services' Estimate of Health Care Cost Savings Resulting from the Use of Information Technology*, many studies published in peer-reviewed journals show "substantial improvement in clinical processes" when physicians use EHRs:

> The effects of EHRs include reducing laboratory and radiology test ordering by 9 to 14% [Bates, 1999; Tierney, 1990; Tierney, 1987], lowering ancillary test charges by up to 8% [Tierney, 1988], reducing hospital admissions, costing an average of $17,000 each, by 2 to 3% [Jha, 2001], and reducing excess medication usage by 11% [Wang, 2003; Teich, 2000] [GAO, 2005].

41.3.2 Incentives

As can be seen from the literature, CDSS can improve quality and in many cases reduce resources needed for treatment. The benefit of this resource reduction, however, goes most frequently to health plans under cost-based reimbursement of providers. The provider of care may need additional incentives to adopt CDSS as part of the regular work process. Otherwise, any added time needed to access and respond to the CDSS prompts and alerts may reduce the provider's personal productivity (see the extra 5.5 min per patient noted earlier) without offsetting compensation. These additional incentives may be funds for purchase of the hardware and software needed for CDSS, payment for using CDSS, or payment for reporting evidence of improved quality of care or cost reduction.

41.3.3 Evaluation

Clinical decision support systems should be evaluated according to how well they enhance performance in the user's environment [Nykanen et al., 1992]. If CDSS use is to become widespread and supported, society should judge CDSSs not only on enhanced provider performance, but also on whether patient outcomes are improved and system-wide healthcare costs are contained. Evaluation of information systems is extremely difficult because so many changes occur when they are introduced into an existing work flow. Attributing changes in productivity, costs of care, and patient outcomes to the introduction and use of a CDSS or a CPRS is difficult when the work patterns and the culture of the workplace are also changing. Many clinical information system applications have been self-evaluated in their original development site. While impressive findings have been published, the generalizability of those findings needs verification.

41.3.4 CDSS Hurdles

In a review of medical diagnostic decision support systems, Miller [1994] examines the development of CDSSs over the past 40 years, and identifies several hurdles to be overcome before large-scale, generic CDSSs grow to widespread use. These hurdles include determining (1) how to support medical knowledge base construction and maintenance over time, (2) the amount of reasoning power and detailed representation of medical knowledge required (e.g., how strong a match of medical terms is needed to join medical concepts with appropriate information), (3) how to integrate CDSSs into the clinical environment to reduce the costs of patient data capture, and (4) how to provide flexible user interfaces and environments that adjust to the abilities and desires of the user (e.g., with regard to typing expertise and pointing devices).

41.3.5 Research Databases

Computer-based patient records can have great value for developing research databases, medical knowledge, and quality assurance information that would otherwise require an inordinate amount of manual resources to obtain in their absence. An example of CPR use in research is found in a study undertaken at Latter Day Saints (LDS) Hospital. Using the HELP CPR system to gather data on 2847 surgical patients, this study found that administering antibiotics prophylactically during the 2-h window before surgery (as opposed to earlier or later within a 48-h window) minimized the chance of surgical-wound infection. It also reduced the surgical infection rate for this time category to 0.59%, compared to the 1.5% overall infection rate for all the surgical patients under study [Classen et al., 1992].

The same system was used at LDS Hospital to link the clinical information system data (including a measure of nursing acuity) with the financial systems' data. Using clinical data to adjust for the severity of patient illness, Evans et al. [1994] measured the effect of adverse drug events due to hospital drug administration on hospital length of stay and cost. The difference attributable to adverse drug events among similar patients was estimated to be an extra 1.94 patient days and $1939 in costs.

41.3.6 Telemedicine

A CPR may hold and exchange radiological and pathological images of the patient taken or scanned in digital form. The advantage is that digital images may be transferred long distances without a reduction in quality of appearance. This allows patients to receive proficient medical advice even when they and their local family practitioners are far from the consulting physicians. It also allows health managers to move such clinical work to take advantage of excess radiology and pathology capacity elsewhere in the system. Further, joint telemedicine consults in real time can also add to the ability of local physicians to become better at diagnosing and treating some conditions (such as those requiring expertise in dermatology) by learning from the long-distance specialist as he or she treats their patients. Nevertheless, while telemedicine has been shown to work in actual practice, the scientific literature does not present definitive findings of cost effectiveness or efficacy [Hersch, 2001a,b].

When personally identifiable healthcare data are transported electronically across state borders for telemedicine uses, the applicability of state laws and policies regarding the confidentiality and privacy of this data is not often obvious to the sender or receiver. This uncertainty raises legal questions for organizations that wish to move this data over national networks for patient management, business, or analytical reasons. When vendors of CPRS plan nationwide distribution of their products, they must consider among other things variation in state laws regarding the validity of electronic information for use in official medical records, the length of time for retention of medical record information, and liability for the consequences of EHR failure. The users of systems that exchange patient information across state borders for treatment must also consider the appropriate state licensing requirements.

41.4 Federal Initiatives for Health Information System Interoperability and Connectivity

For years, the Department of Veterans Affairs (VISTA — Veterans Health Information Systems and Technology Architecture) and the Department of Defense (CHCS-II — Composite Health Care System) have invested in the development of CPRS to improve care for veterans and active servicemen. The Indian Health Service has adopted and modified the VA's VISTA system for its own CPRS use. HHS has a history of undertaking national terminology development [Humphreys et al., 1998] and medical informatics research [Fitzmaurice et al., 2002]. In 2004, however, the federal government began to take the initiative to lay the foundation for improving the coordination of these efforts and promoting the interoperability and connectivity of health information systems across the country.

During the 2004–2005 period, the President placed a greater federal emphasis on using health information technology to improve patient safety and the quality of health care. President George Bush in his 2004 State of the Union message said, "By computerizing health records, we can avoid dangerous medical mistakes, reduce costs, and improve care." And on April 26, 2004, "Within ten years, every American must have a personal electronic medical record. That's a good goal for the country to achieve" [Bush, 2004].

As recommended by NCVHS [2001] and others, the President created the Office of the National Coordinator for Health Information Technology [Bush, 2004], and on May 6, 2004, the Secretary of

Health and Human Services (1) appointed David Brailer, M.D, Ph.D., as the first National Coordinator, and (2) announced that the medical vocabulary known as SNOMED CT (Systematized Nomenclature for Medicine — Clinical Terms, a clinical reference language standard created by the College of American Pathologists) could be downloaded free for use in the United States through HHS' National Library of Medicine [US DHHS, 2004]. By July 21, 2004, the National Coordinator produced a strategic framework to guide the nationwide implementation of health information technology in both the public and private sectors. This plan has four major goals to be pursued for the vision of improved health care. They are to build an interoperable national health information system that will:

1. Inform clinical practice
2. Interconnect clinicians
3. Personalize care
4. Improve population health [ONCHIT, 2004]

Through its Transforming Healthcare Quality Through Health Information Technology Program, the AHRQ in September 2004 awarded 100 grants ($139 Million over 3 years), 5 state demonstration contracts ($25 million over 5 years), a National Health Information Technology Resource Center contract ($18.4 million over 5 years), and initiated a data standards program ($10 million in 1 year). AHRQ's research program embodies the vision of the federal strategic framework for promoting and invests federal research funds to build a knowledge base of how regional and local information technology networks and applications can improve quality of care and patient safety [AHRQ, 2004].

Recognizing the benefits of CPRS and the exchange of clinical information, the President in his State of the Union message to the American people on February 3, 2005, called for additional investment, saying, "I ask Congress to move forward on … improved information technology to prevent medical error and needless costs …" The federal government has begun to devote resources to support its vision of national networks for the exchange of clinical information to benefit patient care. The vision is one of interoperable clinical applications of health information technology applications over a national set of regional networks and of connectivity to all health providers to these networks.

Health information systems are beginning to rely on intranet networks within the health enterprise to link the information created by disparate applications currently in use (e.g., to link existing [legacy] applications such as laboratory, radiology, and pharmacy information systems), and on private networks to exchange patient information among health providers. These systems use web browsers, object-oriented technology, and document formatting languages, including hypertext markup language (HTML) and extensible markup language (XML). Indeed, the structure of HL7's Version 3 of its suite of clinical message standards for health institutions' electronic clinical messages employs this technology [HL7, 2001], as does ASTM's proposed Continuity of Care Record standard [ASTM, 2005].

Currently, the health industry does not have acceptable standards for encrypting clinical message exchanges and for electronic signatures that are in widespread use. Although the confidentiality of subjects of personal health information is considered sufficiently protected and the authentication of the sender and receiver sufficiently assured for those providers who currently exchange clinical information through fax and telephone, pilot tests of electronic prescribing conducted by the Medicare Program should provide additional information on ways to improve protection and authentication for clinical exchanges through electronic networks. These 2006 pilots are mandated by the Medicare Prescription Drug, Improvement, and Modernization Act of 2003 (MMA) [Public Law, 2003].

41.4.1 PITAC

The President's Information Technology Advisory Committee (PITAC) is a private-sector member committee chartered originally in 1998 to provide the president with independent expert advice on maintaining America's preeminence in advanced information technology. In its 2004 report, *Revolutionizing Health Care Through Information Technology*, PITAC recommended Federal leadership in developing a national framework containing four essential elements. These elements are: electronic health records, computer-assisted clinical decision support, computerized provider order entry, and “secure, private, interoperable, electronic health information exchange, including both highly specific standards for capturing new data and tools for capturing non-standards-compliant electronic information from legacy systems” [PITAC, 2004]. In 2005, the President folded PITAC into the President’s Council of Advisors on Science and Technology (PCAST).

41.5 Private Sector Initiatives

The Markle Foundation with support from The Robert Wood Johnson Foundation under its Connecting for Health program and in collaboration with the eHealth Initiative organizes working groups representing the public and private sectors to tackle the barriers to the development of an interconnected health information infrastructure. These working groups have produced papers On Linking Health Care Information (February 2005), Achieving Electronic Connectivity in Health Care (July 2004), Connecting Americans to Their Healthcare (July 2004), and Financial, Legal and Organizational Approaches to Achieving Electronic Connectivity in Healthcare (October 2004) and Connecting Healthcare in the Information Age (June 5, 2003). The Markle Foundation [Markle Foundation, 2005] convenes recognized experts and health sector stakeholders to reach consensus on how specific barriers should be tackled and preparing roadmaps for action.

The Healthcare Information and Management Systems Society (HIMSS) is a membership organization that focuses on providing leadership for the optimal use of healthcare information technology and management systems to better human health. One of its projects, Integrating the Healthcare Enterprise (IHE), is a multi-year initiative that has as its goal to create “the framework for passing vital health information seamlessly — from application to application, system to system, and setting to setting — across the entire healthcare enterprise.” HIMSS, the Radiological Society of North America (RSNA), and the American College of Cardiology (ACC) work collaboratively with the aim “to improve the way computer systems in healthcare share critical information.” In 2005, at the HIMSS Annual Conference, the IHE Connect-a-thon and Interoperability Showcase demonstrated the communication of documents containing patient care information found in ASTM’s. Continuity of Care Record standard across the products of 32 health information system and application vendors using existing health data standards [HIMSS, IHE, 2005, http://www.himss.org/ASP/topics_ihe.asp.]. Public demonstrations of the applications of health data standards are invaluable for learning what works in the electronic exchange of patient care data and how it works. Essentially, what is learned is how to make health data standards work for specific health care applications and how to make them better.

41.6 Driving Forces for CPRS

41.6.1 Patient Safety

Patient safety is a real concern in the U.S. healthcare system but is not well understood. The publication of the IOM study *To Err is Human* in 1999, informed the American public that between 44,000 and 98,000 people died of medical errors in hospitals [IOM, 1999]. In 2003, Zhan and Miller estimated that complications of often preventable injuries and complications in hospitals in the United States lead to more than 32,000 deaths, 2.4 million extra days of care, and costs exceeding $9B annually [Zahn and Miller, 2003]. Among the conditions studied were accidental puncture and laceration, anesthesia complications, postoperative infections and bedsores, surgical wounds reopening, and obstetric traumas during childbirth. Health information technology is clearly part of the remedy. In a study of 36 hospitals, Barker et al., found that 19% of the doses were in error and that “the percentage of [drug] errors rated potentially harmful was 7%, or more than 40 per day in a typical 300-patient facility” [Barker, 2003]. Bates et al., found that the rate of serious medication errors dropped by more than half after a large tertiary teaching hospital implemented a computerized physician order entry system [Bates et al., 1998]. IOM recommends that "[t]o reduce the number of medical errors, the nation's health care system must harness available technologies and build an infrastructure for national health information" [IOM, 2003b Patient Safety: Achieving a New Standard for Care, IOM, 2003b].

IOM, which is a foremost advisor to the nation in evaluating scientific evidence and obtaining professional opinion pertaining to patient safety and quality of care, recommends that: "[t]o reduce the number of medical errors, the nation's healthcare system must harness available technologies and build an infrastructure for national health information." More specifically, IOM recommends a seamless national network that requires EHRs, secure platforms for exchange of info among providers and patients, and data standards that would make health information understandable by the information systems of different providers. Further, healthcare organizations must adopt information technology systems that are able to collect and share essential health information on patients and their care [IOM, 2003b].

41.6.2 Quality of Care

In *Crossing the Quality Chasm: A New Health System for the 21st Century*, [IOM, 2001], IOM noted that a chasm exists between current practice and the best we can do and urged the United States (1) to adopt six attributes of quality care: safe, effective, patient-centered, timely, efficient, and equitable and (2) to use information technology to improve the quality of care.

Quality of care deficiencies are widespread in the United States and, judging from one study, evenly distributed across metropolitan areas [McGlynn et al., 2003], from a random sample of health care experiences of adults living in 12 metropolitan areas in the United States over a 2-year period found that study participants received 54.9% of recommended care. They evaluated health system performance on 439 indicators of quality of care for 30 acute and chronic conditions as well as preventive care, with the participants receiving 53.5, 56.1, and 54.9% of the recommended care, respectively for the three categories. In the 12 metropolitan areas studied, Seattle, Washington, received the recommended care 59% of the time (the highest), Little Rock, Arkansas, 51% (the lowest) [Kerr, 2004].

The contribution of CPRSs for improving quality of care is to deliver medical knowledge and appropriate patient information to the healthcare decision makers — especially the physician and patient — at the time such information is needed and to aggregate clinical entries to obtain quality measures more efficiently than by combing through paper medical records. Once in place, CPRS can also reduce the cost of obtaining standardized quality measures, compared with manual abstractions of paper medical records.

41.6.2.1 Rising Healthcare Spending

Healthcare expenditures in the United States totaled $1.7 trillion in 2003 ($5,670 per capita) and 15.3% of our gross domestic product. This increase was a 7.7% increase over 2002 expenditures and exceeded the rate of inflation of the Consumers Price Index (2.3%) threefold [Smith, 2005; and Bureau of Labor Statistics, 2005]. Two major concerns over rising health spending are matters of (1) obtaining value for the dollar spent on health care and (2) achieving productive competitiveness in the U.S. A study by Hussey et al., showed that the United States spends more per capita on health care than four other comparable countries: Australia, Canada, New Zealand, and England. However, the United States underperforms on such measures as: breast cancer deaths, leukemia deaths, asthma deaths, suicide rates, and cancer screening. The implication is that even though the United States spends more, outcomes are not necessarily better.

41.6.2.2 Competition in the U.S. Economy

Firms in the United States are concerned that many goods produced in the United States are more expensive than in other parts of the world in part because of the higher costs of health care in the United States, and the larger portion of healthcare costs they incur as part of their labor costs. Also, if the U.S. healthcare system itself is not as productive as are the healthcare systems of other countries, our workers will spend more time obtaining health care and recovering from illness, and less time working. As a result, U.S. companies are encouraging health plans to improve the quality of care provided to their employees and to contract with lower cost providers. The Leapfrog Group and other employer groups are promoting a better health system by encouraging employer purchasers to buy healthcare services from providers using:

- Computer physician order entry to permit computer-generated prompts, alerts, and reminders to inform treatment decisions.
- ICU physician staffing — the use of board-certified hospital intensivists, hospital-based physicians that would take over a patient's care in the hospital intensive care unit.
- Evidence-based hospital referral — particularly for high-risk surgery and high-risk neonatal intensive care [Birkmeyer, 2001].

41.7 Extended Uses of CPR Data

Data produced by such systems have additional value beyond supporting the care of specific patients. For example, subsets of individual patient care data from CPRs can be used for research purposes, quality assurance purposes, developing and assessing patient care treatment paths (planned sequences of medical services to be implemented after the diagnoses and treatment choices have been made), assessments of treatment strategies across a range of choices, and postmarketing surveillance of drugs and medical device technologies in use in the community after their approval by the Food and Drug Administration. When linked with data measuring patient outcomes, CPR data may be used to help model the results achieved by different treatments, sites of care, and organizations of care.

If patient care data were uniformly defined and recorded, accurately linked, and collected into databases pertaining to particular geographical areas, they would be useful for research into the patient outcomes of alternative medical treatments for specific conditions and for developing information to assist consumers, healthcare providers, health plans, payers, public health officials, and others in making choices about treatments, technologies, sites and providers of care, health plans, and community health needs. This is currently an ambitious vision for research considering the presently limited use of CPRs. There are insufficient incentives for validating, storing, and sharing electronic patient record data, plus improvements are needed that push forward the state of the art in measuring the severity of patient illness so that the outcomes of like patients can be compared. Many healthcare decisions are now based on data of inferior quality or no data at all. The importance of these decisions, however, to the healthcare market is driving higher the demand for uniform, accurate clinical data.

41.8 Federal Programs

Uniform, electronic clinical patient data could be useful to many Federal programs that have responsibility for improving, safeguarding, and financing America's health. For example, the AHRQ is charged "to enhance the quality, appropriateness, and effectiveness of health services, and access to such services, through the establishment of a broad base of scientific research and through the promotion of improvements in clinical and health system practices, including the prevention of diseases and other health conditions" [PL, 1999].

The findings that result from such research should improve patient outcomes of care, quality measurement, and cost and access problems. To examine the influence on patient outcomes of alternative treatments for specific conditions, research needs to account for the simultaneous effects of many patient risk factors, such as diabetes and hypertension. Health insurance claims for payment data do not have sufficient clinical detail for many research, quality assurance, and evaluation purposes. Often, administrative data (such as claims data) must be supplemented with data abstracted from the patients' medical records to be useful. In many cases, the data must be identified and collected prospectively from patients (with their permission) and their providers to ensure availability and uniformity. The use of a CPR could reduce the burden of this data collection, support practice guideline development in the private sector, and support the development, testing, and use of quality improvement measures. Having uniform, computerized patient care data in CPRs would allow disease registries to be developed for many more patient conditions.

Other federal, state, and local health agencies also could benefit from CPR-based data collections. For example, the Food and Drug Administration, which conducts postmarketing monitoring to learn the incidence of unwanted effects of drugs after they are approved, could benefit from analyses of the next 20,000 cases, in which a particular pharmaceutical is prescribed, using data collected in a CPR. Greater confidence in postmarket surveillance could speed approval of new drug applications. The Centers for Medicare and Medicaid Services (CMS) is providing guidance and information to its Quality Improvement Organizations (QIOs) about local and nationwide medical practice patterns founded on analyses of national and regional clinical data about Medicare beneficiaries. Medicare QIOs could analyze more data from provider's CPRs in their own states to provide constructive, quality-enhancing feedback providers of care at less expense. As a further example, the Centers for Disease Control and Prevention with access to locally available (and perhaps anonymized) CPR data on patient care could more quickly and completely monitor the incidence and prevalence of communicable diseases, and engage in real-time surveillance for monitoring bioterrorism threats. State and local public health departments could allocate resources more quickly to address changing health needs with early recognition of community health problems.

Many of these uses require linked data networks and data repositories that communities and patients trust with their health data, or a filter of data flows that searches for events that would trigger a health alert. A national health information network could provide guidance, governance, and principles for the sharing of electronic patient information. Of paramount importance is the protection of the confidentiality of patient information. This may require an approach that gives patients a choice to opt in to such systems of sharing their information to obtain the benefits or to opt out of having the system use their own information.

41.9 Selected Issues

While there are personal and public benefits to be gained from extended use of CPR data beyond direct patient care, the use of personal medical information for these uses, particularly if it contains personal identification, brings with it some requirements and issues that must be faced. Some of the issues that must be addressed by health care policy makers, as well as by private markets, are as follows.

41.9.1 Standards

Standards are needed for the nomenclature, coding, and structure of clinical patient care data; the content of data sets for specific purposes; and the electronic transmission of such data to integrate data efficiently across departmental systems within a hospital and data from the systems of other hospitals and healthcare providers. If benefits are to be realized from rapidly accessing and transmitting patient care data for managing patient care, consulting with experts across long distances, linking physician offices and hospitals, undertaking research, and other applications, data standards are essential [Fitzmaurice, 1994].

The United States has the framework for coordination of U.S. standards developing organizations, development and coordination of the U.S. position on international standards issues, and representation at the technical committee that develops and approves international health data standards. The Healthcare Informatics Standards Board of the American National Standards Institute coordinates the standard developing organizations that work on such standards in the United States, and produces special summary reports on administrative and clinical health data standards [ANSI HISB, 1997, 1998]. The U.S. Technical Advisory Group to the Organization of International Standards (ISO) Technical Committee (TC) 215, Health Informatics, develops and represents U.S. positions on international health data standards issues, new work items, and recommends the U.S. vote on international standards ballots. The ISO TC 215, Health Informatics, was formed in 1998 by over 30 countries to provide a forum for international coordination of health informatics standards.

TABLE 41.2 HIPAA Administrative Simplification Standards

Transactions and Code Sets (TCS) Rule	October 16, 2002/2003
Claims Attachments	Proposed rule 2005
TCS Revisions	Expected 2005
Identifiers	
Employer ID	July 30, 2004
National Provider ID	May 23, 2007
Health Plan ID	Expected 2005
Individual ID	Put on hold by Congress
Security Rule	April 21, 2005
Privacy Rule	April 14, 2003

Note: Small health plans have an additional year before their use of HIPAA standards is mandatory.

Within the United States, administrative health data standards are mandated in the Health Insurance Portability and Accountability Act (HIPAA) of 1996 [Public Law, 1996]. In this law, the Secretary of Health and Human Services (HHS) is directed to adopt standards for nine common health transactions (enrollment, claims, payment, and others) that must be used if those transactions are conducted electronically. Penalties are capped at $100 per violation and a maximum of $25,000 per year for each provision violated. Digital signatures, when adopted by the Secretary, may be deemed to satisfy federal and state statutory requirements for written signatures for HIPAA transactions but there is no industry standard to date. The four categories of HIPAA standards with the dates on which specific standards are mandatory, or the year in which they are expected to be published, are shown in Table 41.2. Published standards are expected to be mandatory about 2 years and 60 days after they are published in final form.

41.9.2 Security

Confidentiality and privacy of individually identifiable patient care and provider data are the most important issues. For most purposes, the HIPAA Privacy Rule [U.S. Department of Health and Human Services, Office of Civil Rights, 2003] is quite stringent with respect to establishing a privacy floor across all states. It creates a fence around the individually identifiable health information it protects, that is, such information that is in the hands of health plans, clearinghouses, and providers who undertake HIPAA transactions (covered entities). Covered entities may use protected health information only for purposes of treatment, payment, or health operations. Without an individual's authorization, there are only 12 ways for a covered entity to legally disclose or use protected health information. Each of these exceptions has requirements of its own.

The HIPAA Privacy Rule is an essential cornerstone for building a national health information infrastructure that eases the way for personal health information to be shared. It gives patients new rights and controls nationwide, including the right to see, obtain a copy of, and add amendments to their health information. For uses and disclosures not permitted by the Privacy Rule, HIPAA-covered entities must obtain the individual's authorization. The penalties for violating the Privacy Rule can be expensive and include imprisonment.

System security and integrity become important as more and more information for patient treatment and other uses is exchanged through national networks. Not only does this issue relate to purposeful violations of privacy, but also to the accuracy of medical knowledge for patient benefit. If the system fails to transmit accurately what was sent to a physician — for example, an MRI, patient history, a practice guideline, or a clinical research finding — and if a physician's judgment and recommendation is based on a flawed image or other misreported medical knowledge — who bears the legal responsibility for a resulting inappropriate patient outcome due to system failure?

TABLE 41.3 Exceptions to the HIPAA Privacy Rule

As required by law
For public health
Victims of abuse
For health oversight activities
For judicial and administrative proceedings
For law enforcement
Disclosures about decedents (coroner, medical examiner)
To facilitate organ transplantation
For research
To avert serious threats to health or safety
For specialized government functions
For workers' compensation

Note: Individual authorization is not required for disclosures and uses of protected health information.

National HIPAA security standards for assuring the confidentiality of electronic protected health information are mandatory as of April 20, 2005. The HIPAA Security Rule addresses the administrative, technical, and physical security procedures that HIPAA-covered entity must use. Some procedure specifications are required; others must be addressed following a risk analysis by the HIPAA-covered entity [Health Insurance Reform, 2003]. This rule supports the Privacy Rule in that it establishes what security protections are reasonable to safeguard electronic health information from impermissible uses and disclosures.

41.9.3 Data Quality

The quality of stored and exchanged clinical data may be questioned in the absence of organized programs and criteria to assess the reliability, validity, and sufficiency of this data. There should be a natural reluctance to use questionable data for making treatment decisions, undertaking research, and for providing useful information to consumers, medical care organizers, and payers. For proper use and analysis, the user should take special care in judging that the information is of sufficient quality to measure and assess the relevant risk factors influencing patient conditions and outcomes. Providers of care may have reluctance even in relying on data supplied by their own patients without some assurance that it is valid.

41.9.4 Varying State Requirements

Electronically stored records in one state may be considered to be legally the same as paper records, but not in another state. In law, regulation, and practice, many states require pen and ink recording and signatures, apparently ruling out electronic records and signatures. To reduce this inconsistency and uncertainty and to provide national guidance, the Electronic Signatures in Global and National Commerce Act was enacted by the U.S. government effective on October 1, 2000. This law gives electronic signatures the same legality as hand-written ones where all parties agree for transactions that are commercial, consumer, or business in nature [Public Law, 2000]. To add to the variability, State privacy laws that (1) conflict with the HIPAA Privacy Rule and (2) are more stringent override the federal Privacy Rule:

Standard unique identifiers for patients, health care providers, institutions and payers are needed to obtain economies and accuracy when linking patient care data at different locations, and patient care data with other relevant data. Under HIPAA, the Secretary of HHS must adopt standards for uniquely identifying providers, health plans, employers, and individuals. Because of national concerns about the confidentiality of personal health information that may be linked using the unique individual health identifier, final implementation of that identifier must await explicit approval by Congress.

Malpractice liability concerns arise as telemedicine and information technology allow physician specialists to give medical advice across state borders electronically to other physicians, other healthcare providers, and patients. Physicians are normally licensed by a state to practice within its own state borders. Does a physician who is active in telemedicine need to obtain a license from each state in which he or she practices medicine from outside the state? If the expert physician outside the patient's state gives bad advice, which state's legal system has jurisdiction for liability considerations?

Benefit–cost analysis methods must be developed and applied to inform investment decision makers about the most productive applications of CPR systems. There is a need for a common approach to measuring the benefits and the costs for comparing alternative information technology applications. Certainly this is difficult since so many things change with the introduction of CPRS. As hard as they are to do well, valid business risk and benefit assessments can advance the development and implementation of commercial CPR applications.

Regional health data repositories and information exchange networks for the benefit of patients, providers, employers, hospital groups, consumers, and state health and service delivery programs raise issues about the ownership of patient care data, the use of identifiable patient care data, and the governance of health data repositories. A study by the IOM [1994] examined the power of regional health data repositories for improving public health, supporting better private health decisions, recognizing medically and cost-effective healthcare providers and health plans, and generally providing the information necessary to improve the quality of healthcare delivery in all settings. Because these data may include personally identifiable data and move outside the environment in which they were created, resolving these issues is of paramount importance for the development of regional health networks.

41.10 Summary

In summary, the benefits of CPRS are becoming better known and accepted. What is unknown are the costs of achieving these benefits in sites other than where the CPRSs were developed and how to successfully overcome institutional obstacles to their implementation. The widespread use of systems that provide clinical decision support depends in good part on the development and use of a common medical terminology or, at least, a reference terminology that contains all the relevant concepts to which different medical terminologies can map. This would enable interoperable electronic health information systems to accurately exchange information about those concepts. Although the HIPAA Privacy Rule gives patients the right to obtain a copy of their health information, research findings are lacking on the benefits of sharing CPR information with the patients themselves.

Strong initiatives by the federal government are leading the private and the government sectors to consider what infrastructure is needed to support patient information exchanges by clinical systems for the care of a patient. Indeed, a patient's CPR may not be a real data repository but a set of links to a patient's data that resides in many diverse electronic medical records — a virtual CPR. In addition, the federal government is making substantial investments in regional, often statewide, health information network demonstrations, and in research that studies local health information technology applications. The purpose of these investments is to learn how these networks can resolve important issues regarding the connectivity of providers to the network, the interoperability of their systems, and how successful they can be for improving patient safety and the quality of care. Many issues have been presented that must be resolved if the vision of a national health information infrastructure to be even partially achieved. The good news is that there is a national will to tackle them.

Acknowledgment

The author is Senior Science Advisor for Information Technology, Agency for Healthcare Research and Quality in the U.S. Department of Health and Human Services.

References

American National Standards Institute, Healthcare Informatics Standards Board (1997). HISB Inventory of Health Care Information Standards Pertaining to the Health Insurance Portability and Accountability Act of 1996, P.L. 104–191. New York.

American National Standards Institute, Healthcare Informatics Standards Board (1998). Inventory of Clinical Information Standards, New York. http://web.ansi.org/rooms/room_41/public/docs.html

Agency for Healthcare Research and Quality (2004). Fact Sheet: The Agency for Healthcare Research and Quality Health Information Technology Programs. Accessed on March 5, 2005, at http://www.ahrq.gov/research/hitfact.htm.

ASTM International (1992). E1460-92: Standard Specifications for Defining and Sharing Modular Health Knowledge Bases (Arden Syntax for Medical Logic Modules). ASTM, Philadelphia, PA.

ASTM International (2005). WK4363 Standard Specification for the Continuity of Care Record (CCR). Accessed on April 8, 2005, at http://www.astm.org/cgi-bin/SoftCart.exe/DATABASE.CART/WORKITEMS/WK4363.htmL+mystore+lghb8081

Barker K.N., Flynn E.A., Pepper G.A., Bates D.W., and Mikeal R.L. Medication errors observed in 36 healthcare facilities. *Arch. Intern. Med.* 2002; 162: 1897–1903.

Bates D.W., Leape L.L., Cullen D.J., Laird N., Petersen L.A., Teich J.M., Burdick E., Hickey M., Kleefield S., Shea B., VanderVliet M., and Seger D.L. Effect of computerized physician order entry and a team intervention on prevention of serious medication errors. *JAMA*, 1998; 280: 1311–1316.

Bates D.W., Kuperman G.J., Rittenberg E., Teich J.M., Fiskio J., Ma'luf N., Onderonk A., Wybenga D., Winkelman J., Brennan T.A., Komeroff A.L., and Tanasijevic M. A randomized trial of a computer-based intervention to reduce utilization of redundant laboratory tests. *Am. J. Med.* 1999; 106: 144–150.

Birkmeyer J.D., Birkmeyer C.M., Wennberg D.E., and Young M. (2000). Leapfrog patient safety standards: the potential benefit of universal adoption. Washington, D.C.: The Leapfrog Group for Patient Safety Accessed on April 8, 2005, at http://www.leapfroggroup.org/media/file/Leapfrog-Launch-Executive_Summary.pdf

Bureau of Labor Statistics (2005) "Table 1A. Consumer Price Index for All Urban Consumers (CPI-U): U.S. city average, by expenditure category and commodity and service group. Accessed on October 11, 2005 at http://www.bls.gov/cpi/cpid03av.pdf.

Bush G.W. (2004a). Incentives for the Use of Health Information Technology and Establishing the Position of the National Health Information Technology Coordinator. Executive Order (April 27, 2004). Accessed at http://www.whitehouse.gov/news/releases/2004/04/20040427-4.html on February 21, 2005.

Bush G.W. (2004b). President Unveils Tech Initiatives for Energy, Health Care, Internet. Remarks by the President at American Association of Community Colleges Annual Convention, Minneapolis Convention Center, Minneapolis, Minnesota (April 26, 2004). Accessed at http://www.whitehouse.gov/news/releases/2004/04/20040426-6.html on February 28, 2005.

Chase C.R., Vacek P.M., Shinozaki T., Giard A.M., and Ashikaga T. Medical information management: Improving the transfer of research results to presurgical evaluation. *Med. Care.* 1983; 21: 410–424.

Cimino, J.J. Saying what you mean and meaning what you say: coupling biomedical terminology and knowledge. *Acad. Med.* 1993; 68: 257–260.

Cimino J.J. Desiderata for controlled medical vocabularies in the twenty-first century. *Meth. Inf. Med.* 1998; 37: 394–403.

Classen D.C., Evans R.S., Pestotnik S.L., Horn S.D., Menlove R. L., and Burke J.P. The timing of prophylactic administration of antibiotics and the risk of surgical wound infection. *NEJM* 1992; 326: 281–285.

Cynthia Smith, Cathy Cowan, Art Sensenig, Aaron Catlin and the Health Accounts Team. Health spending growth slows in 2003. *Health Affairs.* 2005; 24(1): 185–194.

Donaldson M.S. and Lohr K.N. (Eds.) (1994). *Health Data in the Information Age: Use, Disclosure, and Privacy.* National Academy Press, Washington, D.C.

Evans R.S., Pestotnik S.L., Classen D.C., and Burke J.R. Development of an automated antibiotic consultant. *MD Comput.* 1993; 10: 17–22.

Evans R.S., Classen D.C., Stevens M.S., Pestotnik S.L., Gardner R.M., Lloyd J.F., and Burke J.P. Using a health information system to assess the effects of adverse drug events. In *AMIA Proceedings of the 17th Annual Symposium on Computer Applications in Medical Care*, McGraw-Hill, Inc., New York, pp. 161–165.

Fitzmaurice J.M. (1994b). Health Care and the NII. In: Putting the Information Infrastructure to Work: Report of the Information Infrastructure Task Force Committee on Applications and Technology, pp. 41–56. National Institute of Standards and Technology, Gaithersburg, MD.

Fitzmaurice J.M., Adams K., and Eisenberg, J.M. Three decades of research on computer applications in health care. *J. Am. Med. Inform. Assoc.* 2002; 9: 144–160.

General Accounting Office, Health and Human Services' Estimate of Health Care Cost Savings Resulting from the Use of Information Technology. GAO-05-309R (February 17, 2005), pp. 1–9. Accessed on February 20, 2005 at http://www.gao.gov/new.items/d05309r.pdf.

Health Insurance Reform: Security Standards; Final Rule (2003). Federal Register. Rules and Regulations. 45 CFR Parts 160, 162, and 164.68(34); (Feb. 20, 2003): 8334–8381.

Hersh W.R., Wallace J.A., Patterson P.K., Shapiro S.E., Kraemer D.F., Eilers G.M., Chan B.K.S., Greenlick M.R., and Helfand M. (2001a). *Telemedicine for the Medicare Population.* Evidence Report/Technology Assessment No. 24 (Prepared by Oregon Health Sciences University, Portland, OR under Contract No. 290-97-0018). AHRQ Publication No. 01-E012. Rockville (MD): Agency for Healthcare Research and Quality. July 2001. Accessed on April 10, 2005 at http://www.ahrq.gov/clinic/evrptfiles.htm

Hersh W.R., Wallace J.A., Patterson P.K., Kraemer D.F., Nichol W.P., Greenlick M.R., Krages K.P., Helfand M. (2001b). Telemedicine for the Medicare Population: Pediatric, obstetric, and clinician-indirect home interventions. Evidence Report/Technology Assessment No. 24S, Supplement (Prepared by Oregon Health Sciences University, Portland, OR under Contract No. 290-97-0018). AHRQ Publication No. 01-E060. Rockville (MD): Agency for Healthcare Research and Quality. August 2001. Accessed on April 10, 2005 at http://www.ahrq.gov/clinic/evrptfiles.htm

HL7 EHR System Functional Model Draft Standard for Trial Use, July, 2004. Eds: Dickinson G., Fischetti L., and Heard S. Ann Arbor, Michigan: Health Level Seven, Inc., 2004. Accessed on April 7, 2005, at http://www.hl7.org/ehr/downloads/index.asp

HL7 Version 3.0 (Draft), Ann Arbor, Michigan: Health Level Seven. 2001. Accessed on April 8, 2005 at http://www.hl7.org/library/standards.cfm

Humphreys B.L., Lindberg D.A., Schoolman H.M., and Barnett G.O. The Unified Medical Language System: an informatics research collaboration. *J. Am. Med. Inform. Assoc.* 1998; 5: 1–11.

Hussey P.S., Anderson G.F., Osborn R., Feek C., McLaughlin V., Millar J., and Epstein A. How does the quality of care compare in five countries? *Health Affairs* 2004; 23: 89–99.

Institute of Medicine (1991). Revised edition 1997. The Computer-Based Patient Record: An Essential Technology for Health Care, Detmer D.E., Dick R.S., and Steen E.B. (Eds.). *Committee on Improving the Patient Record*, National Academy Press, Washington, D.C.

Institute of Medicine, Committee on Data Standards for Patient Safety. 2003a. *Key Capabilities of an Electronic Health Record System. Letter Report.* 2003. Washington, D.C.:The National Academies Press. Accessed at http://books.nap.edu/html/ehr/NI000427.pdf on February 20, 2005.

Institute of Medicine (1994). *Health Data In the Information Age: Use, Disclosure, and Privacy.* Donaldson M.S. and Lohr K.N. (Eds.), National Academy Press, Washington, D.C.

Institute of Medicine (1999). *To Err Is Human: Building a Safer Health System.* Kohn L.T., Corrigan J., and Donaldson M.S. (Eds.). Committee on Quality of Health Care in America. National Academy Press, Washington, D.C.

Institute of Medicine (2001). *Crossing the Quality Chasm: A New Health System for the 21st Century.* National Academy Press, Washington, D.C.

Institute of Medicine (2003b). *Patient Safety: Achieving a New Standard for Care.* Aspden P., Corrigan J.M., Wolcott J., and Erickson S.M. (Eds.). Committee on Data Standards for Patient Safety. National Academies Press, Washington, D.C.

Jha A.K., Kuperman G.J., Rittenberg E., Teich J.M., and Bates D.W. Identifying hospital admissions due to adverse drug events using a computer-based monitor. *Pharmacoepidemiol. Drug Safety*, 2001; 10: 113–119.

Johnston M.E., Langton K.B., Haynes R.B., and Mathieu A. Effects of computer-based clinical decision support systems on clinician performance and patient outcome. *Ann. Int. Med.* 1994; 120: 135–142.

Kerr E.A., McGlynn E.A., Adams J., Keesey J., and Asch S.M. Profiling the quality of care in twelve communities: results from the CQI study. *Health Affairs* 2004; 23: 247–256.

Korpman R.A. Patient care automation; the future is now. Part 2. The current paper system — can it be made to work? *Nurs. Econ.* 1990; 8: 263–267.

Korpman, R.A. Patient care automation; the future is now, Part 8. Does reality live up to the promise? *Nurs. Econ.* 1991; 9: 175–179.

Markle Foundation, Connecting for Health, General Resources (2005). Accessed on April 7, 2005 at http://www.connectingforhealth.org/resources/generalresources.html

McDonald C.J., Hui S.J., Smith D.M., Tierney W.M., Cohen S.J., Weinberger M., et al. Reminders to physicians from an introspective computer medical record. A two-year randomized trial. *Ann. Int. Med.* 1984; 100: 130–138.

McGlynn E.A., Asch S.M., Adams J., Keesey J., Hicks J., DeCristofaro A., and Kerr E.A. The Quality of Health Care Delivered to Adults in the United States, *N. Engl. J. Med.* 2003; 348: 2635–2645.

Miller, R.A. Medical diagnostic decision support systems — past, present, and future. *J. Am. Med. Inform. Assoc.* 1994; 1: 8–27.

National Committee On Vital and Health Statistics (2000). Report to the Secretary of the U.S. Department of Health and Human Services on Uniform Data Standards for Patient Medical Record Information. Department of HHS: July 6, 2000, Accessed at http://www.ncvhs.hhs.gov/hipaa000706.pdf on February 28, 2005.

National Committee on Vital and Health Statistics. Information for Health — A Strategy for Building the National Health Information Infrastructure: Report and Recommendations From the NCVHS. Washington, D.C., Department of Health and Human Services, November 15, 2001. Accessed on April 9, 2005, at http://www.ncvhs.hhs.gov/nhiilayo.pdf.

National Coordination Office for High Performance Computing and Communication (1994). HPCC FY 1995 Implementation Plan. Executive Office of The President, Washington, D.C.

National Library of Medicine (2005a). Unified Medical Language System, Section 2, Metathesaurus. Accessed on April 9, 2005 at http://www.nlm.nih.gov/research/umls/meta2.html

National Library of Medicine (2005b). Unified Medical Language System, Metathesaurus. Accessed on April 9, 2005 at http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

Nykanen, P., Chowdhury, S., and Wiegertz, O. Evaluation of decision support systems in medicine. *Comput Methods Programs Biomed*; 1991; 34(2-3): 229–238.

Office of the National Coordinator for Health Information Technology (ONCHIT) (2004). *Strategic Framework: The Decade of Health Information Technology: Delivering Consumer-centric and Information-rich Health Care.* July 21, 2004, Washington, D.C., Dept. of Health and Human Services. Accessed at http://www.hhs.gov/healthit/frameworkchapters.html on February 21, 2005.

President's Information Technology Advisory Committee (2004). Revolutionizing health care through information technology. Arlington, VA. NCO for ITRD National Coordinating Office for Information Technology Research and Development, June 2004.

Pryor, T. Allan and Hripcsak, George. (1994). Sharing MLMs: An Experiment Between Columbia-Presbyterian and LDS Hospital. In *AMIA Proceedings of the 17th Annual Symposium on Computer Applications in Medical Care*, McGraw-Hill, Inc., New York, pp. 399–403.

Public Law 104–191, 1996. The Health Insurance Portability and Accountability Act of 1996, August 21, 1996.

Public Law 105–277, Department of Transportation and Related Agencies Appropriations Act, October 21, 1998.

Public Law 106–129, Healthcare Research and Quality Act of 1999, December 6, 1999. Accessed on April 8, 2005, at http://www.ahrq.gov/ hrqa99a.htm

Public Law 106–229. June 30, 2000. Electronic signatures in global and National commerce act. Accessed on March 5, 2005, at http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=106_cong_public_laws&docid=f:publ229.106.pdf

Public Law 108–173, Medicare Prescription Drug, Improvement, and Modernization Act of 2003, December 8, 2003. Accessed on April 8, 2005, at http://www.cms.hhs.gov/medicarereform/MMAactFullText.pdf

Teich J.M, Merchia P.R., Schmiz J.L., Kuperman G.J., Spurr C.D., and Bates D.W. Effects of computerized physician order entry on prescribing practices. *Arch. Intern. Med.* 2000; 160: 2741–2747.

Tierney W.M., McDonald C.J., Martin D.K, and Rogers M.P. Computerized display of past test results. Effect on outpatient testing. *Ann. Intern. Med.* 1987; 107: 569–574.

Tierney, W.M., McDonald, C.J., Hui S.J., and Martin, D.K. Computer predictions of abnormal test results. Effects on outpatient testing. *JAMA* 1988; 259: 1194–11988.

Tierney W.M., Miller M.E., and McDonald C.J. The effect on test ordering of informing physicians of the charges for outpatient diagnostic tests. *NEJM* 1990; 322: 1499–1504.

Tierney W.N. and McDonald C.M. Practice Databases and their uses in clinical research. *Statist. Med.* 1991; 10: 541–557.

Tierney W.M., Miller M.E., Overhage J.M., and McDonald C.J. Physician inpatient order writing on microcomputer workstations. *JAMA* 1993; 269: 379–383.

Tuttle M.S., Sherertz D.D., Fagan L.M., Carlson R.W., Cole W.G., Shipma P.B., and Nelson S.J. 1994. Toward an interim standard for patient-centered knowledge-access. In *AMIA Proceedings of the 17th Annual Symposium on Computer Applications in Medical Care*, McGraw-Hill, Inc., New York, pp. 564–568.

U.S. Department of Health and Human Services, Office of Civil Rights (2003). Standards for Privacy of Individually Identifiable Health Information; Security Standards for the Protection of Electronic Protected Health Information; General Administrative Requirements Including, Civil Money Penalties: Procedures for Investigations, Imposition of Penalties, and Hearings. Regulation Text (Unofficial Version) (45 CFR Parts 160 and 164); December 28, 2000 as amended: May 31, 2002, August 14, 2002, February 20, 2003, and April 17, 2003. (August 2003). Accessed on April 9, 2005, at http://www.hhs.gov/ocr/combinedregtext.pdf

U.S. Department of Health and Human Services Press Release. May 6, 2004. Secretary Thompson, Seeking Fastest Possible Results, Names First Health Information Technology Coordinator. Accessed on March 5, 2005, at http://www.hhs.gov/news/press/2004pres/20040506.html.

U.S. Senate (1999). Healthcare Research and Quality Act of 1999, Senate Bill.580. Introduced January 6, 1999, signed into law on December 6, 1999. Accessed on March 5, 2005, at http://www.ahrq.gov/hrqa99a.htm.

Wang S.J., Middleton B., Prosser L.A., Bardon C.G., Spurr C.D., Carchidi P.J., Kittler A.F., Goldzer R.C., Fairchild D.G., Sussman A.J., Kuperman G.J., and Bates D.W. A cost–benefit analysis of electronic medical records in primary care. *Am. J. Med.* 2003; 114: 397–403.

Wang D., Peleg M., Tu S.W., Boxwala A.A., Ogunyemi O., Zeng Q., Greenes R.A., Patel V.L., and Shortliffe E.H. Design and implementation of the GLIF3 guideline execution engine. *Biomed. Inform.* 2004; 37: 305–18.

Zhan C. and Miller M.R. Excess length of stay, charges, mortality are attributable to medical injuries during hospitalization. *JAMA* 2003; 290: 1868–1874

42

Overview of Standards Related to the Emerging Health Care Information Infrastructure

Jeffrey S. Blair
IBM Health Care Solutions

As the cost of health care has become a larger percentage of the gross domestic product of many developed nations, the focus on methods to improve health care productivity and quality has increased. To address this need, the concept of a health care information infrastructure has emerged. Major elements of this concept include patient-centered care facilitated by computer-based patient record systems, continuity of care enabled by the sharing of patient information across information networks, and outcomes measurement aided by greater availability and specificity of health care information.

The creation of this health care information infrastructure will require the integration of existing and new architectures, products, and services. To make these diverse components work together, health care information standards (classifications, guides, practices, terminology) will be required [ASTM, 1994]. This chapter will give you an overview of the major existing and emerging health care information standards, and the efforts to coordinate, harmonize, and accelerate these activities. It is organized into the major topic areas of:

- Identifier standards
- Communications (message format) standards
- Content and structure standards
- Clinical data representations (codes)
- Confidentiality, data security, and authentication
- Quality indicators and data sets
- International Standards
- Coordinating and promotion organizations
- Summary

42.1 Identifier Standards

There is a universal need for health care identifiers to uniquely specify each patient, provider, site of care, and product; however, there is no universal acceptance or satisfaction with these systems.

42.1.1 Patient Identifiers

The social security number (SSN) is widely used as a patient identifier in the United States today. However, critics point out that it is not an ideal identifier. They say that not everyone has an SSN; several individuals may use the same SSN; and the SSN is so widely used for other purposes that it presents an exposure to violations of confidentiality. These criticisms raise issues that are not unique to the SSN. A draft document has been developed by the American Society for Testing and Materials (ASTM) E31.12. Subcommittee to address these issues. It is called the "Guide for the Properties of a Universal Health Care Identifier" (UHID). It presents a set of requirements outlining the properties of a national system creating a UHIC, includes critiques of the SSN, and creates a sample UHD [ASTM E31.12, 1994]. Despite the advantages of a modified/new patient identifier, there is not yet a consensus as to who would bear the cost of adopting a new patient identifier system.

42.1.2 Provider Identifiers

The Health Care Financing Administration (HCFA) has created a widely used provider identifier known as the Universal Physician Identifier Number (UPIN) [Terrell et al., 1991]. The UPIN is assigned to physicians who handle Medicare patients, but it does not include nonphysician caregivers. The National Council of Prescription Drug Programs (NCPDP) has developed the standard prescriber identification number (SPIN) to be used by pharmacists in retail settings. A proposal to develop a new national provider identifier number has been set forth by HCFA [1994]. If this proposal is accepted, then HCFA would develop a national provider identifier number which would cover all caregivers and sites of care, including Medicare, Medicaid, and private care. This proposal is being reviewed by various state and federal agencies. It has also been sent to the American National Standards, Institute's Health Care Informatics Standards Planning Panel (ANSI HISPP) Task Force on Provider Identifiers for review.

42.1.3 Site-of-Care Identifiers

Two site-of-care identifier systems are widely used. One is the health industry number (HIN) issued by the Health Industry Business Communications Council (HIBCC). The HIN is an identifier for health care facilities, practitioners, and retail pharmacies. HCFA has also defined provider of service identifiers for Medicare usage.

42.1.4 Product and Supply Labeling Identifiers

Three identifiers are widely accepted. The labeler identification code (LIC) identifies the manufacturer or distributor and is issued by HIBCC [1994]. The LIC is used both with and without bar codes for products and supplies distributed within a health care facility. The universal product code (UPC) is maintained by the Uniform Code Council and is typically used to label products that are sold in retail settings. The national drug code is maintained by the Food and Drug Administration and is required for reimbursement by Medicare, Medicaid, and insurance companies. It is sometimes included within the UPC format.

42.2 Communications (Message Format) Standards

Although the standards in this topic area are still in various stages of development, they are generally more mature than those in most of the other topic areas. They are typically developed by committees within standards organizations and have generally been accepted by users and vendors. The overviews of these standards given below were derived from many sources, but considerable content came from the Computer-based Patient Record Institute's (CPRI) "Position Paper on Computer-based Patient Record Standards" [CPRI, 1994] and the Agency for Health Care Policy and Research's (AHCPR) "Current Activities of Selected Health Care Informatics Standards Organizations" [Moshman Associates, 1994].

42.2.1 ASC X12N

This committee is developing message format standards for transactions between payers and providers. It is rapidly being accepted by both users and vendors. It defines the message formats for the following transaction types [Moshman Associates, 1994]:

- 834 — enrollment
- 270 — eligibility request
- 271 — eligibility response
- 837 — health care claim submission
- 835 — health care claim payment remittance
- 276 — claims status request
- 277 — claims status response
- 148 — report of injury or illness

ASC X12N is also working on the following standards to be published in the near future:

- 257, 258 — Interactive eligibility response and request. These transactions are an abbreviated form of the 270/271.
- 274, 275 — patient record data response and request. These transactions will be used to request and send patient data (tests, procedures, surgeries, allergies, etc.) between a requesting party and the party maintaining the database.
- 278, 279 — health care services (utilization review) response and request. These transactions will be used to initiate and respond to a utilization review request.

ASC X12N is recognized as an accredited standards committee (ASC) by the American National Standards Institute (ANSI).

42.2.2 American Society for Testing and Materials

42.2.2.1 Message Format Standards

The following standards were developed within American Society for Testing and Materials (ASTM) Committee E31. This committee has applied for recognition as an ASC by ANSI:

1. ASTM E1238 standard specification for transferring clinical observations between independent systems. E1238 was developed by ASTM Subcommittee E31.11. This standard is being used by most of

the largest commercial laboratory vendors in the United States to transmit laboratory results. It has also been adopted by a consortium of 25 French laboratory system vendors. Health level seven (HL7), which is described later in this topic area, has incorporated E1238 as a subset within its laboratory results message format [CPRI, 1994].

2. ASTM E1394 standard specification for transferring information between clinical instruments. E1394 was developed by ASTM Subcommittee E31.14. This standard is being used for communication of information from laboratory instruments to computer systems. This standard has been developed by a consortium consisting of most U.S. manufacturers of clinical laboratory instruments [CPRI, 1994].

3. ASTM 1460 specification for defining and sharing modular health knowledge bases (Arden Syntax). E1460 was developed by ASTM Subcommittee E31.15. The Arden Syntax provides a standard format and syntax for representing medical logic and for writing rules and guidelines that can be automatically executed by computer systems. Medical logic modules produced in one site-of-care system can be sent to a different system within another site of care and then customized to reflect local usage [CPRI, 1994].

4. ASTM E1467 specification for transferring digital neurophysical data between independent computer systems. E1467 was developed by ASTM Subcommittee E31.17. This standard defines codes and structures needed to transmit electrophysiologic signals and results produced by electroencephalograms and electromyograms. The standard is similar in structure to ASTM E1238 and HL7; and it is being adopted by all the EEG systems manufacturers [CPRI, 1994].

42.2.3 Digital Imaging and Communications

This standard is developed by the American College of Radiology — National Electronic Manufacturers' Association (ACR-NEMA). It defines the message formats and communications standards for radiologic images. Digital imaging and communications (DICOM) is supported by most radiology picture archiving and communications systems (PACS) vendors and has been incorporated into the Japanese Image Store and Carry (ISAC) optical disk system as well as Kodak's PhotoCD. ACR-NEMA is applying to be recognized as an accredited organization by ANSI [CPRI, 1994].

42.2.4 Health Level Seven (HL7)

HL7 is used for intra-institution transmission of orders; clinical observations and clinical data, including test results; admission, transfer, and discharge records; and charge and billing information. HL7 is being used in more than 300 U.S. health care institutions including most leading university hospitals and has been adopted by Australia and New Zealand as their national standard. HL7 is recognized as an accredited organization by ANSI [Hammond, 1993; CPRI, 1994].

42.2.5 Institute of Electrical and Electronics Engineers, Inc. P1157

42.2.5.1 Medical Data Interchange Standard

Institute of Electrical and Electronics Engineers, Inc. (IEEE) Engineering in Medicine and Biology Society (EMB) is developing the medical data interchange standard (MEDIX) standards for the exchange of data between hospital computer systems [Harrington, 1993; CPRI, 1994]. Based on the International Standards Organization (ISO) standards for all seven layers of the OSI reference model, MEDIX is working on a framework model to guide the development and evolution of a compatible set of standards. This activity is being carried forward as a joint working group under ANSI HISPP's Message Standards Developers Subcommittee (MSDS). IEEE is recognized as an accredited organization by ANSI.

IEEE P1073 Medical Information Bus (MIB): This standard defines the linkages of medical instrumentation (e.g., critical care instruments) to point-of-care information systems [CPRI, 1994].

National Council for Prescription Drug Programs (NCPDP): These standards developed by NCPDP are used for communication of billing and eligibility information between community pharmacies and third-party payers. They have been in use since 1985 and now serve almost 60% of the nation's community pharmacies. NCPDP has applied for recognition as an accredited organization by ANSI [CPRI, 1994].

42.3 Content and Structure Standards

Guidelines and standards for the content and structure of computer-based patient record (CPR) systems are being developed within ASTM Subcommittees E31.12 and E31.19. They have been recognized by other standards organizations (e.g., HL7); however, they have not matured to the point where they are generally accepted or implemented by users and vendors.

A major revision to E1384, now called a standard description for content and structure of the computer-based patient record, has been made within Subcommittee E31.19 [ASTM, 1994]. This revision includes work from HISPP on data modeling and an expanded framework that includes master tables and data views by user.

Companion standards have been developed within E31.19. They are E1633, A Standard Specification for the Coded Values Used in the Automated Primary Record of Care [ASTM, 1994], and E1239–94, A Standard Guide for Description of Reservation/Registration-A/D/T Systems for Automated Patient Care Information Systems [ASTM, 1994]. A draft standard is also being developed for object-oriented models for R-A/D/T functions in CPR systems. Within the E31.12 Subcommittee, domain specific guidelines for nursing, anesthesiology, and emergency room data within the CPR are being developed [Moshman Associates, 1994; Waegemann, 1994].

42.4 Clinical Data Representations (Codes)

Clinical data representations have been widely used to document diagnoses and procedures. There are over 150 known code systems. The codes with the widest acceptance in the United States include:

1. International Classification of Diseases (ICD) codes, now in the ninth edition (ICD-9), are maintained by the World Health Organization (WHO) and are accepted worldwide. In the United States, HCFA and the National Center for Health Statistics (NCHS) have supported the development of a clinical modification of the ICD codes (ICD-9-CM). WHO has been developing ICD-10; however, HCFA projects that it will not be available for use within the United States for several years. Payers require the use of ICD-9-CM codes for reimbursement purposes, but they have limited value for clinical and research purposes due to their lack of clinical specificity [Chute, 1991].

2. Current Procedural Terminology (CPT) codes are maintained by the American Medical Association (AMA) and are widely used in the United States for reimbursement and utilization review purposes. The codes are derived from medical specialty nomenclatures and are updated annually [Chute, 1991].

3. The systematized nomenclature of medicine (SNOMED) is maintained by the College of American Pathologists and is widely accepted for describing pathologic test results. It has a multiaxial (11 fields) coding structure that gives it greater clinical specificity than the ICD and CPT codes, and it has considerable value for clinical purposes. SNOMED has been proposed as a candidate to become the standardized vocabulary for computer-based patient record systems [Rothwell et al., 1993].

4. Digital imaging and communications (DICOM) is maintained by the American College of Radiology — National Electronic Manufacturers' Association (ACR-NEMA). It sets forth standards for indices of radiologic diagnoses as well as for image storage and communications [Cannavo, 1993].

5. Diagnostic and Statistical Manual of Mental Disorders (DSM), now in its fourth edition (DSM-IV), is maintained by the American Psychiatric Association. It sets forth a standard set of codes and descriptions for use in diagnoses, prescriptions, research, education, and administration [Chute, 1991].

6. Diagnostic Related Groups (DRGs) are maintained by HCFA. They are derivatives of ICD-9-CM codes and are used to facilitate reimbursement and case-mix analysis. They lack the clinical specificity to be of value in direct patient care or clinical research [Chute, 1991].

7. Unified Medical Language System (UMLS) is maintained by the National Library of Medicine (NLM). It contains a metathesaurus that links clinical terminology, semantics, and formats of the major clinical coding and reference systems. It links medical terms (e.g., ICD, CPT, SNOMED, DSM, CO-STAR, and D-XPLAIN) to the NLM's medical index subject headings (MeSH codes) and to each other [Humphreys, 1991; Cimino et al., 1993].

8. The Canon Group has not developed a clinical data representation, but it is addressing two important problems: clinical data representations typically lack clinical specificity and are incapable of being generalized or extended beyond a specific application. "The Group proposes to focus on the design of a general schema for medical-language representation including the specification of the resources and associated procedures required to map language (including standard terminologies) into representations that make all implicit relations "visible," reveal "hidden attributes," and generally resolve "ambiguous references" [Evans et al., 1994].

42.5 Confidentiality, Data Security, and Authentication

The development of computer-based patient record systems and health care information networks have created the opportunity to address the need for more definitive confidentiality, data security, and authentication guidelines and standards. The following activities address this need:

1. During 1994, several bills were drafted in Congress to address health care privacy and confidentiality. They included the Fair Health Information Practices Act of 1994 (H.R. 4077), the Health Care Privacy Protection Act (S. 2129), and others. Although these bills were not passed as drafted, their essential content is expected to be included as part of subsequent health care reform legislation. They address the need for uniform comprehensive federal rules governing the use and disclosure of identifiable health and information about individuals. They specify the responsibilities of those who collect, use, and maintain health information about patients. They also define the rights of patients and provide a variety of mechanisms that will allow patients to enforce their rights.

2. ASTM Subcommittee E31.12 on Computer-based Patient Records is developing Guidelines for Minimal Data Security Measures for the Protection of Computer-based patient Records [Moshman Associates, 1994].

3. ASTM Subcommittee E31.17 on Access, Privacy, and Confidentiality of Medical Records is working on standards to address these issues [Moshman Associates, 1994].

4. ASTM Subcommittee E31.20 is developing standard specifications for authentication of health information [Moshman Associates, 1994].

5. The Committee on Regional Health Data Networks convened by the Institute of Medicine (IOM) has completed a definitive study and published its findings in a book entitled Health Data in the Information Age: Use, Disclosure, and Privacy [Donaldson and Lohr, 1994].

6. The Computer-based Patient Record Institute's (CPRI) Work Group on Confidentiality, Privacy, and Legislation has completed white papers on "Access to Patient Data" and on "Authentication," and a publication entitled "Guidelines for Establishing Information Security: Policies at Organizations using Computer-based Patient Records" [CPRI, 1994].

7. The Office of Technology Assessment has completed a two-year study resulting in a document entitled "Protecting Privacy in Computerized Medical Information." It includes a comprehensive review of system/data security issues, privacy information, current laws, technologies used for protection, and models.

8. The U.S. Food and Drug Administration (FDA) has created a task force on Electronic/Identification Signatures to study authentication issues as they relate to the pharmaceutical industry.

42.6 Quality Indicators and Data Sets

The Joint Commission on Accreditation of Health Care Organizations (JCAHO) has been developing and testing obstetrics, oncology, trauma, and cardiovascular clinical indicators. These indicators are intended to facilitate provider performance measurement. Several vendors are planning to include JCAHO clinical indicators in their performance measurement systems [JCAHO, 1994].

The health employers data and information set (HEDIS) version 2.0 has been developed with the support of the National Committee for Quality Assurance (NCQA). It identifies data to support performance measurement in the areas of quality (e.g., preventive medicine, prenatal care, acute and chronic disease, and mental health), access and patient satisfaction, membership and utilization, and finance. The development of HEDIS has been supported by several large employers and managed care organizations [NCQA, 1993].

42.7 International Standards

The ISO is a worldwide federation of national standards organizations. It has 90 member countries. The purpose of ISO is to promote the development of standardization and related activities in the world. ANSI was one of the founding members of ISO and is representative for the United States [Waegemann, 1994].

ISO has established a communications model for open systems interconnection (OSI). IEEE/MEDIX and HL7 have recognized and built upon the ISO/OSI framework. Further, ANSI HISPP has a stated objective of encouraging compatibility of U.S. health care standards with ISO/OSI. The ISO activities related to information technology take place within the Joint Technical Committee (JTC) 1.

The Comite Europeen de Noramalisation (CEN) is a European standards organization with 16 technical committees (TCs). Two TCs are specifically involved in health care: TC 251 (Medical Informatics) and TC 224 WG12 (Patient Data Cards) [Waegemann, 1994].

The CEN TC 251 on Medical Informatics includes work groups on: Modeling of Medical Records; Terminology, Coding, Semantics, and Knowledge Bases; Communications and Messages; Imaging and Multimedia; Medical Devices; and Security, Privacy, Quality, and Safety. The CEN TC 251 has established coordination with health care standards development in the United States through ANSI/HISPP.

In addition to standards developed by ISO and CEN, there are two other standards of importance. United Nations (U.N.) EDIFACT is a generic messaging-based communications standard with health-specific subsets. It parallels X12 and HL7, which are transaction-based standards. It is widely used in Europe and in several Latin American countries. The READ Classification System (RCS) is a multiaxial medical nomenclature used in the United Kingdom. It is sponsored by the National Health Service and has been integrated into computer-based ambulatory patient record systems in the United Kingdom [CAMS, 1994].

42.8 Standards Coordination and Promotion Organizations

In the United States, two organizations have emerged to assume responsibility for the coordination and promotion of health care standards development: the ANSI Health Care Informatics Standards Planning Panel (HISPP) and the Computer-based Patient Record Institute (CPRI). The major missions of an ANSI HISPP are:

1. To coordinate the work of the standards groups for health care data interchange and health care informatics (e.g., ACR/NEMA, ASTM, HL7, IEEE/MEDIX) and other relevant standards groups (e.g., X3, X12) toward achieving the evolution of a unified set of nonredundant, nonconflicting standards.
2. To interact with and provide input to CEN TC 251 (Medical Informatics) in a coordinated fashion and explore avenues of international standards development. The first mission of coordinating

standards is performed by the Message Standards Developers Subcommittee (MSDS). The second mission is performed by the International and Regional Standards Subcommittee. HISPP also has four task groups (1) Codes and Vocabulary, (2) Privacy, Security, and Confidentiality, (3) Provider Identification Numbering Systems, and (4) Operations. Its principal membership is composed of representatives of the major health care standards development organizations (SDOs), government agencies, vendors, and other interested parties. ANSI HISPP is by definition a planning panel, not an SDO [Hammond, 1994; ANSI HISPP, 1994].

The CPRI's mission is to promote acceptance of the vision set forth in the Institute of Medicine Study report "The Computer-based Patient Record: An Essential Technology for Health Care." CRPI is a nonprofit organization committed to initiating and coordinating activities to facilitate and promote the routine use of computer-based patient records. The CPRI takes initiatives to promote the development of CPR standards, but it is not an SDO itself. CPRI members represent the entire range of stakeholders in the health care delivery system. Its major work groups are the: (1) Codes and Structures Work Group; (2) CPR Description Work Group; (3) CPR Systems Evaluation Work Group; (4) Confidentiality, Privacy, and Legislation Work Group; and (5) Professional and Public Education Work Group [CPRI, 1994].

Two work efforts have been initiated to establish models for principal components of the emerging health care information infrastructure. The CPR Description Work Group of the CPRI is defining a consensus-based model of the computer-based patient record system. A joint working group to create a common data description has been formed by the MSDS Subcommittee of ANSI HISPP and IEEE/MEDIX. The joint working group is an open standards effort to support the development of a common data model that can be shared by developers of health care informatics standards [IEEE, 1994].

The CPRI has introduced a proposal defining a public/private effort to accelerate standards development for computer-based patient record systems [CPRI, 1994]. If funding becomes available, the project will focus on obtaining consensus for a conceptual description of a computer-based patient record system; addressing the need for universal patient identifiers; developing standard provider and sites-of-care identifiers; developing confidentiality and security standards; establishing a structure for and developing key vocabulary and code standards; completing health data interchange standards; developing implementation tools; and demonstrating adoptability of standards in actual settings. This project proposes that the CPRI and ANSI HISPP work together to lead, promote, coordinate, and accelerate the work of SDOs to develop health care information standards.

The Workgroup on Electronic Data Interchange (WEDI) is a voluntary, public/private task force which was formed in 1991 as a result of the call for health care administrative simplification by the director of the Department of Health and Human Services, Dr. Louis Sullivan. They have developed an action plan to promote health care EDI which includes: promotion of EDI standards, architectures, confidentiality, identifiers, health cards, legislation, and publicity [WEDI, 1993].

42.9 Summary

This chapter has presented an overview of major existing and emerging health care information infrastructure standards and the efforts to coordinate, harmonize, and accelerate these activities. Health care informatics is a dynamic area characterized by changing business and clinical processes, functions, and technologies. The effort to create health care informatics standards is therefore also dynamic. For the most current information on standards, refer to the "For More Information" section at the end of this chapter.

References

American National Standards Institute's Health Care Informatics Standards Planning Panel (1994). Charter statement. New York.

American Society for Testing and Materials (ASTM) (1994). Guide for the properties of a universal health care identifier. ASTM Subcommittee E31.12, Philadelphia.

American Society for Testing and Materials (ASTM) (1994). Membership information packet: ASTM Committee E31 on computerized systems, Philadelphia.

American Society for Testing and Materials (ASTM) (1994). A standard description for content and structure of the computer-based patient record, E1384–91/1994 revision. ASTM Subcommittee E31.19, Philadelphia.

American Society for Testing and Materials (ASTM) (1994). Standard guide for description of reservation registration-admission, discharge, transfer (R-ADT) systems for automated patient care information systems, E1239–94. ASTM Subcommittee E31.19, Philadelphia.

American Society for Testing and Materials (ASTM) (1994). A standard specification for the coded values used in the automated primary record of care, E1633. ASTM Subcommittee E31.19, Philadelphia.

Cannavo M.J. (1993). The last word regarding DEFF & DICOM. Healthcare Informatics 32.

Chute C.G. (1991). Tutorial 19: Clinical data representations. Washington, DC, Symposium on Computer Applications in Medical Care.

Cimino J.J., Johnson S.B., Peng P. et al. (1993). *From ICD9-CM to MeSH Using the UMLS: A How-to-Guide*, SCAMC, Washington, DC.

Computer Aided Medical Systems Limited (CAMS) (1994). CAMS News 4: 1.

Computer-based Patient Record Institute (CPRI) (1994). CPRI-Mail 3: 1.

Computer-based Patient Record Institute (CPRI) (1994). Position paper computer-based patient record standards. Chicago.

Computer-based Patient Record Institute (CPRI) (1994). Proposal to accelerate standards development for computer-based patient record systems. Version 3.0, Chicago.

Donaldson M.S. and Lohr K.N. (Eds.) (1994). *Health Data in the Information Age: Use, Disclosure, and Privacy*. Washington, DC, Institute of Medicine, National Academy Press.

Evans D.A., Cimino J.J., Hersh W.R. et al. (1994). Toward a medical-concept representation language. *J. Am. Med. Inform. Assoc.* 1: 207.

Hammond W.E. (1993). Overview of health care standards and understanding what they all accomplish. *HIMSS Proceedings*, Chicago, American Hospital Association.

Hammond W.E., McDonld C., Beeler G. et al. (1994). Computer standards: Their future within health care reform. *HIMSS Proceedings*, Chicago, Health Care Information and Management Systems Society.

Harrington J.J. (1993). *IEEE P1157 MEDIX: A Standard for Open Systems Medical Data Interchange.* New York, Institute of Electrical and Electronic Engineers.

Health Care Financing Administration (HCFA) (1994). Draft issue papers developed by HCFA's national provider identifier/national provider file workgroups. Baltimore.

Health Industry Business Communications Council (HIBCC) (1994). Description of present program standards activity. Phoenix.

Humphreys B. (1991). Tutorial 20: Using and assessing the UMLS knowledge sources. Symposium on Computer Applications in Medical Care, Washington, DC.

Institute of Electrical and Electronics Engineers (IEEE) (1994). Trial-use standard for health care data interchange — Information model methods: Data model framework. IEEE Standards Department, New York.

Joint Commission on Accreditation of Health Organizations (JCAHO) (1994). The Joint Commission Journal on Quality Improvement. Oakbrook Terrace, IL.

Moshman Associates, Inc. (1994). Current activities of selected health care informatics standards organizations. Bethesda, MD. Office of Science and Data Development, Agency for Health Care Policy and Research.

National Committee for Quality Assurance (1993). Hedis 2.0: Executive summary. Washington, DC.

Rothwell D.J., Cote R.A., Cordeau J.P. et al. (1993). Developing a standard data structure for medical language — The SNOMED proposal. Washington, DC, SCAMC.

Terell S.A., Dutton B.L., Porter L. et al. (1991). In search of the denominator: Medicare physicians — How many are there? Baltimore, Health Care Financing Administration.

Waegemann C.P. (1994). Draft — 1994 resource guide: Organizations involved in standards and development work for electronic health record systems. Newton, Mass, Medical Records Institute.

Workgroup for Electronic Data Interchange (WEDI) (1993). WEDI report: October 1993. Convened by the Department of Health and Human Services, Washington, DC.

Further Reading

For copies of standards accredited by ANSI, you can contact the American National Standards Institute, 11 West 42d St., NY, NY 10036 (212), 642–4900. For information on ANSI Health Care Informatics Standards Planning Panel (HISPP), contact Steven Cornish (212) 642–4900.

For copies of individual ASTM standards, you can contact the American Society for Testing and Materials, 1916 Race Street, Philadelphia, PA 19103–1187 (215), 299–5400.

For copies of the "Proposal to Accelerate Standards Development for Computer-based Patient Record Systems," contact the Computer-based Patient Record Institute (CPRI), Margaret Amatayakul, 1000 E. Woodfield Road, Suite 102, Schaumburg, IL 60173 (708) 706–6746.

For information on provider identifier standards and proposals, contact the Health Care Financing Administration (HCFA), Bureau Program Operations, 6325 Security Blvd., Baltimore, MD 21207 (410), 966–5798. For information on ICD-9-CM codes, contact HCFA, Medical Coding, 401 East Highrise Bldg. 6325 Security Blvd., Baltimore, MD 21207 (410), 966–5318.

For information on site-of-care and supplier labeling identifiers, contact the Health Industry Business Communications Council (HIBCC), 5110 N. 40th Street, Suite 250, Phoenix, AZ 85018 (602), 381–1091.

For copies of standards developed by Health Level 7, you can contact HL7, 3300 Washtenaw Avenue, Suite 227, Ann Arbor, MI 48104 (313), 665–0007.

For copies of standards developed by the Institute of Electrical and Electronic Engineers/Engineering in Medicine and Biology Society, in New York City, call (212) 705–7900. For information on IEEE/MEDIX meetings, contact Jack Harrington, Hewlett-Packard, 3000 Minuteman Rd., Andover, MA 01810 (508), 681–3517.

For more information on clinical indicators, contact the Joint Commission on Accreditation of Health Care Organizations (JCAHO), Department of Indicator Measurement, One Renaissance Blvd., Oakbrook Terrace, IL 60181 (708), 916–5600.

For information on pharmaceutical billing transactions, contact the National Council for Prescription Drug Programs (NCPDP), 2401 N. 24th Street, Suite 365, Phoenix, AZ 85016 (602), 957–9105.

For information on HEDIS, contact the National Committee for Quality Assurance (NCQA), Planning and Development, 1350 New York Avenue, Suite 700, Washington, D.C. 20005 (202), 628–5788.

For copies of ACR/NEMA DICOM standards, contact David Snavely, National Equipment Manufacturers Association (NEMA), 2101 L. Street N.W., Suite 300, Washington, D.C. 20037 (202), 457–8400.

For information on standards development in the areas of computer-based patient record concept models, confidentiality, data security, authentication, and patient cards, and for information on standards activities in Europe, contact Peter Waegemann, Medical Records Institute (MRI), 567 Walnut, P.O. Box 289, Newton, MA 02160 (617), 964–3923.

43

Introduction to Informatics and Nursing

Kathleen A. McCormick
SAIC

Joyce Sensmeier
HIMSS

Connie White Delaney
The University of Minnesota

Carol J. Bickford
American Nurses Association

43.1 Introduction

Everyday in hospitals, critical care environments, ambulatory clinics, academic environments, and vendor corporations, nurses are working side-by-side with colleagues in engineering to develop, evaluate, and maintain information systems and other technology solutions in healthcare. Together these professional teams have evolved from a domain specific focus to the development and implementation of integrated systems with decision support that enhance patient care, verify if outcomes are met, assure quality safe care, and document resource consumption. This chapter provides an overview of the demography of the profession, definitions of nursing informatics, current nursing informatics activities, and examples of bridging healthcare informatics and nursing.

43.2 Demography

Registered nurses comprise the largest healthcare provider group in the United States. Their work environments include traditional healthcare settings such as hospitals, ambulatory care clinics, private practice settings, schools, correctional facilities, home health, community health, and public health environments. Nurses are the most frequent providers of healthcare services to homeless populations, faith communities, and other disenfranchised, underserved, and uninsured populations in the less traditional health related settings. In each setting, the registered nurse serves as the unrecognized knowledge worker, addressing the data, information, and knowledge needs of both patients and healthcare providers. Nurses' record keeping and written documentation provide integral support for the necessary communication activities between nurses and other healthcare providers and therefore must reflect the chronology of findings, decision-making processes, activities, evaluations, and outcomes.

The definition of nursing has evolved over the years. In 2003 the American Nurses Association published this contemporary definition that reflects the holistic and health focus of registered nurses in the United States:

> Nursing is the protection, promotion, and optimization of health and abilities, prevention of illness and injury, alleviation of suffering through the diagnosis and treatment of human response, and advocacy in the care of individuals, families, communities, and populations [ANA, 2003].

Like other professions, after initial educational preparation and licensure, the registered nurse may continue studying for preparation in a specialty practice, such as pediatrics, gerontology, cardiology, women's health, oncology, and perioperative nursing. Graduate preparation in clinical specialties may lead to designation as an advanced practice registered nurse or APRN, nurse anesthetist, or midwife. Others may be interested in preparing for a role specialty, such as administration, education, case management, or informatics.

In studies since the 1980s, nurses have been identified as the largest users of information systems in healthcare and play a critical role in creating an effective healthcare information infrastructure.

43.3 Nurses — Largest Group Using Computers and Implementing Information Systems

The 2000 National Sample Survey of Registered Nurses projects that 8406 or 0.4% of the approximately 2.7 million registered nurses in the United States identify nursing informatics as their nursing specialty (http://bhpr.hrsa.gov/healthworkforce/). Because of the increased national interest in healthcare informatics and implementation of healthcare information systems, coupled with the increased numbers of graduate nursing informatics educational programs, the 2004 National Sample Survey of Registered Nurses is expected to report a greater prevalence of informatics nurses.

Nearly three-quarters of nurse informaticists are currently developing or implementing clinical documentation systems according to a first-of-its-kind survey performed by the Healthcare Information and Management Systems Society (HIMSS) and sponsored by Omnicell, Inc. [HIMSS, 2004a]. A total of 537 responses were received to the web-based survey, which was performed to gain a better understanding of the background of nurse informaticists, the issues they address and the tools they use to perform their jobs. Two-thirds of respondents reported that systems implementation is their top job responsibility.

Drawing on their extensive clinical background, another three-quarters are involved with their organization's clinical information systems implementation. Over half are involved in the development or implementation of computerized provider order entry (CPOE), and 48% are developing or implementing an electronic medical record (EMR). Nurse informaticists are also actively involved in developing or implementing Bar Coded Medication Management, ICU, Master Patient Index (MPI) and Picture Archival and Communication Systems (PACS).

Approximately half of the respondents indicated that they were involved with five or fewer areas of development or implementation. Conversely, 12% of respondents are involved in ten or more areas. Many of the nurses (46%) indicated that they have been involved with the removal/replacement of at least one system.

43.3.1 Expanding Nursing Informatics Roles

Like the expanding roles of other registered nurses, the role of the informatics nurse reflects significant diversity and expertise. Consider the listing provided in the Scope and Standards of Nursing Informatics Practice [ANA, 2001b]: project manager, consultant, educator, researcher, product developer, decision support/outcomes manager, advocate/policy developer, entrepreneur, chief information officer, and business owner. Some of these experts are purchasers of information systems in hospitals, outpatient settings, community and home care nursing environments. Information system vendors have begun designating a senior executive level nurse, Chief Nursing Officer (CNO) to direct the informatics nurse contingent and patient care software development components of the organization, quite like the CNO role in a hospital or multifacility healthcare enterprise. Recent job announcements posted at the HIMSS, the American Medical Informatics Association (AMIA), and the Capitol Area Roundtable in Nursing Informatics (CARING) Web sites sought individuals for systems analyst, database administrator, and implementation specialist positions.

43.4 Definition of "Nursing Informatics"

Health informatics comprises multiple discipline-specific informatics practices; nursing informatics is one of these specialties. Nursing informatics, an applied science, is defined by the American Nurses Association [2001] as a specialty that:

> integrates nursing science, computer science, and information science to manage and communicate data, information, and knowledge in nursing practice. Nursing informatics facilitates the integration of data, information, and knowledge to support patients, nurses, and other providers in their decision-making in all roles, and settings. This support is accomplished through the use of information structures, information processes, and information technology (p. 17).

43.4.1 Informatics Nurses Have a United Voice

Increasing numbers of local and regional networking groups of informatics nurses prompted the Nursing Informatics Working Group of the American Medical Informatics Association (AMIA), the professional nurses represented by the Healthcare Information and Management Systems Society (HIMSS), and the American Nurses Association (ANA) to foster and support the recent development of the Alliance for Nursing Informatics (ANI). This new entity represents more than 2000 nurses and brings together 18 distinct nursing informatics groups (see Table 43.1) in the United States that function separately at local, regional, national, and international levels and have established programs, publications, and organizational structures for their members [HIMSS, 2004c]. The basic objectives of the Alliance are to:

1. Provide a consolidated forum for the informatics community
2. Provide input to a national nursing informatics research agenda
3. Facilitate the dissemination of nursing informatics best practices
4. Present the collective voice of the nursing informatics specialty in national public policy initiatives and standards activities

TABLE 43.1 Eighteen Distinct Nurses Informatics Groups in the U.S. That Have Come Together as an Alliance for Nursing Informatics (ANI) 2004

AMIA	American Medical Informatics Association Nursing Informatics Working Group
ANA	American Nurses Association (Liaison)
ANIA	American Nursing Informatics Association
BANIC	Boston Area Nursing Informatics Consortium
CARING	Capital Area Roundtable on Informatics in Nursing
CSRA-CIN	Central Savannah River Area Clinical Informatics Network
CHIN	Connecticut Healthcare Informatics Network
DVNCN	Delaware Valley Nursing Computer Network
HINJ	Health Informatics of New Jersey
HIMSS	Healthcare Information and Management Systems Society Nursing Informatics Community Iowa HIMSS Nursing Informatics Committee
INFO	Informatics Nurses From Ohio
MNIN	Michigan Nursing Informatics Network
MINING	Minnesota Nursing Informatics Group
CONI	North Carolina State Nurses Association Council on NI
NISCNE	Nursing Information Systems Council of New England
PISUG	Perinatal Information Systems User Group
PSNI	Puget Sound Nursing Informatics
SCINN	South Carolina Informatics Nursing Network
UNIN	Utah Nursing Informatics Network

In one of its first joint efforts, the nurses represented by the ANI provided testimony to the President's Information Technology Advisory Committee (PITAC) during an open meeting on April 13, 2004. In part, this testimony focused on the benefits of creating an effective health care information infrastructure in all settings for all healthcare providers.

43.5 Nursing Process

Most nurses have been prepared in their educational programs to use the nursing process as a framework to guide thinking and professional practice. Assessment, diagnosis or problem/issue definition, planning, implementation, and evaluation comprise the steps in the nursing process. Employers value the demonstrated expertise and critical thinking skills of the informatics nurse who uses the nursing process. The nursing process serves as the foundation for the *Scope and Standards of Nursing Informatics Practice* [ANA, 2001b], which provides specific standards of practice and standards of professional performance statements that assist the informatics nurse in practice. The content can be used when developing position descriptions and performance appraisals, and also provides a structure for informatics curriculum development for educators and a research agenda for nursing and interdisciplinary groups such as bioengineers and nurses.

43.5.1 Ethics and Regulation

Registered nurses have a long tradition of concern about ethics, patient advocacy, safety, and quality of care. Beginning with the first clinical experience, the registered nurse must know the differences and necessary practice associated with privacy, confidentiality, and security. Just as for registered nurse colleagues, the Code of Ethics for Nurses With Interpretive Statements [ANA, 2001a] provides a framework for the informatics nurse. Although primarily focused on support activities for the healthcare environment, the informatics nurse has the obligation to be concerned about issues of confidentiality, security, and privacy surrounding the patient, clinician, and enterprise and the associated data, information, and knowledge.

The federal government's current focus on establishing the National Health Information Network (NHIN), electronic health record (EHR), personal health record, and regional health information

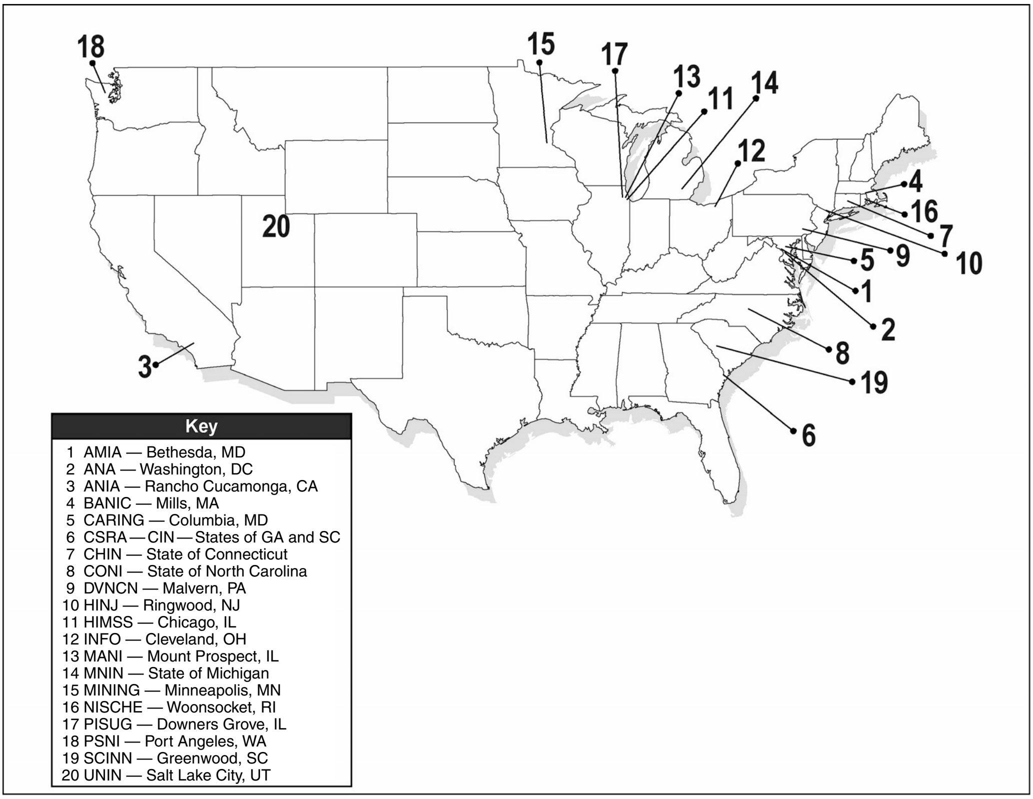

18
15
17
13
11
14
12
4
16
7
10
9
5
1
2
8
19
6
3
20
Key
1 AMIA — Bethesda, MD
2 ANA — Washington, DC
3 ANIA — Rancho Cucamonga, CA
4 BANIC — Mills, MA
5 CARING — Columbia, MD
6 CSRA — CIN — States of GA and SC
7 CHIN — State of Connecticut
8 CONI — State of North Carolina
9 DVNCN — Malvern, PA
10 HINJ — Ringwood, NJ
11 HIMSS — Chicago, IL
12 INFO — Cleveland, OH
13 MANI — Mount Prospect, IL
14 MNIN — State of Michigan
15 MINING — Minneapolis, MN
16 NISCHE — Woonsocket, RI
17 PISUG — Downers Grove, IL
18 PSNI — Port Angeles, WA
19 SCINN — Greenwood, SC
20 UNIN — Salt Lake City, UT

FIGURE 43.1 The figure shows a U.S. map with representative locations for some of these regional groups. The figure also lists the nursing informatics organizations that are affiliating with ANI.

organizations (RHIO) provides numerous ethical and regulatory issues [Thompson and Brailer, 2004]. For example, the current U.S. healthcare environment has yet to resolve the problem of clinical practice and licensure across state lines for individuals working with nurse call centers, telehealth applications, and electronic prescriptions. Another example is related to unequal distribution of resources, or the digital divide which is characterized by those without a working personal computer and high-speed access to the Internet at home.

Consider the ethical issues associated with the data and information management of genetic databases. Data integrity, appropriate database structuring, standardized terminologies and indexing processes, and correct representation of complex multidimensional structures and images pose new ethical issues. What assurances must be in place to prevent public dissemination of an individual's adverse genetic profile? Will that prevent these individuals from securing health insurance, cause termination from a job, or lead to discrimination in schools or the hiring process? Similarly, the move toward increased patient participation in clinical decision making continues to create tensions in the decision support and information systems development arenas. Informatics nurses join their colleagues in biomedical engineering and in clinical practice to identify and resolve such ethical questions.

43.6 Standards in Vocabularies and Data Sets

Nursing has been developing nomenclatures for over 24 years to address the nursing process components of diagnosis, interventions, and outcomes. Table 43.2 lists the ANA recognized terminologies supporting nursing practice in 2004. More recently the International Organization for Standardization (ISO) of

TABLE 43.2 ANA Recognized Terminologies Supporting Nursing Practice, 2004

ABC codes
Clinical Care Classification (CCC) [Formerly Home Health Care Classification]
International Classification for Nursing Practice (ICNP®)
Logical Observation Identifiers Names and Codes (LOINC®)
NANDA-Nursing Diagnoses, Definitions, and Classification
Nursing Outcomes Classification (NOC)
Nursing Management Minimum Data Set (NMMDS)
Nursing Interventions Classification System (NIC)
Nursing Minimum Data Set (NMDS)
Omaha System
Patient Care Data Set (PCDS)
PeriOperative Nursing Data Set (PNDS)
SNOMED CT®

Geneva, Switzerland, recently published an international standard for nursing. The standard — Health Informatics: Integration of a Reference Terminology Model for Nursing — is the first step toward creating comparable nursing data across settings, organizations, and countries. Such data assists in identifying and implementing "best nursing practices" or determining how scarce nursing resources should be spent. The International Medical Informatics Association (IMIA) Nursing Informatics Working Group developed the standard in collaboration with the International Council of Nurses (ICN) [Saba et al., 2003]. This collaboration was accomplished as a result of the international network of nurses who in turn were developing standards for nomenclature and classifications within their respective countries. This group recognized the value of an international collaborative effort to compare quality, efficiencies, and outcomes of care resulting from nursing care that is delivered internationally.

In addition, numerous nurses have been working on other ISO TC-215 and Health Level Seven (HL7) standards, Logical Observation Identifiers Names and Codes (LOINC), and Systematized Nomenclature of Medicine (SNOMED) committees. Wherever health standards are being developed and applied, nurses are present on those committees and working groups to provide input that often includes consideration of professional nursing standards.

43.7 Clinical Information Systems

Healthcare information systems that adequately support nursing practice have not emerged over the past decades. Figure 43.2 identifies some of the system components, influencing factors, and relationships that have not been fully considered in describing the complexity of nursing, a profession that relies so heavily on evidence, knowledge, and critical thinking. Consequently the requisite detailed analysis and design processes have never begun or have failed to generate the appropriate and diverse information system components necessary for successful support for nurses and nursing practice.

Recent research has begun to identify the relationships of computer and information literacy of nurses and the use of information systems [McNeil, 2003]. Nurses' experience with different computer applications corresponds with higher confidence in using a new information system, according to a recent study [Dillon et al., 2003]. The study, which surveyed 139 nurses at a 450-bed regional hospital center, also found that this "self-efficacy" is higher, on average, for younger nurses and those with more advanced degrees.

Using a computer at home and having general computer skills, for example, were associated with higher levels of self-efficacy according to the researchers. Nurses' self-assessed ability to use word processing, conduct Internet searches, and use e-mail also corresponded with higher self-efficacy.

The study also found that "a younger age, a higher level of education, and a more positive attitude toward the new information system had a slight association with self-efficacy." "Improvement of nurses'

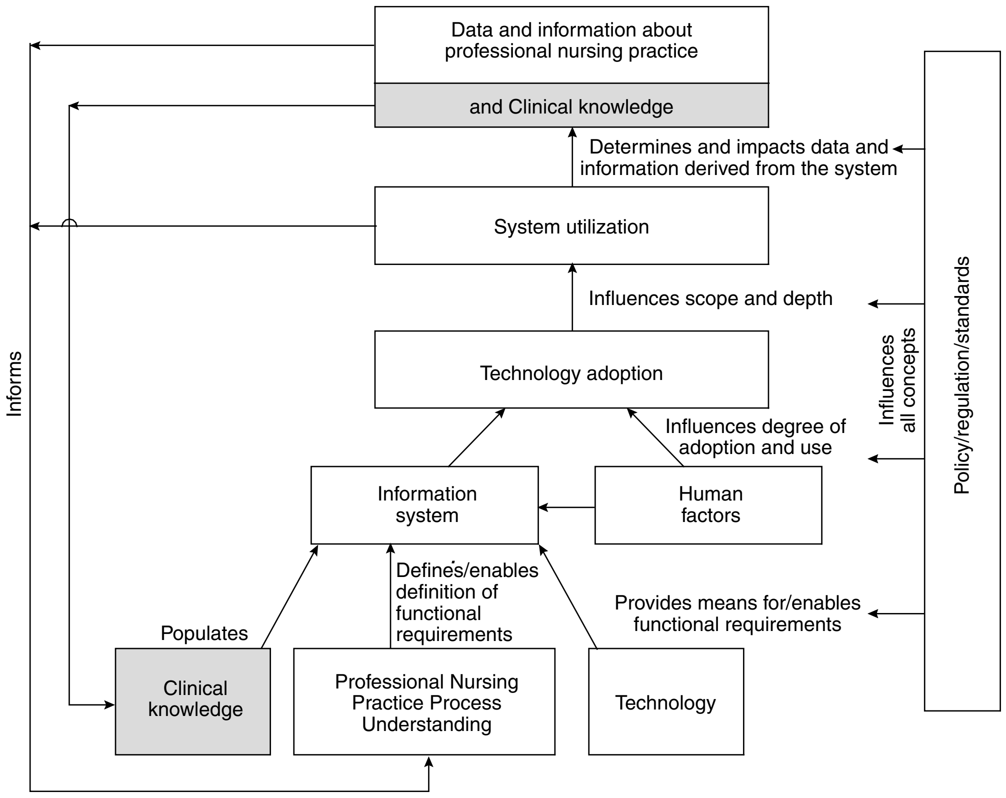

Data and information about professional nursing practice
and Clinical knowledge
Determines and impacts data and information derived from the system
System utilization
Influences scope and depth
Technology adoption
Influences degree of adoption and use
Information system
Human factors
Defines/enables definition of functional requirements
Provides means for/enables functional requirements
Populates
Clinical knowledge
Professional Nursing Practice Process Understanding
Technology
Informs
Influences all concepts
Policy/regulation/standards

FIGURE 43.2 The organizing framework for clinical information systems: critical knowledge as the critical factor.

self-efficacy toward the system will not guarantee a successful implementation," the study concluded, "but it is expected that with a more organized and strategic approach to implementation, the adoption of a new information system will be enhanced."

43.8 Bridging Nursing and Engineering Specialties: The Bioinformatics Partnership of Nurses and Engineers

Bioinformatics by definition focuses on "how information is represented and transmitted in biological systems, starting at the molecular level" [Shortliffe et al., 2001]. Accepting the central goal of nursing informatics as improving the health of populations, communities, families, and individuals by optimizing information management and communication, it is clear that the intersection of these two fields — the biological systems of human beings and the applied/clinical practice of nursing — creates a synergy related to direct provision of care, and administrative, educational, and research priorities.

Bioinformatics focuses on the representation, communication, and management of information related to basic biological sciences and biological processes. It is readily dependent upon using integrative models to expand and disseminate the science as well as providing clinically relevant contributions. Although nursing predominantly focuses on aspects of clinical care, characteristics of nursing, especially in decision making, particularly position this discipline and its nursing informatics specialty as productive partners with bioinformatics. These nursing characteristics can contribute to the bridging of bench science with input on care needs, as well as the clinical significance of innovations put into real world practice.

Nursing contributes to the data, information, and knowledge supporting patients, families, and communities across all settings. Two key threads are significant to information related to the biological aspects of people (1) nursing is positioned to assess not only individual information, but also information related to families and communities and (2) this assessment (information access) occurs across all settings. The Human Genome Project is a key revolution illustrating the importance of information from the individual and family levels across all settings of care to the aggregate. Consider, for example, genetic diseases, pharmaceutical discoveries, and healthy aging.

Moreover, collaboration with nurse scientists offers additional information not commonly the focus of the medical record. Nursing by its very nature is context dependent and consequently has collected information related to the patient/family context, domain, as well as the environment and domain of healthcare delivery. The knowledge discovery in databases research strengthens informatics capacity to consider the interrelationships of cellular societal systems.

This collaborative capacity extends beyond the cellular focus of bioinformatics. Addressing the complexity of healthcare systems and process can benefit from the collaboration with nursing. As the central information broker across all settings, nursing can be a pivotal partner in examining and designing efficient safe care systems. Studying and extending the science of complex adaptive systems from cellular to societal levels is one example.

Nursing is well positioned to partner with bioinformatics and engineering to address the essential demands of the healthcare system. A substantial cadre of nursing informatics scientists has been prepared, some even with Ph.D.s in engineering. Specifically, there are a growing number of scientists in nursing informatics with knowledge discovery expertise, encompassing multiple Knowledge Discovery in Databases (KDD) methods, including complex adaptive systems in patient care and with the consumer. Nursing has maintained crucial expertise in knowledge representation and information messaging, both essential to bioinformatics, nursing, and the interface area between the two disciplines. And significantly, nursing informatics has established a national network to support bridging the disciplines.

43.8.1 Benefits of Using Information Systems in Patient Care

Two types of data are available related to the benefits of utilizing information systems in the delivery of patient care. Early in the last decade, researchers began assessing the value of computer terminals at the bedside or at the point of care. In one study of the impact of bedside terminals on the quality of nursing documentation, Marr et al. [1993] found that comprehensiveness of documentation measured by the presence or absence of components of the record was better with bedside terminals. They also found that timeliness of documentation was improved when performed closer to the actual time that care was delivered. Other benefits of nurses using computers to document practice were (1) the integration of care plans with nursing interventions, (2) calculation of specific acuity, and (3) automatic bills for nursing services.

The Nicholas E. Davies Award of Excellence for Electronic Health Records recognizes excellence in the implementation of EHRs in healthcare organizations and primary care practices. Established in 1995 this program has recognized 19 hospitals in the past 9 years. As part of the application process, organizations must document the financial impact of their implemented EHR. A 2001 winner, the University of Illinois at Chicago Medical Center, documented that during a two year period, $1.2 million of nurse time was reallocated from manual documentation tasks to direct hands-on patient care [CPRI-HOST, 2001]. Registered nurses in the charge nurse role gained 2.75 h per shift in the medication administration process.

Data are also available on the computerization of evidence-based practice recommendations [Saba and McCormick, 2005]. Although these data are sparse, existing publications cover a diverse range of topics from the integration of information technology with outcomes management, coding, and taxonomy issues relevant to outcomes, including standardized language and other issues tied to the nursing minimum data set, and the development of nursing-sensitive outcome measures from nursing care and interventions. Former studies focusing on outcomes suggest that nurses should serve on multidisciplinary teams and collaborate with others in building IT systems to improve organizational learning, use of evidence, and quality [McCormick, 2005].

Other reported studies of bedside terminals have documented the ease of use, elimination of redundant data, system support at the bedside, availability, currency of data, and access to expert systems. Use of clinical information systems have been shown to provide soft benefits related to improvements in patient safety, care provider communication, and workflow enhancements. Historically there have been few studies to show hard benefits of dollars saved, or a substantial decrease in care hours. However, as more nursing-focused software is implemented a positive return on investment (ROI) has been noted [Curtis, 2004]. Nursing and engineering must work together to articulate the shared benefits achieved with successful systems implementations.

43.9 Barriers to Creating an Effective Healthcare Information Infrastructure

The survey of nurse informaticists identified financial resources as the largest barrier to success in their role of implementing healthcare information infrastructures. This is consistent with the responses of chief information officers (CIOs) as reflected in the 15th Annual Leadership Survey [HIMSS, 2004]. Lack of user acceptance or administrative support, and software design that ignores current workflow processes were also described as barriers. Healthcare CIOs in this survey have recognized the importance of a clinical champion by identifying the need to increase IT staff to address clinical issues.

A 1997 review of information technology use by nurses suggested that barriers include: lack of integration of nursing systems within hospital information systems, the need for a unified nursing language, and lack of point-of-care terminals [Bowles, 1997]. However, more recent reviews describe that these barriers are slowly being resolved [Androwich et al., 2003]. Additional barriers include lack of a standard design for clinical systems that would address the inconsistency of entering or extracting data from the end user's perspective [Hermann, 2004].

In another study recently undertaken by the Interagency Council on Information Resources for Nurses (ICIRN), the information literacy of nurses was surveyed Tanner et al. [2004]. Information literacy was identified as a nursing informatics competency for the nurse. Information literacy has been found to be an essential element in the application and use of evidence based practice (EBP). The study identified the gaps in knowledge and skills for identifying, accessing, retrieving, evaluating, and utilizing research evidence to provide best practice for patients. The study also reported that over 64% of nurses regularly need information, but 43% rated workplace information resources as totally inadequate or less than adequate. The three primary organizational constraints in the practice settings were identified as (1) the presence of other goals of higher priority, (2) difficulty recruiting and retaining staff, and (3) organizational budget for acquisition of information resources. Three personal barriers were identified (1) lack of understanding of organization or structure of electronic databases, (2) difficulty accessing information, and (3) lack of skills to use and synthesize evidence into practice.

Findings from another study indicated that nurses at every level and role exhibit large gaps in knowledge and competencies at each step in the information literacy process — from lack of awareness that they need information to lack of access or ability to successfully search and utilize information needed for practice, particularly in an electronic format [Pravikoff et al., 2003]. However, significantly more nurses who received their most recent nursing degree after 1990 acknowledged successful searches of evidence in the National Library of Medicine (NLM) Medline and Cumulative Index to Nursing and Allied Health Literature (CINAHL) when compared with those who graduated before 1990 [Tanner, 2000]. Thus, training in computer skills has been a barrier that may diminish with more nurses becoming computer literate in high school and college [Pierce, 2000]. The "tipping point" may be a generation gap.

Most nurses did not have training regarding computer use or typing skills in their nursing or college curriculum unless they returned to school for further education after 1990 [Gloe, 2004]. Therefore, the practicing nurse should be provided with opportunities for learning basic “keyboarding,” that is, typing skills and computer basics such as how to use the mouse. This would lessen the anxiety for computer use. Nurses are caring for patients with greater acuity and requiring more documentation, thus creating higher stress than ever before. Adding the stress of computerization when one is totally unfamiliar with the computer can be a major challenge. Incorporating this challenge into their current workload may seem like an insurmountable task to nurses and could be a large barrier for EHR implementation. Thus, providing educational opportunities as well as designing user friendly applications for use by the nursing staff can lessen the stress and remove barriers.

Summarizing an Agency for Healthcare Research and Quality (AHRQ) conference examining quality research, the attendees asked if the primary barriers to achieving the NHIN are more political than technical. The participants identified the need for standards to govern the infrastructure, the lack of broad-based agreement among stakeholders of a system concept, and the legal concerns about privacy and confidentiality [Lang and Mitchell, 2004]. Vahey et al. [2004] summarized the need for additional research on the barriers to quality improvement that they defined as: lack of standardized measures, inadequate information systems to collect data, inadequate resources to pay for data collection and translation, and technological issues. As stated by others at the same conference [Lamb et al., 2004] the necessary demand and incentives are currently not in place to assure the development of information infrastructures that will support quality improvement.

Finally, in a monograph sponsored by AMIA in collaboration with the ANA, the major constraints to full implementation of the information infrastructure for nursing were identified as lack of: policy, regulation and standards, technology, information systems, human factors, technology adoption, and system utilization [Androwich et al., 2003]. Other barriers include lack of a positive ROI for the use of clinical documentation systems, and limited availability of applications that specifically support the work of nurses.

Many reports, such as those recently published by the Institute of Medicine (IOM), Committe on Quality in Health Care support the use of technology in reducing medical errors and encourage implementation of evidence-based healthcare practice. The results of these recent studies identify gaps and barriers that limit effecting these prescribed practices among nurses, the largest number of health care professionals who provide the greatest percentage of direct contact, time, and intensity with the consumer/patient.

43.9.1 Opportunities to Create an Effective Healthcare Information Infrastructure

Patient safety is a well-documented priority for healthcare organizations. This focus provides an opportunity for healthcare organizations to evaluate the use of information technology and the related infrastructure to deliver safe and effective patient care. Computerized provider order entry (CPOE), clinical information systems, and bar coded medication management are three top applications for healthcare organizations in the next several years as reported in the survey of CIOs [HIMSS, 2004b]. Nurses play a critical role, as almost three-quarters of respondents are involved with the implementation of their organization’s clinical information system, 52% with the implementation of CPOE software, and 48% with implementation of an EMR. The extensive clinical background of nurse informaticists is valuable, as nurses have an intimate understanding of the workflow, environment, and procedures that are necessary to achieve success.

Another opportunity that could also be a potential barrier for creating an effective healthcare information infrastructure, relates to leadership and a clear strategic vision [Kennedy, 2004]. The complexity of creating a healthcare information infrastructure is immense and can only be developed once it has been fully defined. Several efforts have been initiated in this direction one at the conceptual level (i.e., IOM’s definition of the Computer-based Patient Record) and the other at the detailed data level (i.e., HL7’s Reference Data Model). Yet the missing link may be the discussion about how to distill this information into practical, usable models that can be applied to improve the work environment of the nurse and enhance patient safety. An innovative approach is needed to create a cohesive agenda across a very dynamic, complex healthcare delivery organization. Once an "effective" healthcare information infrastructure is defined, it could be recognized and promoted from both a nursing and collaborative health care model.

43.10 Research and Development Needs

The nursing profession has both a critical mass of nurses involved in information technology and experience with implementing technology and related systems. Yet there is a need to educate nurses involved with implementing and utilizing information infrastructures. The focus has shifted nationally from developing an information infrastructure, to using an information infrastructure to assure safe, effective, patient-centered, timely, efficient, and equitable care. Information technology is the critical tool to be used in the redesign of systems supporting the delivery of care to achieve the type of quality care recommended by the current federal leaders and the recent IOM reports.

The nursing profession is a participant in the delivery of care in the hospital and outpatient environment, and is centrally placed in community and home care. There is a need to create centers for evaluation of the barriers and benefits of information infrastructures. These centers should be located in academic centers or developed in cooperation with commercial developers of information systems. These centers should be multidisciplinary and nursing focused. Nurses must be involved in these types of research and evaluation efforts.

The National Institute on Nursing Research (NINR) developed priorities for research in nursing informatics in 1993. These included (a) formalization of nursing vocabularies, (b) design and management of databases for nursing information, (c) development of technologies to support nursing practice, (d) use of telecommunications technologies in nursing, (e) patient use of information, (f) identification of nurses' information needs, and (g) systems modeling and evaluation [NINR, 1993]. While NINR has funded a number of studies in these areas and achieved outcomes in advancing nursing informatics, the Institute has never received sufficient funds to disseminate the findings, translate them to practice, or expand the individual studies to the support of Centers of Excellence.

Other research has been recommended in the areas of (1) prototyping methodology to explore specific ways to realize innovations, (2) pilot tests of technology based innovations and new workflow processes, (3) analyses of successful and unsuccessful outcomes and change processes in the implementation of systems, and (4) demonstrations of the return on investment from nursing documentation on such areas as patient safety, errors, quality, effectiveness, and efficiencies [Androwich et al., 2003].

The Health Resources and Services Administration (HRSA) Division of Nursing has funded training in nursing informatics. Enhanced budgets would provide the necessary funds to expand those programs with a focus on the literacy gaps in nursing at both the academic levels and in practice areas.

According to a workgroup report to the American Academy of Nursing Technology and Work Force Conference, the ideal nursing care-delivery system enables staff nurses to increase their productivity, job satisfaction, and the quality of care by increasing the time spent on direct care activities [Sensmeier et al., 2002]. The report recommends that this system must include information technology that replaces the paper-based, administrative tasks with a paperless, point-of-care, computer-based patient record imbedded with intelligent, rules-based capabilities that automate the manual workflow processes, policies, and procedures, and that support the nurses' critical thinking. Research to explore these recommendations is needed.

Other target areas for practice-based research identified by the Boston Area Nursing Informatics Consortium (BANIC) [Kennedy, 2004] include, workflow and workplace design for point-of-care applications, creating a model for systems value at the point of care, clinical systems implementation, clinical documentation and utilization of standardized nursing language, clinical systems evaluation tools, consumer health systems, web-based education, and methods for evaluating applications, specifically CPOE and medication management.

The AHRQ has also funded research in nursing informatics and the impact on quality and evidence-based practice. One such study demonstrated that installing a computerized medical information management system in hospital intensive care units can significantly reduce the time spent by ICU nurses on documentation, giving them more time for direct patient care [AHRQ, 2003]. Follow-up studies are needed to validate these findings.

The AHRQ hosted a conference where nurses helped the Agency define research priorities. The conferees concurred that the sole reason to design and implement clinical information systems was to use the information to track, interpret, and improve quality of care. They identified information systems technology as the crucial bridge to translate research into practice [Lamb et al., 2004]. They suggested that a research emphasis should be placed on reducing the barriers to getting timely and credible information to the nurses. The group identified the gap between collecting data to measure quality, and using the data to improve the quality of care. They identified this as the greatest opportunity for achieving the ROI for developing systems. Information technology supported by evidence-based practice and patient safety initiatives is key to a quality agenda. However, the information technology is available, so the barriers to achieving its benefits must be evaluated and solutions developed so that direct providers and decision makers can implement systems [Lamb et al., 2004].

Conference participants recommended the following research goals relative to nursing informatics: (1) standardized quality indicators need development and measurement including nurse-sensitive measures from nursing care and interventions measures, and quality indicators to link care across settings, (2) improved risk adjustment methodologies, and (3) a national health information structure to inform quality improvement efforts [Vahey et al., 2004]. They further cited the need for integration and collaboration among stakeholders in conducting research on the information systems' needs to measure quality, improvements in patient safety, and risk and error reduction.

Since nursing practice is often absent in databases and systems of reimbursement from private and public sources, other recommended research has focused on the inclusion of nursing-sensitive quality indicators achieved from nursing care and interventions, in conjunction with the analysis of workforce and contributions of advanced practice nurses (nurse practitioners) [Brooten et al., 2004]. Lamb et al. [2004] stressed that the critical research question in the quality initiative involves the complex analysis of the interplay between information infrastructure systems, organization, financial and clinical practice features. Not only did this group recommend that AHRQ expand their research agenda for nursing sensitive areas, but they also recommended broader dissemination of the results of the impact of nursing staffing on achieving quality outcomes.

The Center for Disease Control and Prevention (CDC) has sponsored research in public health and biodefense information infrastructure. While nurses have been involved in implementing these programs, the impact of nursing research on these areas could benefit from additional funding.

The Center for Medicare and Medicaid Services (CMS) provides data from which health services research nurses have evaluated the impact of nursing care on quality, effectiveness, and efficiencies. Researchers have identified that quality indicators of care from the largest group of health care workers, namely nurses, are absent in the databases and subsequently the systems of reimbursement from CMS [Brooten et al., 2004]. The utilization of cooperative agreements and contracts for evaluation of the ROI for nursing information systems embedded in health care information systems has not been widespread. The pilots and demonstration studies recommended above could be facilitated by CMS contracts and grants. Models of incentives could be studied to determine how CMS could include nursing documentation data in health care records to evaluate quality, outcomes, and cost impacts. Constructing nursing-sensitive quality indicators from existing databases and establishing their validity needs further research. Medicare and Medicaid data are incomplete representations of nursing's contributions to quality outcomes. Lamb et al. further recommended that systems be created and maintained to assure that the quality data can be captured at the point of care and translated into useful clinical information to be applied by nurses and other health professionals [Lamb et al., 2004].

Finally, the National Library of Medicine has sponsored research that has helped advance the literacy and impact of nursing informatics. Further targeted funding for nursing informatics and consumer informatics would be required to accelerate the national health information infrastructure in the United States.

43.11 Summary

Development of an increased awareness of nursing activities in informatics has been the major objective in preparing this chapter. This awareness thereby promotes the establishment of better partnerships between engineers, biomedical engineers, and the nurses working in informatics as well as in practice, administration, research, and education. Wherever nurses are participating in health care, they can team with engineers to develop better solutions to improve health care processes and outcomes.

References

AHRQ (2003). Case Study Finds Computerized ICU Information System Can Significantly Reduce Time Spent by Nurses on Documentation. Press Release Date: October 10, 2003. http://www.ahrq.gov/news/press/pr2003/compicupr.htm

American Nurses Association (2001a). *Code of Ethics for Nurses with Interpretive Statements*. Washington, DC: American Nurses Publishing.

American Nurses Association (2001b). *Scope and Standards of Nursing Informatics Practice*. Washington, DC: American Nurses Publishing.

American Nurses Association (2003). *Nursing's Social Policy Statement*, 2nd ed. Washington, DC: American Nurses Publishing.

Androwich, I.M., Bickford, C.J., Button, P.S., Hunter, K.M., Murphy, J., and Sensmeier, J. (2003). *Clinical Information Systems: A Framework for Reaching the Vision*. Washington, DC: American Nurses Publishing.

Bowles, K.W. (1997). The benefits and barriers to nursing information systems. *Computers in Nursing* 15, 191–196.

Brooten, D., Youngblut, J.M., Kutcher, J., and Bobo, C. (2004). Quality and the nursing workforce: APNs, patient outcomes, and health care costs. *Nursing Outlook* 52, 45–52.

Committee on Quality of Health Care in America, Institute of Medicine. *Crossing the Quality Chasm: A New Health System for the 21st Century* (2001). Washington, DC: National Academy Press.

CPRI-HOST (2001). The Seventh Annual Nicholas E. Davies Award Proceedings, November.

Curtis, K. Personal correspondence, March 29, 2004.

Dillon, T.W., Lending, D., Crews, T.R., and Blankenship, R. (2003). Nursing self-efficacy of an integrated clinical and administrative information system. CIN: Computers, Informatics, Nursing, 21, 198–205.

Gloe, D. Personal correspondence, March 24, 2004.

Hermann, B. Personal correspondence, March 30, 2004.

HIMSS (2004a). HIMSS Nursing Informatics Survey, February 23, 2004. http://www.himss.org/content/files/nursing_info_survey2004.pdf

HIMSS (2004b). 15th Annual HIMSS Leadership Survey. http://www.himss.org/2004survey/ASP/index.asp

HIMSS (2004c). Nursing Informatics Groups form Alliance through HIMSS and AMIA to Provide Unified Structure. Press Release: Date October 19, 2004. AMIA and HIMSS. http://www.himss.org

Kennedy, M. Personal correspondence, March 30, 2004.

Lamb, G.S., Jennings, B.M., Mitchell, P.H., and Lang, N.M. (2004). Quality agenda: Priorities for action-recommendations of the American academy of nursing conference on health care quality. *Nursing Outlook* 52: 60–65.

Lang, N.M. and Mitchell, P.H. (2004). Guest editorial: Quality as an enduring an encompassing concept. *Nursing Outlook* 52: 1–2.

Marr, P., Duthie, E., and Glassman, K. (1993). Bedside terminals and quality of nursing documentation. *Computers in Nursing* 11, 176–182, 1993.

McNeil, B., Elfrink, V., Bickford, C., Pierce, S., Beyea, S., Averill, C., and Klappenback, C. (2003). Nursing information technology knowledge, skills, and preparation of student nurses, nursing faculty, and clinicians: a U.S. survey. *Journal of Nursing Education* 42: 341–359.

McCormick, K.A. (2005). Translating Evidence into Practice: Evidence, Clinical Practice Guidelines, and Automated Implementation Tools. In *Essential of Nursing Informatics*, 4th ed. Saba, V.K. and McCormick, K.A. New York: McGraw-Hill.

NINR. Report on Nursing Information, US PHS, 1993.

Pierce, S. (2000). Readiness for evidence-based practice: Information literacy needs of nursing faculty and students in a southern U.S. state. DAI, 62(12B), 5645. Accession no. AAI3035514.

Pravikoff, D., Pierce, S., and Tanner, A. (2003). Are nurses ready for evidence-based practice? *American Journal of Nursing* 103, 95–96.

Saba, V.K. and McCormick, K.A. (2005). *Essentials of Nursing Informatics*, 4th ed. New York: McGraw-Hill.

Saba, V., Coenen, A., McCormick, K., and Bakken, S. (2003). Nursing Language: Terminology Model for Nursing. *ISO Bulletin* September. http://www.iso.ch/iso/en/commcentre/isobulletin/articles/2003/pdf/terminology03-09.pdf

Sensmeier, J., Raiford, R., Taylor, S., and Weaver, C. (2002) Using Innovative Technology to Enhance Patient Care Delivery. American Academy of Nursing Technology and Workforce Conference, Washington, DC July 12–14 last accessed at: http://www.himss.org/content/files/AANNsgSummitHIMSSFINAL_18770.pdf

Shortliffe, E., Perreault, L., Wiederhold, G., and Fagan, L. (2001). *Medical Informatics: Computer Applications in Health Care and Biomedicine*. 2nd ed. New York: Springer-Verlag.

Tanner, A., Pierce, S., and Pravikoff, D. (2004) Readiness for Evidence-Based Practice: Information Literacy Needs of Nurses in the US. MedInfo 2004. 11 (Part 2) 936–940.

Tanner, A. (2000). Readiness for evidence-based practice: Information literacy needs of nursing faculty and students in a southern U.S. state. DAI 62(12B), 5647. Accession no. AAI303551.

Thompson, T.G. and Brailer, D.J. (2004). The Decade of Health Information Technology: Delivering Consumer-centric and Information-rich Health care. Framework for Strategic Action. U.S. Department of Health and Human Services.

Vahey, D.C., Swan, B.A., Lang, N.M., and Mitchell, P.H. (2004). Measuring and improving health care quality: Nursing's contribution to the state of the art. *Nursing Outlook* 52: 6–10.

44

Non-AI Decision Making

Ron Summers
Loughborough University

Derek G. Cramp
Ewart R. Carson
City University

Non-AI decision making can be defined as those methods and tools used to increase information content in the context of some specific clinical situation without having cause to refer to knowledge embodied in a computer program. Theoretical advances in the 1950s added rigor to this domain when Meehl argued that many clinical decisions could be made by statistical rather than intuitive means [1]. Evidence of this view was supported by Savage [2], whose theory of choice under uncertainty is still the classical and most elegant formulation of subjective Bayesian decision theory, and was very much responsible for reintroducing Bayesian decision analysis to clinical medicine. Ledley and Ludsted [3] provided further evidence that medical reasoning could be made explicit and represented in decision theoretic ways. Decision theory also provided the means for Nash to develop a "Logoscope," which might be considered as the first mechanical diagnostic aid [4].

An information system developed using non-AI decision-making techniques may comprise procedural or declarative knowledge. Procedural knowledge maps the decision-making process into the methods by which the clinical problems are solved or clinical decisions made. Examples of techniques that form a procedural knowledge base are those that are based on algorithmic analytical models, clinical algorithms, or decision trees. Information systems based on declarative knowledge comprise what can essentially be termed a database of facts about different aspects of a clinical problem; the causal relationships between these facts form a rich network from which explicit (say) cause–effect pathways can be determined. Semantic networks and causal probabilistic networks are perhaps the best examples of information systems based on declarative knowledge. There are other types of clinical decision aids, based purely on statistical methods applied to patient data, for example, classification analyses based on 1ogistic regression, relative frequencies of occurrence, pattern-matching algorithms, or neural networks.

The structure of this chapter mirrors to some extent the different methods and techniques of non-AI decision making mentioned earlier. It is important to distinguish between analytical models based on quantitative or qualitative mathematical representations and decision theoretic methods typified by the use of clinical algorithms, decision trees, and set theory. Most of the latter techniques add to an information base by way of procedural knowledge. It is then that advantage can be taken of the many techniques that have statistical decision theoretic principles as their underpinning.

This section begins with a discussion of simple linear regression models and pattern recognition, but then more complex statistical techniques are introduced, for example, the use of Bayesian decision analysis, which leads to the introduction of causal probabilistic networks. The majority of these techniques add information by use of declarative knowledge. Particular applications are used throughout to illustrate the extent to which non-AI decision making is used in clinical practice.

44.1 Analytical Models

In the context of this chapter, the analytical models considered are qualitative and quantitative mathematical models that are used to predict future patient state based on present state and a historical representation of what has passed. Such models could be representations of system behavior that allow test signals to be used so that response of the system to various disturbances can be studied, thus making predictions of future patient state.

For example, Leaning et al. [5,6] produced a 19-segment quantitative mathematical model of the blood circulation to study the short-term effects of drugs on the cardiovascular system of normal, resting patients. The model represented entities such as the compliance, flow, and volume of model segments in what was considered a closed system. In total, the quantitative mathematical model comprised 61 differential equations and 159 algebraic equations. Evaluation of the model revealed that it was fit for its purpose in the sense of heuristic validity, that is, it could be used as a tool for developing explanations for cardiovascular control, particularly in relation to the central nervous system (CNS).

Qualitative models investigate time-dependent behavior by representing patient state trajectory in the form of a set of connected nodes, the links between the nodes reflecting transitional constraints placed on the system [7]. The types of decision making supported by this type of model are assessment and therapy planning. In diagnostic assessment, the precursor nodes and the pathway to the node (decision) of interest define the causal mechanisms of the disease process. Similarly, for therapy planning, the optimal plan can be set by investigation of the utility values associated with each link in the disease–therapy relationship. These utility values refer to a cost function, where cost can be defined as the monetary cost of providing the treatment and cost benefit to the patient in terms of efficiency, efficacy, and effectiveness of alternative treatment options. Both quantitative [8] and qualitative [9] analytical models can be realized in other ways to form the basis of rule-based systems; however that excludes their analysis in this chapter.

44.2 Decision Theoretic Models

44.2.1 Clinical Algorithms

The clinical algorithm is a procedural device that mimics clinical decision making by structuring the diagnostic or therapeutic decision processes in the form of a classification tree. The root of the tree represents some initial state, and the branches yield the different options available. For the operation of the clinical algorithm the choice points are assumed to follow branching logic with the decision function being a yes/no (or similar) binary choice. Thus, the clinical algorithm comprises a set of questions that must be collectively exhaustive for the chosen domain and the responses available to the clinician at each branch point must be mutually exclusive. These decision criteria pose rigid constraints on the type of medical problem that can be represented by this method, as the lack of flexibility is appropriate only for a certain set of well-defined clinical domains. Nevertheless, there is rich literature available; examples include the use of the clinical algorithm for acid–base disorders [10] and diagnosis of mental disorders [11]. A comprehensive guide to clinical algorithms can be found on the website of the American Academy of Family Physicians [12].

44.2.2 Decision Trees

A more rigorous use of classification tree representations than the clinical algorithm can be found in decision tree analysis. Although from a structural perspective, decision trees and clinical algorithms are similar in appearance, for decision tree analysis the likelihood and cost benefit for each choice are also calculated in order to provide a quantitative measure for each option available. This allows the use of optimization procedures to gauge the probability of success for the correct diagnosis being made or for a beneficial outcome from therapeutic action being taken. A further difference between the clinical algorithm and decision tree analysis is that the latter has more than one type of decision node (branch point): at decision nodes the clinician must decide on which choice (branch) is appropriate for the given clinical scenario; at chance nodes the responses available have no clinician control, for example, the response may be due to patient specific data; and outcome nodes define the chance nodes at the "leaves" of the decision tree. That is, they summarize a set of all possible clinical outcomes for the chosen domain.

The possible outcomes from each chance node must obey the rules of probability and sum to unity; the probability assigned to each branch reflects the frequency of that event occurring in a general patient population. It follows that these probabilities are dynamic, with accuracy increasing, as more evidence becomes available. A utility value can be added to each of the outcome scenarios. These utility measures reflect a trade-off between competing concerns, for example, survivability and quality of life, and may be assigned heuristically.

When the first edition of this chapter was written in 1995 it was noted that although a rich literature describing potential applications existed [13], the number of practical applications described was limited. The situation has changed and there has been an explosion of interest in applying decision analysis to clinical problems. Not only is decision analysis methodology well described [14–17] but there are also numerous articles appearing in mainstream medical journals, particularly *Medical Decision Making*. An important driver for this acceleration of interest has been the desire to contain costs of medical care, while maintaining clinical effectiveness and quality of care. Cost-effectiveness analysis is an extension of decision analysis and compares the outcome of decision options in terms of the monetary cost per unit of effectiveness. Thus, it can be used to set priorities for the allocation of resources and to decide between one or more treatment or intervention options. It is most useful when comparing treatments for the same clinical condition. Cost-effectiveness analysis and its implications are described very well elsewhere [18,19]. One reason for the lack of clinical applications using decision trees is that the underpinning software technologies remain relatively underdeveloped. Babič et al. [20] address this point by comparing decision tree software with other non-AI decision-making methods.

44.2.3 Influence Diagrams

In the 1960s researchers at Stanford Research Institute (SRI) proposed the use of influence diagrams as representational models when developing computer programs to solve decision problems. However, it was recognized somewhat later by decision analysts at SRI [21] that such diagrams could be used to facilitate communication with domain experts when eliciting information about complex decision problems. Influence diagrams are a powerful mode of graphic representation for decision modeling. They do not replace but complement decision trees and it should be noted that both are different graphical representations of the same mathematical model and operations. Recently, two exciting papers have been published that make the use of influence diagrams accessible to those interested in medical decision making [22,23].

44.3 Statistical Models

44.3.1 Database Search

Interrogation of large clinical databases yields statistical evidence of diagnostic value and in some representations form the basis of rule induction used to build expert systems [24]. These systems will not be discussed here. However, the most direct approach for clinical decision making is to determine the relative frequency of occurrence of an entity, or more likely group of entities, in the database of past cases. This enables a prior probability measure to be estimated [25]. A drawback of this simple, direct approach to problem solving is the apparent tautology of more evidence available leading to fewer matches in the database being found; this runs against common wisdom that more evidence leads to an increase in probability of a diagnosis being found. Further, the method does not provide a weight for each item of evidence to gauge those that are more significant for patient outcome.

With the completion of the Human Genome sequence there has been renewed interest in database search methods for finding data (e.g., single nucleotide polymorphisms — or more simply SNPs) in the many genetic database resources that are distributed throughout the world [26]. Vyas and Summers [27] provide both a summary of the issues surrounding the use of metadata to combine these dispersed data resources and suggest a solution via a semantic web-based knowledge architecture. It is clear that such methods of generating data will have an increasing impact on the advent of molecular medicine.

44.3.2 Regression Analysis

Logistic regression analysis is used to model the relationship between a response variable of interest and a set of explanatory variables. This is achieved by adjusting the regression coefficients, the parameters of the model, until a "best fit" to the data set is achieved. This type of model improves upon the use of relative frequencies, as 1ogistic regression explicitly represents the extent to which elements of evidence are important in the value of the regression coefficients. An example of clinical use can be found in the domain of gastroenterology [28].

44.3.3 Statistical Pattern Analysis

The recognition of patterns in data can be formulated as a statistical problem of classifying the results of clinical findings into mutually exclusive but collectively exhaustive decision regions. In this way, not only can physiologic data be classified but also the pathology that they give rise to and the therapy options available to treat the disease. Titterington [29] describes an application in which patterns in a complex data set are recognized to enhance the care of patients with head injuries. Pattern recognition is also the cornerstone of computerized methods for cardiac rhythm analysis [30]. The methods used to distinguish patterns in data rely on discriminant analysis. In simple terms, this refers to a measure of separability between class populations.

In general, pattern recognition is a two-stage process as shown in Figure 44.1. The pattern vector, P, is an n-dimensional vector derived from the data set used. Let Ω_p be the pattern space, which is the set of all possible values P may assume, then the pattern recognition problem is formulated as finding a way of dividing Ω_p into mutually exclusive and collectively exhaustive regions. For example, in the analysis of the electrocardiogram the complete waveform may be used to perform classifications of diagnostic value. A complex decision function would probably be required in such cases. Alternatively (and if appropriate), the pattern vector can be simplified to investigation of sub features within a pattern. For cardiac arrhythmia analysis, only the R–R interval of the electrocardiogram is required, which allows a much simpler decision function to be used. This may be a linear or nonlinear transformation process:

$$X = \tau P$$

where X is termed the feature vector and τ is the transformation process.

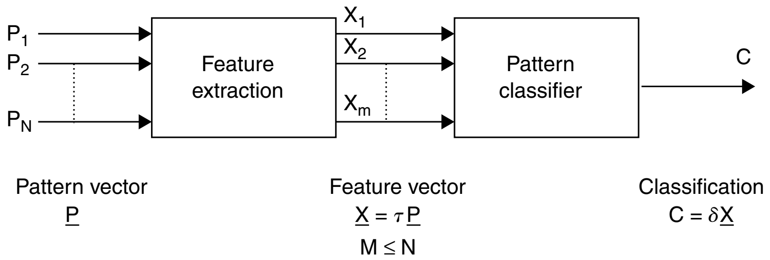


FIGURE 44.1 Pattern recognition.

Just as the pattern vector P belongs to a pattern space Ω_p, so the feature vector X belongs to a feature space Ω_X. As the function of feature extraction is to reduce the dimensionality of the input vector to the classifier, some information is lost. Classification of Ω_X can be achieved using numerous statistical methods including: discriminant functions (linear and polynomial), kernel estimation, k-nearest neighbor, cluster analysis, and Bayesian analysis.

44.3.4 Bayesian Analysis

Ever since their reinvestigation by Savage in 1954 [2], Bayesian methods of classification have provided one of the most popular approaches used to assist in clinical decision making. Bayesian classification is an example of a parametric method of estimating class-conditional probability density functions. Clinical knowledge is represented as a set of prior probabilities of diseases to be matched with conditional probabilities of clinical findings in a patient population with each disease. The classification problem becomes one of a choice of decision levels, which minimizes the average rate of misclassification or to minimize the maximum of the conditional average loss function (the so-called minmax criterion) when information about prior probabilities is not available. Formally, the optimal decision rule that minimizes the average rate of misclassification is called the Bayes rule; this serves as the inference mechanism that allows the probabilities of competing diagnoses to be calculated when patient specific clinical findings become available.

The great advantage of Bayesian classification is that a large clinical database of past cases is not required, thus allowing the time taken to reach a decision to be faster compared with other database search techniques; furthermore, classification errors due to the use of inappropriate clinical inferences are quantifiable. However, a drawback of this approach to clinical decision making is that the disease states are considered as complete and mutually exclusive, whereas in real life neither assumption may be true.

Nevertheless, Bayesian decision analysis functions as a basis for differential diagnosis and has been used successfully, for example, in the diagnosis of acute abdominal pain [31]. De Dombal first described this system in 1972, but it took another 20 years or so for it to be accepted via a multicenter multinational trial. The approach has been exploited in ILIAD; this is a commercially available [32] computerized diagnostic decision support system with some 850 to 990 frames in its knowledge base. As it is a Bayesian system, each frame has the prevalence of a disease for its prior probability. There is the possibility however, that the prevalence rates may not have general applicability. This highlights a very real problem, namely the validity of relating causal pathways in clinical thinking and connecting such pathways to a body of validated (true) evidence. Ideally, such evidence will come from randomized controlled clinical or epidemiological trials. However, such studies may be subject to bias.

To overcome this, Eddy et al. [33] devised the Confidence Profile Method. This is a set of quantitative techniques for interpreting and displaying the results of individual studies (trials), exploring the effects of any biases that might affect the internal validity of the study, adjusting for external validity, and, finally, combining evidence from several sources. This meta-analytical approach can formally incorporate experimental evidence and, in a Bayesian fashion, also the results of previous analytical studies or subjective judgments about specific factors that might arise when interpreting evidence. Influence diagram representations play an important role in linking results in published studies and estimates of probabilities and statements about causality.

Currently, much interest is being generated as to how, what is perceived as the Bayesian action-oriented approach can be used in determining health policy, where the problem is perceived to be a decision problem rather than a statistical problem, see for instance Lilford and Braunholz [34].

Bayesian analysis continues to be used in a wide range of clinical applications either as a single method or as part of a multimethod approach, for example, for insulin sensitivity [35], for understanding incomplete data sets [36], and has been used extensively for the analysis of clinical trials (e.g., see References 37 and 38).

Bayesian decision theory also provides a valuable framework for health care technology assessment. Bayesian methods have also become increasingly visible in places where they are being applied to the analysis of economic models [39], being applied particularly to two decision problems commonly encountered in pharmaco-economics and health technology assessment generally, namely: adoption and allocation.

Acceptability curves generated from Bayesian cost effectiveness analyses can be interpreted as the probability that the new intervention is cost effective at a given level of willingness-to-pay. For Bayesian methods applied to clinical trials with cost as well as efficacy data see O'Hagan and Stevens [40]. This probabilistic interpretation of study findings provides information that is more relevant and more transparent to decision makers. The Bayesian value of information analysis offers a decision-analytic framework to explore the conceptually separate decisions of whether a new technology should be adopted from the question of whether more research is required to inform this choice in the future [41,42]. Thus, it is a useful analytical framework for decision-makers who wish to achieve allocative efficacy.

The Bayesian approach has several advantages over the frequentist approach. First, it allows accumulation and updating of knowledge by using the prior distribution. Second, it yields more flexible inferences and emphasizes predictions rather than hypothesis testing. Third, probabilities involving multiple endpoints are relatively simply to estimate. Finally, it provides a solid theoretical framework for decision analyses.

44.3.5 Dempster–Shafer Theory

One way to overcome the problem of mutually exclusive disease states is to use an extension to Bayesian classification put forward by Dempster [43] and Shafer [44]. Here, instead of focusing on a single disorder, the method can deal with combinations of several diseases. The key concept used is that the set of all possible diseases is partitioned into *n*-tuples of possible disease state combinations.

A simple example will illustrate this concept. Suppose there is a clinical scenario in which four disease states describe the whole hypothesis space. Each new item of evidence will impact on all the possible subsets of the hypothesis space and is represented by a function, the basic probability assignment. This measure is a belief function that must obey the law of probability and sum to unity across the subsets impacted upon. In the example, all possible subsets comprise: one that has all four disease states in it; four, which have three of the four diseases as members; six, which have two diseases as members; and finally, four subsets that have a single disease as a member. Thus, when new evidence becomes available in the form of a clinical finding, only certain hypotheses, represented by individual subsets, may be favored.

44.3.6 Syntactic Pattern Analysis

As demonstrated earlier, a large class of clinical problem solving using statistical methods involves classification or diagnosis of disease states, selection of optimal therapy regimes, and prediction of patient outcome. However, in some cases the purpose of modeling is to reconstruct the input signal from the data available. This cannot be done by methods discussed thus far. The syntactic approach to pattern recognition uses a hierarchical decomposition of information and draws upon an analogy to the syntax of language. Each input pattern is described in terms of more simple sub-patterns, which themselves are

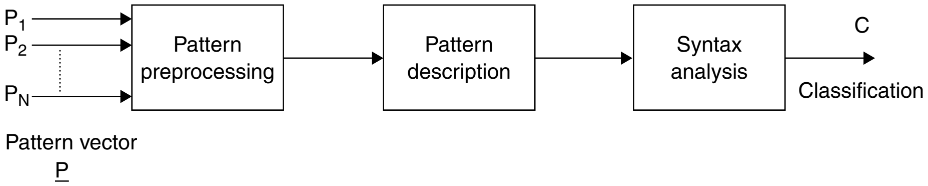


FIGURE 44.2 Syntactic pattern recognition system.

decomposed into simpler subunits, until the most elementary subpatterns, termed the pattern primitives, are reached. The pattern primitives should be selected so that they are easy to recognize with respect to the input signal. Rules that govern the transformation of pattern primitives back (ultimately) to the input signal are termed the grammar.

In this way a string grammar, G, which is easily representable in computer-based applications, can be defined:

$$G = \{V_T, V_N, S, P\}$$

where, V_T are the terminal variables (pattern primitives); V_N are the non-terminal variables; S is the start symbol; and P is the set of production rules, which specify the transformation between each level of the hierarchy. It is an important assumption that in set theoretic terms, the union of V_T and V_N is the total vocabulary of G, and the intersection of V_T and V_N is the null (empty) set.

A syntactic pattern recognition system therefore comprises three functional subunits (Figure 44.2): a preprocessor — this manipulates the input signal, P, into a form that can be presented to the pattern descriptor, the pattern descriptor that assigns a vocabulary to the signal, and the syntax analyzer that classifies the signal accordingly. This type of system has been used successfully to represent the electrocardiogram [45,46] and the electroencephalogram [47] and for representation of the carotid pulse wave [48].

44.3.7 Causal Modeling

A causal probabilistic network (CPN) is an acyclic multiply-connected graph, which at a qualitative level comprises nodes and arcs [49]. Nodes are the domain objects and may represent, for example, clinical findings, pathophysiologic states, diseases, or therapies. Arcs are the causal relationships between successive nodes and are directed links. In this way the node and arc structure represents a model of the domain. Quantification is expressed in the model by a conditional probability table being associated with each arc, allowing the state of each node to be represented as a binary value or more frequently as a continuous probability distribution.

In root nodes the conditional probability table reduces to a probability distribution of all its possible states.

A key concept of CPNs is that computation is reduced to a series of local calculations, using only one node and those that are linked to it in the network. Any node can be instantiated with an observed value; this evidence is then propagated through the CPN via a series of local computations. Thus, CPNs can be used in two ways: to instantiate the leaf nodes of the network with known patterns for given disorders to investigate expected causal pathways; or to instantiate the root nodes or nodes in the graphical hierarchy with, for example, test results to obtain a differential diagnosis. The former method has been used to investigate respiratory pathology [50], and the latter method has been read to obtain pathologic information from electromyography [51].

44.3.8 Artificial Neural Networks

Artificial neural networks (ANNs) mimic their biologic counterparts, although at the present time on a much smaller scale. The fundamental unit in the biological system is the neuron. This is a specialized cell that, when activated, transmits a signal to its connected neighbors. Both activation and transmission involve chemical transmitters, which cross the synaptic gap between neurons. Activation of the neuron takes place only when a certain threshold is reached. This biologic system is modeled in the representation of an artificial neural network. It is possible to identify three basic elements of the neuron model: a set of weighted connecting links that form the input to the neuron (analogous to neurotransmission across the synaptic gap), an adder for summing the input signals, and an activation function that limits the amplitude of the output of the neuron to the range (typically) −1 to +1. This activation function also has a threshold term that can be applied externally and forms one of the parameters of the neuron model. Many books are available which provide a comprehensive introduction to this class of model (e.g., see Reference 52).

ANNs can be applied to two categories of problems: prediction and classification. It is the latter that has caught the imagination of biomedical engineers for its similarity to diagnostic problem solving. For instance, the conventional management of patients with septicemia requires a diagnostic strategy that takes up to 18 to 24 h before initial identification of the causal microorganism. This can be compared to a method in which an ANN is applied to a large clinical database of past cases; the quest becomes one of seeking an optimal match between present clinical findings and patterns present in the recorded data. In this application, pattern matching is a nontrivial problem as each of the 5000 past cases has 51 data fields. It has been shown that for this problem the ANN method outperforms other statistical methods such as k-nearest neighbor [53].

The use of ANNs in clinical decision making is becoming widespread. A further example of their use in critical care medicine is given by Yamamura et al. [54]. ANNs are used in chronic and acute clinical episodes. An example of the former is their use in cancer survival predictions [55] and an example of the latter is their use in the emergency room to detect early onset of myocardial infarction [56].

44.4 Summary

This chapter has reviewed what are normally considered to be the major categories of approach available to support clinical decision making, which do not rely on what is classically termed artificial intelligence (AI). They have been considered under the headings of analytical, decision theoretic, and statistical models, together with their corresponding subdivisions. It should be noted, however, that the division into non-Al approaches and AI approaches that is adopted in this volume (see the Chapter entitled Expert Systems: Methods and Tools) is not totally clear-cut. In essence the range of approaches can in many ways be regarded as a continuum. There is no unanimity as to where the division should be placed and the separation adopted; here is but one of a number that is feasible. It is therefore desirable that the reader should consider these two chapters together and choose an approach that is relevant to the particular clinical context.

References

[1] Meehl, R. 1954. *Clinical versus Statistical Prediction.* Minnesota, University of Minnesota Press.

[2] Savage L.I. 1954. *The Foundations of Statistics.* New York, John Wiley & Sons.

[3] Ledley R.S. and Ludsted L.B. 1959. Reasoning foundations of medical diagnosis. *Science* 130: 9.

[4] Nash F.A. 1954. Differential diagnosis: an apparatus to assist the logical faculties. *Lancet* 4: 874.

[5] Leaning M.S., Pullen H.E., Carson E.R. et al. 1983. Modelling a complex biological system: the human cardiovascular system: 1. Methodology and model description. *Trans. Inst. Meas. Contr.* 5: 71.

[6] Leaning M.S., Pullen H.E., Carson E.R. et al. 1983. Modelling a complex biological system: the human cardiovascular system: 2. Model validation, reduction and development. *Trans. Inst. Meas. Contr.* 5: 87.

[7] Kuipers B.J. 1986. Qualitative simulation. *Artif. Intel.* 29: 289.

[8] Furukawa T., Tanaka H., and Hara S. 1987. FLUIDEX: A microcomputer-based expert system for fluid therapy consultations. In M.K. Chytil and R. Engelbrecht (Eds.), *Medical Expert Systems*. Wilmslow, Sigma Press, pp. 59–74.

[9] Bratko I., Mozetic J., and Lavrac N. 1988. In Michie D. and Bratko I. (Eds.), *Expert Systems: Automatic Knowledge Acquisition*. Reading, Mass. Addison-Wesley, pp. 61–83.

[10] Bleich H.L. 1972. Computer-based consultations: Electrolyte and acid–base disorders. *Am. J. Med.* 53: 285.

[11] McKenzie D.P., McGary P.D., Wallac et al. 1993. Constructing a minimal diagnostic decision tree. *Meth. Inform. Med.* 32: 161.

[12] http://www.aafp.org/x19449.xml (accessed February 2005).

[13] Pauker S.G. and Kassirer J.P. 1987. Decision analysis. *N. Engl. J. Med.* 316: 250.

[14] Weinstein M.C. and Fineberg H.V. 1980. *Clinical Decision Analysis*. London, W.B. Saunders.

[15] Watson S.R. and Buede D.M. 1994. *Decision Synthesis*. Cambridge, Cambridge University Press.

[16] Sox H.C., Blatt M.A., Higgins M.C., and Marton K.I. 1988. *Medical Decision Making*. Boston, Butterworth Heinemann.

[17] Llewelyn H. and Hopkins A. 1993. *Analysing How We Reach Clinical Decisions*. London, Royal College of Physicians.

[18] Gold M.R., Siegel J.E., Russell L.B., and Weinstein M.C. (Eds.) 1996. *Cost-Effectiveness in Health and Medicine*. New York, Oxford University Press.

[19] Sloan F.A. (Ed.) 1996. *Valuing Health Care*. Cambridge, Cambridge University Press.

[20] Babič S.H., Kokol P., Podgorelec V., Zorman M., Šprogar M., and Štiglic M.M. 2000. The art of building decision trees. *J. Med. Syst.* 24: 43–52.

[21] Owen D.L. 1984. The use of influence diagrams in structuring complex decision problems. In Howard R.A. and Matheson J.E. (Eds.), *Readings on the Principles and Applications of Decision Analysis*, Vol. 2. Menlo Park, CA, Strategic Decisions Group, pp. 763–772.

[22] Owens D.K., Shachter R.D., and Nease R.F. 1997. Representation and analysis of medical decision problems with influence diagrams. *Med. Decis. Mak.* 17: 241.

[23] Nease R.F. and Owens D.K. 1997. Use of influence diagrams to structure medical decisions. *Med. Decis. Mak.* 17: 263.

[24] Quinlan J.R. 1979. Rules by induction from large collections of examples. In D. Michie (Ed.), *Expert Systems in the Microelectronic Age*. Edinburgh, Edinburgh University Press.

[25] Gammerman A. and Thatcher A.R. 1990. Bayesian inference in an expert system without assuming independence. In M.C. Golumbic (Ed.), *Advances in Artificial Intelligence*. New York, Springer-Verlag, pp. 182–218.

[26] Goble C.A., Stevens R., and Ng S. 2001. Transparent access to multiple bioinformatics information sources. *IBM Syst. J.* 40: 532–551.

[27] Vyas H. and Summers R. 2004. Impact of semantic web on bioinformatics. In *Proceedings of the International Symposium of Santa Caterina on Challenges in the Internet and Interdisciplinary Research (SSCCII)*, CD-ROM Proceedings.

[28] Spiegelhalter D.J. and Knill-Jones R.P. 1984. Statistical and knowledge-based approaches to clinical decision-support systems with an application in gastroenterology. *J. Roy. Stat. Soc.* A 147: 35.

[29] Titterington D.M., Murray G.D., Murray L.S. et al. 1981. Comparison of discriminant techniques applied to a complex set of head injured patients. *J. Roy. Stat. Soc. A* 144: 145.

[30] Morganroth J. 1984. Computer recognition of cardiac arrhythmias and statistical approaches to arrhythmia analysis. *Ann. NY Acad. Sci.* 432: 117.

[31] De Dombal F.T., Leaper D.J., Staniland J.R. et al. 1972. Computer-aided diagnosis of acute abdominal pain. *Br. Med. J.* 2: 9.

[32] *Applied Medical Informatics*. Salt Lake City, UT.

[33] Eddy D.M., Hasselblad V., and Shachter R. 1992. *Meta-Analysis by the Confidence Profile Methods*. London, Academic Press.

[34] Lilford R.J. and Braunholz D. 1996. The statistical basis of public policy: a paradigm shift is overdue. *Br. Med. J.* 313: 603.

[35] Agbaje O.F., Luzio S.D., Albarrak A.I.S., Lunn D.J., Owens D.R., and Hovorka R. 2003. Bayesian hierarchical approach to estimate insulin sensitivity by minimal model. *Clin. Sci.* 105: 551–560.

[36] Crawford S.L., Tennstedt S.L., and McKinlay J.B. 1995. Longitudinal care patterns for disabled elders: A Bayesian analysis of missing data. In Gatsonis C., Hodges J., and Kass R.E. (Eds.), *Case Studies in Bayesian Statistics*, Vol. 2. New York: Springer-Verlag, pp. 293–308.

[37] Lewis R.J. and Wears R.L. 1993. An introduction to the Bayesian analysis of clinical trials. *Ann. Emerg. Med.* 22: 1328–1336.

[38] Spiegelhalter D.J., Freedman L.S., and Parmar M.K.B. 1994. Bayesian approaches to randomised trials. *J. Roy. Stat. Soc.* A 157: 357–416.

[39] Parmigiani G. 2002. *Modeling in Medical Decision Making: A Bayesian Approach.* Chichester, Wiley.

[40] O'Hagan A. and Stevens J.W. 2003. Bayesian methods for design and analysis of cost-effectiveness trials in the evaluation of health care technologies. *Stat. Meth. Med. Res.* 11: 469–490.

[41] Claxton K. and Posnett J. 1996. An economic approach to clinical trial design and research priority setting. *Health Econ.* 5: 513–524.

[42] Claxton K. 1999. The irrelevance of inference: A decision making approach to the stochastic evaluation of health care technologies. *J. Health Econ.* 18: 341–364.

[43] Dempster A. 1967. Upper and lower probabilities induced by multi-valued mapping. *Ann. Math. Stat.* 38: 325.

[44] Shafer G. 1976. *A Mathematical Theory of Evidence.* Princeton, NJ, Princeton University Press.

[45] Belforte G., De Mori R., and Ferraris E. 1979. A contribution to the automatic processing of electrocardiograms using syntactic methods. *IEEE Trans. Biomed. Eng.* BME 26: 125.

[46] Birman K.P. 1982. Rule-based learning for more accurate ECG analysis. *IEEE Trans. Pat. Anal. Mach. Intell.* PAMI 4: 369.

[47] Ferber G. 1985. Syntactic pattern recognition of intermittant EEG activity. *Meth. Inf. Med.* 24: 79.

[48] Stockman G.C. and Kanal L.N. 1983. Problem reduction in representation for the linguistic analysis of waveforms. *IEEE Trans. Pat. Anal. Mach. Intel.* PAMI 5: 287.

[49] Andersen S.K., Jensen F.V., and Olesen K.G. 1987. *The HUGIN Core-Preliminary Considerations on Inductive Reasoning: Managing Empirical Information in AI Systems.* Riso, Denmark.

[50] Summers R., Andreassen S., Carson E.R. et al. 1993. A causal probabilistic model of the respiratory system. In *Proceedings of the IEEE 15th Annual Conference of the Engineering in Medicine and Biology Society.* New York, IEEE, pp. 534–535.

[51] Jensen F.V., Andersen S.K., Kjaerulff U. et al. 1987. MUNIN: On the case for probabilities in medical expert systems — a practical exercise. In Fox J., Fieschi M., and Engelbrecht R. (Eds.), *Proceedings of the Ist Conference European Society for AI in Medicine.* Heidelberg, Springer-Verlag, pp. 149–160.

[52] Haykin S. 1994. *Neural Networks: A Comprehensive Foundation.* New York, Macmillan.

[53] Worthy P.J., Dybowski R., Gransden W.R., et al. 1993. Comparison of learning vector quantisation and nearest neighbour for prediction of microorganisms associated with septicaemia. In: *Proceedings of the IEEE 15th Annual Conference of the Engineering in Medicine and Biology Society*, New York, IEEE, pp. 273–274.

[54] Yamamura S., Takehira R., Kawada K., Nishizawa K., Katayama S., Hirano M., and Momose Y. 2003. Application of artificial neural network modelling to identify severely ill patients whose aminoglycoside concentrations are likely to fall below therapeutic concentrations. *J. Clin. Pharm. Ther.* 28: 425–432.

[55] Burke H.B., Goodman P.H., and Rosen D.B. 1997. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer* 79: 857–862.

[56] Baxt W.G., Shofer F.S., Sites F.D., and Hollander J.E. 2002. A neural computational aid to the diagnosis of acute myocardial infarction. *Ann. Emerg. Med.* 39: 366–373.

45

Medical Informatics and Biomedical Emergencies: New Training and Simulation Technologies for First Responders

Joseph M. Rosen
Christopher Swift
Eliot B. Grigg
Matthew F. McKnight
Susan McGrath
Dennis McGrath
Peter Robbie
C. Everett Koop
Dartmouth-Hitchcock Medical Center

Delivering detection, diagnostic, and treatment information to first responders remains a central challenge in disaster management. This is particularly true in biomedical emergencies involving highly infectious agents. Adding inexpensive, established information technologies to existing response system will produce beneficial outcomes. In some instances, however, emerging technologies will be necessary to enable an immediate, continuous response. This article identifies and describes new training, education, and simulation technologies that will help first responders cope with bioterrorist events. The September 11 Commission report illuminated many of the errors leading to Al Qaeda's dramatic attacks in New York and Washington, DC Among them was a well-documented failure to coordinate intelligence within the federal bureaucracy. Yet the greatest failure noted by the Commissioners was a failure of policy — a failure of imagination [1]. Federal, state, and local officials must avoid similar myopia in working to secure the homeland. Responding to future attacks with Chemical, Biological, Radiological, Nuclear or Explosive (CBRNE) agents is one area where imagination and innovation will be necessary.

September 11 marked a watershed in Americans' understanding of the world and their place in it. Al Qaeda's attacks were audacious, innovative, and cruel. Yet they were also highly conventional. The use of civilian aircraft as guided missiles mimicked the fuel air bombs that the U.S. forces used to attack Osama bin Laden's training camps in 1998. September 11 was also a tactical strike. Notwithstanding the scale of suffering and destruction, it undermined neither the capacity of the United States to defend itself militarily, nor the public's will to wage defensive war [2].

From securing government buildings to screening airline passengers, much of the political debate stemming from the September 11 Commission report now focuses on preventing similar tactical strikes. This goal is laudable, but may not be practically attainable. U.S. intelligence and law enforcement agencies are neither omniscient nor omnipotent. Robust search and seizure capabilities will inevitably be restrained in a society that values civil rights and individual freedom. Even Russia, whose government retains extensive domestic intelligence gathering and surveillance powers, proved unable to interdict armed hostage takers at Moscow's Dubrovka theater and the Beslan primary school.

Against that backdrop, a reflexive focus on preventing *any* form of terrorism may detract attention and divert funding from preparations necessary to respond to a *strategic* terrorist attack. Policy makers and the American public should not lose sight of the big picture. It may not always be possible to preempt, much less prevent future terrorist attacks. It is possible, however, to ameliorate or mitigate the strategic effects of future attack through proper planning, resources, and training. With that object in mind, training local first responders for a full spectrum of CBRNE contingencies must remain foremost among the reforms considered in the wake of the September 11 Commission report. This examination of emerging training and simulation technologies endeavors to inform that debate.

This investigation proceeds in four stages. First, it addresses the political implications of CBRNE agents, with a particular focus on Biological Warfare (BW). Second, it compares and contracts the incremental U.S. National Disaster Medical System (NDMS) with the flexible, net-centric disaster management model employed by German emergency personnel. Third, it examines the role of Information Technology (IT) and medical informatics in disaster response. Finally, it assesses the role of simulation technologies in developing and disseminating CBRNE disaster training tools across both legal and geographic jurisdictions.

45.1 Asymmetric Warfare and Bioterrorism

Epidemics played a major role in destabilizing great empires. Between 250 and 650 AD, the Roman empire "was assaulted by successive waves of pandemics that reduced the population by at least one-quarter . . ." [3]. Naturally occurring phenomena carried strategic implications. As the number of infections rose, the result was significantly reduced economic productivity, a declining agricultural and tax base, and severe shortages in military manpower. "Rome's greatest enemy cannot from within," observes medieval historian Norman F. Cantor, "but from biomedical plagues that the Romans could not possibly understand or combat" [3]. Germs, not the Goths, brought the Roman civilization to its knees.

A similar civilization collapse is improbable today. Contemporary scientific knowledge is infinitely more advanced that in the classical period. Economies are now digital and global, relying less and less on physical human labor. Had imperial Rome possessed an institution comparable to the Centers for Disease Control (CDC), its response to epidemiological crises may have proved more robust. Modern society's capacity to detect and contain epidemics does not render them harmless, however. Indeed, the very characteristics that make biological agents unwieldy and unpredictable as battlefield weapons — silence, incubation time, and uncontrollability — render them especially effective for terrorist attacks [4].

Four characteristics make BW preferable to nuclear, chemical, or even radiological weapons. First, they are easily concealed. Small containers and host carriers can carry enough pathogen to infect hundreds, if not thousands. Second, incubation periods enable silent dispersion and subsequent transmission within

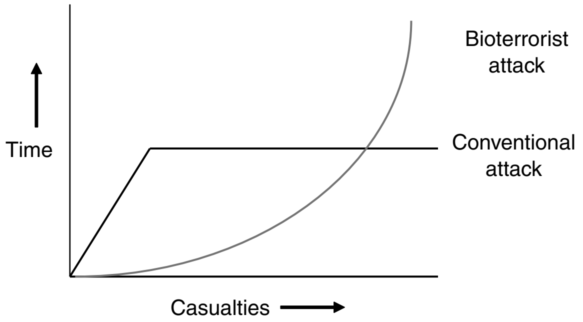


FIGURE 45.1 Conventional and bioterrorist attacks.

target populations. Initial cases would be almost indistinguishable from naturally occurring phenomena. Third, BW attacks threaten the very public health personnel necessary for effective detection, containment, and treatment — especially in high-density metropolitan areas. Those responsible for diagnosing and containing an agent are most likely to be infected by it. Finally, the longer a BW attack remains undetected, the more difficult it becomes to prevent continued proliferation. Unlike a static, conventional strike, the damage caused by a bioterrorist attack can grow geometrically [4] (see Figure 45.1).

Serious vulnerabilities exist. The 2000 TOPOFF exercises in Denver, Colorado, together with simulations run by Dartmouth College in 2001, indicated that BW attacks might threaten vital U.S. security interests [4]. Properly executed by a competent adversary, such an event could swiftly assume strategic dimensions. In 1993, for example, the Congressional Office of Technology Assessment (OTA) estimated that the aerosolized release of one hundred kilograms of anthrax in the Washington, DC area would produce between 130,000 and 3 million fatalities [5]. Although that analysis does not account for reforms undertaken during the last decade, it is still worth noting that scale of lethality reported by OTA is comparable to that from a thermonuclear bomb [5].

These simulations also revealed that existing local, state, and federal capabilities are ill suited for BW. Time lags in resources allocation and failure to distribute resources effectively within the disaster area proved highly problematic [6]. Such failings only exacerbate existing vulnerabilities. Swift detection, containment, and treatment are essential in mitigating both the tactical effects and strategic implications of a bioterrorist attack. Failure would concede control of the operational tempo to the contagion and those employing it as a weapon.

45.2 Disaster Management Paradigms

Although these threats are recognized within the U.S. Government, the current structure of the federal emergency response system remains insufficient. The situation is even more pronounced when one considers existing barriers to coordination between federal authorities and first responders at the state and local level. Recent studies by the Federation of American Scientists revealed that "physicians. Nurses, emergency medical workers, police and fire officials feel unprepared for a WMD emergency — particularly at the level of cities and counties" [7]. To address these apparent shortcomings, we must first rethink our approach to disaster response, as well as the manner in which we train first responders.

The U.S. Government established the National Disaster Management System (NDMS) in 1983. Drawing resources from the Federal Emergency Management Agency (FEMA), as well as the Departments of Defense (DOD), Health and Human Services (HHS), and Veterans Affairs (VA), the system concentrated sophisticated equipment and highly trained personnel in regional Disaster Management Assistance Teams (DMATs). HHS oversaw medical stockpiles. DOD provided transportation for medical evacuation. The VA would open regional medical centers, when needed, to mobilize DMAT teams [2] (see Figure 45.2).

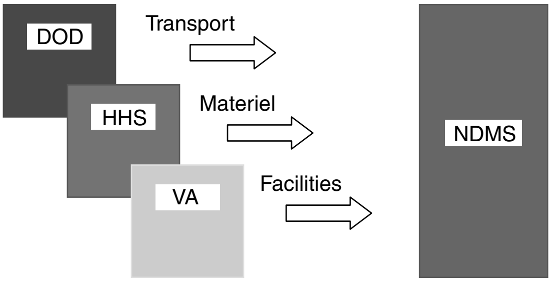
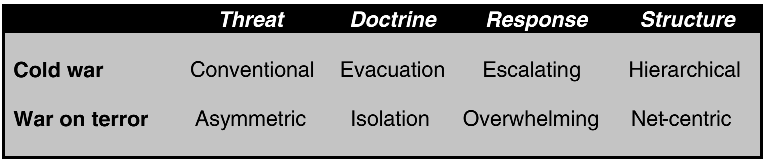


FIGURE 45.2 Cold war NDMS structure.



	Threat	*Doctrine*	*Response*	*Structure*
Cold war	Conventional	Evacuation	Escalating	Hierarchical
War on terror	Asymmetric	Isolation	Overwhelming	Net-centric

FIGURE 45.3 NDMS transformation.

NDMS was a product of the Cold War. Though designed for natural and transportation disasters, the infrastructure was also intended to support mass casualties evacuated from military operations outside the Continental United States (OCONUS). Were the Soviet Union to invade western Europe, the United States could accommodate the anticipated mass of NATO casualties. Unfortunately, NDMS was not designed with homeland security in mind, much less bioterrorism. In an age where intercontinental thermonuclear war represented the primary threat to U.S. security, there was little strategic value in responding to an attack in which most if not all the civilian casualties would be killed.

NDMS's Cold War roots are evident in three characteristics. First, the role of the federal government is to supplement state and local emergency response capabilities by providing triage, austere medical care, and casualty staging. Second, NDMS is an incremental, echelon-based system. The larger the casualty pool, the more DMATs deployed. Third, the chief mission of first responders is stabilization and evacuation. As currently configured, DMATs treat only 10% of all casualties on-scene, transporting the remaining 90% from the incident site to remote medical facilities operated by the VA and others.

This system is poorly configured for the challenges of catastrophic terrorism. The first problem is doctrinal. BW and other CBREN attacks will generally require swift isolation and decontamination. Though useful in earthquakes and airline disasters, a Cold War doctrine emphasizing swift patient evacuation could exacerbate crisis conditions. In the case of bioterrorism, delayed onset places greater emphasis on vaccination and restricted movement. New doctrines must bring the hospital to the patient, rather than the patient to the hospital (see Figure 45.3).

The second problem is architecture. In the current NDMS system, local first responders often encounter a disaster with minimal external support. State and federal resources become available only after the declaration of a state of emergency — a process that can take hours, if not days. The result is often a delayed operations tempo: by "the time a crisis is detected, the scale of it is appreciated, and federal resources are put into play, it may be too late" [2]. Staging disaster response may unintentionally cede the initiative to the adversary, allowing an event to escalate beyond the incident area, or to acquire public resonances several orders of magnitude greater than the scope of the actual event. Though "federal personnel [might] deal with the horrendous aftermath, [they] would not be involved in the direct response" [8].

The need for a continuous, integrated response is particularly evident in BW scenarios. At the tactical level, casualties must be isolated and the incident site contained. All should be treated on-site or screened before evacuation to remote medical facilities. In many instances, hospitals are likely to be the locus of pathogen detection, as well as a nexus for further infection. Absent adequate staging of personnel, equipment, and supplies, the confluence of treatment facility and incident cite could lead to broader contamination while compromising a node in the local public health network. Failure to maintain strict containment regimes will only exacerbate a pathogen's potential physical, psychological, and political effects.

Although some doctrinal reforms are underway, the NDMS's hierarchical architecture and incremental approach may prove unable to keep pace with anthropomorphic epidemics occurring simultaneously in multiple population centers. To ensure effective consequence management, U.S. emergency response paradigms must move beyond echelon-based care to a continuous care system in which "patients are followed by single or distinct groups or providers throughout the system" [2]. Ultimately, the goal must be "to have all resources, both system performance and human resources, available immediately wherever the bioattack is detected" [8].

45.3 Disaster Response Paradigms

A terrorist attack of national scope will inevitably initiate "a multi-agency operation requiring sophisticated (and sometimes chaotic) communications and coordination" [2]. Enhancing interoperability across function and geographic jurisdictions is now a primary objective for first responders and the federal agencies supporting them. Efforts to improve communications networks and enhance incident Command and Control (C^2) are now underway. Chief among them is the adoption of new NDMS doctrines for cite isolation and treatment. Also notable is the DOD's creation of a Northern Command (NORTHCOM), which plays a leading role in coordination capabilities among relevant federal, state, and local agencies.

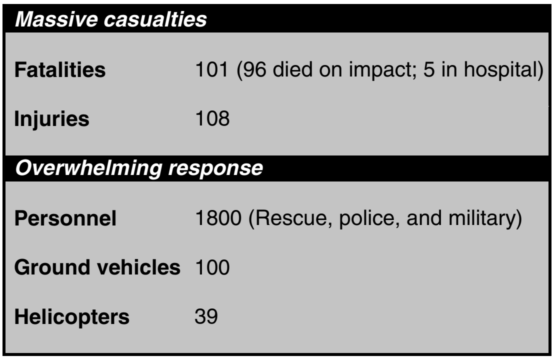
Despite these efforts, however, first responder training often falls to a diverse collection of agencies possessing disparate sources of authority, funding, and information. The Department of Energy (DOE), for example, oversees training for nuclear and radiological attacks. HHS and the Centers for Disease Control (CDC) oversee bioterrorism programs. The Departments of Justice (DOJ) and Homeland Security (DHS) each provide a broad spectrum of grants and training programs for fire, law enforcement, and emergency medical personnel. The result is an unwieldy and extraordinarily complex bureaucratic web.

Congressional oversight displays similar stovepipes, with 79 committees and subcommittees, all 100 senators, "and at least 412 of the 435 House members" sharing "some degree of responsibility for homeland security operations" [9]. This broad balkanization presents two major challenges. First, it obstructs coordination and cooperation, both within the federal government and between Washington and the states. Second, it contributes to political confusion and policy failure. Absent a broad view of the entire homeland security apparatus, legislators are likely to respond to parochial interests rather than "develop a broad overview of homeland security priorities" [9].

Significant challenges also exist at the state and local level. Some 80% of U.S. first responders are unpaid volunteers [7]. Most training occurs under the auspices of local departments, sometimes either supported by state or regional academies. As such, the "level of preparedness often varies from jurisdiction to jurisdiction and from agency to agency" [10]. The result is a patchwork quilt that covers some areas, but leaves others woefully bare. "Absent better coordination and approaches to the dissemination of training materials," warns the Federation of American Scientists, "much of the investment [in Homeland Security] is likely to be wasted and decades could pass before the need is met" [7].

For a new disaster management paradigm to emerge, we must first identify the objectives sought and the characteristics necessary to achieve them. The 1998 crash of the high-speed InterCity Express (ICE) train 884 near Eschede, Germany provides a valuable model. Like the United States, Germany is a federal republic. Political authority is distributed among federal, state, and local officials. Disaster management is an interdepartmental endeavor involving numerous agencies with varied capabilities and legal jurisdictions. There is no centralized hierarchy responsible for emergency planning or response. Given these similar systems, one might also expect similar outcomes.

The German response to ICE 884 proved otherwise. Emergency personnel responded within 4 min of the train derailing. Within 8 min, first responders arrived on the scene, declared a disaster and issued



Massive casualties	
Fatalities	101 (96 died on impact; 5 in hospital)
Injuries	108
Overwhelming response	
Personnel	1800 (Rescue, police, and military)
Ground vehicles	100
Helicopters	39

FIGURE 45.4 InterCity Express 884.

mutual aid requests to neighboring emergency command centers. Within 15 min, 14 additional emergency physicians were en route to the incident site via rescue helicopters. Military units cleared landing zones. Commanders mobilized heavy-duty rescue equipment. The rescue operation took four short hours, during which time more than 1,800 personnel from various agencies evacuated all of the 108 injured victims. Only five died in hospital [11] (see Figure 45.4).

What then, are the primary differences between the U.S. and German disaster management paradigms? The first is organizational. Where the U.S. system is bifurcated and bureaucratized, the Germans employ an integrated, multifaceted approach drawing on prepositioned assets and previously agreed mutual-aid covenants. Where the U.S. approach stresses hierarchy, incremental response, and echelon-based care, the German system favors an immediate, overwhelming response in which all players know their assigned role and are trained to act in concert. Where the U.S. system stresses specialized skills, the German system trains first responders to operate in concert with colleagues from various professional and experiential backgrounds.

The second major difference is operational: namely, the extensive use of Health Information Systems (HIS). In Germany, each element in the disaster response system shares standardized disaster incident information. The result is a net-centric C^2 architecture in which emergency personnel, incident commanders, and remote medical facilities can accurately track both the dispatch of resources to the incident site, as well as anticipate care required following the evacuation of trauma victims. This ability to swiftly share and coordinate diagnostic, geographic, and logistical information across all of the responding agencies helps create a seamless and highly adaptable disaster response system while simultaneously reducing confusion and duplication. Incident commanders can command, rather than improvize.

45.4 Medical Informatics and Emergency Response

The use of medical informatics can be as simple as reporting as patient's condition to physicians en route to the hospital, or as complex as managing triage in a mass casualty, multiple-incident terrorist attack. In both instances, first responders collect, categorize, and communicate casualty information to other personnel in the disaster management system. In the latter scenario, however, the nature and scope of the events introduces a high degree of complexity and sensitivity, requiring careful management of patients and resources alike. Both instances use IT to significantly improve the richness and reach of available information (see Figure 45.5).

As evident in the case of ICE 884, the systematized collection, dissemination, and application of casualty information is necessary for triage and treatment in any conventional disaster. This is true even when information density is relatively low, as in a motor vehicle accident. In a major terrorist attack, however, medical informatics can assume strategic importance. Understanding the nature and location of the incident, the current condition of the victims, and the likely cause will each be critical prerequisites for

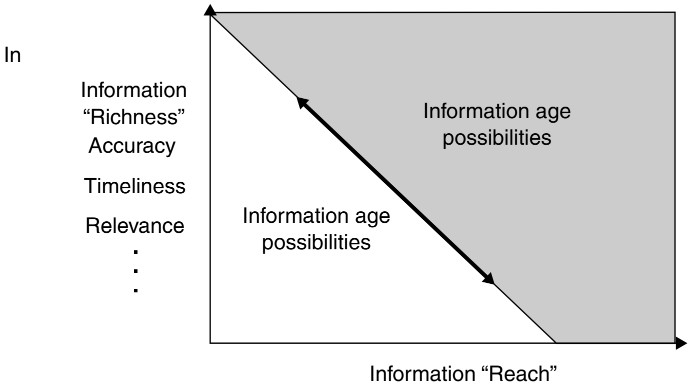


FIGURE 45.5 Information Richness and Reach. (Taken from D.S. Alberts, J.J. Gartska, R.E. Hayes, and D.A. Signori, *CCRO Publication Series*, 2001.)

successful consequence management. Providing that information in the face of significant information density may require a brute force scaling of current HIS systems, together with significant increases in telecommunications bandwidth [13]. It will also require significant investments in public health. Existing care systems must "be enhanced, rather than simply being augmented by DMATs that drop in after an attack is discovered" [2].

In large-scale, long-duration events, diagnostic information alone will not be sufficient. Also vital is geographic information regarding the location of casualties and incident sites, together with the position of both material and personnel. Ideally, a dynamic, integrated HIS system would be combining Global Positioning System (GPS) and Geographical Information System (GIS) technologies to provide first responders and incident commanders alike with a seamless data source. Likewise, logistical and environmental data would prove invaluable in tracking not only casualties, but also the potential spread of a biological, chemical, or radiological plume beyond the initial incident zone.

From GIS to HIS, effectively integrating and employing medical informatics will involve both hardware and software issues — issues that must be addressed in establishing interoperability and coordination across a broad spectrum of probable disaster situations. These new technologies will also require properly configured "wetware." To establish a net-centric disaster response domain, emergency personnel must first know how to use the systems in question, and be able to develop doctrines that codify best practices. Training, education, and simulation (TES) will each play a critical role in bringing new systems online.

It is axiomatic that highly trained personnel "are an essential element in delivering effective emergency medical care" [14]. Less obvious is the high degree of variation in training regimes, even within highly structured and closely regulated professions. In the case of clinical surgical education, for example, quality is often "quite unpredictable and depends mainly on the instructor and the particular cases to which the surgeon is exposed during his or her training" [15]. Similar patterns exist in other fields. Prior to September 11, the National Standard Curricula for Emergency Medical Technicians (EMTs) maintained by the U.S. Department of Transportation lacked detailed discussion of the resources, strategies, and techniques necessary in a CBRNE attack. Even in the wake of recent reforms, the complexity of the current training programs, together with the absence of a central clearinghouse for best practices, provides few reliable mechanisms for identifying, updating, and communicating response procedures [7]. Despite widespread recognition, federal initiatives to correct this oversight remain illusive.

The state of civilian emergency training stands in stark contrast with innovative programs in the military sphere. By 2007, Army medics training under the recently created "Healthcare occupational specialty will be certified at a civilian emergency medical technician level . . . and have the ability to treat chemical, biological and nuclear exposed casualties" [7]. Army Reserve and National Guard troops must meet the same requirements by 2009. While questions remain as to whether these reforms will promote the acquisition of true core competency, they could substantially increase the number of qualified EMTs available in both the civilian and military emergency response systems. This is a critical first step in mounting an instantaneous, overwhelming disaster response.

45.5 Advanced Technologies

Against this backdrop, the object in any first responder training program "must be to provide a core set of skills that should be useful to the broad set of people who may become involved in responding to a terrorist incident..." [7]. The potential cohort is as broad as it is diverse. First responders may be firefighters, police officers, EMTs and, at least in some instances, even military personnel. Each profession brings its own preconceptions and procedures to an incident site. Absent agreed frameworks for cooperation, commanders could find themselves managing those functional and jurisdictional differences in addition to the disaster itself.

There are three primary challenges in preparing first responders for Homeland Security missions. The first is establishing a uniform training regimen for personnel that may respond to terrorist strikes. The second is integrating medical informatics into training and response, so that each element in the disaster response system possesses the situational awareness necessary to contain events, treat victims, and minimize unnecessary casualties among emergency response personnel. The third challenge is training for interoperability, using the U.S. military's proven doctrine of "train as you fight, fight as you train."

TES technologies can play a central role in meeting each of these challenges. In recent years, both Fortune 500 Companies and the U.S. Department of Defense adopted Computer-Based Training (CBT), which now provides the most cost-effective method for standardizing the acquisition of basic cognitive knowledge. Trainees can develop more advanced skills through interactive videoconferencing and simulations, which reduce the time necessary for hands-on training and improve the knowledge they bring to drills in the field. By using standardized, computer-based curricula, small training cadres from the federal government could supplement local and regional programs whenever necessary (see Figure 45.6).

Simulation technologies could also play a central role. Just as war games are invaluable in military planning and training, simulated crisis scenarios will likely prove instrumental in teaching first responders to cope with a broad spectrum of disaster situations. In addition to providing cost-effective technologies for regular drills, Virtual Reality (VR) and Augmented Reality (AR) trainers would enable personnel from various professional and jurisdictional backgrounds to simulate different operational roles. Frontline firefighters or police officers might play incident commanders, while state and federal decision makers could be tasked with first responder duties. Likewise, fire, police, and medical personnel might swap duties, thus providing valuable opportunities for cross training while elevating each individual's understanding of the manner in which the various elements of the disaster management system operate.

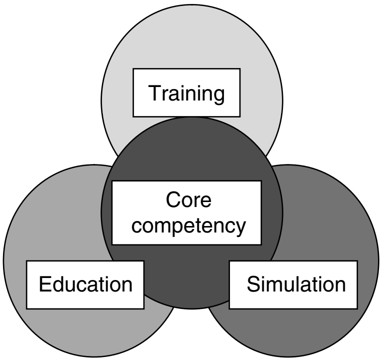


FIGURE 45.6 Training, Education and Simulation (TES).

Finally, these same technologies may also improve the use of medical informatics in crisis situations. The very hardware and software used for instruction purposes could carry real-time information regarding the condition, location, and status of victims in multiple casualty event. Combined with GPS and GIS technology, VR programs might link physicians in remote hospitals with first responders at the incident site — a particularly valuable interface in BW attacks and other CBRNE scenarios requiring swift treatment and effective containment [8].

Foremost among the beneficiaries of TES technologies would be large urban areas where daily operation tempos may limit the time available for continuing education. Rural and geographically isolated areas would also benefit, particularly from the dissemination of CBT software, which might be mailed to local fire and police stations on compact disks [16]. TES holds great promise at the state and federal level as well, particularly in intergovernmental planning and operations. Combined with remote communications platforms, databases, and sensor platforms, the same systems used to train and drill first responders could be employed within and among the various government agencies supporting personnel at the incident site. Broad dissemination of these C^2 nodes would dramatically enhance the flow and coordination of information in a multi-site, multi-event terrorist attack [8].

45.6 Conclusions

The objects of simulation-based training, drilling, and war-gaming are twofold. First, they will improve the "jointness" of disaster response and consequence management. Second, they will enhance the breadth and depth of existing capabilities, allowing local emergency personnel to respond in a much wider spectrum of potential crises.

As the threat of global terrorism grows, so does the need for a corps of fire, police, and EMTs trained to take the initiative in chaotic, dynamic events. The ideal first responder must be well rounded, an expert in his or her field, prepared for an eventuality and capable, when necessary, of working alone. These characteristics may be innate, but can also be learned. The human dimension is essential. Enabling local decision making is not only desirable, but a necessary first step in creating a net-centric crisis response paradigm in which other disaster information flows horizontally among all components of the Homeland Security architecture.

A centralized, federally controlled emergency management system is the ideal model. Instead, the optimal solution is to establish an operation milieu in which all elements of the Homeland Security architecture can most effectively employ their assets, information, and personnel. Establishing such a system demands more than just time and money. It will also require an institutional and technological transformation every bit as powerful and profound as the Revolution in Military Affairs (RMA) that transformed the U.S. military in the 1990s. That paradigm shift could take years — years that policy makers and the public may not have. While deeper reflection and greater imagination will be necessary to secure the homeland, implementing TES technologies and improving the use of medical informatics represent important near-term solutions.

References

[1] "The 9/11 Commission Report," National Commission on Terrorist Attacks Upon the United States, Washington, DC, July 22, 2004.

[2] J.M. Rosen, E. Grigg, S. McGrath, S. Lillibridge, and C.E. Koop, "Cybercare NDMS: An Improved Strategy for Biodefense Using Information Technologies," in *Integration of Health Telematics into Medical Practice*, M. Nerlich and U. Schaechinger, Eds. Amsterdam: IOS Press, 2003, pp. 95–114.

[3] N.F. Cantor, *In the Wake of the Plague: The Black Death and World It Made*. New York, NY: Free Press, 2001.

[4] J.M. Rosen, C.E. Koop, and E.B. Grigg, "Cybercare: A System for Confronting Bioterrorism," *The Bridge*, 32, 34–50, 2002.

[5] OTA, Proliferation of Weapons of Mass Destruction: Assessing the Risk, Government Printing Office, Washington, DC, 1999.

[6] J.M. Rosen, R. Gougelet, M. Mughal, and R. Hutchinson, "Conference Report of the Medical Disaster Conference," Dartmouth College, Hanover, NH, June 13–15, 2001.

[7] H. Kelly, V. Blackwood, M. Roper, G. Higgins, G. Klein, J. Tyler, D. Fletcher, H. Jenkins, A. Chisolm, and K. Squire, "Training Technology against Terror: Using Advanced Technology to Prepare America's Emergency Medical Personnel and First Responders for a Weapons of Mass Destruction Attack," Federation of American Scientists, Washington, DC, September 9, 2002.

[8] J.M. Rosen, E.B. Grigg, M.F. McKnight, C.E. Koop, S. Lillibridge, B.L. Kindberg, L. Hettinger, and R. Hutchinson, "Transforming Medicine for Biodefense and Healthcare Delivery: Developing a Dual-Use Doctrine that Utilizes Information Superiority and Network-Based Organization," *IEEE Eng. Med. Biol.*, 23, 89–101, 2004.

[9] "Homeland Security Oversight," in *The Washington Post.* Washington, DC, 2004, p. A18.

[10] M.F. Murphy, "Emergency Medical Services in Disaster," pp. 90–103.

[11] D. Langley, S. Michael Lockhardt, J. Michael Lockhardt, J.M. Rosen, and M.F. McKnight, "ICE 884: Response to Disaster," Durham, NH: Team Hill Studios, 2004.

[12] D.S. Alberts, J.J. Gartska, R.E. Hayes, and D.A. Signori, "Understanding Information Age Warfare," *CCRO Publication Series*, 2001.

[13] The authors wish to thank Jon Bowersox for this insight.

[14] S.L. Delp, P. Loan, C. Basdogan, and J.M. Rosen, "Surgical Simulation: An Emerging Technology for Training in Emergency Medicine," *Presence*, 6, 147–159, 1997.

[15] T. Lange, D.J. Indelicato, and J.M. Rosen, "Virtual Reality in Surgical Training," *Surg. Techn. Outcomes*, 9, 61–79, 2000.

[16] The prototype Virtual Terrorism Response Academy (VTRA) developed by Joseph Henderson and the Interactive Medical Laboratory at Dartmouth College operates on this principle.

V

Biomedical Sensors

Michael R. Neuman
Michigan Technological University

SENSORS CONVERT SIGNALS OF ONE type of quantity such as hydrostatic fluid pressure into an equivalent signal of another type of quantity, for example, an electrical signal. Biomedical sensors take signals representing biomedical variables and convert them into what is usually an electrical signal. As such, the biomedical sensor serves as the interface between a biologic and an electronic system and must function in such a way as to not adversely affect either of these systems. In considering biomedical sensors, it is necessary to consider both sides of the interface: the biologic and the electronic, since both biologic and electronic factors play an important role in sensor performance.

TABLE V.1 Classification of Biomedical Sensors

Physical sensors
Geometric
Mechanical
Thermal
Hydraulic
Electric
Optical
Chemical sensors
Gas
Electrochemical
Photometric
Other physical and chemical methods
Bioanalytic

Many different types of sensors can be used in biomedical applications. Table V.1 gives a general classification of these sensors. It is possible to categorize all sensors as being either physical or chemical. In the case of physical sensors, quantities such as geometric, mechanical, thermal, and hydraulic variables are measured. In biomedical applications these can include things such as muscle displacement, blood pressure, core body temperature, blood flow, cerebrospinal fluid pressure, and bone growth. Two types of physical sensors deserve special mention with regard to their biomedical application: sensors of electrical phenomena in the body, usually known as electrodes, play a special role as a result of their diagnostic and therapeutic applications. The most familiar of these are sensors used to obtain the electrocardiogram, an electrical signal produced by the heart. The other type of physical sensor that finds many applications in biology and medicine is the optical sensor. These sensors can use light to collect information, and, in the case of fiber optic sensors, light is the signal transmission medium as well.

The second major classification of sensing devices is chemical sensors. In this case the sensors are concerned with measuring chemical quantities such as identifying the presence of particular chemical compounds, detecting the concentrations of various chemical species, and monitoring chemical activities in the body for diagnostic and therapeutic applications. A wide variety of chemical sensors can be classified in many ways. One such classification scheme is illustrated in Table V.1 and is based upon the methods used to detect the chemical components being measured. Chemical composition can be measured in the gas phase using several techniques, and these methods are especially useful in biomedical measurements associated with the pulmonary system. Electrochemical sensors measure chemical concentrations or, more precisely, activities based on chemical reactions that interact with electrical systems. Photometric chemical sensors are optical devices that detect chemical concentrations based upon changes in light transmission, reflection, or color. The familiar litmus test is an example of an optical change that can be used to measure the acidity or alkalinity of a solution. Other types of physical chemical sensors such as the mass spectrometer use various physical methods to detect and quantify chemicals associated with biologic systems.

Although they are essentially chemical sensors, bioanalytic sensors are often classified as a separate major sensor category. These devices incorporate biologic recognition reactions such as enzyme–substrate, antigen–antibody, or ligand-receptor to identify complex biochemical molecules. The use of biologic reactions gives bioanalytic sensors high sensitivity and specificity in identifying and quantifying biochemical substances.

One can also look at biomedical sensors from the standpoint of their applications. These can be generally divided according to whether a sensor is used for diagnostic or therapeutic purposes in clinical medicine and for data collection in biomedical research. Sensors for clinical studies such as those carried out in the clinical chemistry laboratory must be standardized in such a way that errors that could result in an incorrect diagnosis or inappropriate therapy are kept to an absolute minimum. Thus these sensors must

TABLE V.2 Types of Sensor-Subject Interfaces

Noncontecting (noninvasive)
Skin surface (contacting)
Indwelling (minimally invasive)
Implantable (invasive)

not only be reliable themselves, but appropriate methods must exist for testing the sensors that are a part of the routine use of the sensors for making biomedical measurements.

One can also look at biomedical sensors from the standpoint of how they are applied to the patient or research subject. Table V.2 shows the range of general approaches to attaching biomedical sensors. At the top of the list we have the method that involves the least interaction with the biologic object being studied; the bottom of the list includes sensors that interact to the greatest extent. Clearly if a measurement can be made equally well by a sensor that does not contact the subject being measured or by one that must be surgically implanted, the former is by far the most desirable. However, a sensor that is used to provide information to help control a device already surgically placed in the body to replace or assist a failing organ should be implanted, since this is the best way to communicate with the internal device.

You will notice in reading this section that the majority of biomedical sensors are essentially the same as sensors used in other applications. The unique part about biomedical sensors is their application. There are, however, special problems that are encountered by biomedical sensors that are unique to them. These problems relate to the interface between the sensor and the biologic system being measured. The presence of foreign materials, especially implanted materials, can affect the biologic environment in which they are located. Many biologic systems are designed to deal with foreign materials by making a major effort to eliminate them. The rejection reaction that is often discussed with regard to implanted materials or transplanted tissues is an example of this. Thus, in considering biomedical sensors, one must worry about this rejection phenomenon and how it will affect the performance of the sensor. If the rejection phenomenon changes the local biology or chemistry around the sensor, this can result in the sensor measuring phenomena associated with the reaction that it has produced as opposed to phenomena characteristic of the biologic system being studied.

Biologic systems can also affect sensor performance. This is especially true for indwelling and implanted sensors. Biologic tissue represents a hostile environment which can degrade sensor structure and performance. In addition to many corrosive ions, body fluids contain enzymes that break down complex molecules as a part of the body's effort to rid itself of foreign and toxic materials. These can attack the materials that make up the sensor and its package, causing the sensor to lose calibration or fail.

Sensor packaging is an especially important problem. The package must not only protect the sensor from the corrosive environment of the body, but it must allow that portion of the sensor that performs the actual measurement to communicate with the biologic system. Furthermore, because it is frequently desirable to have sensors be as small as possible, especially those that are implanted and indwelling, it is important that the packaging function be carried out without significantly increasing the size of the sensor structure. There are many measurements that can now be made on biological specimens ranging from molecules through cells and larger structures. These measurements involve specialized sensors and instrumentation systems that are necessary for these types of measurements and their ultimate application in diagnostic medicine. This section concludes with a chapter devoted to describing some of the more common of these measurements. Although there have been many improvements in sensor packaging, this remains a major problem in biomedical sensor research. High-quality packaging materials that do not elicit major foreign body responses from the biologic system are still being sought.

Another problem that is associated with implanted sensors is that once they are implanted, access to them is very limited. This requires that these sensors be highly reliable so that there is no need to repair or replace them. It is also important that these sensors be highly stable, since in most applications it is not possible to calibrate the sensor *in vivo*. Thus, sensors must maintain their calibration once they are implanted, and for applications such as organ replacement, this can represent a potentially long time, the remainder of the patient's life.

In the following sections we will look at some of the sensors described above in more detail. We will consider physical sensors with special sections on biopotential electrodes and optical sensors. We will also look at chemical sensors, including bioanalytic sensing systems.[1] Although it is not possible to cover the field in extensive detail in a handbook such as this, it is hoped that these sections can serve as an introduction to this important aspect of biomedical engineering and instrumentation.

[1] There are many measurements that can now be made on biological specimens ranging from molecules through cells and larger structures. These measurements involve specialized sensors and instrumentation systems that are necessary for these types of measurements and their ultimate application in diagnostic medicine. This section concludes with a chapter devoted to describing some of the more common of these measurements.

46

Physical Measurements

Michael R. Neuman
Michigan Technological University

Physical variables associated with biomedical systems are measured by a group of sensors known as physical sensors. Although many specific physical variables can be measured in biomedical systems, these can be categorized into a simple list as shown in Table 46.1. Sensors for these variables, whether they are measuring biomedical systems or other systems, are essentially the same. Thus, sensors of linear displacement can frequently be used equally well for measuring the displacement of the heart muscle during the cardiac cycle or the movement of a robot arm. There is, however, one notable exception regarding the similarity of these sensors: the packaging of the sensor and attachment to the system being measured. Although physical sensors used in nonbiomedical applications need to be packaged so as to be protected from their environment, few of these sensors have to deal with the harsh environment of biologic tissue, especially with the mechanisms inherent in this tissue for trying to eliminate the sensor as a foreign body. Another notable exception to this similarity of sensors for measuring physical quantities in biologic and nonbiologic systems are the sensors used for fluidic measurements such as pressure and flow. Special needs for these measurements in biologic systems have resulted in special sensors and instrumentation systems for these measurements that can be quite different from systems for measuring pressure and flow in nonbiologic environments.

In this chapter, we will attempt to review various examples of sensors used for physical measurement in biologic systems. Although it would be beyond the scope of this chapter to cover all these in detail, the principal sensors applied for biologic measurements will be described. Each section will include a brief description of the principle of operation of the sensor and the underlying physical principles, examples of some of the more common forms of these sensors for application in biologic systems, methods of signal processing for these sensors where appropriate, and important considerations for when the sensor is applied.

TABLE 46.1 Physical Variables and Sensors

Physical quantity	Sensor	Variable sensed
Geometric	Strain gauge	Strain
	LVDT	Displacement
	Ultrasonic transit time	Displacement
Kinematic	Velocimeter	Velocity
	Accelerometer	Acceleration
Force–Torque	Load cell	Applied force or torque
Fluidic	Pressure transducer	Pressure
	Flow meter	Flow
Thermal	Thermometer	Temperature
	Thermal flux sensor	Heat flux

TABLE 46.2 Comparison of Displacement Sensors

Sensor	Electrical variable	Measurement circuit	Sensitivity	Precision	Range
Variable resistor	Resistance	Voltage divider, ohmmeter, bridge, current source	High	Moderate	Large
Foil strain gauge	Resistance	Bridge	Low	Moderate	Small
Liquid metal strain gauge	Resistance	Ohmmeter, bridge	Moderate	Moderate	Large
Silicon strain gauge	Resistance	Bridge	High	Moderate	Small
Mutual inductance coils	Inductance	Impedance bridge, inductance meter	Moderate to high	Moderate to low	Moderate to large
Variable reluctance	Inductance	Impedance bridge, inductance meter	High	Moderate	Large
LVDT	Inductance	Voltmeter	High	High	High
Parallel plate capacitor	Capacitance	Impedance bridge, capacitance meter	Moderate to high	Moderate	Moderate to large
Sonic/ultrasonic	Time	Timer circuit	High	High	Large

46.1 Description of Sensors

46.1.1 Linear and Angular Displacement Sensors

A comparison of various characteristics of displacement sensors described in detail below is outlined in Table 46.2.

46.1.1.1 Variable Resistance Sensor

One of the simplest sensors for measuring displacement is a variable resistor similar to the volume control on an audio electronic device [1]. The resistance between two terminals on this device is related to the linear or angular displacement of a sliding tap along a resistance element. Precision devices are available that have a reproducible, linear relationship between resistance and displacement. These devices can be connected in circuits that measure resistance such as an ohmmeter or bridge, or they can be used as a part of a circuit that provides a voltage that is proportional to the displacement. Such circuits include the voltage divider (as illustrated in Figure 46.1a) or driving a known constant current through the resistance and measuring the resulting voltage across it. This sensor is simple and inexpensive and can be used for measuring relatively large displacements.

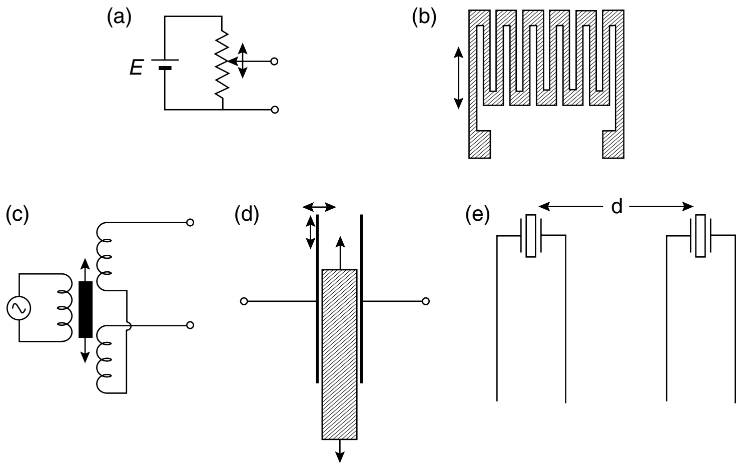


FIGURE 46.1 Examples of displacement sensors: (a) variable resistance sensor, (b) foil strain gauge, (c) linear variable differential transformer (LVDT), (d) parallel plate capacitive sensor, and (e) ultrasonic transit time displacement sensor.

There are some things to keep in mind when applying this type of displacement sensor. Since the circuit is a simple voltage divider, it is important that the electrical load on the output is very small such that there is very little current in the slider circuit. Significant current will introduce nonlinearities in the voltage vs. displacements characteristics. The sensor requires mechanical attachment to the structure being displaced and to some reference point, this can present difficulties in some biologic situations. Furthermore because the slider must move along the resistance element, this can introduce some friction that may alter the displacement.

46.1.1.2 Strain Gauge

Another displacement sensor based on an electrical resistance change is the strain gauge [2]. If a long narrow electrical conductor such as a piece of metal foil or a fine gauge wire is stretched within its elastic limit, it will increase in length and decrease in cross-sectional area. Because the electric resistance between both ends of this foil or wire can be given by

$$R = \rho \frac{l}{A} \tag{46.1}$$

where ρ is the electrical resistivity of the foil or wire material, l is its length, and A is its cross-sectional area, this stretching will result in an increase in resistance. The change in length can only be very small for the foil or wire to remain within its elastic limit, so the change in electric resistance will also be small. The relative sensitivity of this device is given by its gauge factor, γ, which is defined as

$$\gamma = \frac{\Delta R/R}{\Delta l/l} \tag{46.2}$$

where ΔR is the change in resistance when the structure is stretched by an amount Δl. Foil strain gauges are the most frequently applied and consist of a structure such as shown in Figure 46.1b. A piece of metal foil that is bonded to an insulating polymeric film such as polyimide that has a much greater compliance than the foil itself is chemically etched into the pattern shown in Figure 46.1b. When a strain is applied in the sensitive direction, the long direction of the individual elements of the strain gauge, the length of the gauge will be slightly increased, and this will result in an increase in the electrical resistance seen between the terminals. Since the displacement or strain that this structure can measure is quite small for it to remain within its elastic limit, it can only be used to measure small displacements such as occur as

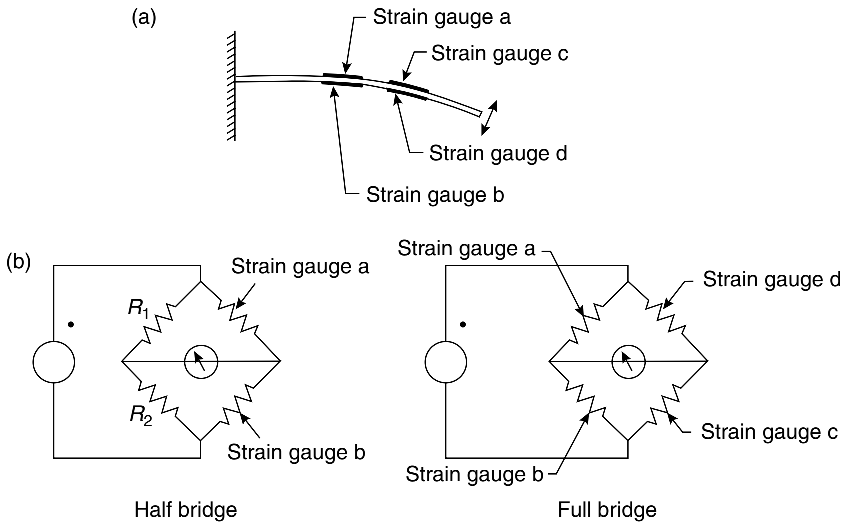


FIGURE 46.2 Strain gauges on a cantilever structure to provide temperature compensation: (a) cross-sectional view of the cantilever and (b) placement of the strain gauges in a half bridge or full bridge for temperature compensation and enhanced sensitivity.

loads are applied to structural beams. If one wants to increase the range of a foil strain gauge, one has to attach it to some sort of a mechanical impedance converter such as a cantilever beam. If the strain gauge is attached to one surface of the beam as shown in Figure 46.2a, a fairly large displacement at the unsupported end of the beam can be translated to a relatively small displacement on the beam's surface. It is possible for this structure to be used to measure larger displacements at the cantilever beam tip using a strain gauge bonded on the beam surface.

Because the electric resistance changes for a strain gauge are quite small, the measurement of this resistance change can be challenging. Generally, Wheatstone bridge circuits are used. It is important to note, however, that changes in temperature can also result in electric resistance changes that are of the same order of magnitude or even larger than the electric resistance changes due to the strain. Thus, it is important to temperature-compensate strain gauges in most applications. A simple method of temperature compensation is to use a double or quadruple strain gauge and a bridge circuit for measuring the resistance change. This is illustrated in Figure 46.2. If one can use the strain gauge in an application such as the cantilever beam application described above, one can place one or two of the strain gauge structures on the concave side of the beam and the other one or two on the convex side of the beam. Thus, as the beam deflects, the strain gauge on the convex side will experience tension, and that on the concave side will experience compression. By putting these gauges in adjacent arms of the Wheatstone bridge, their effects can double the sensitivity of the circuit in the case of the double strain gauge and quadruple it in the case where the entire bridge is made up of strain gauges on a cantilever. In addition to increased sensitivity, the bridge circuit minimizes temperatures effects on the strain measurements. Placing strain gauges that are on opposite sides of the beam in adjacent arms of the bridge results in a change in bridge output voltage when the beam is deflected. This occurs because the strain gauges on one side of the beam will increase in resistance while those on the other side will decrease in resistance when the beam is deflected. On the other hand, when the temperature of the beam changes, all of the strain gauges will have the same change in resistance, and this change will not affect the bridge output voltage.

In some applications it is not possible to place strain gauges so that one gauge is undergoing tension while the other is undergoing compression. In this case, the second strain gauge used for temperature compensation can be oriented such that its sensitive axis is in a direction where strain is minimal. Thus, it is still possible to have the temperature compensation by having two identical strain gauges at the same temperature in adjacent arms of the bridge circuit, but the sensitivity improvement described in the previous paragraph is not seen.

Another constraint imposed by temperature is that the material to which the strain gauge is attached and the strain gauge both have temperature coefficients of expansion. Thus, even if a gauge is attached to a structure under conditions of no strain, if the temperature is changed, the strain gauge could experience some strain due to the different expansion that it will have compared to the structure to which it is attached. To avoid this problem, strain gauges have been developed that have identical temperature coefficients of expansion to various common materials. In selecting a strain gauge, one should choose a device with thermal expansion characteristics as close as possible to those of the object upon which the strain is to be measured.

A more compliant structure that has found applications in biomedical instrumentation is the liquid metal strain gauge [3]. Instead of using a solid electric conductor such as the wire or metal foil, mercury confined to a compliant, thin wall, narrow bore elastomeric tube is used. The compliance of this strain gauge is determined by the elastic properties of the tube. Since only the elastic limit of the tube is of concern, this sensor can be used to detect much larger displacements than conventional strain gauges. Its sensitivity is roughly the same as a foil or wire strain gauge, but it is not as reliable. The mercury can easily become oxidized or small air gaps can occur in the mercury column. These effects make the sensor's characteristics noisy and sometimes results in complete failure.

Another variation on the strain gauge is the semiconductor strain gauge. These devices are frequently made out of pieces of silicon with strain gauge patterns formed using semiconductor microelectronic technology. The principal advantage of these devices is that their gauge factors can be more than 50 times greater than that of the solid and liquid metal devices. They are available commercially, but they are a bit more difficult to handle and attach to structures being measured due to their small size and brittleness.

46.1.2 Inductance Sensors

46.1.2.1 Mutual Inductance

The mutual inductance between two coils is related to many geometric factors, one of which is the separation of the coils. Thus, one can create a very simple displacement sensor by having two coils that are coaxial but with different separation. By driving one coil with an ac signal and measuring the voltage signal induced in the second coil, this voltage will be related to how far apart the coils are from one another. When the coils are close together, the mutual inductance will be relatively high, and so a higher voltage will be induced in the second coil; when the coils are more widely separated, the mutual inductance will be lower as will the induced voltage. The relationship between voltage and separation will be determined by the specific geometry of the coils and in general will not be a linear relationship with separation unless the change of displacement is relatively small. Nevertheless, this is a simple method of measuring separation that works reasonably well provided the coils remain coaxial. If there is movement of the coils transverse to their axes, it is difficult to separate the effects of transverse displacement from those of displacement along the axis.

46.1.2.2 Variable Reluctance

A variation on this sensor is the variable reluctance sensor wherein a single coil or two coils remain fixed on a form which allows a high reluctance material such as piece of iron to move into or out of the center of the coil or coils along their axis. Since the position of this core material determines the number of flux linkages through the coil or coils, this can affect the self-inductance or mutual inductance of the coils. In the case of the mutual inductance, this can be measured using the technique described in the previous paragraph, whereas self-inductance changes can be measured using various instrumentation circuits used for measuring inductance. This method is also a simple method for measuring displacements, but the characteristics are generally nonlinear, and the sensor often has only moderate precision.

46.1.2.3 Linear Variable Differential Transformer

By far the most frequently applied displacement transducer based upon inductance is the linear variable differential transformer (LVDT) [4]. This device is illustrated in Figure 46.1c and is essentially a three-coil variable reluctance transducer. The two secondary coils are situated symmetrically about and coaxial with the primary coil and connected such that the induced voltages in each secondary oppose each other. When a high-reluctance core is located in the center of the structure equidistant from each secondary coil, the voltage induced in each secondary will be the same. Since these voltages oppose one another, the output voltage from the device will be zero. As the core is moved closer to one or the other secondary coils, the voltages in each coil will no longer be equal, and there will be an output voltage proportional to the displacement of the core from the central, zero-voltage position. Because of the symmetry of the structure, this voltage is linearly related to the core displacement. When the core passes through the central, zero point, the phase of the output voltage from the sensor changes by 180°. Thus, by measuring the phase angle as well as the voltage, one can determine the position of the core. The circuit associated with the LVDT not only measures the voltage but often measures the phase angle as well.

Linear variable differential transformers are available commercially in many sizes and shapes. Depending on the configuration of the coils, they can measure displacements ranging from tens of micrometers through several centimeters.

46.1.3 Capacitive Sensors

Displacement sensors can be based upon measurements of capacitance as well as inductance. The fundamental principle of operation is the capacitance of a parallel plate capacitor as given by

$$C = e\frac{A}{d} \tag{46.3}$$

where e is the dielectric constant of the medium between the plates, d is the separation between the plates, and A is the cross-sectional area of the plates. Each of the quantities in Equation 46.3 can be varied to form a displacement transducer. By moving one of the plates with respect to the other, Equation 46.3 shows us that the capacitance will vary inversely with respect to the plate separation. This will give a hyperbolic capacitance–displacement characteristic. However, if the plate separation is maintained at a constant value and the plates are displaced laterally with respect to one another so that the area of overlap changes, this can produce a capacitance–displacement characteristic that can be linear, depending on the shape of the actual plates.

The third way that a variable capacitance transducer can measure displacement is by having a fixed parallel plate capacitor with a slab of dielectric material having a dielectric constant different from that of air that can slide between the plates (Figure 46.1d). The effective dielectric constant for the capacitor will depend on how much of the slab is between the plates and how much of the region between the plates is occupied only by air. This, also, can yield a transducer with linear characteristics.

The electronic circuitry used with variable capacitance transducers, is essentially the same as any other circuitry used to measure capacitance. As with the inductance transducers, this circuit can take the form of a bridge circuit or specific circuits that measure capacitive reactance.

46.1.4 Sonic and Ultrasonic Sensors

If the velocity of sound in a medium is constant, the time it takes a short burst of that sound energy to propagate from a source to a receiver will be proportional to the displacement between the two transducers. This is given by

$$d = cT \tag{46.4}$$

where c is the velocity of sound in the medium, T is the transit time, and d is the displacement. A simple system for making such a measurement is shown in Figure 46.1e [5]. A brief sonic or ultrasonic pulse is generated at the transmitting transducer and propagates through the medium. It is detected by the receiving transducer at time T after the burst was initiated. The displacement can then be determined by applying Equation 46.4.

In practice, this method is best used with ultrasound, since the wavelength is shorter, and the device will neither produce annoying sounds nor respond to extraneous sounds in the environment. Small piezoelectric transducers to generate and receive ultrasonic pulses are readily available. The electronic circuit used with this instrument carries out three functions (1) generation of the sonic or ultrasonic burst, (2) detection of the received burst, and (3) measurement of the time of propagation of the ultrasound. An advantage of this system is that the two transducers are coupled to one another only sonically. There is no physical connection as was the case for the other sensors described in this section.

46.1.5 Velocity Measurement

Velocity is the time derivative of displacement, and so all the displacement transducers mentioned above can be used to measure velocity if their signals are processed by passing them through a differentiator circuit. There are, however, two additional methods that can be applied to measure velocity directly.

46.1.5.1 Magnetic Induction

If a magnetic field that passes through a conducting coil varies with time, a voltage is induced in that coil that is proportional to the time-varying magnetic field. This relationship is given by

$$v = N\frac{\mathrm{d}\phi}{\mathrm{d}t} \tag{46.5}$$

where v is the voltage induced in the coil, N is the number of turns in the coil, and ϕ is the total magnetic flux passing through the coil (the product of the flux density and area within the coil). Thus a simple way to apply this principle is to attach a small permanent magnet to an object whose velocity is to be determined, and attach a coil to a nearby structure that will serve as the reference against which the velocity is to be measured. A voltage will be induced in the coil whenever the structure containing the permanent magnet moves, and this voltage will be related to the velocity of that movement. The exact relationship will be determined by the field distribution for the particular magnet and the orientation of the magnet with respect to the coil.

46.1.5.2 Doppler Ultrasound

When the receiver of a signal in the form of a wave such as electromagnetic radiation or sound is moving at a nonzero velocity with respect to the emitter of that wave, the frequency of the wave perceived by the receiver will be different than the frequency of the transmitter. This frequency difference, known as the Doppler shift, is determined by the relative velocity of the receiver with respect to the emitter and is given by

$$f_{\mathrm{d}} = \frac{f_{\mathrm{o}}u}{c} \tag{46.6}$$

where f_{d} is the Doppler frequency shift, f_{o} is the frequency of the transmitted wave, u is the relative velocity between the transmitter and receiver, and c is the velocity of sound in the medium. This principle can be applied in biomedical applications as a Doppler velocimeter. A piezoelectric transducer can be used as the ultrasound source with a similar transducer as the receiver. When there is no relative movement between the two transducers, the frequency of the signal at the receiver will be the same as that at the emitter, but when there is relative motion, the frequency at the receiver will be shifted according to Equation 46.6.

The ultrasonic velocimeter can be applied in the same way that the ultrasonic displacement sensor is used. In this case the electronic circuit produces a continuous ultrasonic wave and, instead of detecting

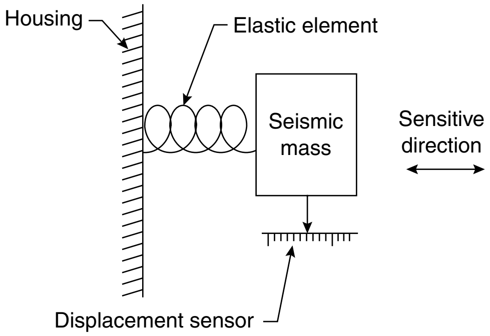


FIGURE 46.3 Fundamental structure of an accelerometer.

the transit time of the signal, now detects the frequency difference between the transmitted and received signals. This frequency difference can then be converted into a signal proportional to the relative velocity between the two transducers.

46.1.6 Accelerometers

Acceleration is the time derivative of velocity and the second derivative with respect to time of displacement. Thus, sensors of displacement and velocity can be used to determine acceleration when their signals are appropriately processed through differentiator circuits. In addition, there are direct sensors of acceleration based upon Newton's second law and Hooke's law. The fundamental structure of an accelerometer is shown in Figure 46.3. A known seismic mass is attached to the housing by an elastic element. As the structure is accelerated in the sensitive direction of the elastic element, a force is applied to that element according to Newton's second law. This force causes the elastic element to be distorted according to Hooke's law, which results in a displacement of the mass with respect to the accelerometer housing. This displacement is measured by a displacement sensor. The relationship between the displacement and the acceleration is found by combining Newton's second law and Hooke's law

$$a = \frac{k}{m}x \tag{46.7}$$

where x is the measured displacement, m is the known mass, k is the spring constant of the elastic element, and a is the acceleration. Any of the displacement sensors described above can be used in an accelerometer. The most frequently used displacement sensors are strain gauges or the LVDT. One type of accelerometer uses a piezoelectric sensor as both the displacement sensor and the elastic element. A piezoelectric sensor generates an electric signal that is related to the dynamic change in shape of the piezoelectric material as a force is applied. Thus, piezoelectric materials can only directly measure time varying forces. A piezoelectric accelerometer is, therefore, better for measuring changes in acceleration than for measuring constant accelerations. A principal advantage of piezoelectric accelerometers is that they can be made very small, which is useful in many biomedical applications. Very small and relatively inexpensive accelerometers are now made on a single silicon chip using microelectromechanical systems (MEMS) technology. An example is shown in Figure 46.4. A small piece of silicon is etched to give the paddle-like structure that is attached to the silicon frame at one end and is free to move with respect to the frame by flexing the "handle" of the paddle. This movement results in strains being induced on the surfaces of the "handle," and a strain gauge integrated into this handle structures detects the strain and converts it to an electrical signal. The paddle, itself serves as the seismic mass, and the "handle" is the elastic element.

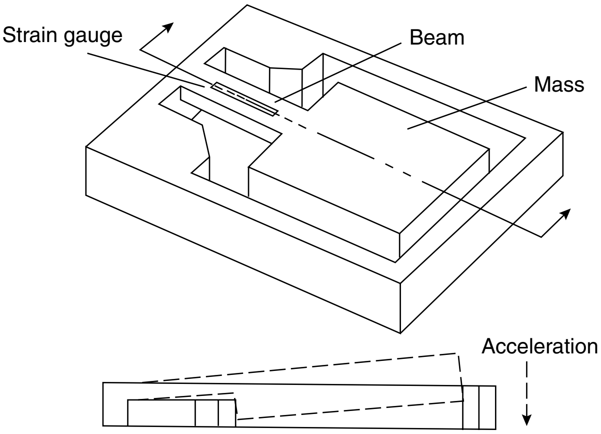


FIGURE 46.4 Example of a silicon chip accelerometer fabricated using MEMS technology. The lower figure shows the upward deflection of the seismic mass with a downward acceleration.

46.1.7 Force

Force is measured by converting the force to a displacement and measuring the displacement with a displacement sensor. The conversion takes place as a result of the elastic properties of a material. Applying a force to the material distorts the material's shape, and this distortion can be measured by a displacement sensor. For example, the cantilever structure shown in Figure 46.2a could be a force sensor. Applying a vertical force at the tip of the beam will cause the beam to deflect according to its elastic properties. This deflection can be detected using a displacement sensor such as a strain gauge as described previously.

A common form of force sensor is the load cell. This consists of a block of material with known elastic properties that has strain gauges attached to it. Applying a force to the load cell stresses the material, resulting in a strain that can be measured by the strain gauge. Applying Hooke's law, one finds that the strain is proportional to the applied force. The strain gauges on a load cell are usually in a half-bridge or full-bridge configuration to minimize the temperature sensitivity of the device. Load cells come in various sizes and configurations, and they can measure a wide range of forces.

46.1.8 Measurement of Fluid Dynamic Variables

The measurement of the fluid pressure and flow in both liquids and gases is important in many biomedical applications. These two variables, however, often are the most difficult variables to measure in biologic applications because of interactions with the biologic system and stability problems. Some of the most frequently applied sensors for these measurements are described in the following paragraphs.

46.1.8.1 Pressure Measurement

Sensors of pressure for biomedical measurements such as blood pressure [6] consist of a structure such as shown in Figure 46.5. In this case a fluid coupled to the fluid to be measured is housed in a chamber with a flexible diaphragm making up a portion of the wall, with the other side of the diaphragm at atmospheric pressure. When a pressure exists across the diaphragm, it will cause the diaphragm to deflect. This deflection is then measured by a displacement sensor. In the example in Figure 46.5, the displacement sensor consists of four fine-gauge wires drawn between a structure attached to the diaphragm and the housing of the pressure sensor so that these wires serve as strain gauges. When pressure causes the diaphragm to deflect, two of the fine-wire strain gauges will be extended by a small amount, and the other two will contract by the same amount. By connecting these wires into a bridge circuit, a voltage proportional to the deflection of the diaphragm and hence the pressure can be obtained.

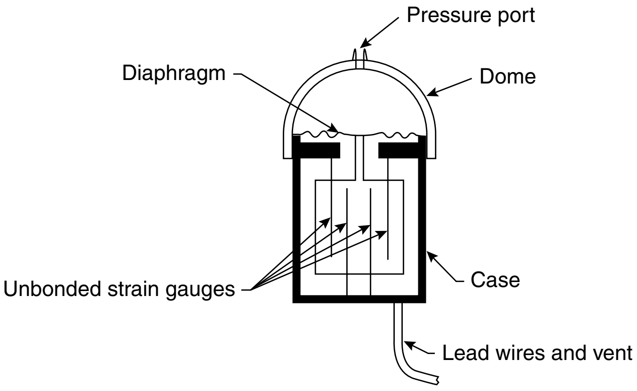


FIGURE 46.5 Structure of an unbonded strain gauge pressure sensor. (Reproduced from Neuman M.R. 1993. In R.C. Dorf (Ed.), *The Electrical Engineering Handbook,* Boca Raton, FL, CRC Press. With permission).

Semiconductor technology has been applied to the design of pressure transducers such that the entire structure can be fabricated from silicon. A portion of a silicon chip can be formed into a diaphragm and semiconductor strain gauges incorporated directly into that diaphragm to produce a small, inexpensive, and sensitive pressure sensor. Such sensors can be used as disposable, single-use devices for measuring blood pressure without the need for additional sterilization before being used on the next patient. This minimizes the risk of transmitting blood-borne infections in the cases where the transducer is coupled directly to the patient's blood for direct blood pressure measurement.

In using this type of sensor to measure blood pressure, it is necessary to couple the chamber containing the diaphragm to the blood or other fluids being measured. This is usually done using a small, flexible plastic tube known as a catheter, that can have one end placed in an artery of the subject while the other is connected to the pressure sensor. This catheter is filled with a physiologic saline solution so that the arterial blood pressure is coupled to the sensor diaphragm. This external blood-pressure-measurement method is used quite frequently in the clinic and research laboratory, but it has the limitation that the properties of the fluid in the catheter and the catheter itself can affect the measurement. For example, both ends of the catheter must be at the same vertical level to avoid a pressure offset due to hydrostatic effects. Also, the compliance of the tube will affect the frequency response of the pressure measurement. Air bubbles in the catheter or obstructions due to clotted blood or other materials can introduce distortion of the waveform due to resonance and damping. These problems can be minimized by utilizing a miniature semiconductor pressure transducer that is located at the tip of a catheter and can be placed in the blood vessel rather than being positioned external to the body. Such internal pressure sensors are available commercially and have the advantages of a much broader frequency response, no hydrostatic pressure error, and generally clearer signals than the external system.

Although it is possible to measure blood pressure using the techniques described above, this remains one of the major problems in biomedical sensor technology. Long-term stability of pressure transducers is not very good. This is especially true for pressure measurements of venous blood, cerebrospinal fluid, or fluids in the gastrointestinal tract, where pressures are usually relatively low. Long-term changes in baseline pressure for most pressure sensors require that they be frequently adjusted to be certain of zero pressure. Although this can be done relatively easily when the pressure transducer is located external to the body, this can be a major problem for indwelling or implanted pressure transducers. Thus, these transducers must be extremely stable and have low baseline drift to be useful in long-term applications. The packaging of the pressure transducer is also a problem that needs to be addressed, especially when the transducer is in contact with blood for long periods. Not only must the package be biocompatible, but it also must allow the appropriate pressure to be transmitted from the biologic fluid to the diaphragm. Thus, a material that is mechanically stable under corrosive and aqueous environments in the body is needed.

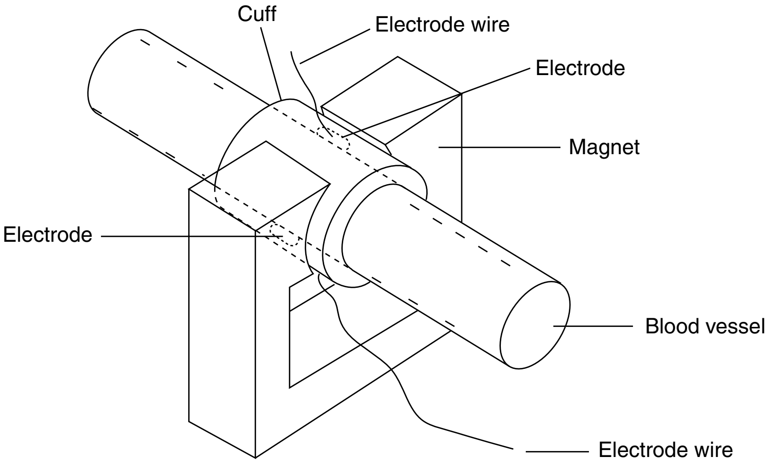


FIGURE 46.6 Fundamental structure of an electromagnetic flowmeter. (Reproduced from Neuman M.R. 1986. In J.D. Bronzino (Ed.), *Biomedical Engineering and Instrumentation: Basic Concepts and Applications,* Boston, PWS Publishers. With permission.)

46.1.8.2 Measurement of Flow

The measurement of true volummetric flow in the body represents one of the most difficult problems in biomedical sensing [7]. The sensors that have been developed measure velocity rather than volume flow, and they can only be used to measure flow if the velocity is measured for a tube of known cross-section. Thus, most flow sensors constrain the vessel to have a specific cross-sectional area.

The most frequently used flow sensor in biomedical systems is the electromagnetic flow meter illustrated in Figure 46.6. This device consists of a means of generating a magnetic field transverse to the flow vector in a vessel. A pair of very small biopotential electrodes are attached to the wall of the vessel such that the vessel diameter between them is at right angles to the direction of the magnetic field. As the blood flows in the structure, ions in the blood deflect in the direction of one or the other electrodes due to the magnetic field, and this results in a voltage across the electrodes that is given by

$$v = Blu \tag{46.8}$$

where B is the magnetic field, l is the distance between the electrodes, and u is the average instantaneous velocity of the fluid across the vessel. If the sensor constrains the blood vessel to have a specific diameter, then its cross-sectional area will be known, and multiplying this area by the velocity will give the volume flow.

Although d.c. flow sensors have been developed and are available commercially, the most desirable method is to use ac excitation of the magnetic field so that offset potential effects from the biopotential electrodes do not generate errors in this measurement.

Small ultrasonic transducers can also be attached to a blood vessel to measure flow as illustrated in Figure 46.7. In this case the transducers are oriented such that one transmits a continuous ultrasound signal that illuminates the blood. Cells within the blood diffusely reflect this signal in the direction of the second sensor so that the received signal undergoes a Doppler shift in frequency that is proportional to the velocity of the blood. By measuring the frequency shift and knowing the cross-sectional area of the vessel, it is possible to determine the flow.

Another method of measuring flow that has had biomedical application is the measurement of cooling of a heated object by convection. The object is usually a thermistor (see Section 46.1.8.3) placed either in a blood vessel or in tissue, and the thermistor serves as both the heating element and the temperature sensor. In one mode of operation, the amount of power required to maintain the thermistor at a temperature

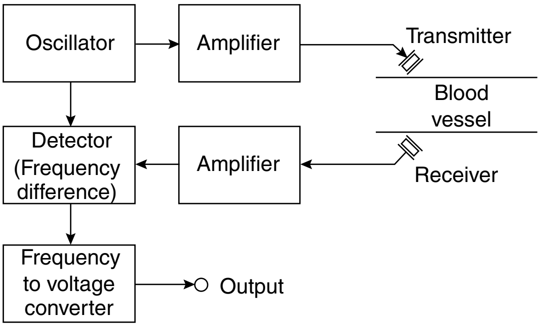


FIGURE 46.7 Structure of an ultrasonic Doppler flowmeter with the major blocks of the electronic signal processing system. The oscillator generates a signal that, after amplification, drives the transmitting transducer. The oscillator frequency is usually in the range of 1 to 10 MHz. The reflected ultrasound from the blood is sensed by the receiving transducer and amplified before being processed by a detector circuit. This block generates the frequency difference between the transmitted and received ultrasonic signals. This difference frequency can be converted into a voltage proportional to frequency, and hence flow velocity, by the frequency to voltage converter circuit.

TABLE 46.3 Properties of Temperature Sensors

Sensor	Form	Sensitivity	Stability	Range (°C)
Metal resistance thermometer	Coil of fine platinum wire	Low	High	−100 to 700
Thermistor	Bead, disk, chip, or rod	High	Moderate	−50 to 150
Thermocouple	Pair of wires	Low	High	−100 to >1500
Mercury in glass thermometer	Column of Hg in glass capillary	Moderate	High	−50 to 400
Silicon *p*–*n* diode	Electronic component	Moderate	High	−50 to 150

slightly above that of the blood upstream is measured. As the flow around the thermistor increases, more heat is removed from the thermistor by convection, and so more power is required to keep it at a constant temperature. Relative flow is then measured by determining the amount of power supplied to the thermistor.

In a second approach the thermistor is heated by applying a current pulse and then measuring the cooling curve of the thermistor as the blood flows across it. The thermistor will cool more quickly as the blood flow increases. Both these methods are relatively simple to achieve electronically, but both also have severe limitations. They are essentially qualitative measures and strongly depend on how the thermistor probe is positioned in the vessel being measured. If the probe is closer to the periphery or even in contact with the vessel wall, the measured flow will be different than if the sensor is in the center of the vessel.

46.1.8.3 Temperature

There are many different sensors of temperature [8], but three find particularly wide application to biomedical problems. Table 46.3 summarizes the properties of various temperature sensors, and these three, including metallic resistance thermometers, thermistors, and thermocouples, are described in the following paragraphs.

TABLE 46.4 Temperature Coefficient of Resistance for Common Metals and Alloys

Metal or alloy	Resistivity at 20°C $\mu\Omega$-cm	Temperature coefficient of resistance (%/°C)
Platinum	9.83	0.3
Gold	2.22	0.368
Silver	1.629	0.38
Copper	1.724	0.393
Constantan (60% Cu, 40% Ni)	49.0	0.0002
Nichrome (80% Ni, 20% Cr)	108.0	0.013

Source: Pender H. and McIlwain K. 1957. *Electrical Engineers' Handbook*, 4th ed., New York, John Wiley & Sons.

46.1.8.4 Metallic Resistance Thermometers

The electric resistance of a piece of metal or wire generally increases as the temperature of that electric conductor increases. A linear approximation to this relationship is given by

$$R = R_0[1 + \alpha(T - T_0)] \tag{46.9}$$

where R_0 is the resistance at temperature T_0, α is the temperature coefficient of resistance, and T is the temperature at which the resistance is being measured. Most metals have temperature coefficients of resistance of the order of 0.1 to 0.4%/°C, as indicated in Table 46.4. The noble metals are preferred for resistance thermometers, since they do not corrode easily and, when drawn into fine wires, their cross-section will remain constant, thus avoiding drift in the resistance over time which could result in an unstable sensor. It is also seen from Table 46.4 that the noble metals, gold and platinum, have some of the highest temperature coefficients of resistance of the common metals.

Metal resistance thermometers are often fabricated from fine-gauge insulated wire that is wound into a small coil. It is important in doing so to make certain that there are not other sources of resistance change that could affect the sensor. For example, the structure should be utilized in such a way that no external strains are applied to the wire, since the wire could also behave as a strain gauge. Metallic films and foils can also be used as temperature sensors, and commercial products are available in the wire, foil, or film forms. The electric circuits used to measure resistance, and hence the temperature, are similar to those used with the wire or foil strain gauges. A bridge circuit is the most desirable, although ohmmeter circuits can also be used. It is important to make sure that the electronic circuit does not pass a large current through the resistance thermometer for that would cause self-heating due to the Joule conversion of electric energy to heat.

46.1.9 Thermistors

Unlike metals, semiconductor materials have an inverse relationship between resistance and temperature. This characteristic is very nonlinear and cannot be characterized by a linear equation such as for the metals. The thermistor is a semiconductor temperature sensor. Its resistance as a function of temperature is given by

$$R = R_0 e^{\beta\left[\frac{1}{T} - \frac{1}{T_0}\right]} \tag{46.10}$$

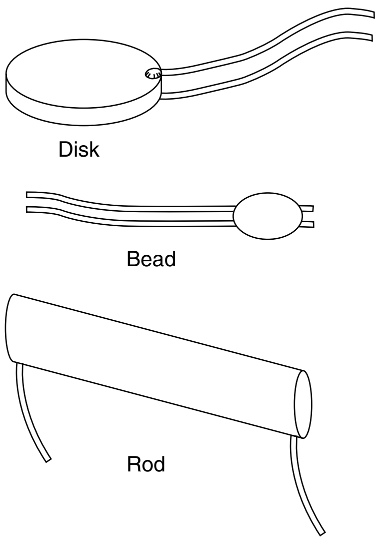


FIGURE 46.8 Common forms of thermistors.

where β is a constant determined by the materials that make up the thermistor. Thermistors can take a variety of forms and cover a large range of resistances. The most common forms used in biomedical applications are the bead, disk, or rod forms of the sensor as illustrated in Figure 46.8. These structures can be formed from a variety of different semiconductors ranging from elements such as silicon and germanium to mixtures of various semiconducting metallic oxides. Most commercially available thermistors are manufactured from the latter materials, and the specific materials as well as the process for fabricating them are closely held industrial secrets. These materials are chosen not only to have high sensitivity but also to have the greatest stability, since thermistors are generally not as stable as the metallic resistance thermometers. However, thermistors can be close to an order of magnitude more sensitive.

46.1.10 Thermocouples

When different regions of an electric conductor or semiconductor are at different temperatures, there is an electric potential between these regions that is directly related to the temperature differences. This phenomenon, known as the Seebeck effect, can be used to produce a temperature sensor known as a thermocouple by taking a wire of metal or alloy A and another wire of metal or alloy B and connecting them as shown in Figure 46.9. One of the junctions is known as the sensing junction, and the other is the reference junction. When these junctions are at different temperatures, a voltage proportional to the temperature difference will be seen at the voltmeter when metals A and B have different Seebeck coefficients. This voltage is roughly proportional to the temperature difference and can be represented over the relatively small temperature differences encountered in biomedical applications by the linear equation

$$V = S_{\mathrm{AB}}(T_{\mathrm{s}} - T_{\mathrm{r}}) \tag{46.11}$$

where S_{AB} is the Seebeck coefficient for the thermocouple made up of metals A and B. Although this equation is a reasonable approximation, more accurate data are usually found in tables of actual voltages as a function of temperature difference. In some applications the voltmeter is located at the reference junction, and one uses some independent means such as a mercury in glass thermometer to measure the reference junction temperature. Where precision measurements are made, the reference junction is often placed in an environment of known temperature such as an ice bath. Electronic measurement of reference junction temperature can also be carried out and used to compensate for the reference junction

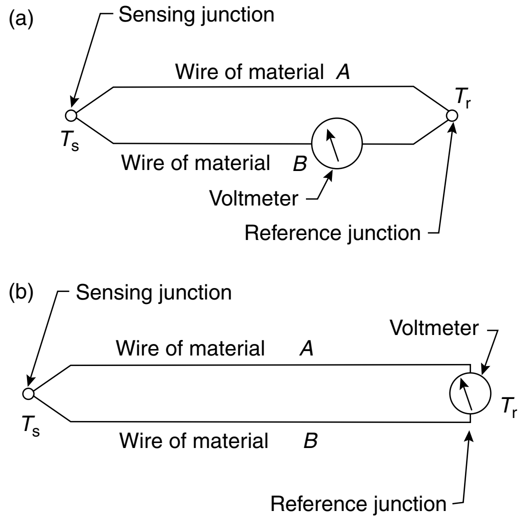


FIGURE 46.9 Circuit arrangement for a thermocouple showing the voltage-measuring device, the voltmeter, interrupting one of the thermocouple wires (a) and at the cold junction (b).

TABLE 46.5 Common Thermocouples

Type	Materials	Seebeck coefficient, μV/°C[a]	Temperature range (°C)
S	Platinum/platinum 10% rhodium	6	0 to 1700
T	Copper/constantan	50	−190 to 400
K	Chromel/alumel	41	−200 to 1370
J	Iron/constantan	53	−200 to 760
E	Chromel/constantan	78	−200 to 970

[a] Seebeck coefficient value is at a temperature of 25°C.

temperature so that the voltmeter reads a signal equivalent to what would be seen if the reference junction were at 0°C. This electronic reference junction compensation is usually carried out using a metal resistance temperature sensor to determine reference junction temperature.

The voltages generated by thermocouples used for temperature measurement are generally quite small being on the order of tens of microvolts per °C. Thus, for most biomedical measurements where there is only a small difference in temperature between the sensing and reference junction, very sensitive voltmeters or amplifiers must be used to measure these potentials. Thermocouples have been used in industry for temperature measurement for many years. Several standard alloys to provide optimal sensitivity and stability of these sensors have evolved. Table 46.5 lists these common alloys, the Seebeck coefficient for thermocouples of these materials at room temperature, and the full range of temperatures over which these thermocouples can be used.

Thermocouples can be fabricated in many different ways depending on their applications. They are especially suitable for measuring temperature differences between two structures, since the sensing junction can be placed on one while the other has the reference junction. Higher-output thermocouples or thermopiles can be produced by connecting several thermocouples in series. Thermocouples can be made from very fine wires that can be implanted in biologic tissues for temperature measurements, and it is also possible to place these fine-wire thermocouples within the lumen of a hypodermic needle to make short-term temperature measurements in tissue. Microfabrication technology has been made it possible to make thermocouples small enough to fit within a living cell.

46.2 Biomedical Applications of Physical Sensors

Just as it is not possible to cover the full range of physical sensors in this chapter, it is also impossible to consider the many biomedical applications that have been reported for these sensors. Instead, some representative examples will be given. These are summarized in Table 46.6 and will be briefly described in the following paragraphs.

Liquid metal strain gauges are especially useful in biomedical applications, because they are mechanically compliant and provide a better mechanical impedance match to most biomedical tissues than other types of strain gauges. By wrapping one of these strain gauges around a circumference of the abdomen, it will stretch and contract with the abdominal breathing movements. The signal from the strain gauge can then be used to monitor breathing in patients or experimental animals. The advantage of this sensor is its compliance so that it does not interfere with the breathing movements or substantially increase the required breathing effort.

One of the original applications of the liquid metal strain gauge was in limb plethysmography [3]. One or more of these sensors are wrapped around an arm or leg at various points and can be used to measure changes in circumference that are related to the cross-sectional area and hence the volume of the limb at those points. If the venous drainage from the limb is occluded, the limb volume will increase as it fills with blood. Releasing the occlusion allows the volume to return to normal. The rate of this decrease in volume can be monitored using the liquid metal strain gauges, and this can be used to identify venous blockage when the return to baseline volume is too slow.

Breathing movements, although not volume, can be seen using a simple magnetic velocity detector. By placing a small permanent magnet on the anterior side of the chest or abdomen and a flat, large-area coil on the posterior side opposite from the magnet, voltages are induced in the coil as the chest of abdomen moves during breathing. The voltage itself can be used to detect the presence of breathing movements, or it can be electronically integrated to give a signal related to displacement.

The LVDT is a displacement sensor that can be used for more precise applications. For example, it can be used in studies of muscle physiology where one wants to measure the displacement of a muscle or where one is measuring the isometric force generated by the muscle (using a load cell) and must ensure that there is no muscle movement. It can also be incorporated into other physical sensors such as a pressure sensor or a tocodynamometer, a sensor used to electronically "feel" uterine contractions of patients in labor or those at risk of premature labor and delivery.

TABLE 46.6 Examples of Biomedical Applications of Physical Sensors

Sensor	Application	Signal range	Reference
Liquid metal strain gauge	Breathing movement	0–0.05 (strain)	
	Limb plethysmography	0–0.02 (strain)	3
Magnetic displacement sensor	Breathing movement	0–10 mm	10
LVDT	Muscle contraction	0–20 mm	
	Uterine contraction sensor	0–5 mm	11
Load cell	Electronic scale	0–440 lbs (0–200 kg)	12
Accelerometer	Subject activity	0–20 m/sec^2	13
Miniature silicon pressure sensor	Intra-arterial blood pressure	0–50 Pa (0–350 mmHg)	
	Urinary bladder pressure	0–10 Pa (0–70 mmHg)	
	Intrauterine pressure	0–15 Pa (0–100 mmHg)	14
Electromagnetic flow sensor	Cardiac output (with integrator)	0–500 ml/min	
	Organ blood flow	0–100 ml/min	15

In addition to studying muscle forces, load cells can be used in various types of electronic scales for weighing patients or study animals. The simplest electronic scale consists of a platform placed on top of a load cell. The weight of any object placed on the platform will produce a force that can be sensed by the load cell. In some critical care situations in the hospital, it is important to carefully monitor the weight of a patient. For example, this is important in watching water balance in patients receiving fluid therapy. The electronic scale concept can be extended by placing a load cell under each leg of the patient's bed and summing the forces seen by each load cell to get the total weight of the patient and the bed. Since the bed weight remains fixed, weight changes seen will reflect changes in patient weight.

Accelerometers can be used to measure patient or research subject activity. By attaching a small accelerometer to the individual being studied, any movements can be detected. This can be useful in sleep studies where movement can help to determine the sleep state. Miniature accelerometers and recording devices can also be worn by patients to study activity patterns and determine effects of disease or treatments on patient activity [9].

Miniature silicon pressure sensors are used for the indwelling measurement of fluid pressure in most body cavities. The measurement of intra-arterial blood pressure is the most frequent application, but pressures in other cavities such as the urinary bladder and the uterus are also measured. The small size of these sensors and the resulting ease of introduction of the sensor into the cavity make these sensors important for these applications.

The electromagnetic flow sensor has been a standard method in use in the physiology laboratory for many years. Its primary application has been for measurement of cardiac output and blood flow to specific organs in research animals. New miniature inverted electromagnetic flow sensors make it possible to temporarily introduce a flow probe into an artery through its lumen to make clinical measurements.

The measurement of body temperature using instruments employing thermistors as the sensor has greatly increased in recent years. Rapid response times of these low-mass sensors make it possible to quickly assess patients' body temperatures so that more patients can be evaluated in a given period. This can then help to reduce health care costs. The rapid response time of low-mass thermistors makes them a simple sensor to be used for sensing breathing. By placing small thermistors near the nose and mouth, the elevated temperature of exhaled air can be sensed to document a breath [10].

The potential applications of physical sensors in medicine and biology are almost limitless. To be able to use these devices, however, scientists must first be familiar with the underlying sensing principles. It is then possible to apply these in a form that addresses the problems at hand.

References

[1] Doebelin E.O. 2003. *Measurement Systems: Applications and Design*, New York, McGraw-Hill.

[2] Dechow P.C. 1988. Strain gauges. In J. Webster (Ed.), *Encyclopedia of Medical Devices and Instrumentation*, pp. 2715–2721, New York, John Wiley & Sons.

[3] Whitney R.J. 1949. The measurement of changes in human limb-volume by means of a mercury-in-rubber strain gauge. *J. Physiol.* 109: 5.

[4] Schaevitz H. 1947. The linear variable differential transformer. *Proc. Soc. Stress Anal.* 4: 79.

[5] Stegall H.F., Kardon M.B., Stone H.L. et al. 1967. A portable simple sonomicrometer. *J. Appl. Physiol.* 23: 289.

[6] Geddes L.A. 1991. *Handbook of Blood Pressure Measurement*, Totowa, NJ, Humana.

[7] Roberts V.C. 1972. *Blood Flow Measurements*, Baltimore, Williams & Wilkins.

[8] Herzfeld C.M. (Ed). 1962. *Temperature: Its Measurement and Control in Science and Industry*, New York, Reinhold.

[9] Patterson S.M., Krantz D.S., Montgomery L.C. et al. 1993. Automated physical activity monitoring: validation and comparison with physiological and self-report measures. *Psychophysiology* 30: 296.

[10] Sekey, A. and Seagrave, C. 1981. Biomedical subminiature thermistor sensor for analog control by breath flow, *Biomater. Med. Dev. Artif. Org.* 9: 73–90.

[11] Angelsen B.A. and Brubakk A.O. 1976. Transcutaneous measurement of blood flow velocity in the human aorta. *Cardiovasc. Res.* 10: 368.

[12] Rolfe P. 1971. A magnetometer respiration monitor for use with premature babies. *Biomed. Eng.* 6: 402.

[13] Reddy N.P. and Kesavan S.K. 1988. Linear variable differential transformers. In J. Webster (Ed.), *Encyclopedia of Medical Devices and Instrumentation*, pp. 1800–1806, New York, John Wiley & Sons.

[14] Roe F.C. 1966. New equipment for metabolic studies. *Nurs. Clin. N. Am.* 1: 621.

[15] Fleming D.G., Ko W.H., and Neuman M.R. (Eds). 1977. *Indwelling and Implantable Pressure Transducers*, Cleveland, CRC Press.

[16] Wyatt D.G. 1971. Electromagnetic blood flow measurements. In B.W. Watson (Ed.), *IEE Medical Electronics Monographs*, London, Peregrinus.

[17] Bently, J.P. 2005. *Principles of Measurement Systems* 4th ed., Englewood Cliffs, N.J., Pearson Prentice-Hall.

[18] Webster J.G. 1999. *Mechanical Variables Measurement — Solid, Fluid, and Thermal*, Boca Raton, CRC Press.

[19] Michalski, L., Eckersdorf, K., Kucharski, J., and Mc Ghee, J. 2001. *Temperature Measurement*, 2nd ed., New York, John Wiley & Sons.

[20] Childs, P.R.N. 2001. *Practical Temperature Measurement*, Oxford, Butterworth-Heinemann.

Further Information

Good overviews of physical sensors are found in these books: Doebelin E.O. 1990. *Measurement Systems: Application and Design*, 4th ed., New York, McGraw-Hill; Harvey, G.F. (Ed.). 1969. *Transducer Compendium*, 2nd ed., New York, Plenum. One can also find good descriptions of physical sensors in chapters of two works edited by John Webster. Chapters 2, 7, and 8 of his textbook (1998) *Medical Instrumentation: Application and Design*, 3rd ed., New York, John Wiley & Sons, and several articles in his *Encyclopedia on Medical Devices and Instrumentation*, published by Wiley in 1988, cover topics on physical sensors.

Although a bit old, the text *Transducers for Biomedical Measurements* (New York, John Wiley & Sons, 1974) by Richard S.C. Cobbold, remains one of the best descriptions of biomedical sensors available. By supplementing the material in this book with recent manufacturers' literature, the reader can obtain a wealth of information on physical (and for that matter chemical) sensors for biomedical application.

The journals *IEEE Transactions on Biomedical Engineering and Medical and Biological Engineering and Computing* are good sources of recent research on biomedical applications of physical sensors. The journals *Physiological Measurement and Sensors and Actuators* are also good sources for this material as well as papers on the sensors themselves. The IEEE sensors Journal covers many different types of sensors, but biomedical devices are included in its scope.

47

Biopotential Electrodes

Michael R. Neuman
Michigan Technological University

Biologic systems frequently have electric activity associated with them. This activity can be a constant d.c. electric field, a constant flux of charge-carrying particles or current, or a time-varying electric field or current associated with some time-dependent biologic or biochemical phenomenon. Bioelectric phenomena are associated with the distribution of ions or charged molecules in a biologic structure and the changes in this distribution resulting from specific processes. These changes can occur as a result of biochemical reactions, or they can emanate from phenomena that alter local anatomy.

One can find bioelectric phenomena associated with just about every organ system in the body. Nevertheless, a large proportion of these signals are associated with phenomena that are at the present time not especially useful in clinical medicine and represent time-invariant, low-level signals that are not easily measured in practice. There are, however, several signals that are of diagnostic significance or that provide a means of electronic assessment to aid in understanding biologic systems. These signals, their usual abbreviations, and the systems they measure are listed in Table 47.1. Of these, the most familiar is the electrocardiogram, a signal derived from the electric activity of the heart. This signal is widely used in diagnosing disturbances in cardiac rhythm, signal conduction through the heart, and damage due to cardiac ischemia and infarction. The electromyogram is used for diagnosing neuromuscular diseases, and the electroencephalogram is important in identifying brain dysfunction and evaluating sleep. The other signals listed in Table 47.1 are currently of lesser diagnostic significance but are, nevertheless, used for studies of the associated organ systems.

Although Table 47.1 and the above discussion are concerned with bioelectric phenomena in animals and these techniques are used primarily in studying mammals, bioelectric signals also arise from plants [1]. These signals are generally steady-state or slowly changing, as opposed to the time-varying signals listed in Table 47.1. An extensive literature exists on the origins of bioelectric signals, and the interested reviewer is referred to the text by Plonsey and Barr for a general overview of this area [2].

TABLE 47.1 Bioelectric Signals Sensed by Biopotential Electrodes and Their Sources

Bioelectric signal	Abbreviation	Biologic source
Electrocardiogram	ECG	Heart — as seen from body surface
Cardiac electrogram	—	Heart — as seen from within
Electromyogram	EMG	Muscle
Electroencephalogram	EEG	Brain
Electrooptigram	EOG	Eye dipole field
Electroretinogram	ERG	Eye retina
Action potential	—	Nerve or muscle
Electrogastrogram	EGG	Stomach
Galvanic skin reflex	GSR	Skin

47.1 Sensing Bioelectric Signals

The mechanism of electric conductivity in the body involves ions as charge carriers. Thus, picking up bioelectric signals involves interacting with these ionic charge carriers and transducing ionic currents into electric currents required by wires and electronic instrumentation. This transducing function is carried out by electrodes that consist of electrical conductors in contact with the aqueous ionic solutions of the body. The interaction between electrons in the electrodes and ions in the body can greatly affect the performance of these sensors and requires that specific considerations be made in their application.

At the interface between an electrode and an ionic solution redox (oxidation–reduction), reactions need to occur for a charge to be transferred between the electrode and the solution. These reactions can be represented in general by the following equations:

$$C1C^{n+} + ne^{-} \tag{47.1}$$

$$A^{m-}1\,A + me^{-} \tag{47.2}$$

where n is the valence of cation material C, and m is the valence of anion material, A. For most electrode systems, the cations in solution and the metal of the electrodes are the same, so the atoms C are oxidized when they give up electrons and go into solution as positively charged ions. These ions are reduced when the process occurs in the reverse direction. In the case of the anion reaction, Equation 47.2, the directions for oxidation and reduction are reversed. For best operation of the electrodes, these two reactions should be reversible, that is, it should be just as easy for them to occur in one direction as the other.

The interaction between a metal in contact with a solution of its ions produces a local change in the concentration of the ions in solution near the metal surface. This causes charge neutrality not to be maintained in this region, which can result in causing the electrolyte surrounding the metal to be at a different electrical potential from the rest of the solution. Thus, a potential difference known as the half-cell potential is established between the metal and the bulk of the electrolyte. It is found that different characteristic potentials occur for different materials and different redox reactions of these materials. Some of these potentials are summarized in Table 47.2. These half-cell potentials can be important when using electrodes for low frequency or d.c. measurements.

The relationship between electric potential and ionic concentrations or, more precisely, ionic activities is frequently considered in electrochemistry. Most commonly two ionic solutions of different activity are separated by an ion-selective semipermeable membrane that allows one type of ion to pass freely through the membrane. It can be shown that an electric potential E will exist between the solutions on either side of the membrane, based upon the relative activity of the permeable ions in each of these solutions.

TABLE 47.2 Half-Cell Potentials for Materials and Reactions Encountered in Biopotential Measurement

Metal and reaction	Half-cell potential, V
$Al \rightarrow Al^{3+} + 3e^-$	−1.706
$Ni \rightarrow Ni^{2+} + 2e^-$	−0.230
$H_2 \rightarrow 2H^+ + 2e^-$	0.000 (by definition)
$Ag + Cl^- \rightarrow AgCl + e^-$	+0.223
$Ag \rightarrow Ag^+ + e^-$	+0.799
$Au \rightarrow Au^+ + e^-$	+1.680

This relationship is known as the Nernst equation

$$E = -\frac{RT}{nF} \ln\left(\frac{a_1}{a_2}\right) \tag{47.3}$$

where a_1 and a_2 are the activities of the ions on either side of the membrane, R is the universal gas constant, T is the absolute temperature, n is the valence of the ions, and F is the Faraday constant. More detail on this relationship can be found in Chapter 48.

When no electric current flows between an electrode and the solution of its ions or across an ion-permeable membrane, the potential observed should be the half-cell potential or the Nernst potential, respectively. If, however, there is a current, these potentials can be altered. The difference between the potential at zero current and the measured potentials while current is passing is known as the over voltage and is the result of an alteration in the charge distribution in the solution in contact with the electrodes or the ion-selective membrane. This effect is known as polarization and can result in diminished electrode performance, especially under conditions of motion. There are three basic components to the polarization over potential: the ohmic, the concentration, and the activation over potentials. More details on these over potentials can be found in electrochemistry or biomedical instrumentation texts [3].

Perfectly polarizable electrodes pass a current between the electrode and the electrolytic solution by changing the charge distribution within the solution near the electrode. Thus, no actual current crosses the electrode–electrolyte interface. Nonpolarized electrodes, however, allow the current to pass freely across the electrode–electrolyte interface without changing the charge distribution in the electrolytic solution adjacent to the electrode. Although these types of electrodes can be described theoretically, neither can be fabricated in practice. It is possible, however, to come up with electrode structures that closely approximate their characteristics.

Electrodes made from noble metals such as platinum are often highly polarizable. A charge distribution different from that of the bulk electrolytic solution is found in the solution close to the electrode surface. Such a distribution can create serious limitations when movement is present and the measurement involves low frequency or even d.c. signals. If the electrode moves with respect to the electrolytic solution, the charge distribution in the solution adjacent to the electrode surface will change, and this will induce a voltage change in the electrode that will appear as motion artifact in the measurement. Thus, for most biomedical measurements, nonpolarizable electrodes are preferred to those that are polarizable.

The silver–silver chloride electrode is one that has characteristics similar to a perfectly nonpolarizable electrode and is practical for use in many biomedical applications. The electrode (Figure 47.1a) consists of a silver base structure that is coated with a layer of the ionic compound silver chloride. Some of the silver chloride when exposed to light is reduced to metallic silver, so a typical silver–silver chloride electrode has finely divided metallic silver within a matrix of silver chloride on its surface. Since the silver chloride is relatively insoluble in aqueous solutions, this surface remains stable. Because there is minimal polarization associated with this electrode, motion artifact is reduced compared to polarizable electrodes such as the

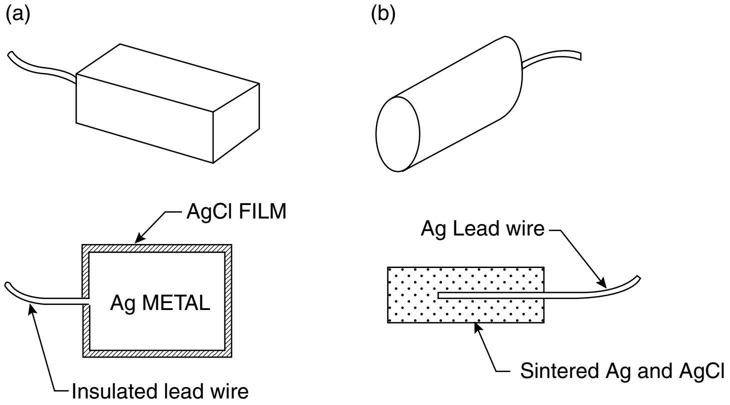


FIGURE 47.1 Silver–silver electrodes for biopotential measurements: (a) metallic silver with a silver chloride surface layer and (b) sintered electrode structure. The lower views show the electrodes in cross-section.

platinum electrode. Furthermore, due to the reduction in polarization, there is also a smaller effect of frequency on electrode impedance, especially at low frequencies.

Silver–silver chloride electrodes of this type can be fabricated by starting with a silver base and electrolytically growing the silver chloride layer on its surface [3]. Although an electrode produced in this way can be used for most biomedical measurements, it is not a robust structure, and pieces of the silver chloride film can be chipped away after repeated use of the electrode. A structure with greater mechanical stability is the sintered silver–silver chloride electrode in Figure 47.1b. This electrode consists of a silver lead wire surrounded by a sintered cylinder made up of finely divided silver and silver-chloride powder pressed together.

In addition to its nonpolarizable behavior, the silver–silver chloride electrode exhibits less electrical noise than the equivalent polarizable electrodes. This is especially true at low frequencies, and so silver–silver chloride electrodes are recommended for measurements involving very low voltages for signals that are made up primarily of low frequencies. A more detailed description of silver–silver chloride electrodes and methods to fabricate these devices can be found in Janz and Ives [5] and biomedical instrumentation textbooks [4].

47.2 Electric Characteristics

The electric characteristics of biopotential electrodes are generally nonlinear and a function of the current density at their surface. Thus, having the devices represented by linear models requires that they be operated at low potentials and currents.[1] Under these idealized conditions, electrodes can be represented by an equivalent circuit of the form shown in Figure 47.2. In this circuit R_d and C_d are components that represent the impedance associated with the electrode–electrolyte interface and polarization at this interface. R_s is the series resistance associated with interfacial effects and the resistance of the electrode materials themselves. The battery E_{hc} represents the half-cell potential described above. It is seen that the impedance of this electrode will be frequency dependent, as illustrated in Figure 47.3. At low frequencies the impedance is dominated by the series combination of R_s and R_d, whereas at higher frequencies C_d bypasses the effect of R_d so that the impedance is now close to R_s. Thus, by measuring the impedance of an electrode at high and low frequencies, it is possible to determine the component values for the equivalent circuit for that electrode.

[1] Or at least at an operating point where the voltage and current is relatively fixed.

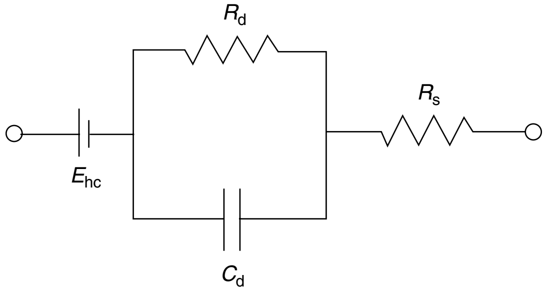


FIGURE 47.2 The equivalent circuit for a biopotential electrode.

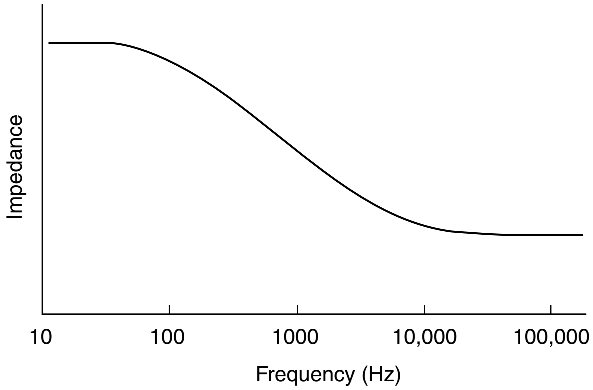


FIGURE 47.3 An example of biopotential electrode impedance as a function of frequency. Characteristic frequencies will be somewhat different for electrode different geometries and materials.

TABLE 47.3 The Effect of Electrode Properties on Electrode Impedance

Property	Change in property	Changes in electrode impedance
Surface area	↑	↓
Polarization	↑	↑ At low frequencies
Surface roughness	↑	↓
Radius of curvature	↑	↓
Surface contamination	↑	↑

↑ — increase in quantity; ↓ — decrease in property.

The electrical characteristics of electrodes are affected by many physical properties of these electrodes. Table 47.3 lists some of the more common physical properties of electrodes and qualitatively indicates how these can affect electrode impedance.

47.3 Practical Electrodes for Biomedical Measurements

Many different forms of electrodes have been developed for different types of biomedical measurements. To describe each of these would go beyond the constraints of this article, but some of the more commonly used electrodes are presented in this section. The reader is referred to the monograph by Geddes for more details and a wider selection of practical electrodes [6].

47.3.1 Body-Surface Biopotential Electrodes

This category includes electrodes that can be placed on the body surface for recording bioelectric signals. The integrity of the skin is not compromised when these electrodes are applied, and they can be used for short-term diagnostic recording such as taking a clinical electrocardiogram or long-term chronic recording such as occurs in cardiac monitoring.

47.3.1.1 Metal Plate Electrodes

The basic metal plate electrode consists of a metallic conductor in contact with the skin with a thin layer of an electrolyte gel between the metal and the skin to establish this contact. Examples of metal plate electrodes are seen in Figure 47.4a. Metals commonly used for this type of electrode include German silver (a nickel–silver alloy), silver, gold, and platinum. Sometimes these electrodes are made of a foil of the metal so as to be flexible, and sometimes they are produced in the form of a suction electrode (Figure 47.4b) to make it easier to attach the electrode to the skin to make a measurement and then move it to another point to repeat the measurement. These types of electrodes are used primarily for diagnostic recordings of biopotentials such as the electrocardiogram or the electroencephalogram. Metal disk electrodes with a gold surface in a conical shape such as shown in Figure 47.4c are frequently used for EEG recordings. The apex of the cone is open so that electrolyte gel or paste can be introduced to both make good contact between the electrode and the head and to allow this contact medium to be replaced should it dry out during its use. These types of electrodes were the primary types used for obtaining diagnostic electrocardiograms for many years. Today, disposable electrodes such as described in the next section are frequently used. These do not require as much preparation or strapping to the limbs as the older electrodes did, and since they are disposable, they do not need to be cleaned between applications to patients. Because they are usually silver–silver chloride electrodes, they have less noise and motion artifact than the metal electrodes.

47.3.1.2 Electrodes for Chronic Patient Monitoring

Long-term monitoring of biopotentials such as the electrocardiogram as performed by cardiac monitors places special constraints on the electrodes used to pick up the signals. These electrodes must have a stable interface between them and the body, and frequently nonpolarizable electrodes are, therefore, the best for this application. Mechanical stability of the interface between the electrode and the skin can help to reduce motion artifact, and so there are various approaches to reduce interfacial motion between the electrode and the coupling electrolyte or the skin. Figure 47.4d is an example of one approach to reduce motion artifact by recessing the electrode in a cup of electrolytic fluid or gel. The cup is then securely fastened to the skin surface using a double-sided adhesive ring. Movement of the skin with respect to the electrode may affect the electrolyte near the skin–electrolyte interface, but the electrode–electrolyte interface can be several millimeters away from this location, since it is recessed in the cup. The fluid movement is unlikely to affect the recessed electrode–electrolyte interface as compared to what would happen if the electrode was separated from the skin by just a thin layer of electrolyte.

The advantages of the recessed electrode can be realized in a simpler design that lends itself to mass production through automation. This results in low per-unit cost so that these electrodes can be considered disposable. Figure 47.4e illustrates such an electrode in cross section. The electrolyte layer now consists of an open-celled sponge saturated with a thickened (high-viscosity) electrolytic solution. The sponge serves the same function as the recess in the cup electrodes and is coupled directly to a silver–silver chloride electrode. Frequently, the electrode itself is attached to a clothing snap through an insulating-adhesive disk that holds the structure against the skin. This snap serves as the point of connection to a lead wire. Many commercial versions of these electrodes in various sizes are available, including electrodes with a silver–silver chloride interface or ones that use metallic silver as the electrode material.

A modification of this basic monitoring electrode structure is shown in Figure 47.4f. In this case the metal electrode is a silver foil with a surface coating of silver chloride. The foil gives the electrode increased flexibility to fit more closely over body contours. Instead of using the sponge, a hydrogel film (really a sponge on a microscopic level) saturated with an electrolytic solution and formed from materials that

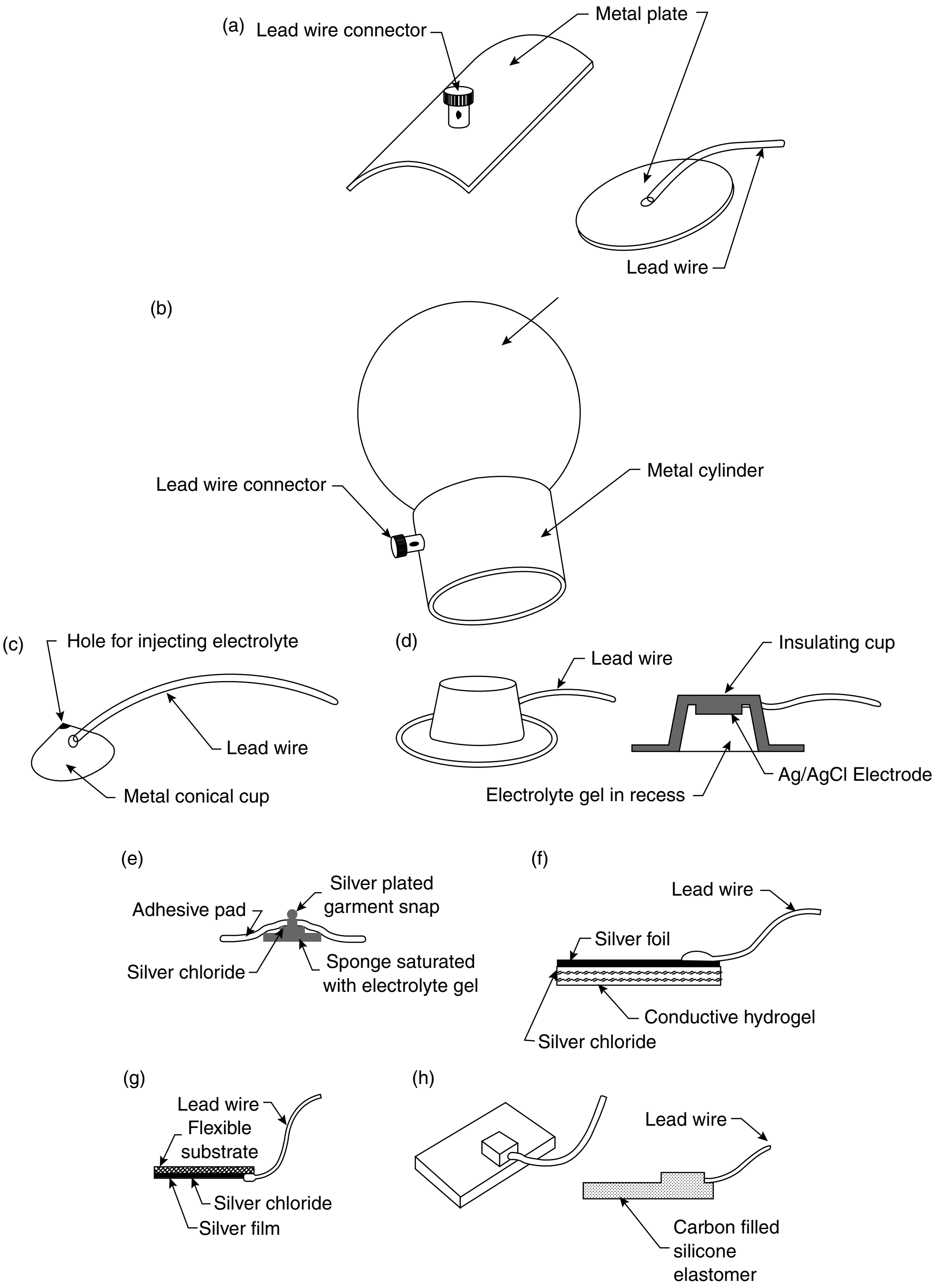


FIGURE 47.4 Examples of different skin electrodes: (a) metal plate electrodes, (b) suction electrode for ECG, (c) metal cup EEG electrode, (d) recessed electrode, (e) disposable electrode with electrolyte-impregnated sponge (shown in cross-section), (f) disposable hydrogel electrode (shown in cross-section), (g) thin-film electrode for use with neonates (shown in cross-section), (h) carbon-filled elastomer dry electrode.

are very sticky is placed over the electrode surface. The opposite surface of the hydrogel layer can be attached directly to the skin, and since it is very sticky, no additional adhesive is needed. The mobility and concentration of ions in the hydrogel layer is generally lower than for the electrolytic solution used in the sponge or the cup. This results in an electrode that has a higher source impedance as compared to these other structures. An important advantage of this structure is its ability to have the electrolyte stick directly on the skin. This greatly reduces interfacial motion between the skin surface and the electrolyte, and hence there is a smaller amount of motion artifact in the signal. This type of hydrogel electrode is, therefore, especially valuable in monitoring patients who move a great deal or during exercise.

Thin-film flexible electrodes such as shown in Figure 47.4g have been used for monitoring neonates. They are basically the same as the metal plate electrodes; only the thickness of the metal in this case is less than a micrometer. These metal films need to be supported on a flexible plastic substrate such as polyester or polyimide. The advantage of using only a thin metal layer for the electrode lies in the fact that these electrodes are x-ray transparent. This is especially important in infants where repeated placement and removal of electrodes, so that x-rays may be taken, can cause substantial skin irritation.

Electrodes that do not use artificially applied electrolyte solutions or gels and, therefore, are often referred to as dry electrodes have been used in some monitoring applications. These sensors as illustrated in Figure 47.4h can be placed on the skin and held in position by an elastic band or tape. They are made up of a graphite or metal-filled polymer such as silicone. The conducting particles are ground into a fine powder, and this is added to the silicone elastomer before it cures so to produce a conductive material with physical properties similar to that of the elastomer. When held against the skin surface, these electrodes establish contact with the skin without the need for an electrolytic fluid or gel. In actuality such a layer is formed by sweat under the electrode surface. For this reason these electrodes tend to perform better after they have been left in place for an hour or two so that this layer forms. Some investigators have found that placing a drop of physiologic saline solution on the skin before applying the electrode accelerates this process. This type of electrode has found wide application in home infant cardiorespiratory monitoring because of the ease with which it can be applied by untrained caregivers.

Dry electrodes are also used on some consumer products such as stationary exercise bicycles and treadmills to pick up an electrocardiographic signal to determine heart rate. When a subject grabs the metal contacts, there is generally enough sweat to establish good electrical contact so that a Lead I electrocardiogram can be obtained and used to determine the heart rate. The signals, however, are much noisier than those obtained from other electrodes described in this section.

47.3.2 Intracavitary and Intratissue Electrodes

Electrodes can be placed within the body for biopotential measurements. These electrodes are generally smaller than skin surface electrodes and do not require special electrolytic coupling fluid, since natural body fluids serve this function. There are many different designs for these internal electrodes, and only a few examples are given in the following paragraphs. Basically these electrodes can be classified as needle electrodes, which can be used to penetrate the skin and tissue to reach the point where the measurement is to be made, or they are electrodes that can be placed in a natural cavity or surgically produced cavity in tissue. Figure 47.5 illustrates some of these internal electrodes.

A catheter tip or probe electrode is placed in a naturally occurring cavity in the body such as in the gastrointestinal system. A metal tip or segment on a catheter makes up the electrode. The catheter or, in the case where there is no hollow lumen, probe, is inserted into the cavity so that the metal electrode makes contact with the tissue. A lead wire down the lumen of the catheter or down the center of the probe connects the electrode to the external circuitry.

The basic needle electrode shown in Figure 47.5b consists of a solid needle, usually made of stainless steel, with a sharp point. An insulating material coats the shank of the needle up to a millimeter or two of the tip so that the very tip of the needle remains exposed. When this structure is placed in tissue such as skeletal muscle, electrical signals can be picked up by the exposed tip. One can also make needle electrodes by running one or more insulated wires down the lumen of a standard hypodermic needle. The electrode

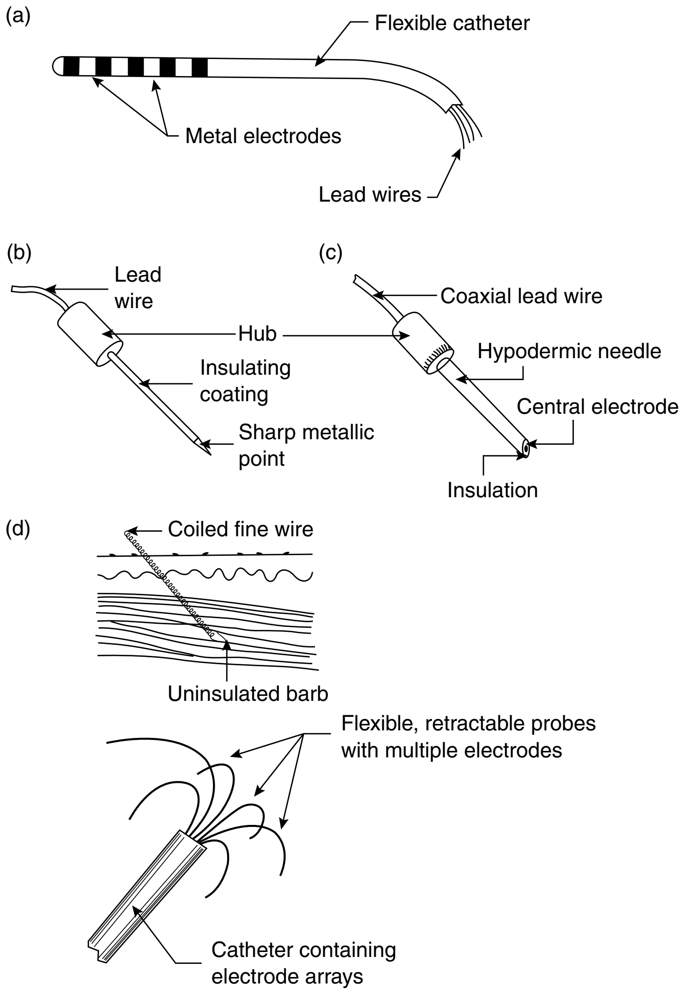


FIGURE 47.5 Examples of different internal electrodes: (a) catheter or probe electrode, (b) needle electrode, (c) coaxial needle electrode, (d) coiled wire electrode. (Reprinted with permission from Webster J.G. (Ed.). 1992. *Medical Instrumentation: Application and Design*, John Wiley & Sons, New York.)

as shown in Figure 47.5c is shielded by the metal of the needle and can be used to pick up very localized signals in tissue.

Fine wires can also be introduced into tissue using a hypodermic needle, which is then withdrawn. This wire can remain in tissue for acute or chronic measurements. Caldwell and Reswick [7] and Knutson et al. [12] have used fine coiled wire electrodes in skeletal muscle for several years without adverse effects.

The advantage of the coil is that it makes the electrode very flexible and compliant. This helps it and the lead wire to endure the frequent flexing and stretching that occurs in the body without breaking.

The relatively new clinical field of cardiac electrophysiology makes use of electrodes that can be advanced into the heart to identify aberrant regions of myocardium that cause life-threatening arrhythmias. These electrodes may be similar to the multiple electrode probe or catheter shown in Figure 47.5a or they might be much more elaborate such as the "umbrella" electrode array in Figure 47.5e. In this case the electrode array with multiple electrodes on each umbrella rib is advanced into the heart in collapsed form through a blood vessel in the same way as a catheter is passed into the heart. The umbrella is then opened in the heart such that the electrodes on the ribs contact the endocardium and are used to record and map intracardiac electrograms. Once the procedure is finished, the umbrella is collapsed and withdrawn through the blood vessel. A similar approach can be taken with an electrode array on the surface of a balloon. The collapsed balloon is advanced into one of the chambers of the heart and then distended. Simultaneous recordings are made from each electrode of the array, and then the balloon is collapsed and withdrawn [16,17].

47.3.3 Microelectrodes

The electrodes described in the previous paragraphs have been applied to studying bioelectric signals at the organism, organ, or tissue level but not at the cellular level. To study the electric behavior of cells, electrodes that are themselves smaller than the cells being studied need to be used. Three types of electrodes have been described for this purpose: etched metal electrodes, micropipette electrodes, and metal-film-coated micropipette electrodes. The metal microelectrode is essentially a subminiature version of the needle electrode described in the previous section (Figure 47.6a). In this case, a strong metal wire such as tungsten is used. One end of this wire is etched electrolytically to give tip diameters on the order of a few micrometers. The structure is insulated up to its tip, and it can be passed through the membrane of a cell to contact the cytosol. The advantage of this type of electrode is that it is both small and robust and can be used for neurophysiologic studies. Its principal disadvantage is the difficulty encountered in its fabrication and high source impedance.

The second and most frequently used type of microelectrode is the glass micropipette. This structure, as illustrated in Figure 47.6b consists of a fine glass capillary drawn to a very narrow point and filled with an electrolytic solution. The point can be as narrow as a fraction of a micrometer, and the dimensions of this electrode are strongly dependent on the skill of the individual drawing the tip. The electrolytic solution in the lumen serves as the contact between the interior of the cell through which the tip has been impaled and a larger conventional electrode located in the shank of the pipette. These electrodes also suffer from high source impedances and fabrication difficulty.

A combined form of these two types of electrodes can be achieved by depositing a metal film over the outside surface of a glass micropipette as shown in Figure 47.6c. In this case, the strength and smaller dimensions of the micropipette can be used to support films of various metals that are insulated by an additional film up to a point very close to the actual tip of the electrode structure. These electrodes have been manufactured in quantity and made available as commercial products. Since they combine

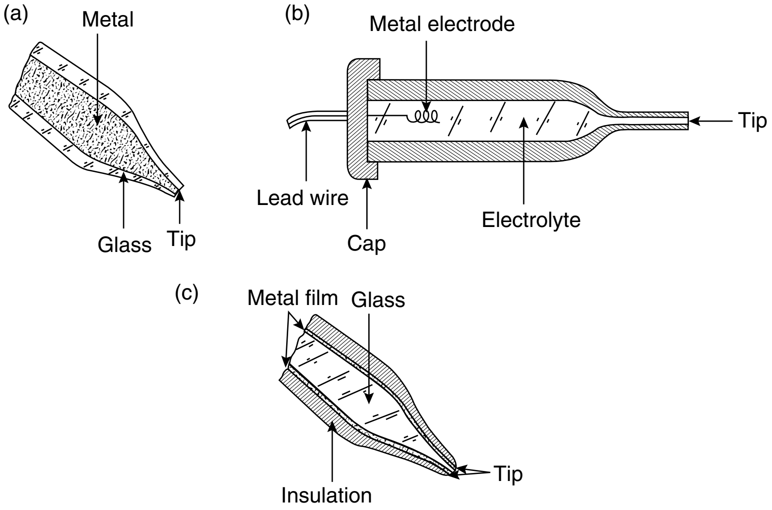


FIGURE 47.6 Microelectrodes: (a) metal, (b) micropipette, (c) thin metal film on micropipette. (Reprinted with permission from Webster J.C. (Ed.). 1992. *Medical Instrumentation: Application and Design*, Houghton Mifflin, Boston.)

the features of both the metal and the micropipette electrodes, they also suffer from many of the same limitations. They do, however, have the advantage of flexibility due to the capability of being able to make films of different metals on the micropipette surface without having to worry about the strength of the metal, as would be the case if the metal were used alone.

47.3.4 Electrodes Fabricated Using Microelectronic Technology

Modern microelectronic technology can be used to fabricate many different types of electrodes for specific biomedical applications. For example, dry electrodes with high source resistances or microelectrodes with similar characteristics can be improved by incorporating a microelectronic amplifier for impedance conversion right on the electrode itself. In the case of the conventional-sized electrodes, a metal disk 5 to 10 mm in diameter can have a high input impedance microelectronic amplifier configured as a follower integrated into the back of the electrode so that localized processing of the high source impedance signal can produce one of lower, more practical impedance for signal transmission [8]. Single- and multiple-element electrodes can be made from thin-film or silicon technology. Mastrototaro and colleagues have demonstrated probes for measuring intramyocardial potentials using thin, patterned gold films on polyimide or oxidised molybdenum substrates [9]. When electrodes are made from pieces of micromachined silicon, it is possible to integrate an amplifier directly into the electrode [10]. Multichannel amplifiers or multiplexers can be used with multiple electrodes on the same probe. Electrodes for contact with individual nerve fibers can be fabricated using micromachined holes in a silicon chip that are just big enough to pass a single growing axon. Electrical contacts on the sides of these holes can then be used to pick up electrical activity from these nerves [11]. These examples are just a few of the many possibilities that can be realized using microelectronics and three-dimensional micromachining technology to fabricate specialized electrodes.

47.4 Biomedical Applications

Electrodes can be used to perform a wide variety of measurements of bioelectric signals. An extensive review of this would be beyond the scope of this chapter, but some typical examples of applications are highlighted in Table 47.4. The most popular application for biopotential electrodes is in obtaining the electrocardiogram for diagnostic and patient-monitoring applications. A substantial commercial market exists for various types of electrocardiographic electrodes, and many of the forms described in the previous section are available commercially. Other electrodes for measuring bioelectric potentials for application in diagnostic medicine are indicated in Table 47.4. Research applications of biopotential electrodes are

TABLE 47.4 Examples of Applications of Biopotential Electrodes

Application	Biopotential	Type of electrode
Cardiac monitoring	ECG	Ag/AgCl with sponge Ag/AgCl with hydrogel
Infant cardiopulmonary monitoring	ECG impedance	Ag/AgCl with sponge Ag/AgCl with hydrogel Thin-film Filled elastomer dry
Sleep encephalography	EEG	Gold cups Ag/AgCl cups Active electrodes
Diagnostic muscle activity	EMG	Needle
Cardiac electrograms	Electrogram	Intracardiac probe
Implanted telemetry of biopotentials	ECG EMG	Stainless steel wire loops Platinum disks
Eye movement	EOG	Ag/AgCl with hydrogel

highly varied and specific for individual studies. Although a few examples are given in Table 47.4, the field is far too broad to be completely covered here.

Biopotential electrodes are one of the most common biomedical sensors used in clinical medicine. Although their basic principle of operation is the same for most applications, they take on many forms and are used in the measurement of many types of bioelectric phenomena. They will continue to play an important role in biomedical instrumentation systems.

References

[1] Yoshida T., Hayashi K., and Toko K. (1988). The effect of anoxia on the spatial pattern of electric potential formed along the root. *Ann. Bot.* 62: 497.

[2] Plonsey R. and Barr R.C. (1988). *Bioelectricity*, New York, Plenum Press.

[3] Weast R.C. (Ed.) (1974). *Handbook of Chemistry and Physics*, 55th ed., Boca Raton, FL, CRC Press.

[4] Webster J.G. (Ed.) (1992). *Medical Instrumentation: Application and Design*, Boston, Houghton Mifflin.

[5] Janz G.I. and Ives D.J.G. (1968). Silver–silver chloride electrodes. *Ann. NY Acad. Sci.* 148: 210.

[6] Geddes L.A. (1972). *Electrodes and the Measurement of Bioelectric Events*, New York, John Wiley & Sons.

[7] Caldwell C.W. and Reswick J.B. (1975). A percutaneous wire electrode for chronic research use. *IEEE Trans. Biomed. Eng.* 22: 429.

[8] Ko W.H. and Hynecek J. (1974). Dry electrodes and electrode amplifiers. In H.A. Miller and D.C. Harrison (Eds.), *Biomedical Electrode Technology*, pp. 169–181, New York, Academic Press.

[9] Mastrototaro J.J., Massoud H.Z., Pilkington T.C. et al. (1992). Rigid and flexible thin-film microelectrode arrays for transmural cardiac recording. *IEEE Trans. Biomed. Eng.* 39: 271.

[10] Wise K.D., Najafi K., Ji J. et al. (1990). Micromachined silicon microprobes for CNS recording and stimulation. *Proc. Ann. Conf. IEEE Eng. Med. Biol. Soc.* 12: 2334.

[11] Edell D.J. (1986). A peripheral nerve information transducer for amputees: long-term multichannel recordings from rabbit peripheral nerves. *IEEE Trans. Biomed. Eng.* 33: 203.

[12] Knutson J.S., Naples G.G., Peckham P.H., and Keith M.W. (2002). Fracture rates and occurrences of infection and granuloma associated with percutaneous intramuscular electrodes in upper extremity functional electrical simulation applications. *J. Rehab. Res. Dev.* 39: 671–684.

[13] Ives J.R. (2005). New chronic EEG electrode for critical/intensive care unit monitoring. *J. Clin. Neurophysiol.* 22: 119–123

[14] Griss P., Tolvanen-Laakso H.K., Meriläinen P. et al. (2002). Characterization of micromachined spiked biopotential electrodes. *IEEE Trans. Biomed. Eng.* 49: 597–604.

[15] Konings K.T., Kirchhof C.I., Smeets J.R. et al. (1994). High-density mapping of electrically induced atrial fibrillation in humans. *Circulation* 89: 1665–1680.

[16] Rao L., He R., Ding C. et al. (2004). Novel noncontact catheter system for endocardial electrical and anatomical imaging. *Ann. Biomed. Eng.* 32: 573–584.

[17] Chen T.C., Parson I.D., and Downar E. (1991). The construction of endocardial balloon arrays for cardiac mapping. *Pacing. Clin. Electrophysiol.* 14: 470–479.

Further Information

Good overviews of biopotential electrodes are found in Geddes L.A. 1972. *Electrodes and the Measurement of Bioelectric Events*, New York, John Wiley & Sons; and Ferris C.D. 1974. *Introduction to Bioelectrodes*, New York, Plenum. Even though these references are more than 20 years old, they clearly cover the field, and little has changed since these books were written.

Overviews of biopotential electrodes are found in chapters of two works edited by John Webster. Chapter 5 of his textbook, Medical Instrumentation: Application and Design, covers the material of this chapter in more detail, and there is a section on “Bioelectrodes” in his *Encyclopedia on Medical Devices and Instrumentation*, published by Wiley in 1988.

The journals *IEEE Transactions on Biomedical Engineering and Medical and Biological Engineering and Computing* are good sources of recent research on biopotential electrodes.

48

Electrochemical Sensors

Chung-Chiun Liu
Case Western Reserve University

Electrochemical sensors have been used extensively either as a whole or an integral part of a chemical and biomedical sensing element. For instance, blood gas (PO_2, PCO_2, and pH) sensing can be accomplished entirely by electrochemical means. Many important biomedical enzymatic sensors, including glucose sensors, incorporate an enzymatic catalyst and an electrochemical sensing element. The Clark type of oxygen sensor [Clark, 1956] is a well-known practical biomedical sensor based on electrochemical principles, an amperometric device. Electrochemical sensors generally can be categorized as conductivity/capacitance, potentiometric, amperometric, and voltammetric sensors. The amperometric and voltammetric sensors are characterized by their current–potential relationship with the electrochemical system and are less well-defined. Amperometric sensors can also be viewed as a subclass of voltammetric sensors.

Electrochemical sensors are essentially an electrochemical cell which employs a two- or three-electrode arrangement. Electrochemical sensor measurement can be made at steady-state or transient. The applied current or potential for electrochemical sensors may vary according to the mode of operation, and the selection of the mode is often intended to enhance the sensitivity and selectivity of a particular sensor. The general principles of electrochemical sensors have been extensively discussed in many electroanalytic references. However, many electroanalytic methods are not practical in biomedical sensing applications. For instance, dropping mercury electrode polarography is a well-established electroanalytic method, yet its usefulness in biomedical sensor development, particularly for potential *in vivo* sensing, is rather limited. In this chapter, we shall focus on the electrochemical methodologies which are useful in biomedical sensor development.

48.1 Conductivity/Capacitance Electrochemical Sensors

Measurement of the electric conductivity of an electrochemical cell can be the basis for an electrochemical sensor. This differs from an electrical (physical) measurement, for the electrochemical sensor measures the conductivity change of the system in the presence of a given solute concentration. This solute is often the sensing species of interest. Electrochemical sensors may also involve a measuring capacitative impedance resulting from the polarization of the electrodes and the faradaic or charge transfer processes.

It has been established that the conductance of a homogeneous solution is directly proportional to the cross-sectional area perpendicular to the electrical field and inversely proportional to the segment of solution along the electrical field. Thus, the conductance of this solution (electrolyte), G (Ω^{-1}), can be expressed as

$$G = \sigma A/L \tag{48.1}$$

where A is the cross-sectional area (in cm^2), L is the segment of the solution along the electrical field (in cm), and σ (in $\Omega\,\text{cm}^{-1}$) is the specific conductivity of the electrolyte and is related quantitatively to the concentration and the magnitude of the charges of the ionic species. For a practical conductivity sensor, A is the surface of the electrode, and L is the distance between the two electrodes.

Equivalent and molar conductivities are commonly used to express the conductivity of the electrolyte. Equivalent conductance depends on the concentration of the solution. If the solution is a strong electrolyte, it will completely dissociate the components in the solution to ionic forms. Kohlrauch [MacInnes, 1939] found that the equivalent conductance of a strong electrolyte was proportional to the square root of its concentration. However, if the solution is a weak electrolyte which does not completely dissociate the components in the solution to respective ions, the above observation by Kohlrauch is not applicable.

The formation of ions leads to consideration of their contribution to the overall conductance of the electrolyte. The equivalent conductance of a strong electrolyte approaches a constant limiting value at infinite dilution, namely,

$$\Lambda_\text{o} = \Lambda_{\text{lim}\to 0} = \lambda_0^+ + \lambda_0^- \tag{48.2}$$

where Λ_0 is the equivalent conductance of the electrolyte at infinite dilution and λ_0^+ and λ_0^- are the ionic equivalent conductance of cations and anions at infinite dilution, respectively.

Kohlrauch also established the law of independent mobilities of ions at infinite dilution. This implies that Lo at infinite dilution is a constant at a given temperature and will not be affected by the presence of other ions in the electrolytes. This provides a practical estimation of the value of Λ_0 from the values of λ_0^+ and λ_0^-. As mentioned, the conductance of an electrolyte is influenced by its concentration. Kohlrausch stated that the equivalent conductance of the electrolyte at any concentration C in mol/l or any other convenient units can be expressed as

$$\Lambda = \Lambda_0 - \beta \text{C}^{0.5} \tag{48.3}$$

where β is a constant depending on the electrolyte.

In general, electrolytes can be classified as weak electrolytes, strong electrolytes, and ion-pair electrolytes. Weak electrolytes only dissociate to their component ions to a limited extent, and the degree of the dissociation is temperature dependent. However, strong electrolytes dissociate completely, and Equation 48.3 is applicable to evaluate its equivalent conductance. Ion-pair electrolytes can by characterized by their tendency to form ion pairs. The dissociation of ion pairs is similar to that of a weak electrolyte and is affected by ionic activities. The conductivity of ion-pair electrolytes is often nonlinear related to its concentration.

The electrolyte conductance measurement technique, in principle, is relatively straightforward. However, the conductivity measurement of an electrolyte is often complicated by the polarization of the electrodes at the operating potential. Faradaic or charge transfer processes occur at the electrode surface, complicating the conductance measurement of the system. Thus, if possible, the conductivity electrochemical sensor should operate at a potential where no faradaic processes occur. Also, another important consideration is the formation of the double layer adjacent to each electrode surface when a potential is imposed on the electrochemical sensor. The effect of the double layer complicates the interpretation of the conductivity measurement and is usually described by the Warburg impedance. Thus, even in the absence of faradaic processes, the potential effect of the double layer on the conductance of the electrolyte must be carefully assessed. The influence of a faradaic process can be minimized by maintaining a high center constant, L/A, of the electrochemical conductivity sensor, so that the cell resistance lies in the region of 1 to 50 kΩ. This implies the desirable feature of a small electrode surface area and a relatively large distance between the two electrodes. Yet, a large electrode surface area enhances the accuracy of the measurement, since a large deviation from the null point facilitates the balance of the Wheatstone bridge, resulting in improvement of sensor sensitivity. These opposing features can be resolved by using a multiple-sensing electrode configuration in which the surface area of each electrode element is small compared to the distance between the electrodes. The multiple electrodes are connected in parallel, and the output of the sensor represents the total sum of the current through each pair of electrodes. In this mode of measurement, the effect of the double layer is included in the conductance measurement. The effects of both the double layers and the faradaic processes can be minimized by using a high-frequency, low-amplitude alternating current. The higher the frequency and the lower the amplitude of the imposed alternating current, the closer the measured value is to the true conductance of the electrolyte.

48.2 Potentiometric Sensors

When a redox reaction, $\mathrm{Ox} + \mathrm{Ze} = \mathrm{Red}$, takes place at an electrode surface in an electrochemical cell, a potential may develop at the electrode–electrolyte interface. This potential may then be used to quantify the activity (on concentration) of the species involved in the reaction forming the fundamental of potentiometric sensors.

The above reduction reaction occurs at the surface of the cathode and is defined as a half-cell reaction. At thermodynamic equilibrium, the Nernst equation is applicable and can be expressed as:

$$E = E^{\mathrm{o}} + \frac{RT}{ZF}\ln\left(\frac{a_{\mathrm{ox}}}{a_{\mathrm{red}}}\right), \tag{48.4}$$

where E and E^{o} are the measured electrode potential and the electrode potential at standard state, respectively, a_{ox} and a_{red} are the activities of Ox (reactant in this case) and Red (product in this case), respectively; Z is the number of electrons transferred, F the Faraday constant, R the gas constant, and T the operating temperature in the absolute scale. In the electrochemical cell, two half-cell reactions will take place simultaneously. However, for sensing purposes, only one of the two half-cell reactions should involve the species of interest, and the other half-cell reaction is preferably reversible and noninterfering. As indicated in Equation 48.4, a linear relation exists between the measured potential E and the natural logarithm of the ratio of the activities of the reactant and product. If the number of electrons transferred, Z, is one, at ambient temperature (25°C or 298°K) the slope is approximately 60 mV/decade. This slope value governs the sensitivity of the potentiometric sensor.

Potentiometric sensors can be classified based on whether the electrode is inert or active. An inert electrode does not participate in the half-cell reaction and merely provides the surface for the electron transfer or provides a catalytic surface for the reaction. However, an active electrode is either an ion donor or acceptor in the reaction. In general, there are three types of active electrodes: the metal/metal ion, the metal/insoluble salt or oxide, and metal/metal chelate electrodes.

Noble metals such as platinum and gold, graphite, and glassy carbon are commonly used as inert electrodes on which the half-cell reaction of interest takes place. To complete the circuitry for the potentiometric sensor, the other electrode is usually a reference electrode on which a noninterference half-cell reaction occurs. Silver–silver chloride and calomel electrodes are the most commonly used reference electrodes. Calomel consists of $\mathrm{Hg/HgCl_2}$ and is less desirable for biomedical systems in terms of toxicity.

An active electrode may incorporate chemical or biocatalysts and is involved as either an ion donor or acceptor in the half-cell reaction. The other half-cell reaction takes place on the reference electrode and should also be noninterfering.

If more than a single type of ion contributes to the measured potential in Equation 48.4, the potential can no longer be used to quantify the ions of interest. This is the interference in a potentiometric sensor. Thus, in many cases, the surface of the active electrode often incorporates a specific functional membrane which may be ion-selective, ion-permeable, or have ion-exchange properties. These membranes tend to selectivity permit the ions of interest to diffuse or migrate through. This minimizes the ionic interference.

Potentiometric sensors operate at thermodynamic equilibrium conditions. Thus, in practical potentiometric sensing, the potential measurement needs to be made under zero-current conditions. Consequently, a high-input impedance electrometer is often used for measurements. Also, the response time for a potentiometric sensor to reach equilibrium conditions in order to obtain a meaningful reading can be quite long. These considerations are essential in the design and selection of potentiometric sensors for biomedical applications.

48.3 Voltammetric Sensors

The current-potential relationship of an electrochemical cell provides the basis for voltammetric sensors. Amperometric sensors, that are also based on the current-potential relationship of the electrochemical cell, can be considered a subclass of voltammetric sensors. In amperometric sensors, a fixed potential is applied to the electrochemical cell, and a corresponding current, due to a reduction or oxidation reaction, is then obtained. This current can be used to quantify the species involved in the reaction. The key consideration of an amperometric sensor is that it operates at a fixed potential. However, a voltammetric sensor can operate in other modes such as linear cyclic voltammetric modes. Consequently, the respective current potential response for each mode will be different.

In general, voltammetric sensors examine the concentration effect of the detecting species on the current-potential characteristics of the reduction or oxidation reaction involved.

The mass transfer rate of the detecting species in the reaction onto the electrode surface and the kinetics of the faradaic or charge transfer reaction at the electrode surface directly affect the current-potential characteristics. This mass transfer can be accomplished through (a) an ionic migration as a result of an electric potential gradient, (b) a diffusion under a chemical potential difference or concentration gradient, and (c) a bulk transfer by natural or forced convection. The electrode reaction kinetics and the mass transfer processes contribute to the rate of the faradaic process in an electrochemical cell. This provides the basis for the operation of the voltammetric sensor. However, assessment of the simultaneous mass transfer and kinetic mechanism is rather complicated. Thus, the system is usually operated under definitive hydrodynamic conditions. Various techniques to control either the potential or current are used to simplify the analysis of the voltammetric measurement. A description of these techniques and their corresponding mathematical analyses are well documented in many texts on electrochemistry or electroanalysis [Adams, 1969; Bard and Faulkner, 1980; Lingane, 1958; Macdonald, 1977; Murray and Reilley, 1966].

A preferred mass transfer condition is total diffusion, which can be described by Fick's law of diffusion. Under this condition, the cell current, a measure of the rate of the faradaic process at an electrode, usually increases with increases in the electrode potential. This current approaches a limiting value when the rate of the faradaic process at the electrode surface reaches its maximum mass transfer rate. Under this condition, the concentration of the detecting species at the electrode surface is considered as zero and is diffusional mass transfer. Consequently, the limiting current and the bulk concentration of the detecting species can be related by

$$i = ZFkmC^* \tag{48.5}$$

where *km* is the mass transfer coefficient and C^* is the bulk concentration of the detecting species. At the other extreme, when the electrode kinetics are slow compared with the mass transfer rate, the electrochemical system is operated in the reaction kinetic control regime. This usually corresponds to a small overpotential. The limiting current and the bulk concentration of the detecting species can be related as

$$i = ZFkcC^* \tag{48.6}$$

where *kc* is the kinetic rate constant for the electrode process. Both Equation 48.5 and Equation 48.6 show the linear relationship between the limiting current and the bulk concentration of the detecting species. In many cases, the current does not tend to a limiting value with an increase in the electrode potential. This is because other faradaic or nonfaradaic processes become active, and the cell current represents the cumulative rates of all active electrode processes. The relative rates of these processes, expressing current efficiency, depend on the current density of the electrode. Assessment of such a system is rather complicated, and the limiting current technique may become ineffective.

When a voltammetric sensor operates with a small overpotential, the rate of faradaic reaction is also small; consequently, a high-precision instrument for the measurement is needed. An amperometric sensor is usually operated under limiting current or relatively small overpotential conditions. Amperometric sensors operate under an imposed fixed electrode potential. Under this condition, the cell current can be correlated with the bulk concentration of the detecting species (the solute). This operating mode is commonly classified as amperometric in most sensor work, but it is also referred to as the chronosuperometric method, since time is involved.

Voltammetric sensors can be operated in a linear or cyclic sweep mode. Linear sweep voltammetry involves an increase in the imposed potential linearly at a constant scanning rate from an initial potential to a defined upper potential limit. This is the so-called potential window. The current-potential curve usually shows a peak at a potential where the oxidation or reduction reaction occurs. The height of the peak current can be used for the quantification of the concentration of the oxidation or reduction species. Cyclic voltammetry is similar to the linear sweep voltammetry except that the electrode potential returns to its initial value at a fixed scanning rate. The cyclic sweep normally generates the current peaks corresponding to the oxidation and reduction reactions. Under these circumstances, the peak current value can relate to the corresponding oxidation or reduction reaction. However, the voltammogram can be very complicated for a system involving adsorption (nonfaradaic processes) and charge processes (faradaic processes). The potential scanning rate, diffusivity of the reactant, and operating temperature are essential parameters for sensor operation, similar to the effects of these parameters for linear sweep voltammograms. The peak current may be used to quantify the concentration of the reactant of interest, provided that the effect of concentration on the diffusivity is negligible. The potential at which the peak current occurs can be used in some cases to identify the reaction, or the reactant. This identification is based on the half-cell potential of the electrochemical reactions, either oxidation or reduction. The values of these half-cell reactions are listed extensively in handbooks and references.

The described voltammetric and amperometric sensors can be used very effectively to carry out qualitative and quantitative analyses of chemical and biochemical species. The fundamentals of this sensing technique are well established, and the critical issue is the applicability of the technique to a complex, practical environment, such as in whole blood or other biologic fluids. This is also the exciting challenge of designing a biosensor using voltammetric and amperometric principles.

48.4 Reference Electrodes

Potentiometric, voltammetric, and amperometric sensors employ a reference electrode. The reference electrode in the case of potentiometric and amperometric sensors serves as a counter electrode to complete the circuitry. In either case, the reaction of interest takes place at the surface of the working electrode, and this reaction is either an oxidation or reduction reaction. Consequently, the reaction at the counter electrode, that is, the reference electrode, is a separate reduction or oxidation reaction, respectively. It is necessary that the reaction occurring at the reference electrode does not interfere with the reaction at the working electrode. For practical applications, the reaction occurring at the reference electrode should be highly reversible and, as stated, does not contribute to the reaction at the working electrode. In electrochemistry, the hydrogen electrode is universally accepted as the primary standard with which other electrodes are compared. Consequently, the hydrogen electrode serves extensively as a standard reference. A hydrogen reference electrode is relatively simple to prepare. However, for practical applications hydrogen reference electrodes are too cumbersome to be useful in practice.

A class of electrode called the electrode of the second kind, which forms from a metal and its sparingly soluble metal salt, finds use as the reference electrode. The most common electrode of this type includes the calomel electrode, $Hg/HgCl_2$ and the silver–silver chloride electrode, Ag/AgCl. In biomedical applications, particularly in *in vivo* applications, Ag/AgCl is more suitable as a reference electrode.

An Ag/AgCl electrode can be small, compact, and relatively simple to fabricate. As a reference electrode, the stability and reproducibility of an Ag/AgCl electrode is very important. Contributing factors to instability and poor reproducibility of Ag/AgCl electrodes include the purity of the materials used, the aging effect of the electrode, the light effect, and so on. When in use, the electrode and the electrolyte interface contribute to the stability of the reference electrode. It is necessary that a sufficient quantity of Cl^- ions exists in the electrolyte when the Ag/AgCl electrode serves as a reference. Therefore, other silver–silver halides such as Ag/AgBr or Ag/AgI electrodes are used in cases where these other halide ions are present in the electrolyte.

In a voltammetric sensor, the reference electrode serves as a true reference for the working electrode, and no current flows between the working and reference electrodes. Nevertheless, the stability of the reference electrode remains essential for a voltammetric sensor.

48.5 Summary

Electrochemical sensors are used extensively in many biomedical applications including blood chemistry sensors, PO_2, PCO_2, and pH electrodes. Many practical enzymatic sensors, including glucose and lactate sensors, also employ electrochemical sensors as sensing elements. Electrochemically based biomedical sensors have found *in vivo* and *in vitro* applications. We believe that electrochemical sensors will continue to be an important aspect of biomedical sensor development.

References

Adams R.N. 1969. *Electrochemistry at Solid Electrodes*, New York, Marcel Dekker.

Bard A. and Faulkner L.R. 1980. *Electrochemical Methods*, New York, John Wiley & Sons.

Clark L.C. Jr. 1956. Monitor and control of blood and tissue oxygen tissues. *Trans. Am. Soc. Artif. Organs* 2: 41.

Lingane J.J. 1958. *Electroanalytical Chemistry*, New York, London, Interscience.

Macdonald D.D. 1977. *Transient Techniques in Electrochemistry*, New York, Plenum.

MacInnes D.A. 1939. *The Principles of Electrochemistry*, New York, Reinhold.

Murray R.W. and Reilley C.N. 1996. *Electroanalytical Principles*, New York-London, Interscience.

49

Optical Sensors

Yitzhak Mendelson
Worcester Polytechnic Institute

Optical methods are among the oldest and best-established techniques for sensing biochemical analytes. Instrumentation for optical measurements generally consists of a light source, a number of optical components to generate a light beam with specific characteristics and to direct this light to some modulating agent, and a photodetector for processing the optical signal. The central part of an optical sensor is the modulating component, and a major part of this chapter will focus on how to exploit the interaction of an analyte with optical radiation in order to obtain essential biochemical information.

The number of publications in the field of optical sensors for biomedical applications has grown significantly during the past two decades. Numerous scientific reviews and historical perspectives have been published, and the reader interested in this rapidly growing field is advised to consult these sources for additional details. This chapter will emphasize the basic concept of typical optical sensors intended for continuous *in vivo* monitoring of biochemical variables, concentrating on those sensors which have generally progressed beyond the initial feasibility stage and reached the promising stage of practical development or commercialization.

Optical sensors are usually based on optical fibers or on planar waveguides. Generally, there are three distinctive methods for quantitative optical sensing at surfaces:

1. The analyte directly affects the optical properties of a waveguide, such as evanescent waves (electromagnetic waves generated in the medium outside the optical waveguide when light is reflected from within) or surface plasmons (resonances induced by an evanescent wave in a thin film deposited on a waveguide surface).
2. An optical fiber is used as a plain transducer to guide light to a remote sample and return light from the sample to the detection system. Changes in the intrinsic optical properties of the medium itself are sensed by an external spectrophotometer.

3. An indicator or chemical reagent placed inside, or on, a polymeric support near the tip of the optical fiber is used as a mediator to produce an observable optical signal. Typically, conventional techniques, such as absorption spectroscopy and fluorimetry, are employed to measure changes in the optical signal.

49.1 Instrumentation

The actual implementation of instrumentation designed to interface with optical sensors will vary greatly depending on the type of optical sensor used and its intended application. A block diagram of a generic instrument is illustrated in Figure 49.1. The basic building blocks of such an instrument are the light source, various optical elements, and photodetectors.

49.1.1 Light Source

A wide selection of light sources are available for optical sensor applications. These include highly coherent gas and semiconductor diode lasers, broad spectral band incandescent lamps, and narrow-band, 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 because they are small and inexpensive, consume lower power, produce selective wavelengths, and are easy to work with. In contrast, tungsten lamps provide a broader range of wavelengths, higher intensity, and better stability but require a sizable power supply and can cause heating problems inside the apparatus.

49.1.2 Optical Elements

Various optical elements are used routinely to manipulate light in optical instrumentation. These include lenses, mirrors, light choppers, beam splitters, and couplers for directing the light from the light source into the small aperture of a fiber optic sensor or a specific area on a waveguide surface and collecting the light from the sensor before it is processed by the photodetector. For wavelength selection, optical filters, prisms, and diffraction gratings are the most common components used to provide a narrow bandwidth of excitation when a broadwidth light source is utilized.

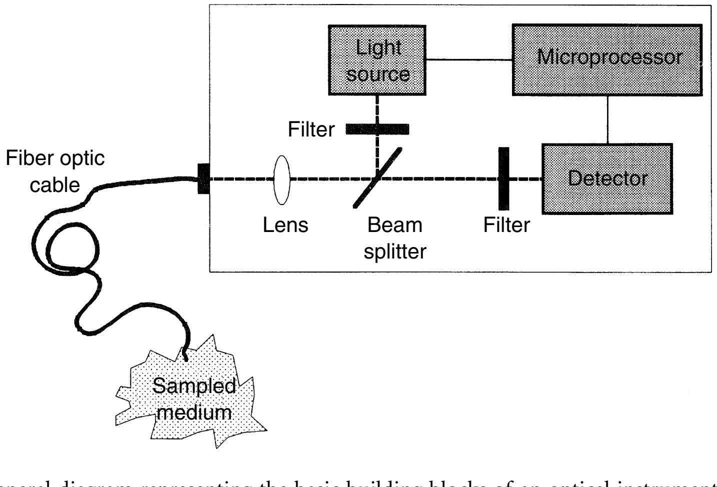


FIGURE 49.1 General diagram representing the basic building blocks of an optical instrument for optical sensor applications.

49.1.3 Photodetectors

In choosing photodetectors for optical sensors, a number of factors must be considered. These include sensitivity, detectivity, noise, spectral response, and response time. Photomultipliers and semiconductor quantum photodetectors, such as photoconductors and photodiodes, are both suitable. The choice, however, is somewhat dependent on the wavelength region of interest. Generally, both types give adequate performance. Photodiodes are usually more attractive because of the compactness and simplicity of the circuitry involved.

Typically, two photodetectors are used in optical instrumentation because it is often necessary to include a separate reference detector to track fluctuations in source intensity and temperature. By taking a ratio between the two detector readings, whereby a part of the light that is not affected by the measurement variable is used for correcting any optical variations in the measurement system, a more accurate and stable measurement can be obtained.

49.1.4 Signal Processing

Typically, the signal obtained from a photodetector provides a voltage or a current proportional to the measured light intensity. Therefore, either simple analog computing circuitry (e.g., a current-to-voltage converter) or direct connection to a programmable gain voltage stage is appropriate. Usually, the output from a photodetector is connected directly to a preamplifier before it is applied to sampling and analog-to-digital conversion circuitry residing inside a computer.

Quite often two different wavelengths of light are utilized to perform a specific measurement. One wavelength is usually sensitive to changes in the species being measured, and 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 fluctuation in instrumentation over time. In other applications, additional discriminations, such as pulse excitation or electronic background subtraction utilizing synchronized lock-in amplifier detection, are useful, allowing improved selectivity and enhanced signal-to-noise ratio.

49.2 Optical Fibers

Several types of biomedical measurements can be made by using either plain optical fibers as a remote device for detecting changes in the spectral properties of tissue and blood or optical fibers tightly coupled to various indicator-mediated transducers. The measurement relies either on direct illumination of a sample through the endface of the fiber or by excitation of a coating on the side wall 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 optic tip.

Optical fibers are based on the principle of total internal reflection. Incident light is transmitted through the fiber if it 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. A typical instrument for performing fiber optic sensing consists of a light source, an optical coupling arrangement, the fiber optic light guide with or without the necessary sensing medium incorporated at the distal tip, and a light detector.

A variety of high-quality optical fibers are available commercially for biomedical sensor applications, depending on the analytic wavelength desired. These include plastic, glass, and quartz fibers which cover the optical spectrum from the UV through the visible to the near IR region. On one hand, plastic optical fibers have a larger aperture and are strong, inexpensive, flexible, and easy to work with but have poor UV transmission below 400 nm. On the other hand, glass and quartz fibers have low attenuation and better transmission in the UV but have small apertures, are fragile, and present a potential risk in *in vivo* applications.

49.2.1 Probe Configurations

There are many different ways 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 configuration offers the most compact and potentially least expensive implementation. However, additional challenges in instrumentation are involved in separating the illuminating signal from the composite signal returning for processing.

The design of intravascular catheters requires special considerations related to the sterility and biocompatibility of the sensor. For example, intravascular fiberoptic sensors must be sterilizable and their material 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 into the peripheral artery or vein via a cut-down and a slow heparin flush is maintained until the device is removed from the blood.

49.2.2 Optical Fiber Sensors

Advantages cited for fiber sensors include their small size and low cost. In contrast to electrical measurements, where the difference of two absolute potentials must be measured, fiber optics are self-contained and do not require an external reference signal. Because the signal is optical, there is no electrical risk to the patient, and there is no direct interference from surrounding electric or magnetic fields. 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.

Despite these advantages, optical fiber sensors exhibit several shortcomings. Sensors with immobilized dyes and other indicators have limited long-term stability, and their shelf life degrades over time. Moreover, ambient light can interfere with the optical measurement unless optical shielding or special time-synchronous gating is performed. As with other implanted or indrolling sensors, organic materials or cells can deposit on the sensor surface due to the biologic response to the presence of the foreign material. All of these problems can result in measurement errors.

49.2.3 Indicator-Mediated Transducers

Only a limited number of biochemical analytes have an intrinsic optical absorption that can be measured with sufficient selectivity directly by spectroscopic methods. Other species, particularly hydrogen, oxygen, carbon dioxide, and glucose, which are of primary interest in diagnostic applications, are not susceptible to direct photometry. Therefore, indicator-mediated sensors have been developed using specific reagents that are properly immobilized on the surface of an optical sensor.

The most difficult aspect of developing an optical biosensor is the coupling of light to the specific recognition element so that the sensor can respond selectively and reversibly to a change in the concentration of a particular analyte. In fiber-optic-based sensors, light travels efficiently to the end of the fiber where it exists and interacts with a specific chemical or biologic recognition element that is immobilized at the tip of the fiber optic. These transducers may include indicators and ionophores (i.e., ion-binding compounds) as well as a wide variety of selective polymeric materials. After the light interacts with the sample, the light returns through the same or a different optical fiber to a detector which correlates the degree of change with the analyte concentration.

Typical indicator-mediated fiber-optic-sensor configurations are shown schematically in Figure 49.2. In (a) the indicator is immobilized directly on a membrane positioned at the end of a fiber. An indicator in the form of a powder can be either glued directly onto a membrane, as shown in (b), or physically retained in position at the end of the fiber by a special permeable membrane (c), a tubular capillary/membrane (d), or a hollow capillary tube (e).

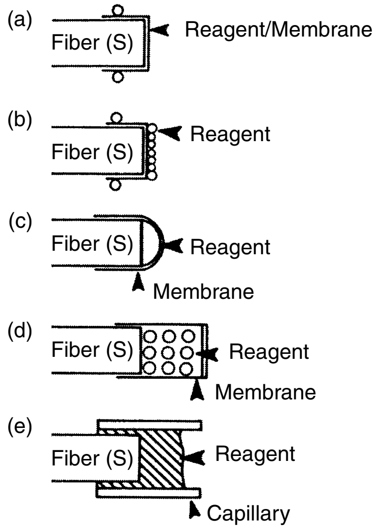


FIGURE 49.2 Typical configuration of different indicator-mediated fiber optic sensor tips. (Taken from Otto S. Wolfbeis, *Fiber Optic Chemical Sensors and Biosensors*, Vol. 1, CRC Press, Boca Raton, 1990.)

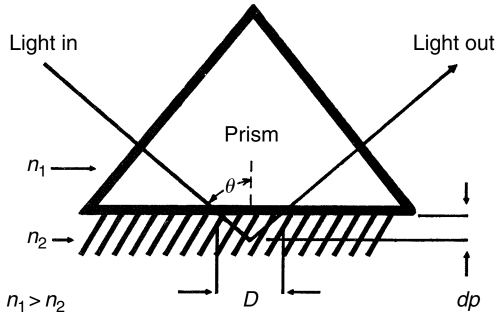


FIGURE 49.3 Schematic diagram of the path of a light ray at the interface of two different optical materials with index of refraction n_1 and n_2. The ray penetrates a fraction of a wave-length (dp) beyond the interface into the medium with the smaller refractive index.

49.3 General Principles of Optical Sensing

Two major optical techniques are commonly available to sense optical changes at sensor interfaces. These are usually based on evanescent wave and surface plasmon resonance principles.

49.3.1 Evanescent Wave Spectroscopy

When light propagates along an optical fiber, it is not confined to the core region but penetrates to some extent into the surrounding cladding region. In this case, an electromagnetic component of the light penetrates a characteristic distance (on the order of one wavelength) beyond the reflecting surface into the less optically dense medium where it is attenuated exponentially according to Beer–Lambert's law (Figure 49.3).

The evanescent wave depends on the angle of incidence and the incident wavelength. This phenomenon has been widely exploited to construct different types of optical sensors for biomedical applications. Because of the short penetration depth and the exponential decay of the intensity, the evanescent wave is absorbed mainly by absorbing compounds very close to the surface. In the case of particularly weak absorbing analytes, 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.

Instead of an absorbing species, a fluorophore can also be used. Light is absorbed by the fluorophore emitting detectable fluorescent light at a higher wavelength, thus providing improved sensitivity. Evanescent wave sensors have been applied successfully to measure the fluorescence of indicators in solution, for pH measurement, and in immunodiagnostics.

49.3.2 Surface Plasmon Resonance

Instead of the dielectric/dielectric interface used in evanescent wave sensors, it is possible to arrange a dielectric/metal/dielectric sandwich layer such that when monochromatic polarized light (e.g., from a laser source) impinges on a transparent medium having a metallized (e.g., Ag or Au) surface, light is absorbed within the plasma formed by the conduction electrons of the metal. This results in a phenomenon known as surface plasmon resonance (SPR). When SPR is induced, the effect is observed as a minimum in the intensity of the light reflected off the metal surface.

As is the case with the evanescent wave, an SPR is exponentially decaying into solution with a penetration depth of about 20 nm. The resonance between the incident light and the plasma wave depends on the angle, wavelength, and polarization state of the incident light and the refractive indices of the metal film and the materials on either side of the metal film. A change in the dielectric constant or the refractive index at the surface causes the resonance angle to shift, thus providing a highly sensitive means of monitoring surface reactions.

The method of 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 absorbed 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 (Figure 49.4). The advantage of this concept is the improved ability to detect the direct interaction between antibody and antigen as an interfacial measurement.

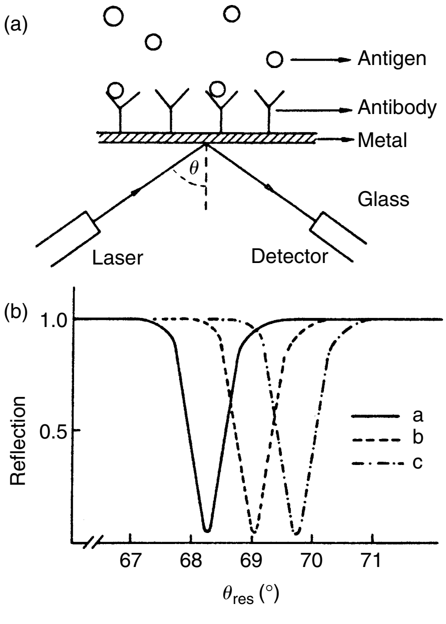


FIGURE 49.4 Surface plasmon resonance at the interface between a thin metallic surface and a liquid (a). A sharp decrease in the reflected light intensity can be observed in (b). The location of the resonance angle is dependent on the refractive index of the material present at the interface.

SPR has been used to analyze immunochemicals and to detect gases. The main limitation of SPR, however, is that the sensitivity depends on the optical thickness of the adsorbed layer, and, therefore, small molecules cannot be measured in very low concentrations.

49.4 Applications

49.4.1 Oximetry

Oximetry refers to the colorimetric measurement of the degree of oxygen saturation of blood, that is, the relative amount of oxygen carried by the hemoglobin in the erythrocytes, by recording 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 of arterial (SaO_2) and mixed venous (SvO_2) blood have been developed, all based on light transmission through, or reflecting from, tissue and blood. The measurement is performed at two specific wavelengths: l1, where there is a large difference in light absorbance between Hb and HbO_2 (e.g., 660 nm red light), and l2, which can be an isobestic wavelength (e.g., 805 nm infrared light), where the absorbance of light is independent of blood oxygenation, or a different wavelength in the infrared region (>805 nm), where the absorbance of Hb is slightly smaller than that of HbO_2.

Assuming for simplicity 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, a simple quantitative relationship can be derived for computing the oxygen saturation of blood:

$$\text{Oxygen saturation} = A - B\left[\frac{\text{OD}(\lambda_1)}{\text{OD}(\lambda_2)}\right] \tag{49.1}$$

where A and B are coefficients which are functions of the specific absorptivities of Hb and HbO_2, and OD is the corresponding absorbance (optical density) of the blood [4].

Since the original discovery of this phenomenon over 50 years ago, there has been progressive development in instrumentation to measure oxygen saturation along three different paths: bench-top oximeters for clinical laboratories, fiber optic catheters for invasive intravascular monitoring, and transcutaneous sensors, which are noninvasive devices placed against the skin.

49.4.1.1 Intravascular Fiber Optic SvO_2 Catheters

In vivo fiberoptic oximeters were first described in the early 1960s by Polanyi and Heir [1]. They demonstrated that in a highly scattering medium such as blood, where a very short path length is required for a transmittance measurement, a reflectance measurement was practical. Accordingly, they showed that a linear relationship exists between oxygen saturation and the ratio of the infrared-to-red (IR/R) light backscattered from the blood

$$\text{oxygen saturation} = a - b(\text{IR/R}) \tag{49.2}$$

where a and b are catheter-specific calibration coefficients.

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. In some commercial instruments (e.g., Oximetrix), automatic compensation for hematocrit is employed utilizing three, rather than two, infrared reference wavelengths. Bornzin et al. [2] and Mendelson et al. [3] described a 5-lumen, 7.5F thermodilution catheter that is comprised of three unequally spaced optical fibers, each fiber 250 mm in diameter, and provides continuous SvO_2 reading with automatic corrections for hematocrit variations (Figure 49.5).

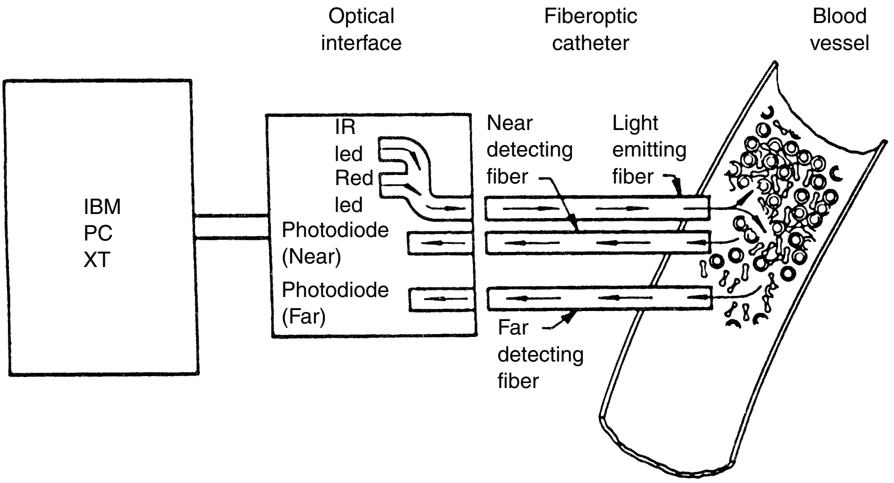


FIGURE 49.5 Principle of a three-fiber optical catheter for SvO_2/HCT measurement. (Taken from Bornzin G.A., Mendelson Y., Moran B.L., et al. 1987. *Proc. 9th Ann. Conf. Eng. Med. Bio. Soc.* pp. 807–809. With permission.)

Intravenous fiberoptic catheters are utilized in monitoring SvO_2 in the pulmonary artery and can be used to indicate the effectiveness of the cardiopulmonary system during cardiac surgery and in the ICU. Several problems limit the wide clinical application of intravascular fiberoptic oximeters. These include the dependence of the individual red and infrared backscattered light intensities and their ratio on hematocrit (especially for SvO_2 below 80%), blood flow, motion artifacts due to catheter tip "whipping" against the blood vessel wall, blood temperature, and pH.

49.4.1.2 Noninvasive Pulse Oximetry

Noninvasive monitoring of SaO_2 by pulse oximetry is a well-established practice in many fields of clinical medicine [4]. The most important advantage of this technique is the capability to provide continuous, safe, and effective monitoring of blood oxygenation at the patient's bedside without the need to calibrate the instrument before each use.

Pulse oximetry, which was first suggested by Aoyagi and colleagues [5] and Yoshiya and colleagues [6], relies on the detection of the time-variant photoplethysmographic signal, caused by changes in arterial blood volume associated with cardiac contraction. SaO_2 is derived by analyzing only the time-variant changes in absorbance caused by the pulsating arterial blood at the same red and infrared wavelengths used in conventional invasive type oximeters. A normalization process is commonly performed by which the pulsatile (a.c.) component at each wavelength, which results from the expansion and relaxation of the arterial bed, is divided by the corresponding nonpulsatile (d.c.) component of the photoplethysmogram, 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 red/infrared ratio which is dependent on SaO_2 but is largely independent of the incident light intensity, skin pigmentation, skin thickness, and tissue vasculature.

Pulse oximeter sensors consist of a pair of small and inexpensive red and infrared LEDs and a single, highly sensitive, silicon photodetector. These components are mounted inside a reusable rigid spring-loaded clip, a flexible probe, or a disposable adhesive wrap (Figure 49.6). The majority of the commercially available sensors are of the transmittance type in which the pulsatile arterial bed, for example, ear lobe, fingertip, or toe, is positioned between the LEDs and the photodetector. Other probes are available for reflectance (backscatter) measurement where both the LEDs and photodetectors are mounted side-by-side facing the skin [7,8].

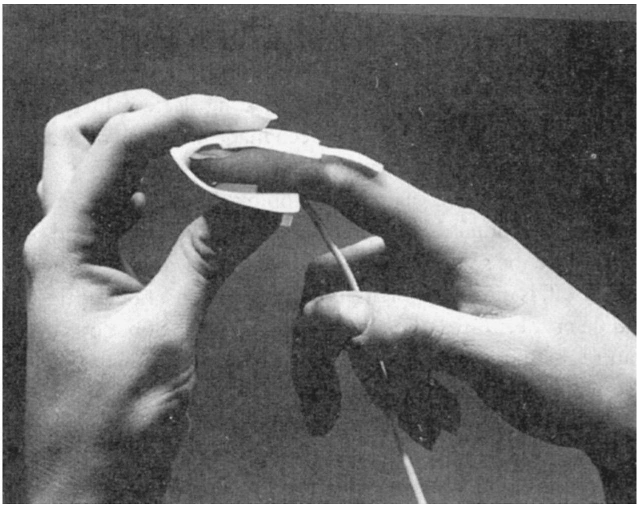


FIGURE 49.6 Disposable finger probe of a noninvasive pulse oximeter.

49.4.1.3 Noninvasive Cerebral Oximetry

Another substance whose optical absorption in the near infrared changes corresponding to its reduced and oxidized state is cytochrome aa3, the terminal member of the respiratory chain. Although the concentration of cytochrome aa3 is considerably lower than that of hemoglobin, advanced instrumentation including time-resolved spectroscopy and differential measurements is being used successfully to obtain noninvasive measurements of hemoglobin saturation and cytochrome aa3 by transilluminating areas of the neonatal brain [9–11].

49.4.2 Blood Gases

Frequent measurement of blood gases, that is, oxygen partial pressure (PO_2), carbon dioxide partial pressure (PCO_2), and pH, is essential to clinical diagnosis and management of respiratory and metabolic problems in the operating room and the ICU. Considerable effort has been devoted over the last two decades to developing disposable extracorporeal and in particular intravascular fiber optic sensors that can be used to provide continuous information on the acid-base status of a patient.

In the early 1970s, Lübbers and Opitz [12] originated what they called optodes (from the Greek, optical path) for measurements of important physiologic gases in fluids and in gases. The principle upon which these sensors was designed was a closed cell containing a fluorescent indicator in solution, with a membrane permeable to the analyte of interest (either ions or gases) constituting one of the cell walls. The cell was coupled by optical fibers to a system that measured the fluorescence in the cell. The cell solution would equilibrate with the PO_2 or PCO_2 of the medium placed against it, and the fluorescence of an indicator reagent in the solution would correspond to the partial pressure of the measured gas.

49.4.2.1 Extracorporeal Measurement

Following the initial feasibility studies of Lübbers and Opitz, Cardiovascular Devices (CDI, USA) developed a GasStat™ extracorporeal system suitable for continuous online monitoring of blood gases *ex vivo* during cardiopulmonary bypass operations. The system consists of a disposable plastic sensor connected inline with a blood loop through a fiber optic cable. Permeable membranes separate the flowing blood from the system chemistry. The CO_2-sensitive indicator consists of a fine emulsion of a bicarbonate buffer in a two-component silicone. The pH-sensitive indicator is a cellulose material to which hydroxypyrene trisulfonate (HPTS) is bonded covalently. The O_2-sensitive chemistry is composed of

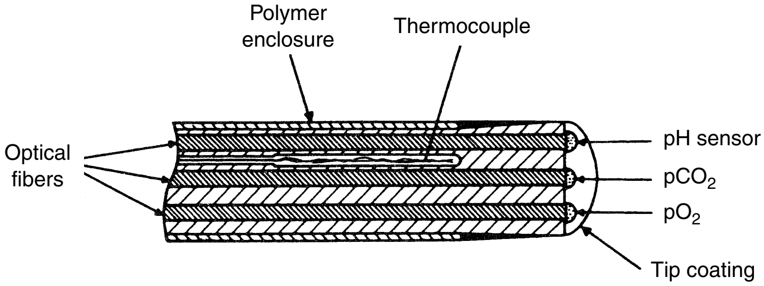


FIGURE 49.7 Structural diagram of an integrated fiber optic blood gas catheter. (Taken from Otto S. Wolfbeis, *Fiber Optic Chemical Sensors and Biosensors*, Vol. 2, CRC Press, Boca Raton, 1990.)

a solution of oxygen-quenching decacyclene in a one-component silicone covered with a thin layer of black PTFE for optical isolation and to render the measurement insensitive to the halothane anesthetic.

The extracorporeal device has two channels, one for arterial blood and the other for venous blood, and is capable of recording the temperature of the blood for correcting the measurements to 37°C. Several studies have been conducted comparing the specifications of the GasStat™ with that of intermittent blood samples analyzed on bench-top blood gas analyzers [13–15].

49.4.2.2 Intravascular Catheters

In recent years, numerous efforts have been made to develop integrated fiber optic sensors for intravascular monitoring of blood gases. Recent literature reports of sensor performance show considerable progress has been made mainly in improving the accuracy and reliability of these intravascular blood gas sensors [16–19], yet their performance has not yet reached a level suitable for widespread clinical application.

Most fiber optic intravascular blood gas sensors employ either a single- or a double-fiber configuration. Typically, the matrix containing the indicator is attached to the end of the optical fiber as illustrated in Figure 49.7. 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 the optode is somewhat slower compared to electrochemical sensors.

Intravascular fiber optic blood gas sensors are normally placed inside a standard 20-gauge catheter, which is sufficiently small to allow 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.

49.4.2.3 pH Sensors

In 1976, Peterson et al. [20] originated the development of the first fiber optic chemical sensor for physiological pH measurement. The basic idea was to contain a reversible color-changing indicator at the end of a pair of optical fibers. The indicator, phenol red, was covalently bound to a hydrophilic polymer in the form of water-permeable microbeads. This technique stabilized the indicator concentration. The indicator beads were contained in a sealed hydrogen-ion-permeable envelope made out of a hollow cellulose tubing. In effect, this formed a miniature spectrophotometric cell at the end of the fibers and represented an early prototype of a fiber optic chemical sensor.

The phenol red dye indicator is a weak organic acid, and the acid form (un-ionized) and base form (ionized) are present in a concentration ratio determined by the ionization constant of the acid and the pH of the medium according to the familiar Henderson-Hasselbalch equation. The two forms of the dye have different optical absorption spectra, so 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 (560 nm) and red (longer than 600 nm) light emerging from the end of one fiber passes through the dye and is reflected back into the other fiber by light-scattering particles. The green light is absorbed by the base form of the indicator. The red light is not absorbed by the indicator and is used as an optical reference. The ratio of green to red light is measured and is related to pH by an S-shaped curve with an approximate high-sensitivity linear region where the equilibrium constant (pK) of the indicator matches the pH of the solution.

The same 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 absorbance intensity. Light in the blue or UV wavelength region excites the fluorescent dye to emit longer wavelength light, and the two forms of the dye may have different excitation or emission spectra to allow their distinction.

The original instrument design for a pH measurement was very simple and consisted of a tungsten lamp for fiber illumination, a rotating filter wheel to select the green and red light returning from the fiber optic sensor, and signal processing instrumentation to give a pH output based on the green-to-red ratio. 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. The sensor was susceptible to ionic strength variation in the order of 0.01 pH unit per 11% change in ionic strength.

Further development of the pH probe for practical use was continued by Markle and colleagues [21]. They designed the fiber optic probe in the form of a 25-gauge (0.5 mm OD) hypodermic needle, with an ion-permeable side window, using 75-μm-diameter plastic optical fibers. The sensor had a 90% response time of 30 s. With improved instrumentation and computerized signal processing and with a three-point calibration, the range was extended to ±3 pH units, and a precision of 0.001 pH units was achieved.

Several reports have appeared suggesting other dye indicator systems that can be used for fiber optic pH sensing [22]. A classic problem with dye indicators is the sensitivity of their equilibrium constant to ionic strength. To circumvent this problem, Wolfbeis and Offenbacher [23] and Opitz and Lübbers [24] demonstrated a system in which a dual sensor arrangement can measure ionic strength and pH and simultaneously can correct the pH measurement for variations in ionic strength.

49.4.2.4 PCO_2 Sensors

The PCO_2 of a sample is typically determined by measuring changes in the pH of a bicarbonate solution that is isolated from the sample by a CO_2-permeable membrane but remains in equilibrium with the CO_2. The bicarbonate and CO_2, as carbonic acid, form a pH buffer system, and, by the Henderson–Hasselbalch equation, hydrogen ion concentration is proportional to the pCO_2 in the sample. This measurement is done with either a pH electrode or a dye indicator in solution.

Vurek [25] demonstrated that the same techniques can also be used with a fiber optic sensor. In his design, one plastic fiber carries light to the transducer, which is made of a silicone rubber tubing about 0.6 mm in diameter and 1.0 mm long, filled with a phenol red solution in a 35-mM bicarbonate. Ambient PCO_2 controls the pH of the solution which 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. The design by Zhujun and Seitz [26] uses a PCO_2 sensor based on a pair of membranes separated from a bifurcated optical fiber by a cavity filled with bicarbonate buffer. The external membrane is made of silicone, and the internal membrane is HPTS immobilized on an ion-exchange membrane.

49.4.2.5 PO_2 Sensors

The development of an indicator system for fiber optic PO_2 sensing is challenging because there are very few known ways to measure PO_2 optically. Although a color-changing indicator would have been desirable, the development of a sufficiently stable indicator has been difficult. The only principle applicable to fiber optics appears to be the quenching effect of oxygen on fluorescence.

Fluorescence quenching is a general property of aromatic molecules, dyes containing them, and some other substances. 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 about 10 nsec (fluorescence), or converted into another excited state with much longer mean lifetime and then reradiated (phosphorescence). Quenching reduces the intensity of fluorescence and is related to the concentration of the quenching molecules, such as O_2.

A fiber optic sensor for measuring PO_2 using the principle of fluorescence quenching was developed by Peterson and colleagues [27]. The dye is excited at around 470 nm (blue) and fluoresces at about 515 nm (green) with an intensity that depends on the PO_2. The optical information is derived from the ratio of green fluorescence to the blue excitation light, which serves as an internal reference signal. The system was chosen for visible light excitation, because plastic optical fibers block light transmission at wavelengths shorter than 450 nm, and glass fibers were not considered acceptable for biomedical use.

The sensor was similar in design to the pH probe continuing the basic idea of an indicator packing in a permeable container at the end of a pair of optical fibers. A dye perylene dibutyrate, absorbed on a macroreticular polystyrene adsorbent, is contained in a oxygen-permeable porous polystyrene envelope. The ratio of green to blue intensity was processed according to the Stren–Volmer equation:

$$\frac{I_0}{I} = 1 + K\mathrm{PO}_2 \tag{49.3}$$

where I and I_0 are the fluorescence emission intensities in the presence and absence of a quencher, respectively, and I is the Stern-Volmer quenching coefficient. This provides a nearly linear readout of PO_2 over the range of 0–150 mmHg (0–20 kPa), with a precision of 1 mmHg (0.13 kPa). The original sensor was 0.5 mm in diameter, but it can be made much smaller. Although its response time in a gas mixture is a fraction of a second, it is slower in an aqueous system, about 1.5 min for 90% response.

Wolfbeis et al. [28] designed a system for measuring the widely used halothane anesthetic which interferes with the measurement of oxygen. This dual-sensor combination had two semipermeable membranes (one of which blocked halothane) so that the probe could measure both oxygen and halothane simultaneously. The response time of their sensor, 15–20 sec for halothane and 10–15 sec for oxygen, is considered short enough to allow gas analysis in the breathing circuit. Potential applications of this device include the continuous monitoring of halothane in breathing circuits and in the blood.

49.4.3 Glucose Sensors

Another important principle that can be used in fiber optic sensors for measurements of high sensitivity and specificity is the concept of competitive binding. This was first described by Schultz et al. [29] to construct a glucose sensor. In their unique sensor, the analyte (glucose) competes for binding sites on a substrate (the lectin concanavalin A) with a fluorescent indicator-tagged polymer [fluorescein isothiocyanate (FITC)-dextran]. The sensor, which is illustrated in Figure 49.8, is 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 (300 m O.D. × 200 m ID) 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 indictor 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, fluorescing the unbound indicator, 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 the glucose concentration.

The response time of the sensor was found to be about 5 min. *In vivo* studies demonstrated fairly close correspondence between the sensor output and actual blood glucose levels. A time lag of about 5 min

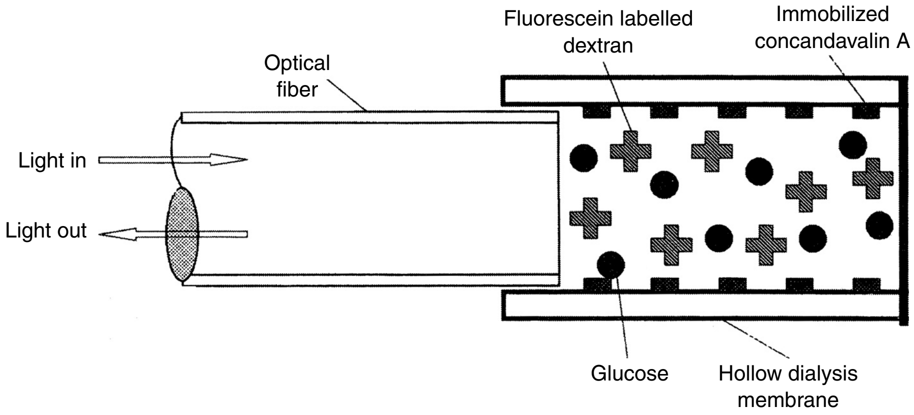


FIGURE 49.8 Schematic diagram of a competitive binding fluorescence affinity sensor for glucose measurement. (Taken from Schultz J.S., Mansouri S., and Goldstein I.J. 1982. *Diabetes Care* 5: 245. With permission.)

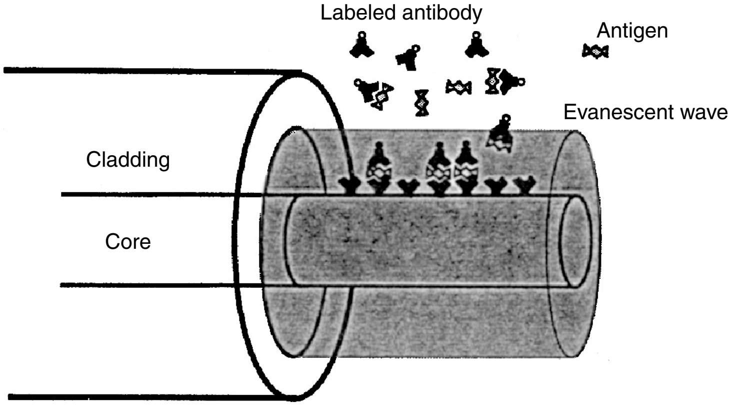


FIGURE 49.9 Basic principle of a fiber optic antigen-antibody sensor. (Taken from Anderson G.P., Golden J.P., and Ligler F.S. 1993. *IEEE Trans. Biomed. Eng.* 41: 578.)

was found and is believed to be due to the diffusion of glucose across the hollow fiber membrane and the diffusion of FTIC-dextran within the tubing.

In principle, the concept of competitive binding can be applied to any analysis for which a specific reaction can be devised. However, long-term stability of these sensors remains the major limiting factor that needs to be solved.

49.4.4 Immunosensors

Immunologic 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 trace levels in complex biologic 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. Therefore, antibody-based immunologic optical systems have been the subject of considerable research [30–34]. In practice, however, the strong binding of antigens to antibodies and vice versa causes difficulties in constructing reversible sensors with fast dynamic responses.

Several immunologic 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 complexes within the evanescent wave as illustrated in Figure 49.9. The fluorescent signal excited within the evanescent wave is then transmitted through the cladded fiber to a fluorimeter for processing.

Experimental studies have indicated that immunologic optical sensors can generally detect micromolar and even picomolar concentrations. However, the major obstacle that must be overcome to achieve high sensitivity in immunologic optical sensors is the nonspecific binding of immobilized antibodies.

References

[1] Polanyi M.L. and Heir R.M. 1962. *In vivo* oximeter with fast dynamic response. *Rev. Sci. Instrum.* 33: 1050.

[2] Bornzin G.A., Mendelson Y., Moran B.L. et al. 1987. Measuring oxygen saturation and hematocrit using a fiberoptic catheter. *Proc. 9th Ann. Conf. Eng. Med. Bio. Soc.* pp. 807–809.

[3] Mendelson Y., Galvin J.J., and Wang Y. 1990. *In vitro* evaluation of a dual oxygen saturation/hematocrit intravascular fiberoptic catheter. *Biomed. Instrum. Tech.* 24: 199.

[4] Mendelson Y. 1992. Pulse oximetry: Theory and application for noninvasive monitoring. *Clin. Chem.* 28: 1601.

[5] Aoyagi T., Kishi M., Yamaguchi K. et al. 1974. Improvement of the earpiece oximeter. *Jpn. Soc. Med. Electron. Biomed. Eng.* 90–91.

[6] Yoshiya I., Shimada Y., and Tanaka K. 1980. Spectrophotometric monitoring of arterial oxygen saturation in the fingertip. *Med. Biol. Eng. Comput.* 18: 27.

[7] Mendelson Y. and Solomita M.V. 1992. The feasibility of spectrophotometric measurements of arterial oxygen saturation from the scalp utilizing noninvasive skin reflectance pulse oximetry. *Biomed. Instrum. Technol.* 26: 215.

[8] Mendelson Y. and McGinn M.J. 1991. Skin reflectance pulse oximetry: *in vivo* measurements from the forearm and calf. *J. Clin. Monit.* 7: 7.

[9] Chance B., Leigh H., Miyake H. et al. 1988. Comparison of time resolved and un-resolved measurements of deoxyhemoglobin in brain. *Proc. Natl Acad. Sci. USA* 85: 4971.

[10] Jobsis F.F., Keizer J.H., LaManna J.C. et al. 1977. Reflection spectrophotometry of cytochrome aa3 *in vivo. Appl. Physiol: Respirat. Environ. Excerc. Physiol.* 43: 858.

[11] Kurth C.D., Steven I.M., Benaron D. et al. 1993. Near-infrared monitoring of the cerebral circulation. *J. Clin. Monit.* 9: 163.

[12] Lübbers D.W. and Opitz N. 1975. 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.* 30C: 532.

[13] Clark C.L., O'Brien J., McCulloch J. et al. 1986. Early clinical experience with GasStat. *J. Extra Corporeal. Technol.* 18: 185.

[14] Hill A.G., Groom R.C., Vinansky R.P. et al. 1985. On-line or off-line blood gas analysis: Cost vs. time vs. accuracy. *Proc. Am. Acad. Cardiovasc. Perfusion* 6: 148.

[15] Siggaard-Andersen O., Gothgen I.H., Wimberley et al. 1988. Evaluation of the GasStat fluorescence sensors for continuous measurement of pH, pCO_2 and pO_3 during CPB and hypothermia. *Scand. J. Clin. Lab. Invest.* 48: 77.

[16] Zimmerman J.L. and Dellinger R.P. 1993. Initial evaluation of a new intra-arterial blood gas system in humans. *Crit. Care Med.* 21: 495.

[17] Gottlieb A. 1992. The optical measurement of blood gases — approaches, problems and trends: Fiber optic medical and fluorescent sensors and applications. *Proc. SPIE* 1648: 4.

[18] Barker S.L. and Hyatt J. 1991. Continuous measurement of intraarterial pHa, $PaCO_2$, and PaO_2 in the operation room. *Anesth. Analg.* 73: 43.

[19] Larson C.P., Divers G.A., and Riccitelli S.D. 1991. Continuous monitoring of PaO_2 and $PaCO_2$ in surgical patients. *Abstr. Crit. Care Med.* 19: 525.

[20] Peterson J.I., Goldstein S.R., and Fitzgerald R.V. 1980. Fiber optic pH probe for physiological use. *Anal. Chem.* 52: 864.

[21] Markle D.R., McGuire D.A., Goldstein S.R. et al. 1981. A pH measurement system for use in tissue and blood, employing miniature fiber optic probes. In D.C. Viano (Ed.), *Advances in Bioengineering*, p. 123, New York, American Society of Mechanical Engineers.

[22] Wolfbeis O.S., Furlinger E., Kroneis H. et al. 1983. Fluorimeter analysis: 1. A study on fluorescent indicators for measuring near neutral (physiological) pH values. *Fresenius' Z. Anal. Chem.* 314: 119.

[23] Wolfbeis O.S. and Offenbacher H. 1986. Fluorescence sensor for monitoring ionic strength and physiological pH values. *Sens. Actuat.* 9: 85.

[24] Opitz N. and Lübbers D.W. 1983. New fluorescence photomatrical techniques for simultaneous and continuous measurements of ionic strength and hydrogen ion activities. *Sens. Actuat.* 4: 473.

[25] Vurek G.G., Feustel P.J., and Severinghaus J.W. 1983. A fiber optic pCO_2 sensor. *Ann. Biomed. Eng.* 11: 499.

[26] Zhujun Z. and Seitz W.R. 1984. A carbon dioxide sensor based on fluorescence. *Anal. Chim. Acta* 160: 305.

[27] Peterson J.I., Fitzgerald R.V., and Buckhold D.K. 1984. Fiber-optic probe for *in vivo* measurements of oxygen partial pressure. *Anal. Chem.* 56: 62.

[28] Wolfbeis O.S., Posch H.E., and Kroneis H.W. 1985. Fiber optical fluorosensor for determination of halothane and/or oxygen. *Anal. Chem.* 57: 2556.

[29] Schultz J.S., Mansouri S., and Goldstein I.J. 1982. Affinity sensor: a new technique for developing implantable sensors for glucose and other metabolites. *Diabetes Care* 5: 245.

[30] Andrade J.D., Vanwagenen R.A., Gregonis D.E. et al. 1985. Remote fiber optic biosensors based on evanescent-excited fluoro-immunoassay: concept and progress. *IEEE Trans. Elect. Devices* ED-32: 1175.

[31] Sutherland R.M., Daehne C., Place J.F. et al. 1984. Optical detection of antibody–antigen reactions at a glass–liquid interface. *Clin. Chem.* 30: 1533.

[32] Hirschfeld T.E. and Block M.J. 1984. Fluorescent immunoassay employing optical fiber in a capillary tube. US Patent No. 4,447,546.

[33] Anderson G.P., Golden J.P., and Ligler F.S. 1993. An evanescent wave biosensor: Part I. Fluorescent signal acquisition from step-etched fiber optic probes. *IEEE Trans. Biomed. Eng.* 41: 578.

[34] Golden J.P., Anderson G.P., Rabbany S.Y. et al. 1994. An evanescent wave biosensor: Part II. Fluorescent signal acquisition from tapered fiber optic probes. *IEEE Trans. Biomed. Eng.* 41: 585.

50

Bioanalytic Sensors

Richard P. Buck
University of North Carolina

50.1 Classification of Biochemical Reactions in the Context of Sensor Design and Development

50.1.1 Introduction and Definitions

Since sensors generate a measurable material property, they belong in some grouping of transducer devices. Sensors specifically contain a recognition process that is characteristic of a material sample at the molecular-chemical level, and a sensor incorporates a transduction process (step) to create a useful signal. Biomedical sensors include a whole range of devices that may be chemical sensors, physical sensors, or some kind of mixed sensor.

Chemical sensors use chemical processes in the recognition and transduction steps. Biosensors are also chemical sensors, but they use particular classes of biological recognition/transduction processes. A pure physical sensor generates and transduces a parameter that does not depend on the chemistry per se, but is a result of the sensor responding as an aggregate of point masses or charges. All these when used in a biologic system (biomatrix) may be considered bioanalytic sensors without regard to the chemical, biochemical, or physical distinctions. They provide an "analytic signal of the biologic system" for some further use.

The chemical recognition process focuses on some molecular-level chemical entity, usually a kind of chemical structure. In classical analysis this structure may be a simple functional group: SiO — in a glass electrode surface, a chromophore in an indicator dye, or a metallic surface structure, such as silver metal that recognizes Ag+ in solution. In recent times, the biologic recognition processes have been better understood, and the general concept of recognition by receptor or chemoreceptor has come into fashion. Although these are often large molecules bound to cell membranes, they contain specific structures that

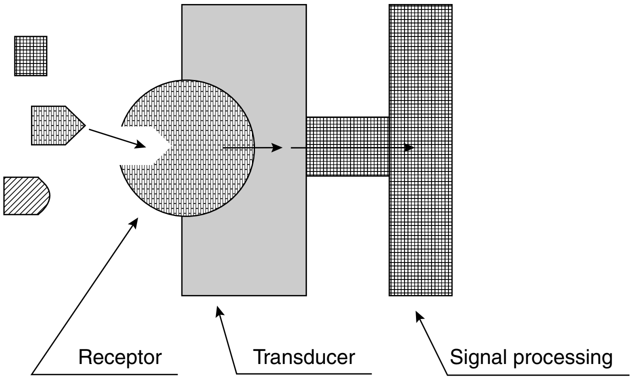


FIGURE 50.1 Generic bioanalytic sensor.

permit a wide variety of different molecular recognition steps including recognition of large and small species and of charged and uncharged species. Thus, chemoreceptor appears in the sensor literature as a generic term for the principal entity doing the recognition. For a history and examples, see References 1 to 6.

Biorecognition in biosensors has especially stressed "receptors" and their categories. Historically, application of receptors has not necessarily meant measurement directly of the receptor. Usually there are coupled chemical reactions, and the transduction has used measurement of the subsidiary products: change of pH, change of dissolved O_2, generation of H_2O_2, changes of conductivity, changes of optical adsorption, and changes of temperature. Principal receptors are enzymes because of their extraordinary selectivity. Other receptors can be the more subtle species of biochemistry: antibodies, organelles, microbes, and tissue slices, not to mention the trace level "receptors" that guide ants, such as pheromones, and other unusual species. A sketch of a generic bioanalytic sensor is shown in Figure 50.1.

50.1.2 Classification of Recognition Reactions and Receptor Processes

The concept of recognition in chemistry is universal. It almost goes without saying that all chemical reactions involved recognition and selection on the basis of size, shape, and charge. For the purpose of constructing sensors, general recognition based on these factors is not usually enough. Frequently in inorganic chemistry a given ion will react indiscriminantly with similar ions of the same size and charge. Changes in charge from unity to two, for example, do change the driving forces of some ionic reactions. By control of dielectric constant of phases, heterogeneous reactions can often be "tailored" to select divalent ions over monovalent ions and to select small versus large ions or vice versa.

Shape, however, has more special possibilities, and natural synthetic methods permit product control. Nature manages to use shape together with charge to build organic molecules, called enzymes, that have acquired remarkable selectivity. It is in the realm of biochemistry that these natural constructions are investigated and catalogued. Biochemistry books list large numbers of enzymes and other selective materials that direct chemical reactions. Many of these have been tried as the basis of selective sensors for bioanalytic and biomedical purposes. The list in Table 50.1 shows how some of the materials can be grouped into lists according to function and to analytic substrate, both organic and inorganic. The principles seem general, so there is no reason to discriminate against the inorganic substrates in favor or the organic substrates. All can be used in biomedical analysis.

50.2 Classification of Transduction Processes — Detection Methods

Some years ago, the engineering community addressed the topic of sensor classification — Richard M. White in *IEEE Trans. Ultra., Ferro., Freq. Control* (UFFC), UFFC-34 (1987) 124, and Wen H. Ko

TABLE 50.1 Recognition Reactions and Receptor Processes

1. Insoluble salt-based sensors	
a. $S^{+} + R^{-}$ 1 (insoluble salt)	
Ion exchange with crystalline SR (homogeneous or heterogeneous crystals)	
chemical signal S^{+n}	receptor R^{-n}
inorganic cations	inorganic anions
examples: $Ag^{+}, Hg_2^{2+}, Pb^{2+}, Cd^{2+}, Cu^{2+}$	$S^{=}, Se^{2=}, SCN^{-}, I^{-}, Br^{-}, Cl^{-}$
b. $S^{-n} + R^{+n}$ 1SR (insoluble salt)	
Ion exchange with crystalline SR (homogeneous or heterogeneous crystals)	
chemical signal S^{-n}	receptor R^{+n}
inorganic anions	inorganic cations
examples: $F^{-}, S^{=}, Se^{2=}, SCN^{-}, I^{-}, Br^{-}, Cl^{-}$	$LaF_2^{+}, Ag^{+}, Hg_2^{2+}, Pb^{2+}, Cd^{2+}, Cu^{2+}$
2. Solid ion exchanges	
a. $S^{+n} + R^{-n}$ (sites) lS^{+n} R^{-n} = SR (in ion exchanger phase)	
Ion exchange with synthetic ion exchangers containing negative fixed sites (homogeneous or heterogeneous, inorganic or organic materials)	
chemical signal S^{+n}	receptor R^{-n}
inorganic and organic ions	inorganic and organic ion sites
examples: H^{+}, Na^{+}, K^{+}	silicate glass Si-0^{-}
H^{+}, Na^{+}, K^{+}, other M^{+n}	synthetic sulfonated, phosphorylated, EDTA-substituted polystyrenes
b. $S^{-n} + R^{+n}$ (sites) 1S^{-n} R^{+n} = SR (in ion exchanger phase)	
Ion exchange with synthetic ion exchangers containing positive fixed sites (homogeneous or heterogeneous, inorganic or organic materials)	
chemical signal S^{-n}	receptor R^{+n}
organic and inorganic ions	organic and inorganic ion sites
examples: hydrophobic anions	quaternized polystyrene
3. Liquid ion exchanger sensors with electrostatic selection	
a. $S^{-n} + R^{-n}$ (sites) lS^{+n} R^{-n} = SR (in ion exchanger phase)	
Plasticized, passive membranes containing mobile trapped negative fixed sites (homogeneous or heterogeneous, inorganic or organic materials)	
chemical signal S^{+n}	receptor R^{-n}
inorganic and organic ions	inorganic and organic ion sites
examples: Ca^{2+}	diester of phosphoric acid or monoester of a phosphonic acid
M^{+n}	dinonylnaphthalene sulfonate and other organic, hydrophobic anions
$R_1, R_2, R_3\ R_4\ N^{+}$ and bis-Quaternary Cations	tetraphenylborate anion or substituted derivatives
cationic drugs tetrasubstituted arsonium^{-}	
b. $S^{-n} + R^{+n}$ (sites) 1S^{-n} R^{+n} = SR (in ion exchanger phase)	
Plasticized, passive membranes containing mobile, trapped negative fixed sites (homogeneous or heterogeneous, inorganic or organic materials)	
chemical signal S^{-n}	receptor R^{+n}
inorganic and organic ions	inorganic and organic sites
examples: anions, simple $Cl^{-}, Br^{-}, C1O_4^{-}$	quaternary ammonium cations: e.g., tridodecylmethyl-ammonium
anions, complex, drugs	quaternary ammonium cations: e.g., tridodecylmethyl-ammonium

(Continued)

TABLE 50.1 (Continued) Recognition Reactions and Receptor Processes

4. Liquid ion exchanger sensors with neutral (or charged) carrier selection	
a. S^{+n} + X and R^{-n} (sites) lS^{+n} X R^{-n} = SXR (in ion exchanger phase)	
Plasticized, passive membranes containing mobile, trapped negative fixed sites (homogeneous or heterogeneous, inorganic or organic materials)	
chemical signal S^{+n}	receptor R^{-n}
inorganic and organic ions	inorganic and organic ion sites
examples: Ca^{2+}	X = synthetic ionophore complexing agent selective to Ca^{2+} R^{-n} usually a substituted tetra phenylborate salt
Na^{+}, K^{+}, H^{+}	X = selective ionophore complexing agent
b. S^{-n} + X and R^{+n} (sites) 1S^{-n} X R^{+n} = SXR (in ion exchanger phase)	
Plasticized, passive membranes containing mobile, trapped negative fixed sites (homogeneous or heterogeneous, inorganic or organic materials)	
chemical signal S^{-n}	receptor R^{+n}
inorganic and organic ions	inorganic and organic ion sites
examples: $HPO_4^{2=}$	R^{+n} = quaternary ammonium salt X = synthetic ionophore complexing agent; aryl organotin compound or suggested cyclic polyamido-polyamines
HCO_3^-	X = synthetic ionophore: trifluoro acetophenone
Cl^-	X = aliphatic organotin compound
5. Bioaffinity sensors based on change of local electron densitites	
S + R 1SR	
chemical signal S	receptor R
protein	dyes
saccharide	lectin
glycoprotein	
susbstrate	enzyme
inhibitor	Transferases Hydrolases (peptidases, esterases, etc.) Lyases Isomerases Ligases
prosthetic group	apoenzyme
antigen	antibody
hormone	"receptor"
substrate analogue	transport system
6. Metabolism sensors based on substrate consumption and product formation	
S + R 1SR → P + R	
chemical signal S	receptor R
substrate	enzyme
examples: lactate (SH_2)	hydrogenases catalyze hydrogen transfer from S to acceptor A (not molecular oxygen!) reversibly pyruvate + NADH +
SH_2 + A 1S + AH_2 lactate + NAD^+	H^+ using lactate dehydrogenase
glucose (SH_2)	
$SH_2 + \frac{1}{2} O_2$ 1S + H_2O or	oxidases catalyze hydrogen transfer to molecular oxygen
$SH_2 + O_2$ 1S + H_2O_2	using glucose oxidase
glucose + O_2 1gluconolactone + H_2O_2	
reducing agents (S)	peroxidases catalyze oxidation of a substrate by H_2O_2 using horseradish peroxidase

(Continued)

TABLE 50.1 (Continued) Recognition Reactions and Receptor Processes

$2S + 2H^+ + H_2O_2 12S^+ + 2H_2O$	
$Fe^{2+} + H_2O_2 + 2H^+ 1Fe^{3+} + 2H_2O$	
reducing agents	oxygenates catalyze substrate oxidations by molecular O_2
L-lactate + O_2 lactate + CO_2 + H_2O	
cofactor	organelle
inhibitor	microbe
activator	tissue slice
enzyme activity	
7. Coupled and hybrid systems using sequences, competition, anti-interference and amplification concepts and reactions.	
8. Biomimetic sensors	
chemical signal S	receptor R
sound	carrier-enzyme
stress	
light	

Source: Adapted from Scheller F., Schubert F. 1989. *Biosensors, #18 in Advances in Research Technologies (Beitrage zur Forschungstec technologies)*, Berlin, Akademie-Verlag, Amsterdam, Elsevier. Cosofret V.V., Buck R.P. 1992. *Pharmaceutical Applications of Membrane Sensors*, Boca Raton, FL, CRC Press.

in IEEE/EMBS Symposium Abstract T.1.1 84CH2068-5 (1984). It is interesting because the physical and chemical properties are given equal weight. There are many ideas given here that remain without embodiment. This list is reproduced as Table 50.2. Of particular interest in this section are "detection means used in sensors" and "sensor conversion phenomena." At present the principle transduction schemes use electrochemical, optical, and thermal detection effects and principles.

50.2.1 Calorimetric, Thermometric, and Pyroelectric Transducers

Especially useful for enzymatic reactions, the generation of heat (enthalpy change) can be used easily and generally. The enzyme provides the selectivity and the reaction enthalpy cannot be confused with other reactions from species in a typical biologic mixture. The ideal aim is to measure total evolved heat, that is, to perform a calorimetric measurement. In real systems there is always heat loss, that is, heat is conducted away by the sample and sample container so that the process cannot be adiabatic as required for a total heat evolution measurement. As a result, temperature difference before and after evolution is measured most often. It has to be assumed that the heat capacity of the specimen and container is constant over the small temperature range usually measured.

The simplest transducer is a thermometer coated with the enzyme that permits the selected reaction to proceed. Thermistors are used rather than thermometers or thermocouples. The change of resistance of certain oxides is much greater than the change of length of a mercury column or the microvolt changes of thermocouple junctions.

Pyroelectric heat flow transducers are relatively new. Heat flows from a heated region to a lower temperature region, controlled to occur in one dimension. The lower temperature side can be coated with an enzyme. When the substrate is converted, the lower temperature side is warmed. The pyroelectric material is from a category of materials that develops a spontaneous voltage difference in a thermal gradient. If the gradient is disturbed by evolution or adsorption of heat, the voltage temporarily changes.

In biomedical sensing, some of the solid-state devices based on thermal sensing cannot be used effectively. The reason is that the sensor itself has to be heated or is heated quite hot by catalytic surface reactions. Thus pellistors (oxides with catalytic surfaces and embedded platinum wire thermometer), chemiresistors, and "Figaro" sensor "smoke" detectors have not found many biologic applications.

TABLE 50.2 Detection Means and Conversion Phenomena Used in Sensors

Detection means		
	Biologic	
	Chemical	
	Electric, magnetic, or electromagnetic wave	
	Heat, temperature	
	Mechanical displacement of wave	
	Radioactivity, radiation	
	Other	
Conversion phenomena		
	Biologic	
		Biochemical transformation
		Physical transformation
		Effect on test organism
		Spectroscopy
		Other
	Chemical	
		Chemical transformation
		Physical transformation
		Electrochemical process
		Spectroscopy
		Other
	Physical	
		Thermoelectric
		Photoelectric
		Photomagnetic
		Magnetoelectric
		Elastomagnetic
		Thermoelastic
		Elastoelectric
		Thermomagnetic
		Thermooptic
		Photoelastic
		Others

50.2.2 Optical, Optoelectronic Transducers

Most optical detection systems for sensors are small, that is, they occupy a small region of space because the sample size and volume are themselves small. This means that common absorption spectrophotometers and photofluorometers are not used with their conventional sample-containing cells, or with their conventional beam-handling systems. Instead light-conducting optical fibers are used to connect the sample with the more remote monochromator and optical readout system. The techniques still remain absorption spectrophotometry, fluorimetry including fluorescence quenching, and reflectometry.

The most widely published optical sensors use a miniature reagent contained or immobilized at the tip of an optical fiber. In most systems a permselective membrane coating allows the detected species to penetrate the dye region. The corresponding absorption change, usually at a sensitive externally preset wavelength, is changed and correlated with the sample concentration. Similarly, fluorescence can be stimulated by the higher-frequency external light source and the lower-frequency emission detected. Some configurations are illustrated in References 1 and 2. Fluorimetric detection of coenzyme A, NAD+/NADH, is involved in many so-called pyridine-linked enzyme systems. The fluorescence of NADH contained or immobilized can be a convenient way to follow these reactions. Optodes, miniature encapsulated dyes, can be placed *in vivo*. Their fluorescence can be enhanced or quenched and used to detect acidity, oxygen, and other species.

A subtle form of optical transduction uses the "peeled" optical fiber as a multiple reflectance cell. The normal fiber core glass has a refractive index greater than that of the exterior coating; there is a range of angles of entry to the fiber so that all the light beam remains inside the core. If the coating is removed and materials of lower index of refraction are coated on the exterior surface, there can be absorption by multiple reflections, since the evanescent wave can penetrate the coating. Chemical reagent can be added externally to create selective layers on the optical fiber.

Ellipsometry is a reflectance technique that depends on the optical constants and thickness of surface layer. For colorless layers, a polarized light beam will change its plane of polarization upon reflection by the surface film. The thickness can sometimes be determined when optical constants are known or approximated by constants of the bulk material. Antibody–antigen surface reaction can be detected this way.

50.2.3 Piezoelectric Transducers

Cut quartz crystals have characteristic modes of vibration that can be induced by painting electrodes on the opposite surfaces and applying a megaHertz ac voltage. The frequency is searched until the crystal goes into a resonance. The resonant frequency is very stable. It is a property of the material and maintains a value to a few parts per hundred million. When the surface is coated with a stiff mass, the frequency is altered. The shift in frequency is directly related to the surface mass for thin, stiff layers. The reaction of a substrate with this layer changes the constants of the film and further shifts the resonant frequency. These devices can be used in air, in vacuum, or in electrolyte solutions.

50.2.4 Electrochemical Transducers

Electrochemical transducers are commonly used in the sensor field. The main forms of electrochemistry used are potentiometry (zero-current cell voltage [potential difference measurements]), amperometry (current measurement at constant applied voltage at the working electrode), and ac conductivity of a cell.

50.2.4.1 Potentiometric Transduction

The classical generation of an activity-sensitive voltage is spontaneous in a solution containing both nonredox ions and redox ions. Classical electrodes of types 1, 2, and 3 respond by ion exchange directly or indirectly to ions of the same material as the electrode. Inert metal electrodes (sometimes called type 0) — Pt, Ir, Rh, and occasionally carbon C — respond by electrons exchange from redox pairs in solution. Potential differences are interfacial and reflect ratios of activities of oxidized to reduced forms.

50.2.4.2 Amperometric Transduction

For dissolved species that can exchange electrons with an inert electrode, it is possible to force the transfer in one direction by applying a voltage very oxidizing (anodic) or reducing (cathodic). When the voltage is fixed, the species will be, by definition, out of equilibrium with the electrode at its present applied voltage. Locally, the species (regardless of charge) will oxidize or reduce by moving from bulk solution to the electrode surface where they react. Ions do not move like electrons. Rather they diffuse from high to low concentration and do not usually move by drift or migration. The reason is that the electrolytes in solutions are at high concentrations, and the electric field is virtually eliminated from the bulk. The field drops through the first 1000 A at the electrode surface. The concentration of the moving species is from high concentration in bulk to zero at the electrode surface where it reacts. This process is called concentration polarization. The current flowing is limited by mass transport and so is proportional to the bulk concentration.

50.2.4.3 Conductometric Transducers

Ac conductivity (impedance) can be purely resistive when the frequency is picked to be about 1000 to 10,000 Hz. In this range the transport of ions is sufficiently slow that they never lose their uniform

TABLE 50.3 Chemical Sensors and Properties Documented in the Literature

I.	General topics including items II–V; selectivity, fabrication, data processing		
II.	Thermal sensors		
III.	Mass sensors		
		Gas sensors	
		Liquid sensors	
IV.	Electrochemical sensors		
	Potentiometric sensors		
		Reference electrodes	
		Biomedical electrodes	
		Applications to cations, anions	
		Coated wire/hybrids	
		ISFETs and related	
		Biosensors	
		Gas sensors	
	Amperometric sensors		
		Modified electrodes	
		Gas sensors	
		Biosensors	
			Direct electron transfer
			Mediated electron transfer
			Biomedical
	Conductimetric sensors		
		Semiconducting oxide sensors	
		Zinc oxide-based	
		Chemiresistors	
		Dielectrometers	
V.	Optical sensors		
		Liquid sensors	
		Biosensors	
		Gas sensors	

concentration. They simply quiver in space and carry current forward and backward each half cycle. In the lower and higher frequencies, the cell capacitance can become involved, but this effect is to be avoided.

50.3 Tables of Sensors from the Literature

The longest and most consistently complete references to the chemical sensor field is the review issue of Analytical Chemistry Journal. In the 1970s and 1980s these appeared in the April issue, but more recently they appear in the June issue. The editors are Jiri Janata and various colleagues [7–10]. Note all possible or imaginable sensors have been made according to the list in Table 50.2. A more realistic table can be constructed from the existing literature that describes actual devices. This list is Table 50.3. Book references are listed in Table 50.4 in reverse time order to about 1986. This list covers most of the major source books and many of the symposium proceedings volumes. The reviews [7–10] are a principal source of references to the published research literature.

50.4 Applications of Microelectronics in Sensor Fabrication

The reviews of sensors since 1988 cover fabrication papers and microfabrication methods and examples [7–10]. A recent review by two of the few chemical sensor scientists (chemical engineers) who also operate a microfabrication laboratory is C. C. Liu, Z.-R. Zhang. 1992. Research and development of chemical sensors using microfabrication techniques. *Selective Electrode* 14: 147.

TABLE 50.4 Books and Long Reviews Keyed to Items in Table 50.3 (Reviewed Since 1988 in Reverse Time Sequence)

I. General Topics

Yamauchi S. (ed). 1992. *Chemical Sensor Technology*, Vol 4, Tokyo, Kodansha Ltd.

Flores J.R., Lorenzo E. 1992. Amperometric biosensors, In M.R. Smyth, J.G. Vos (eds), *Comprehensive Analytical Chemistry*, Amsterdam, Elsevier

Vaihinger S., Goepel W. 1991. Multicomponent analysis in chemical sensing. In W. Goepel, J. Hesse, J. Zemel (eds), *Sensors*, Vol 2, Part 1, pp. 191–237, Weinheim, Germany, VCH Publishers

Wise D.L. (ed). 1991. *Bioinstrumentation and Biosensors*, New York, Marcel Dekker Scheller F., Schubert F. 1989. *Biosensors*, Basel, Switzerland, Birkhauser Verlag, see also [2].

Madou M., Morrison S.R. 1989. *Chemical Sensing with Solid State Devices*, New York, Academic Press.

Janata J. 1989. *Principles of Chemical Sensors*, New York, Plenum Press.

Edmonds T.E. (ed). 1988. *Chemical Sensors*, Glasgow, Blackie.

Yoda K. 1988. Immobilized enzyme cells. *Methods Enzymology*, 137: 61.

Turner A.P.F., Karube I., Wilson G.S. (eds). 1987. *Biosensors: Fundamentals and Applications*, Oxford, Oxford University Press.

Seiyama T. (ed). 1986. *Chemical Sensor Technology*, Tokyo, Kodansha Ltd.

II. Thermal Sensors

There are extensive research and application papers and these are mentioned in books listed under I. However, the up-to-date lists of papers are given in references 7 to 10.

III. Mass Sensors

There are extensive research and application papers and these are mentioned in books listed under I. However, the up-to-date lists of papers are given in references 7 to 10. Fundamentals of this rapidly expanding field are recently reviewed:

Buttry D.A., Ward M.D. 1992. Measurement of interfacial processes at electrode surfaces with the electrochemical quartz crystal microbalance, *Chemical Reviews* 92: 1355.

Grate J.W., Martin S.J., White R.M. 1993. Acoustic wave microsensors, Part 1, *Analyt Chem* 65: 940A; part 2, *Analyt. Chem.* 65: 987A.

Ricco A.T. 1994. SAW Chemical sensors, *The Electrochemical Society Interface Winter:* 38–44.

IVA. Electrochemical Sensors — Liquid Samples

Scheller F., Schmid R.D. (eds). 1992. *Biosensors*: *Fundamentals, Technologies and Applications*, GBF Monograph Series, New York, VCH Publishers.

Erbach R., Vogel A., Hoffmann B. 1992. Ion-sensitive field-effect structures with Langmuir-Blodgett membranes. In F. Scheller, R.D. Schmid (eds). *Biosensors: Fundamentals, Technologies, and Applications*, GBF Monograph 17, pp. 353–357, New York, VCH Publishers.

Ho May Y.K., Rechnitz G.A. 1992. An introduction to biosensors, In R.M. Nakamura, Y. Kasahara, G.A. Rechnitz (eds), *Immunochemical Assays and Biosensors Technology*, pp. 275–291, Washington, DC, American Society Microbiology.

Mattiasson B., Haakanson H. Immunochemically-based assays for process control, 1992. *Advances in Biochemical Engineering and Biotechnology* 46: 81.

Maas A.H., Sprokholt R. 1990. Proposed IFCC Recommendations for electrolyte measurements with ISEs in clinical chemistry, In A. Ivaska, A. Lewenstam, R. Sara (eds), *Contemporary Electroanalytical Chemistry, Proceedings of the ElectroFinnAnalysis International Conference on Electroanalytical Chemistry*, pp. 311–315, New York, Plenum.

Vanrolleghem P., Dries D., Verstreate W. RODTOX: Biosensor for rapid determination of the biochemical oxygen demand, 1990. In C. Christiansen, L. Munck, J. Villadsen (eds), *Proceedings of the 5th European Congress Biotechnology*, Vol 1, pp. 161–164, Copenhagen, Denmark, Munksgaard.

Cronenberg C., Van den Heuvel H., Van den Hauw M., Van Groen B. Development of glucose microelectrodes for measurements in biofilms, 1990. In C. Christiansen, L. Munck, J. Villadsen (eds), *Proceedings of the 5th European Congress Biotechnology*, Vol 1, pp. 548–551, Copenhagen, Denmark, Munksgaard.

Wise D.L. (ed). 1989. *Bioinstrumentation Research, Development and Applications*, Boston, MA, Butterworth-Heinemann.

Pungor E. (ed). 1989. *Ion-Selective Electrodes — Proceedings of the 5th Symposium* [Matrafured, Hungary 1988], Oxford, Pergamon.

Wang J. (ed). 1988. *Electrochemical Techniques in Clinical Chemistry and Laboratory Medicine*, New York, VCH Publishers.

Evans A. 1987. *Potentiometry and lon-seiective Electrodes*, New York, Wiley.

Ngo T.T. (ed). 1987. *Electrochemical Sensors in Immunological Analysis*, New York, Plenum.

(Continued)

TABLE 50.4 (Continued) Books and Long Reviews Keyed to Items in Table 50.3

IVB.	**Electrochemical Sensors — Gas Samples**
	Sberveglieri G. (ed). 1992. *Gas Sensors*, Dordrecht The Netherlands, Kluwer.
	Moseley P.T., Norris J.O.W., Williams D.E. 1991. *Technology and Mechanisms of Gas Sensors*, Bristol, U.K., Hilger.
	Moseley P.T., Tofield B.D. (eds). 1989. *Solid State Gas Sensors*, Philadelphia, Taylor and Francis, Publishers.
V.	**Optical Sensors**
	Coulet P.R., Blum L.J. Luminescence in biosensor design, 1991. In D.L. Wise, L.B. Wingard, Jr (eds). *Biosensors with Fiberoptics*, pp. 293–324, Clifton, N.J., Humana.
	Wolfbeis OS. 1991. Spectroscopic techniques, In O.S. Wolfbeis (ed). *Fiber Optic Chemical Sensors and Biosensors*, Vol 1, pp. 25–60. Boca Raton. FL, CRC Press.
	Wolfbeis O.S. 1987. Fibre-optic sensors for chemical parameters of interest in biotechnology, In R.D. Schmidt (ed). GBF (Gesellschaft fur Biotechnologische Forschung) *Monogr. Series*, Vol 10, pp. 197–206, New York, VCH Publishers.

References

[1] Janata J. 1989. *Principles of Chemical Sensors*, New York, Plenum.

[2] Scheller F. and Schubert F. 1989. *Biosensors, #18 in Advances in Research Technologies (Beitrage zur Forschungstechnologie)*, Berlin, Akademie-Verlag, Amsterdam, Elsevier (English translation).

[3] Turner A.P.F., Karube I., and Wilson G.S. 1987. *Biosensors: Fundamentals and Applications*, Oxford, Oxford University Press.

[4] Hall E.A.H. 1990. *Biosensors*, Milton Keynes, England, Open University Press.

[5] Eddoes M.J. 1990. Theoretical methods for analyzing biosensor performance. In A.E.G. Cass (ed), *Biosensor — A Practical Approach*, Oxford, IRL Press at Oxford University, Ch. 9, pp. 211–262.

[6] Cosofret V.V. and Buck R.P. 1992. *Pharmaceutical Applications of Membrane Sensors*, Boca Raton, FL, CRC Press.

[7] Janata J. and Bezegh A. 1988. Chemical sensors, *Analyt. Chem.* 60: 62R.

[8] Janata J. 1990. Chemical sensors, *Analyt. Chem.* 62: 33R.

[9] Janata J. 1992. Chemical sensors, *Analyt. Chem.* 66: 196R.

[10] Janata J. and Josowicz M., and DeVaney M. 1994. Chemical sensors, *Analyt. Chem.* 66: 207R.

51

Biological Sensors for Diagnostics

Orhan Soykan
Medtronic, Inc.
Michigan Technological University

51.1 Diagnostics Industry

Many biologically relevant molecules can be measured from the samples taken from the body, which constitutes the foundation of the medical diagnostics industry. In 2003, the global clinical diagnostic market was more than U.S. $2 billion. Of that, sales of the laboratory instruments constituted slightly less than half, while the point of care systems and the diagnostic kits made up the rest. Even though this amount accounts for only a few percent of the total spending on health care, it continues to grow, and not surprisingly, a significant number of biomedical engineers are employed in the research, design, and manufacturing of these products.

Utilization of these devices for various functions is shown in Table 51.1.

In this chapter, we will discuss some examples to illustrate the principles and the technologies used for these measurements. They will be categorized in three groups (a) sensors for proteins and enzymes, (b) sensors for nucleic acids, and (c) sensors for cellular processes.

51.2 Diagnostic Sensors for Measuring Proteins and Enzymes

We are all born with the genes that we will carry throughout our lives, and our genetic makeup remains relatively constant except in parts of the immune and reproductive systems. Expression patterns of genes

TABLE 51.1 Diagnostics Industry by Discipline

Discipline	Percentage (%)
Clinical chemistry tests	42.0
Immunodiagnostics	30.6
Hematology/flow cytometry	7.7
Microbiology	6.9
Molecular diagnostics	5.8
Coagulation	3.4
Other	3.5

Source: Simonsen, M., *BBI Newsletter*, 27: pp. 221–228, 2004.

on the other hand are noting but constant. These changes can be due to aging, environmental and physiological conditions we experience, or due to diseases. Hence, many of the diseases can be detected by sensing the presence, or measuring the level of activity of proteins and enzymes in the tissue and blood for diagnostic purposes. In this section, we will review some of these techniques.

51.2.1 Spectrophotometry

Spectrophotometry utilizes the principle of atomic absorption to determine the concentration of a substance in a volume of solution. Transmission of light through a clear fluid containing an analyte has a reciprocal relationship to the concentration of the analyte, as shown in Figure 51.1b [2]. Percent transmission can be calculated as

$$\%T = \frac{I_\mathrm{T}}{I_\mathrm{O}}100$$

where, I_T and I_O are intensities of transmitted and incident light respectively.

Absorption (A) can be defined as $A = -\log(\%T)$, which yields a linear relationship between absorption and the concentration (C) of the solute, as shown in Figure 51.1c.

A schematic diagram of a spectrophotometer is shown in Figure 51.1a. Light from an optical source is first passed through a monochromator, such as a prism or a diffraction grating. Then, a beam splitter produces two light beams where one passes through a cuvette containing the patient sample, and the other through a cuvette containing a reference solution. Intensities of the transmitted light beams are detected and compared to each other to determine the concentration of the analyte in the sample [3].

The exponential form of the Beer–Lambert law can be used to calculate the absorption of light passing through a solution.

$$\frac{I_\mathrm{T}}{I_\mathrm{O}} = \mathrm{e}^{-A}$$

where I_T is the transmitted light intensity, I_O is the incident light intensity, and A is the absorption occurring in the light amplitude as it travels through the media.

Absorption in a cuvette can be calculated as follows:

$$A = abC$$

where a is the absorptivity coefficient, b is the optical path length in the solution, and C is the concentration of the colored analyte of interest.

If A_S and A_R are the absorption in the sample and the reference cuvettes, and the C_S and C_R are the analyte concentrations in the sample and reference cuvettes, then the concentration in the sample cuvette

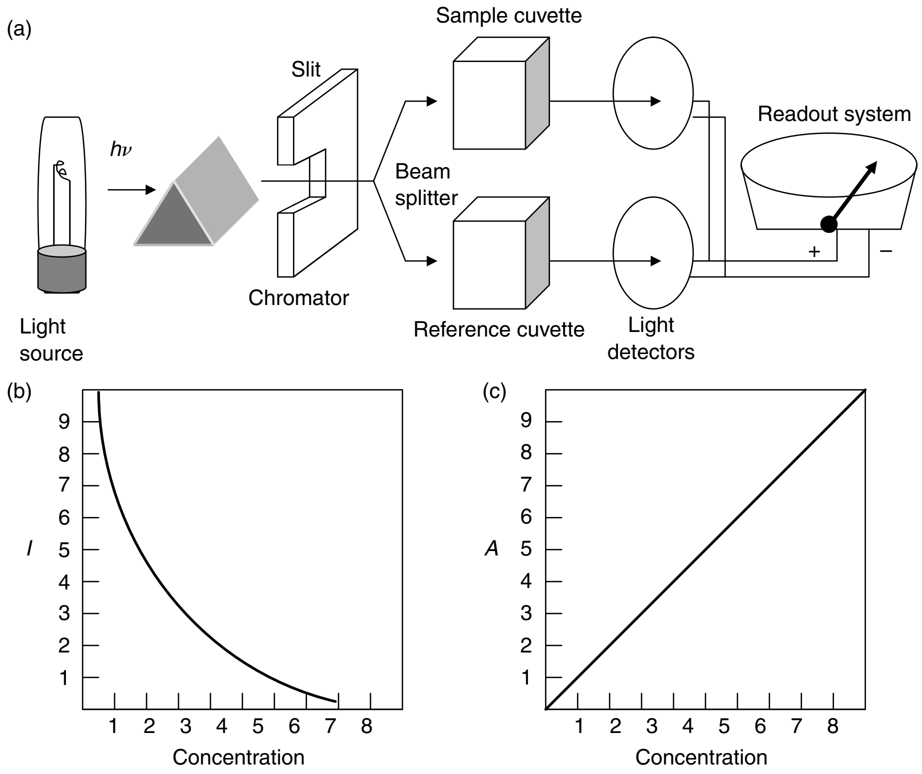


FIGURE 51.1 Principle of spectrophotometry: monochromated light is split into two beams and passed through a sample cuvette as well as a reference solution. Intensities of the transmitted light beams are compared to determine the concentration of the analyte in the sample. Lower two graphs show the percent transmission of light and absorption as a function of the concentration of the analyte of interest in a given solution.

can be calculated as follows:

$$\frac{A_S}{A_R} = \frac{C_S}{C_R} \Rightarrow C_S = \frac{A_S}{A_R} C_R = \frac{\log(I_R)}{\log(I_S)} C_R$$

where I_S and I_R are the intensity of the light transmitted through the cuvettes and detected by the photo-sensors.

A common use of spectrophotometry in clinical medicine is for the measurement of hemoglobin. Hemoglobin is made of heme, an iron compound, and globin, a protein. The iron gives blood its red color and the hemoglobin tests make use of this red color. A chemical is added to a sample of blood to make the red blood cells burst and to release the hemoglobin into the surrounding fluid, coloring it clear red. By measuring this color change using a spectrophotometer, and using the above equations, the concentration of hemoglobin in the blood can be determined [4,5].

For substances that are not colored, one can monitor the absorption at wavelengths that are outside of the visible spectrum, such as infrared and ultraviolet. Additionally, fluorescence spectroscopy can also be utilized.

51.2.2 Immunoassays

When the concentration of the analyte in the biological solution is too low for detection using spectrophotometry, more sensitive methods such as immunoassays are used for the measurement. Immunoassays utilize antibodies developed against the analyte of interest. Since the antigen and the antibody have a very specific interaction and has very high affinity toward each other, the resulting detection system also has a very high sensitivity. A specific example, the enzyme linked immunosorbent assay (ELISA), will be described here.

First, antibodies against the protein to be measured are developed in a host animal. For example, protein can be the human cardiac troponin-T (h-cT), which is a marker of myocardial infarction. Purified h-cT is injected into a rabbit to raise IgG molecules against h-cT, and these antibodies can either be recovered from the blood or produced recombinantly in a bioprocessor. A secondary antibody is also needed, which reacts with the first antibody and provides a colored solution. For example, this secondary antibody can be the goat antirabbit IgG antibody, which is tagged with a coloring compound or a fluorescent molecule. This second antibody can be used for any ELISA test that utilizes rabbit IgGs, regardless of the analyte, and such secondary antibodies are usually available commercially [6].

In the first step, a solution containing the protein of interest is placed on a substrate forming the sensor, such as a polystyrene dish. Then, the first antibody is added to the solution and allowed to react with the analyte. Unbound antibody is washed away and the second antibody is added. After a sufficient time is allowed for the second antibody to react with the first one, a second wash is performed to remove the unbound second antibody. Now the remaining solution contains the complexes formed by the protein of interest as well as the first and the second antibodies. Therefore, the color or the fluorescence produced is a function of the protein concentration. Figure 51.2 shows the steps used in an ELISA process. ELISAs are used for many clinical tests such as determining pregnancy or infectious disease tests such as detecting HIV. Its high sensitivity is due to the extremely high specificity of the antigen–antibody interaction [7,8].

51.2.3 Mass Spectrometry

Sometimes a test for more than one protein is needed and mass spectrometry is the method of choice for that purpose. A good example for this would be the use of tandem mass spectrometry to screen neonates for metabolic disorders such as amino acidemias (e.g., phenylketonuria — PKU), organic acidemias (e.g., propionic acidemia — PPA), and fatty acid oxidation disorders (e.g., Medium-chain acyl-CoA Dehydrogenase deficiency — MCAD) [9]. Although the price of this capital equipment could be high, costs of using it as a sensor is quite low (usually <U.S. $50.00 to screen for more than 20 metabolic disorders), and many states in the United States provide the service to newborns during the first week of life.

A mass spectrometer can be considered as a giant sensor, which measures the mass/charge (m/z) ratio as well as the relative abundance of multiple molecules in a given sample. Mass spectrometers consist of three main components: an ionization source, a physical separation environment, and a detector. Ionization of the molecules in the sample can be done by the deposition of energy from a laser source to remove an electron, a technique known as laser desorption ionization, which is depicted on the left side of Figure 51.3. Alternatively, it is possible to ionize molecules in a fluid flow by applying an electrical voltage to cause them to charge and subsequently spray, a technique known as electrospray ionization.

Following the ionization step, the molecules are sent into a physical separation chamber, where they are separated based on their m/z ratio. This can be achieved by first accelerating them in an electric field to give them a speed of

$$v = \sqrt{\frac{2Uz}{m}}$$

where U is the accelerating potential, z and m are the charge and the mass of the molecule respectively.

Since the velocity is a function of the mass, a determination of the mass of the molecules becomes possible in the physical separation environment. One option is to measure the time of flight in a flight tube, as shown in the middle section of Figure 51.3. This technique is known as the time-of-flight

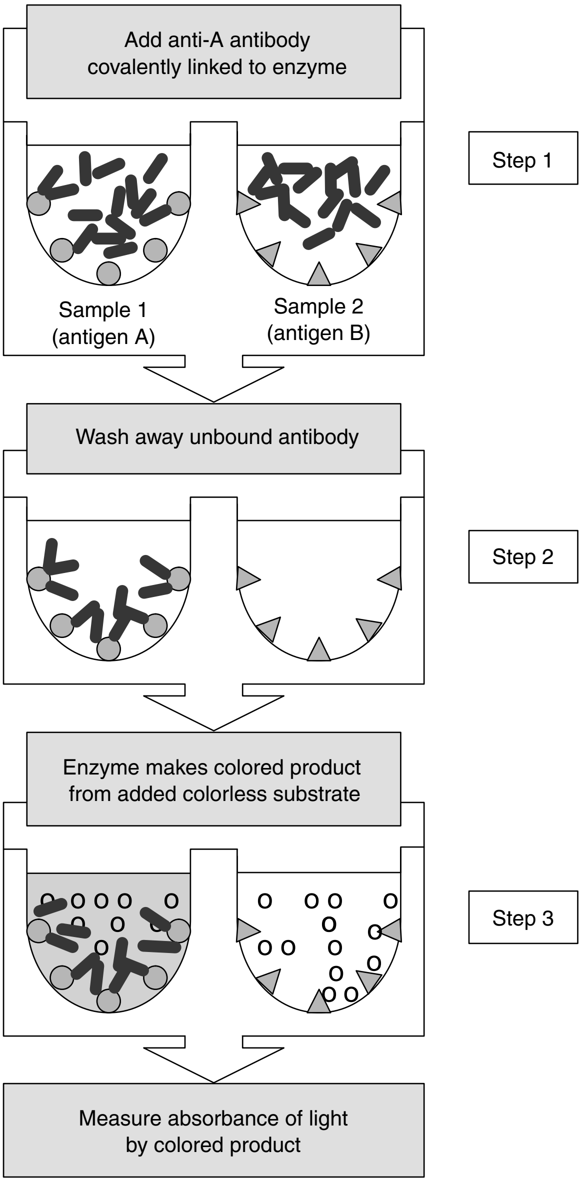

Add anti-A antibody
covalently linked to enzyme
Step 1
Sample 1
(antigen A)
Sample 2
(antigen B)
Wash away unbound antibody
Step 2
Enzyme makes colored product
from added colorless substrate
Step 3
Measure absorbance of light
by colored product

FIGURE 51.2 Steps of ELISA process (1) Specimen containing the target molecule antigen A, shown as round circles, is exposed to the primary antibody, shown as rectangular shapes. (2) Unbound antibody is washed, leaving the plate on the right hand side with no primary antibodies. (3) Secondary antibody depicted with oval shapes is added to the wells. (4) Unbound secondary antibody is also washed away, leaving the primary-secondary antibody complex along with the target molecule in the wells (step not shown). Test on the left plate would give a positive response to this test.

measurement, and makes use of the fact that larger molecules will fly slowly and require more time to cross the flight tube. Molecular mass can be calculated as

$$m = 2\frac{T^2}{L^2}Uz$$

where T is the flight time and L is the length of the flight tube.

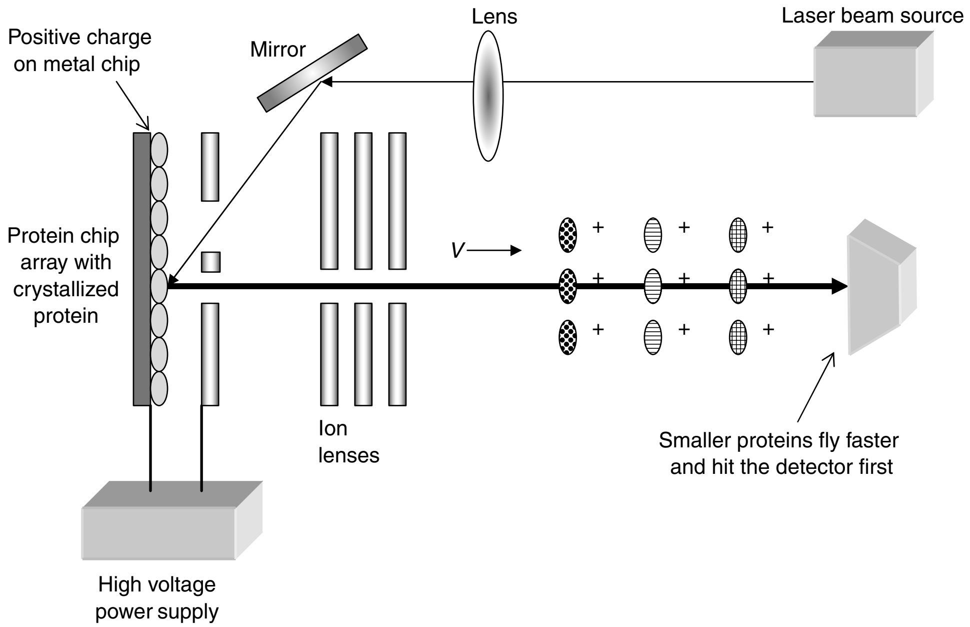

Lens
Laser beam source
Positive charge on metal chip
Mirror
Protein chip array with crystallized protein
V
+
Ion lenses
Smaller proteins fly faster and hit the detector first
High voltage power supply

FIGURE 51.3 Mass spectrophotometry: proteins are released from the chip surface by the application of the laser pulse. A secondary pulse is used to charge the proteins, which are later accelerated in an electric field. A time-of-flight measurement can be used to determine the mass of the protein.

Alternatively, one can apply a fixed magnetic field, perpendicular to the flight path of the ions, and cause a deviation in the flight path of the moving ions to separate them.

Although the mass spectrometer can separate the molecules based on their molecular weight, additional analysis is needed to separate molecules with identical or similar masses. This can be achieved by tandem mass spectrometry, where the ions coming from the first spectrometer enter into an argon collision chamber, and the resulting molecular fragments are catalogued by a secondary mass spectrometer. By studying the fragments, composition of the original mixture and the relative amount of the ions in the solution can be calculated [10–12].

51.2.4 Electrophoresis

Electrophoresis can be used to obtain a rapid separation of the proteins. A typical apparatus used for this purpose is shown in Figure 51.4. Proteins are first exposed to negatively charged sodium dodecyl sulfate, or SDS, which binds to the hydrophobic regions of the proteins, and causes them to unfold. The mixtures are placed into the wells on top of vertically placed gel slabs, as shown in Figure 51.4. Application of the electric potential, usually on the order of hundreds of volts, across the gel creates a force to move the protein molecules downward. Mobility of the proteins in the gel is given by

$$\mu = \frac{Q}{6\pi r\eta}$$

where μ is the electrophoretic mobility (cm/sec), Q, the ionic charge (due to SDS), r, the radius of the protein, and η is the viscosity of the cellulose acetate or polyacrylamide gel.

Therefore, the movement of the ions within the gel is directly proportional to the applied voltage, and inversely proportional to the size of the protein. Gel electrophoresis is run for a fixed amount of time, allowing the small proteins to migrate further than the larger ones, resulting in their separation based on

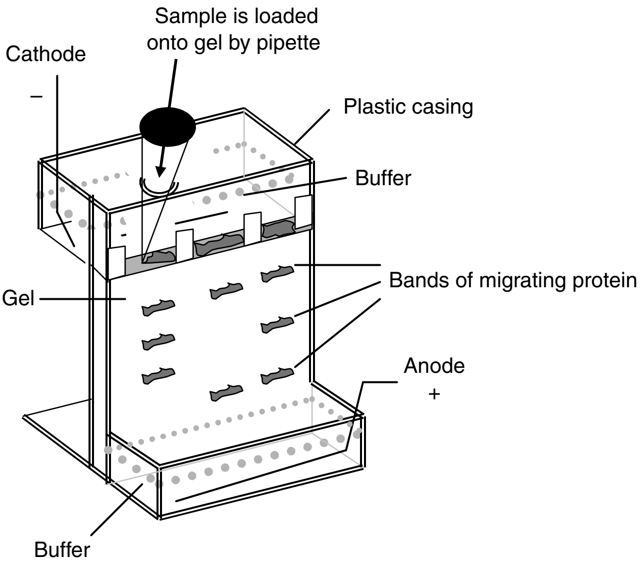


FIGURE 51.4 Gel electrophoresis: proteins loaded on the slab of gel are separated based on their size as they migrate under a constant electric field. Smaller proteins move faster and travel further than the larger ones.

their size. Coomassie blue or silver staining can be used to detect the bands across the gel to confirm the presence of proteins with known molecular masses.

Unlike mass spectroscopy, gel electrophoresis does not provide a quantitative value for the amount of given protein. However, it provides a low cost and relatively rapid method for the analysis of multiple proteins in a specimen, especially when implemented as a capillary electrophoresis system. Therefore, it has been used for the separation of enzymes (e.g., creatinine phosphokinase), mucopolysaccharides, plasma, serum, cerebrospinal fluid, urine, and other bodily fluids [13]. It is also used for quality control applications for the manufacturing of biological compounds to verify the purity or to examine the manufacturing yield [14].

51.2.5 Chromatography

Chromatography is also a simple technique that has applications in the toxicology and serum drug level measurements. In paper chromatography, a solution containing the analytes wicks up an absorbent paper for a period of time, and separation is achieved by the relative position of the analytes while the analyte and the solvent move up in the paper. A more precise technique known as column chromatography uses affinity columns, where the sample is applied to the top of the column and the fractionated molecules are eluted and collected at the bottom of the column (Figure 51.5). Since the smaller molecules with lower affinity to the column material will come out sooner than the larger molecules and ones with the higher affinity to the column, separation is achieved as a function of time. Further analysis can be done in the fractionated samples if desired.

51.3 Sensors for Measuring Nucleic Acids

Nucleic acids present in the body exist in the form of DNA and RNA. Determination of DNA sequences would allow the clinicians to determine the presence to congenital or genetically inherited diseases. On the other hand, measurement of RNA levels would indicate if gene is turned on or off. Discussions in this section focuses on the basics of few of the tools used in practice beginning with some enabling technologies.

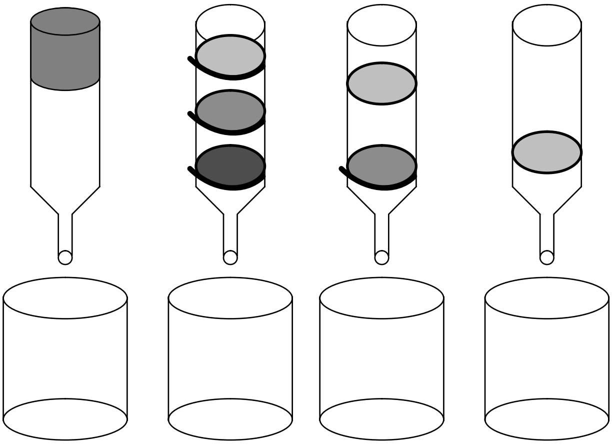


FIGURE 51.5 Affinity chromatography: a chemical column with relatively high affinity to proteins is loaded with a mixture of proteins, and allowed to run downward with the aid of gravity. Eluted fractions are collected in different tubes, each of which would contain different group of proteins.

TABLE 51.2 Genetic Code

		2nd base in codon					
	U	C	A	G			
1st base in codon	U	Phe	Ser	Tyr	Cys	**U**	3rd base in codon
		Phe	Ser	Tyr	Cys	**C**	
		Leu	Ser	**STOP**	**STOP**	**A**	
		Leu	Ser	**STOP**	Trp	**G**	
	C	Leu	Pro	His	Arg	**U**	
		Leu	Pro	His	Arg	**C**	
		Leu	Pro	Gln	Arg	**A**	
		Leu	Pro	Gln	Arg	**G**	
	A	Ile	Thr	Asn	Ser	**U**	
		Ile	Thr	Asn	Ser	**C**	
		Ile	Thr	Lys	Arg	**A**	
		Met	Thr	Lys	Arg	**G**	
	G	Val	Ala	Asp	Gly	**U**	
		Val	Ala	Asp	Gly	**C**	
		Val	Ala	Glu	Gly	**A**	
		Val	Ala	Glu	Gly	**G**	

Amino acids coded by three nucleic acid bases are shown as the entries of the table.

51.3.1 Enabling Technologies — DNA Extraction and Amplification

Cells in the human body contain the genetic material, DNA, which consists of a very long series of nucleic acids. Three nucleic acids in a row in the genome is called a codon, which determines the amino acid to be used when synthesizing a protein. This genetic code is shown in Table 51.2. Table 51.2 has $4^3 = 64$ entries, but since there are only 20 amino acids, many of the codons do code the same amino acid, a fact known as the redundancy of the genetic code. Therefore, a change in the genetic sequence may or may not cause a change in the protein synthesis. If the variation in the genetic sequence causes a change in the amino acid sequence altering the function of the protein, then an altered phenotype may emerge. Although the results of some of these changes are benign, such as different hair and eye colors, others are

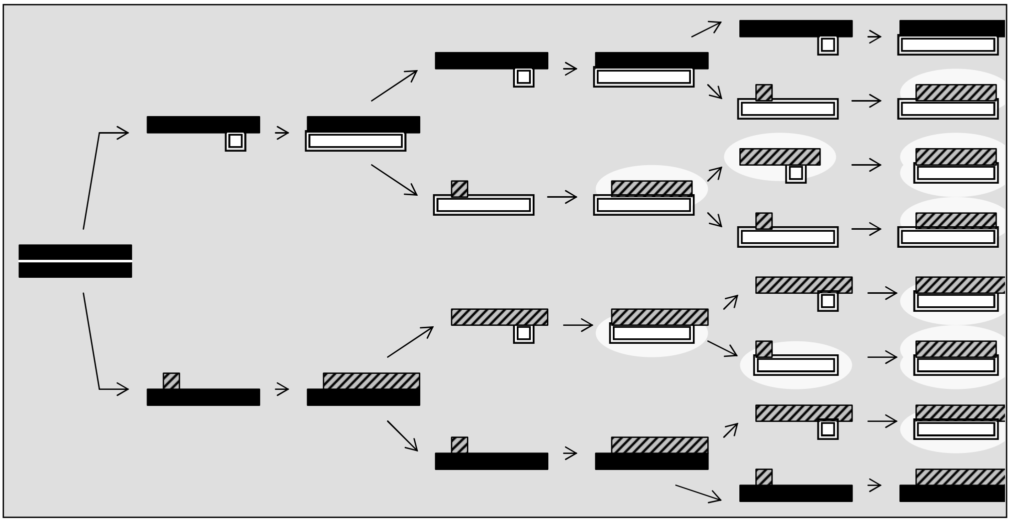


FIGURE 51.6 Polymerase chain reaction: double stranded DNA is first heated to denature the bonds between the two strands. In the second step, primers are allowed to attach to their complementary strands. In the third step, double stranded DNA is formed by the enzyme DNA Polymerase. Process is repeated many times, doubling the amount of DNA at each step.

not, such as increased susceptibility to various diseases. This genetic information can be read from the genes of interest. However, before that can be done, the DNA must be extracted and amplified.

Extraction of DNA from biological samples can be accomplished by precipitation or affinity methods [15]. Amplification of the amount of DNA is also needed before any sequence detection can be done. This can be done by a method known as polymerase chain reaction or PCR in short. This process is depicted in Figure 51.6. Briefly, original double stranded DNA molecules, shown as black rectangles, are heated to more than 90°C for separation. Afterward, DNA primers, shown as squares, as well as nucleic acids are added to the solution to initiate the DNA synthesis, forming two pairs of double stranded DNA at the end of the first cycle. The process is repeated for more than 30 times, doubling the amount of DNA at each step [16]. RNA can also be amplified using a similar process known as reverse transcription-polymerase chain reaction (RT-PCR) [17].

51.3.2 DNA/RNA probes

Gene chip arrays are being utilized as sensors to measure the level of gene expression to discover the genetic causes or to validate the presence of various disorders such as cancer. The most common form of these sensors is the RNA chips that consist of an array of probes. Each spot on the array contains multiple copies of a single genetic sequence immobilized to the sensor surface. These probe sequences are complementary to the RNA sequences being studied. Amplified RNA from the patient is labeled with a fluorescent dye, for example one that fluoresces red in this case. A second set of RNA, the reference RNA, with known concentration is labeled with another fluorescent dye, for example, green in this case. The RNA solutions are mixed and exposed to the sensor. Since sections of RNA from the patient and the reference are complementary to individual probe sequence, there would be a competitive binding during the exposure period (Figure 51.7). Following the incubation period, unbound RNA is removed, and the sensor array is exposed to a light source for the measurement of the fluorescence from each spot on the array. If a spot fluoresces red, the indication would be that the most of the RNA bound to this spot came from the patient, meaning that the patient is expressing this gene at high levels. On the other hand, a green fluorescence would indicate that the binding is from the reference RNA, and the patient is not expressing high levels of the gene. A yellow color would indicate a moderate gene expression, since some

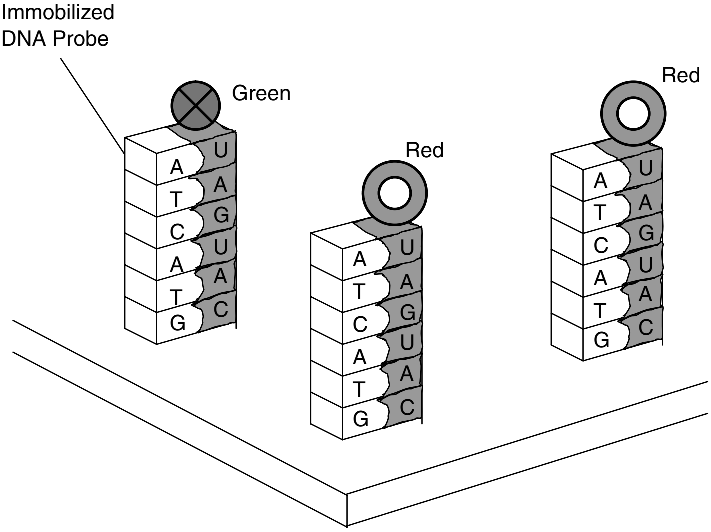


FIGURE 51.7 Gene array: reference RNA labeled with green fluorescence molecules compete with the RNA from the patient labeled with red fluorescent molecules to bind to the immobilized probes on the gene array.

of each type of RNA molecule must bind to the probe to produce the mixed response. The advantage of these sensors is the same as any other sensor array, which is the ability to probe a large number of genes simultaneously [18].

51.3.3 SNP Detection

Single nucleotide polymorphism, or SNP, occurs when only one nucleotide in the genetic sequence is altered. It could be a mutation, and it could be inherited from a parent. Reading a single nucleotide from the genome that had been substituted for another one might stop the reading process of that gene or cause a different protein to be synthesized. Thus detecting a SNP could be important in diagnosing genetically related diseases or a patient's tendency toward being more susceptible to this type of disease.

There are many methods developed for the detection of a SNP, and one of them will be described later, which is illustrated in Figure 51.8. In the first step, a primer consisting of 20 to 50 nucleic acids having a sequence complementary to the gene sequence adjacent to the SNP is synthesized and allowed to anneal to a single strand of the DNA from the patient. In the second step, terminal nucleic acids (A, T, C, and G) with different fluorescent tags are added allowing the double stand to grow by only one base pair. As the third step, the solution is exposed to a laser light for detecting the fluorescence to read the nucleic acid at the SNP location. Since the patient sample will contain two copies for each gene, one from each parent, the sensor might detect one or two nucleic acids for each SNP site [19].

Detection of SNPs can be used as a clinical test to diagnose various diseases. Some examples of these are SNPs for BRAC-1 gene to detect patients with high susceptibility to a type of breast cancer, and long-quantitative trait (QT) genes, which make patients prone to fatal cardiac arrhythmias [20,21].

51.4 Sensors for Cellular Processes

Flow cytometry is used to separate populations of cells in a mixture from one another by means of fluorescently labeled antibodies and DNA specific dyes. Antibodies used are usually against the molecules expressed on the cell surface, such as cell-surface receptors. The labeled cells are diluted in a solution and passed through a nozzle in a single-cell stream while being illuminated by a laser beam. Fluorescence

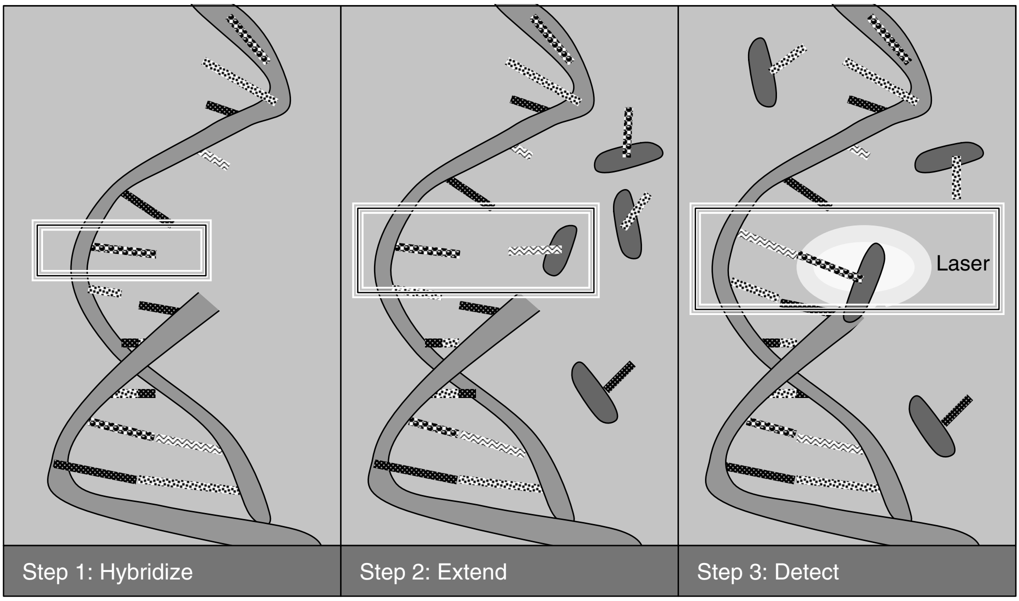


FIGURE 51.8 SNP detection by primer extension method: first the amplified DNA is hybridized to the primers that are specific to the genetic region of interest. Second, a labeled terminating base at the target SNP site extends the primer. Finally, the extended primer is read by fluorescence.

is detected by photomultiplier tubes (PMTs) with optical filters to determine the emission wavelength, which in turn helps to identify the surface receptors (Figure 51.9). Fluorescence data is used to measure the relative proportions of various cell types in the mixture, and to derive the diagnostic information [22].

Flow cytometry is commonly used in the clinical diagnostics, for immunophenotyping leukemia, counting stem cells for optimizing autologous transplants for the treatment of leukemia, counting CD4+/CD8+ lymphocytes in HIV-infected patients to determine the progression of the disease, and the analysis of stained DNA from solid tumors [23].

51.5 Personalized Medicine

Today the personalization of the treatment for the needs of individual patients is done by health care providers. In some cases, clinicians can neither predict reliably the best treatment pathway for a patient, nor anticipate the optimal drug regimen. However, it would be possible to improve the treatment and tailor the therapy to the specific needs of the patients if their genetic information is known. For example, some drugs are known to cause cardiac arrhythmias in patients with certain genotypes and should be avoided. It might be possible that in the future, patients coming to hospitals might be asked to bring their genome cards along with their insurance cards. Knowledge of the genome of individual patients would not only predict their medical vulnerabilities, but also help with the selection of their treatment [24].

Some of these studies have already begun. For example, in the United States, the Food and Drug Administration is already encouraging the pharmaceutical industry to submit the results of genomic tests when seeking approval for new drugs [25]. This new field of research is now being recognized as pharmacogenetics.

Another early application of personalized medicine is the use of microphysiometry, a device that measures the extracellular acidification rate of cells, to establish their sensitivity to chemotherapy [26].

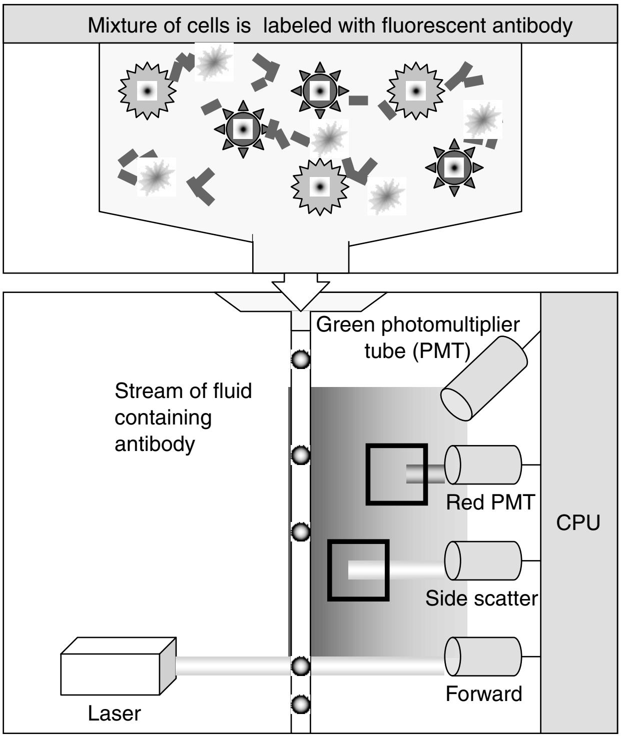

Mixture of cells is labeled with fluorescent antibody
Green photomultiplier tube (PMT)
Stream of fluid containing antibody
Red PMT
CPU
Side scatter
Laser
Forward

FIGURE 51.9 Flow cytometry: fluorescently labeled cells are passed in front of light detectors (labeled as PMTs in the figure) to detect their labeling color, which indicates the phenotype of the cell.

This sensor measures the chemosensitivity by comparing the acidification rate of cells treated with cytostatic agents, such as anticancer drugs, to that of nontreated cells, before a drug is prescribed to the patient.

51.6 Final Comments

As the aging of the population in the Western world, and the increase in the population of the World in general continues, the need for diagnostic procedures will also increase. Due to the need for cost containment, the emphasis will continue to shift from therapeutics to early diagnosis for prevention. Both of these factors will increase the need for diagnostic procedures and additional diagnostic technologies. While the basic methodology for the traditional diagnostics is becoming well established, the techniques needed for personalized medicine are still being developed, and the biomedical engineers will be able to participate in both the research and implementation aspects of these very important areas.

References

[1] Simonsen, M., Nucleic Acid Testing, Proteomics to drive future of diagnostics, *BBI Newsletter*, 27, 221–228, 2004.

[2] Skoog, D.A. and Leary, J.J., *Principles of Instrumental Analysis*, 4th ed., Saunders, Orlando, FL, 1992.

[3] Kellner, R., Mermet, J.-M., Otto, M., and Widmer, H.M., *Analytical Chemistry*, Wiley, Weinheim, GR, 1998.

[4] Kaplan, A., Jack, R., Opheim, K.E., Toivola, B., and Lyon, A.W., *Clinical Chemistry*, 4th ed., Williams & Wilkins, Malvern, PA, 1995.

[5] Burtis, C.A. and Ashwood, E.R., *Tietz Fundamentals of Clinical Chemistry*, 5th ed., W.N. Saunders, Philadelphia, PA, 2001.

[6] Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J.D., *Molecular Biology of the Cell*, 3rd ed., Garland, New York, NY, 1994,

[7] Liu, S., Boyer-Chatenet, L., Lu, H., and Jiang, S., Rapid and automated fluorescence-linked immunosorbent assay for high-throughput screening of HIV-1 fusion inhibitors targeting gp41. *J. Biomol. Screen.* 8, 685–693, 2003.

[8] Bandi, Z.L., Schoen, I., and DeLara, M., Enzyme-linked immunosorbent urine pregnancy tests, *Am. J. Clin. Pathol.*, 87, 236–242, 1987.

[9] Schulze, A., Lindner, M., Kohlmuller, D., Olgemoller, K., Mayatepek, E., and Hoffmann, G.F., Expanded newborn screening for inborn errors of metabolism by electrospray ionization-tandem mass spectrometry: results, outcome, and implications, *Pediatrics*, 111, 1399–1406, 2003.

[10] Liebler, D.C., *Introduction to Proteomics*, Humana Press, Totowa, NJ, 2002.

[11] Pennington, S.R. and Dunn, M.J., *Proteomics*, Springer-Verlag, New York, NY, 2001.

[12] Kambhampati, D., *Protein Microarray Technology*, Wiley, Heppenheim, GR, 2004.

[13] Chen, F.T., Liu, C.M., Hsieh, Y.Z., and Sternberg, J.C., Capillary electrophoresis — a new clinical tool, *Clin. Chem.*, 37, 14–19, 1991.

[14] Reilly, R.M., Scollard, D.A., Wang, J., Monda, H., Chen, P., Henderson, L.A., Bowen, B.M., and Vallis, K.A., A kit formulated under good manufacturing practices for labeling human epidermal growth factor with 111In for radiotherapeutic applications, *J. Nucl. Med.*, 45, 701–708, 2004.

[15] Bowtell, D. and Sambrook, J., *DNA Microarrays: A Molecular Cloning Manual*, Cold Spring Harbor Press, Cold Spring Harbor, NY, 2003.

[16] Malacinski, G.M., *Essentials of Molecular Biology*, 4th ed., Jones and Bartlett, Sudbury, MA, 2003.

[17] Stahlberg, A., Hakansson, J., Xian, X., Semb, H., and Kubista, M., Properties of the reverse transcription reaction in mRNA quantification, *Clin. Chem.*, 50, 509–515, 2004.

[18] Blalock, E., *A Beginner's Guide to Microarrays*, Kluwer, Dordrecht, NL, 2003.

[19] Kwok, P.-Y., *Single Nucleotide Polymorphisms: Methods and Protocols*, Humana, Totowa, NJ, 2003.

[20] Burke, W., Genomic medicine: genetic testing, *N. Engl. J. Med.*, 347, 1867–1875, 2002.

[21] Haack, B., Kupka, S., Ebauer, M., Siemiatkowska, A., Pfister, M., Kwiatkowska, J., Erecinski, J., Limon, J., Ochman, K., and Blin, N., Analysis of candidate genes for genotypic diagnosis in the long QT syndrome, *J. Appl. Genet.*, 45, 375–381, 2004.

[22] Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., Baltimore, D., and Barnell, J., *Molecular Cell Biology*, 4th ed., W.H. Freeman, New York, NY, 2000.

[23] Rose, N.R., Friedman, H., and Fahey, J.L., *Manual of Clinical Laboratory Immunology*, 3rd ed., American Society for Microbiology, Washington, D.C., 1986.

[24] Brown, S.M., *Essentials of Medical Genomics*, Wiley, Hoboken, NJ, 2003.

[25] Guidance for Industry: Pharmacogenomic Data Submissions, U.S. Department of Health and Human Services, Food and Drug Administration, November 2003.

[26] Waldenmaier, D.S., Babarina, A., and Kischkel, F.C., Rapid *in vitro* chemosensitivity analysis of human colon tumor cell lines, *Toxicol. Appl. Pharmacol.*, 192, 237–245, 2003.

VI

Medical Instruments and Devices

Wolf W. von Maltzahn
Rensselaer Polytechnic Institute

NOT TOO LONG AGO, the term *medical instrument* stood for simple hand-held instruments used by physicians for observing patients, examining organs, making simple measurements, or administering medication. These small instruments, such as stethoscopes, thermometers, tongue depressors, and a few surgical tools, typically fit into a physician's hand bag. Today's medical instruments are considerably more complicated and diverse, primarily because they incorporate electronic systems for sensing, transducing, manipulating, storing, and displaying data or information. Furthermore, medical specialists today request detailed and accurate measurements of a vast number of physiologic parameters for diagnosing illnesses and prescribe complicated procedures for treating these. As a result, the number

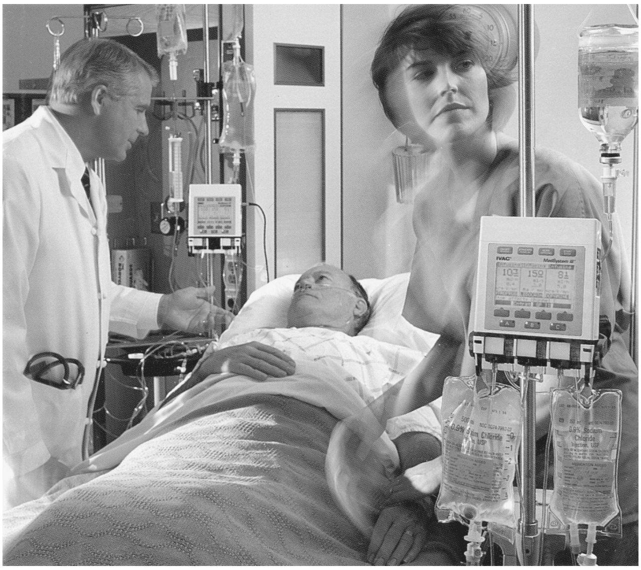


FIGURE VI.1 A typical IV infusion system.

of medical instruments and devices has grown from a few hundred a generation ago to more than 10,000 today, and the complexity of these instruments has grown at the same pace. The description of all these instruments and devices would fill an entire handbook by itself; however, due to the limited space assigned to this topic, only a selected number are described.

While medical instruments acquire and process information and data for monitoring patients and diagnosing illnesses, medical devices use electrical, mechanical, chemical, or radiation energy for achieving a desired therapeutic purpose, maintaining physiologic functions, or assisting a patient's healing process. To mention only a few functions, medical devices pump blood, remove metabolic waste products, destroy kidney stones, infuse fluids and drugs, stimulate muscles and nerves, cut tissue, administer anesthesia, alleviate pain, restore function, or warm tissue. Because of their complexity, medical devices are used mostly in hospitals and medical centers by trained personnel, but some also can be found in private homes operated by patients themselves or their caregivers.

This section on medical instruments and devices neither replaces a textbook on this subject nor presents the material in a typical textbook manner. The authors assume the reader to be interested in but not knowledgeable on the subject. Therefore, each chapter begins with a short introduction to the subject material, followed by a brief description of current practices and principles, and ends with recent trends and developments. Whenever appropriate, equations, diagrams, and pictures amplify and illustrate the topic, while tables summarize facts and data. The short reference section at the end of each chapter points toward further resource materials, including books, journal articles, patents, and company brochures.

The chapters in the first half of this section cover the more traditional topics of bioinstrumentation, such as biopotential amplifiers and noninvasive blood pressure, blood flow, and respiration monitors, while those of the second half focus more on recently developed instruments and devices such as pulse oximeters or home-care monitoring devices. Some of this latter material is new or hard to find elsewhere. A few traditional bioinstrumentation or electroencephalography have been omitted entirely because most textbooks on this subject give excellent introductions and reviews. Transducers, biosensors, and electrodes are covered in other sections of this *handbook*. Thus, this section provides an overview, albeit an incomplete one, of recent developments in the field of medical instruments and devices.

52

Biopotential Amplifiers

Joachim H. Nagel
University of Stuttgart

Biosignals are recorded as potentials, voltages, and electrical field strengths generated by nerves and muscles. The measurements involve voltages at very low levels, typically ranging between 1 μV and 100 mV, with high source impedances and superimposed high level interference signals and noise. The signals need to be amplified to make them compatible with devices such as displays, recorders, or A/D converters for computerized equipment. Amplifiers adequate to measure these signals have to satisfy very specific requirements. They have to provide amplification selective to the physiological signal, reject superimposed noise and interference signals, and guarantee protection from damages through voltage and current surges for both patient and electronic equipment. Amplifiers featuring these specifications are known as *biopotential amplifiers*. Basic requirements and features, as well as some specialized systems, will be presented.

52.1 Basic Amplifier Requirements

The basic requirements that a biopotential amplifier has to satisfy are:

- The physiological process to be monitored should not be influenced in any way by the amplifier
- The measured signal should not be distorted
- The amplifier should provide the best possible separation of signal and interferences
- The amplifier has to offer protection of the patient from any hazard of electrical shock
- The amplifier itself has to be protected against damages that might result from high input voltages as they occur during the application of defibrillators or electrosurgical instrumentation

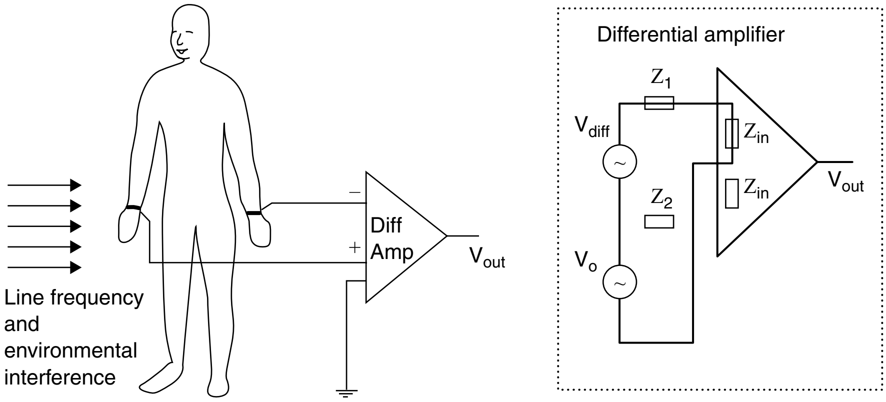


FIGURE 52.1 Typical configuration for the measurement of biopotentials. The biological signal appears between the two measuring electrodes at the right and left arm of the patient and is fed to the inverting and the noninverting inputs of the differential amplifier.

A typical configuration for the measurement of biopotentials is shown in Figure 52.1. Three electrodes, two of them picking up the biological signal and the third providing the reference potential, connect the subject to the amplifier. The input signal to the amplifier consists of five components (1) the desired biopotential, (2) undesired biopotentials, (3) a power line interference signal of 60 Hz (50 Hz in some countries) and its harmonics, (4) interference signals generated by the tissue/electrode interface, and (5) noise. Proper design of the amplifier provides rejection of a large portion of the signal interferences. The main task of the differential amplifier as shown in Figure 52.1 is to reject the line frequency interference that is electrostatically or magnetically coupled into the subject. The desired biopotential appears as a voltage between the two input terminals of the differential amplifier and is referred to as the *differential signal.* The line frequency interference signal shows only very small differences in amplitude and phase between the two measuring electrodes, causing approximately the same potential at both inputs, and thus appears only between the inputs and ground and is called the *common mode signal.* Strong rejection of the common mode signal is one of the most important characteristics of a good biopotential amplifier.

The **common mode rejection ratio (CMRR)** of an amplifier is defined as the ratio of the differential mode gain over the common mode gain. As seen in Figure 52.1, the rejection of the common mode signal in a biopotential amplifier is both a function of the amplifier CMRR and the source impedances Z_1 and Z_2. For the ideal biopotential amplifier with $Z_1 = Z_2$ and infinite CMRR of the differential amplifier, the output voltage is the pure biological signal amplified by G_D, the differential mode gain: $V_{out} = G_D \cdot V_{biol}$. With finite CMRR, the common mode signal is not completely rejected, adding the interference term $G_D \cdot V_c/\text{CMRR}$ to the output signal. Even in the case of an ideal differential amplifier with infinite CMRR, the common mode signal will not completely disappear unless the source impedances are equal. The common mode signal V_c causes currents to flow through Z_1 and Z_2. The related voltage drops show a difference if the source impedances are unequal, thus generating a differential signal at the amplifier input which, of course, is not rejected by the differential amplifier. With amplifier gain G_D and input impedance Z_{in}, the output voltage of the amplifier is:

$$V_{out} = G_D V_{biol} + \frac{G_D V_c}{\text{CMRR}} + G_D V_c \left(1 - \frac{Z_{in}}{Z_{in} + Z_1 - Z_2}\right) \tag{52.1}$$

The output of a real biopotential amplifier will always consist of the desired output component due to a differential biosignal, an undesired component due to incomplete rejection of common mode interference signals as a function of CMRR, and an undesired component due to source impedance unbalance allowing a small proportion of a common mode signal to appear as a differential signal to the amplifier. Since

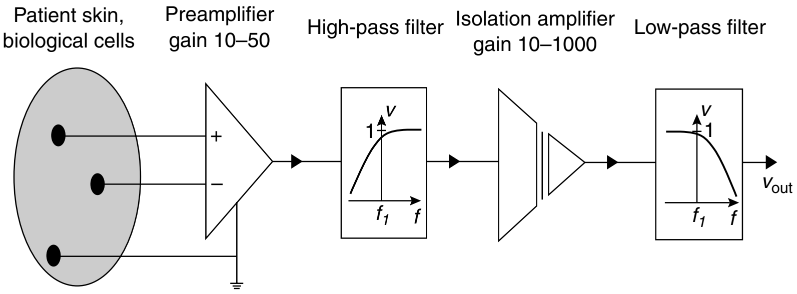


FIGURE 52.2 Schematic design of the main stages of a biopotential amplifier. Three electrodes connect the patient to a preamplifier stage. After removing dc and low-frequency interferences, the signal is connected to an output low-pass filter through an isolation stage which provides electrical safety to the patient, prevents ground loops, and reduces the influence of interference signals.

source impedance unbalances of 5,000 to 10,000 Ω, mainly caused by electrodes, are not uncommon, and sufficient rejection of line frequency interferences requires a minimum CMRR of 100 dB, the input impedance of the amplifier should be at least 10^9 Ω at 60 Hz to prevent source impedance unbalances from deteriorating the overall CMRR of the amplifier. State-of-the-art biopotential amplifiers provide a CMRR of 120 to 140 dB.

In order to provide optimum signal quality and adequate voltage level for further signal processing, the amplifier has to provide a gain of 100 to 50,000 and needs to maintain the best possible signal-to-noise ratio. The presence of high level interference signals not only deteriorates the quality of the physiological signals, but also restricts the design of the biopotential amplifier. Electrode half-cell potentials, for example, limit the gain factor of the first amplifier stage since their amplitude can be several orders of magnitude larger than the amplitude of the physiological signal. To prevent the amplifier from going into saturation, this component has to be eliminated before the required gain can be provided for the physiological signal.

A typical design of the various stages of a biopotential amplifier is shown in Figure 52.2. The electrodes which provide the transition between the ionic flow of currents in biological tissue and the electronic flow of current in the amplifier, represent a complex electrochemical system that is described elsewhere in this handbook. The electrodes determine to a large extent the composition of the measured signal. The preamplifier represents the most critical part of the amplifier itself since it sets the stage for the quality of the biosignal. With proper design, the preamplifier can eliminate, or at least minimize, most of the signals interfering with the measurement of biopotentials.

In addition to electrode potentials and electromagnetic interferences, noise — generated by the amplifier and the connection between biological source and amplifier — has to be taken into account when designing the preamplifier. The total source resistance R_s, including the resistance of the biological source and all transition resistances between signal source and amplifier input, causes thermal voltage noise with a root mean square (rms) value of:

$$E_{\text{rms}} = \sqrt{4kTR_sB} \quad \text{(volt)} \tag{52.2}$$

where k = Boltzmann constant, T = absolute temperature, R_s = resistance in Ω, and B = bandwidth in Hz.

Additionally, there is the inherent amplifier noise. It consists of two frequency-dependent components, the internal voltage noise source e_n and the voltage drop across the source resistance R_s caused by an internal current noise generator i_n. The total input noise for the amplifier with a bandwidth of $B = f_2 - f_1$ is calculated as the sum of its three independent components:

$$E_{\text{rms}}^2 = \int_{f_1}^{f_2} e_n^2 \mathrm{d}f + R_s^2 \int_{f_1}^{f_2} i_n^2 \mathrm{d}f + 4kTR_sB \tag{52.3}$$

High signal-to-noise ratios thus require the use of very low noise amplifiers and the limitation of bandwidth. Current technology offers differential amplifiers with voltage noise of less than 10 nV/$\sqrt{\text{Hz}}$ and current noise less than 1 pA/$\sqrt{\text{Hz}}$. Both parameters are frequency dependent and decrease approximately with the square root of frequency. The exact relationship depends on the technology of the amplifier input stage. Field effect transistor (FET) preamplifiers exhibit about five times the voltage noise density compared to bipolar transistors but a current noise density that is about 100 times smaller.

The purpose of the high pass and low pass filters in Figure 52.2 is to eliminate interference signals like electrode half-cell potentials and preamplifier offset potentials and to reduce the noise amplitude by the limitation of the amplifier bandwidth. Since the biosignal should not be distorted or attenuated, higher order sharp-cutting linear phase filters have to be used. Active Bessel filters are preferred filter types due to their smooth transfer function. Separation of biosignal and interference is in most cases incomplete due to the overlap of their spectra.

The isolation stage serves the galvanic decoupling of the patient from the measuring equipment and provides safety from electrical hazards. This stage also prevents galvanic currents from deteriorating the signal-to-noise ratio especially by preventing ground loops. Various principles can be used to realize the isolation stage. Analog isolation amplifiers use either transformer, optical, or capacitive couplers to transmit the signal through the isolation barrier. Digital isolation amplifiers use a voltage/frequency converter to digitize the signal before it is transmitted easily by optical or inductive couplers to the output frequency/voltage converter. The most important characteristics of an isolation amplifier are low leakage current, isolation impedance, isolation voltage (or mode) rejection (IMR), and maximum safe isolation voltage.

52.1.1 Interferences

The most critical point in the measurement of biopotentials is the contact between electrodes and biological tissue. Both the electrode offset potential and the electrode/tissue impedance are subject to changes due to relative movements of electrode and tissue. Thus, two interference signals are generated as motion artifacts: the changes of the electrode potential and motion-induced changes of the voltage drop caused by the input current of the preamplifier. These motion artifacts can be minimized by providing high input impedances for the preamplifier, usage of non-polarized electrodes with low half-cell potentials such as Ag/AgCl electrodes, and by reducing the source impedance by use of electrode gel. Motion artifacts, interferences from external electromagnetic fields, and noise can also be generated in the wires connecting electrodes and amplifier. Reduction of these interferences is achieved by using twisted pair cables, shielded wires, and *input guarding*.

Recording of biopotentials is often done in an environment that is equipped with many electrical systems which produce strong electrical and magnetic fields. In addition to 60 Hz power line frequency and some strong harmonics, high frequency electromagnetic fields are encountered. At power line frequency, the electric and magnetic components of the interfering fields can be considered separately. Electrical fields are caused by all conductors that are connected to power, even with no flow of current. A current is capacitively coupled into the body where it flows to the ground electrode. If an isolation amplifier is used without patient ground, the current is capacitively coupled to ground. In this case, the body potential floats with a voltage of up to 100 V towards ground. Minimizing interferences requires increasing the distance between power lines and the body, use of isolation amplifiers, separate grounding of the body at a location as far away from the measuring electrodes as possible, and use of shielded electrode cables.

The magnetic field components produce eddy currents in the body. The amplifier, the electrode cable, and the body form an induction loop that is subject to the generation of an interference signal. Minimizing this interference signal requires increasing the distance between the interference source and patient, twisting the connecting cables, shielding of the magnetic fields, and relocating the patient to a place and orientation that offers minimum interference signals. In many cases, an additional narrow band-rejection filter (notch filter) is implemented as an additional stage in the biopotential amplifier to provide sufficient suppression of line frequency interferences.

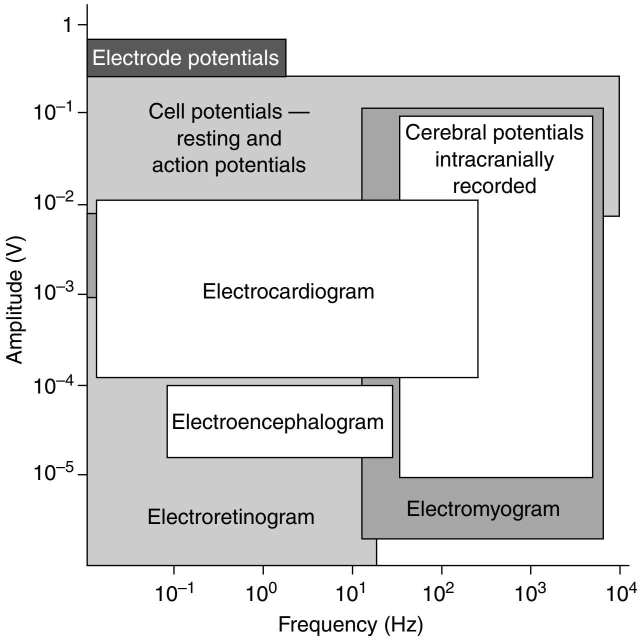


FIGURE 52.3 Amplitudes and spectral ranges of some important biosignals. The various biopotentials completely cover the area from 10^{-5} V to almost 1 V and from dc to 10 kHz.

In order to achieve optimum signal quality, the biopotential amplifier has to be adapted to the specific application. Based on the signal parameters, both appropriate bandwidth and gain factor are chosen. Figure 52.3 shows an overview of the most commonly measured biopotentials and specifies the normal ranges for amplitude and bandwidth.

A final requirement for biopotential amplifiers is the need for calibration. Since the amplitude of the biopotential often has to be determined very accurately, there must be provisions to easily determine the gain or the amplitude range referenced to the input of the amplifier. For this purpose, the gain of the amplifier must be well calibrated. In order to prevent difficulties with calibrations, some amplifiers that need to have adjustable gain use a number of fixed gain settings rather than providing a continuous gain control. Some amplifiers have a standard signal source of known amplitude built in that can be momentarily connected to the input by the push of a button to check the calibration at the output of the biopotential amplifier.

52.2 Special Circuits

52.2.1 Instrumentation Amplifier

An important stage of all biopotential amplifiers is the input preamplifier which substantially contributes to the overall quality of the system. The main tasks of the preamplifier are to sense the voltage between two measuring electrodes while rejecting the common mode signal, and minimizing the effect of electrode polarization overpotentials. Crucial to the performance of the preamplifier is the input impedance which should be as high as possible. Such a differential amplifier cannot be realized using a standard single **operational amplifier** (**op-amp**) design since this does not provide the necessary high input impedance. The general solution to the problem involves voltage followers, or noninverting amplifiers, to attain high input impedances. A possible realization is shown in Figure 52.4a. The main disadvantage of this circuit is that it requires high CMRR both in the followers and in the final op-amp. With the input buffers working at unity gain, all the common-mode rejection must be accomplished in the output amplifier,

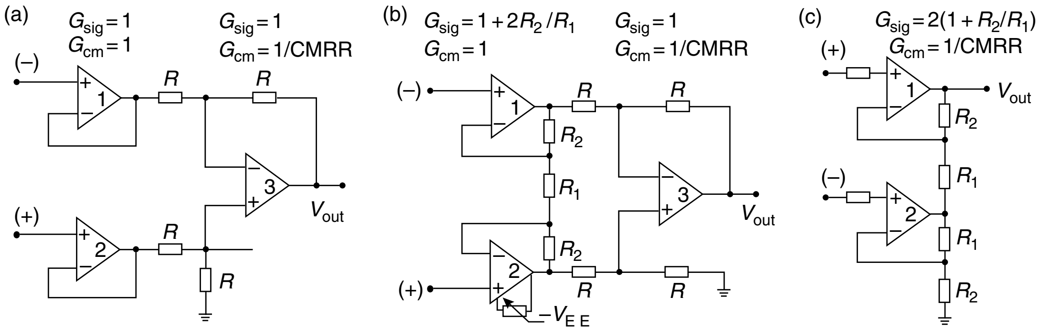


FIGURE 52.4 Circuit drawings for three different realizations of instrumentation amplifiers for biomedical applications. Voltage follower input stage (a), improved, amplifying input stage (b), and 2-op-amp version (c).

requiring very precise resistor matching. Additionally, the noise of the final op-amp is added at a low signal level, decreasing the signal-to-noise ratio unnecessarily. The circuit in Figure 52.4b eliminates this disadvantage. It represents the standard instrumentation amplifier configuration. The two input op-amps provide high differential gain and unity common-mode gain without the requirement of close resistor matching. The differential output from the first stage represents a signal with substantial relative reduction of the common-mode signal and is used to drive a standard differential amplifier which further reduces the common-mode signal. CMRR of the output op-amp as well as resistor matching in its circuit are less critical than in the follower type instrumentation amplifier. Offset trimming for the whole circuit can be done at one of the input op-amps. Complete instrumentation amplifier integrated circuits based on this standard instrumentation amplifier configuration are available from several manufacturers. All components except R_1, which determines the gain of the amplifier, and the potentiometer for offset trimming are contained on the integrated circuit chip. Figure 52.4c shows another configuration that offers high input impedance with only two op-amps. For good CMRR, however, it requires precise resistor matching.

In applications where dc and very low frequency biopotentials are not to be measured, it would be desirable to block those signal components at the preamplifier inputs by simply adding a capacitor working as a passive high-pass filter. This would eliminate the electrode offset potentials and permit a higher gain factor for the preamplifier *and thus a higher CMRR*. A capacitor between electrodes and amplifier input would, however, result in charging effects from the input bias current. Due to the difficulty of precisely matching capacitors for the two input leads, they would also contribute to an increased source impedance unbalance and thus reduce CMRR. Avoiding the problem of charging effects by adding a resistor between the preamplifier inputs and ground as shown in Figure 52.5a also results in a decrease of CMRR due to the diminished and mismatched input impedance. A 1% mismatch for two 1-MΩ resistors can already create a −60 dB loss in CMRR. The loss in CMRR is much greater if the capacitors are mismatched, which cannot be prevented in real systems. Nevertheless, such realizations are used where the specific situation allows. In some applications, a further reduction of the amplifier to a two-electrode amplifier configuration would be convenient, even at the expense of some loss in the CMRR. Figure 52.6 shows a preamplifier design working with two electrodes and providing ac coupling as proposed by Pallás-Areny and Webster [1990].

A third alternative of eliminating dc and low frequencies in the first amplifier stage is a directly coupled quasi-high-pass amplifier design, which maintains the high CMRR of dc coupled high input impedance instrumentation amplifiers [Song et al., 1998]. In this design, the gain determining resistor R_1 (Figure 52.5a) is replaced by a first order high-pass filter consisting of R_1 and a series capacitor C_f. The signal gain of the amplifier is

$$G = 1 + \frac{2R_2}{R_1 + \frac{1}{j\omega C}} \tag{52.4}$$

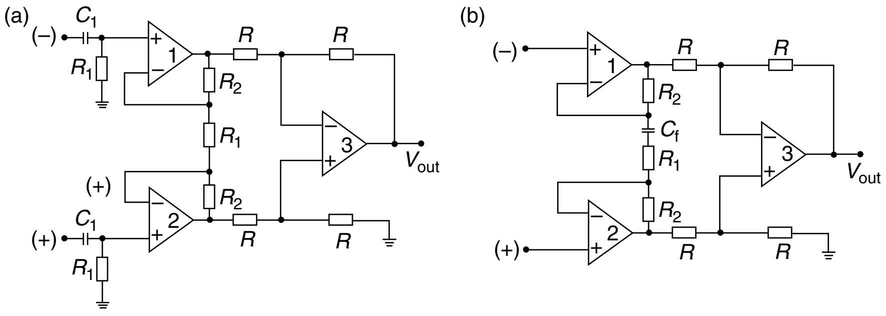


FIGURE 52.5 AC coupled instrumentation amplifier designs. The classical design using an RC high-pass filter at the inputs (a), and a high CMRR "quasi-high-pass" amplifier as proposed by Lu (b).

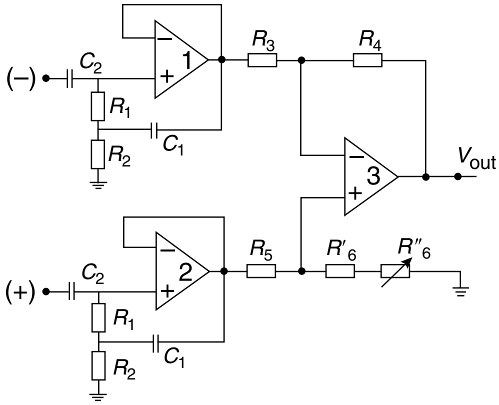


FIGURE 52.6 Composite instrumentation amplifier based on an ac-coupled first stage. The second stage is based on a one op-amp differential amplifier which can be replaced by an instrumentation amplifier.

Thus, dc gain is 1, while the high frequency gain remains at $G = 1 + 2R_2/R_1$. A realization using an off-the-shelf instrumentation amplifier (Burr-Brown INA 118) operates at low power (0.35 mA) with low offset voltage (11 μV typical) and low input bias current (1 nA typical), and offers a high CMRR of 118 dB at a gain of $G = 50$. The very high input impedance (10 GΩ) of the instrumentation amplifier renders it insensitive to fluctuations of the electrode impedance. Therefore, it is suitable for bioelectric measurements using pasteless electrodes applied to unprepared, that is, high impedance skin.

The preamplifier, often implemented as a separate device which is placed close to the electrodes or even directly attached to the electrodes, also acts as an impedance converter which allows the transmission of even weak signals to the remote monitoring unit. Due to the low output impedance of the preamplifier, the input impedance of the following amplifier stage can be low, and still the influence of interference signals coupled into the transmission lines is reduced.

52.2.2 Isolation Amplifier and Patient Safety

Isolation amplifiers can be used to break ground loops, eliminate source ground connections, and provide isolation protection to patient and electronic equipment. In a biopotential amplifier, the main purpose of the isolation amplifier is the protection of the patient by eliminating the hazard of electric shock resulting from the interaction among patient, amplifier, and other electric devices in the patient's environment, specifically defibrillators and electrosurgical equipment. It also adds to the prevention of line frequency interferences.

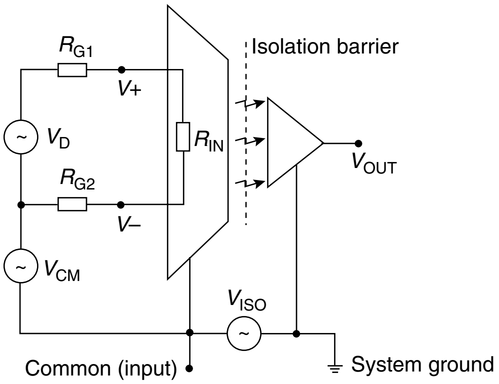


FIGURE 52.7 Equivalent circuit of an isolation amplifier. The differential amplifier on the left transmits the signal through the isolation barrier by a transformer, capacitor, or an opto-coupler.

Isolation amplifiers are realized in three different technologies: transformer isolation, capacitor isolation, and opto-isolation. An isolation barrier provides a complete galvanic separation of the input side, that is, patient and preamplifier, from all equipment on the output side. Ideally, there will be no flow of electric current across the barrier. The isolation-mode voltage is the voltage which appears across the isolation barrier, that is, between the input common and the output common (Figure 52.7). The amplifier has to withstand the largest expected isolation voltages without damage. Two isolation voltages are specified for commercial isolation amplifiers (1) the continuous rating and (2) the test voltage. To eliminate the need for longtime testing, the device is tested at about two times the rated continuous voltage. Thus, for a continuous rating of 2000 V, the device has to be tested at 4000 to 5000 V for a reasonable period of time.

Since there is always some leakage across the isolation barrier, the **isolation mode rejection ratio (IMRR)** is not infinite. For a circuit as shown in Figure 52.7, the output voltage is:

$$V_{\text{out}} = \frac{G}{R_{G1} + R_{G2} + R_{IN}} \left[V_D + \frac{V_{CM}}{\text{CMRR}} \right] + \frac{V_{ISO}}{\text{IMRR}} \tag{52.5}$$

where G is the amplifier gain, V_D, V_{CM}, and V_{ISO} are differential, common mode, and isolation voltages, respectively, and CMRR is the common mode rejection ratio for the amplifier [Burr-Brown, 1994].

Typical values of IMRR for a gain of 10 are 140 dB at dc, and 120 dB at 60 Hz with a source unbalance of 5000 Ω. The isolation impedance is approximately 1.8 pF || 10^{12}Ω.

Transformer coupled isolation amplifiers perform on the basis of inductive transmission of a carrier signal that is amplitude modulated by the biosignal. A synchronous demodulator on the output port reconstructs the signal before it is fed through a Bessel response low-pass filter to an output buffer. A power transformer, generally driven by a 400 to 900 kHz square wave, supplies isolated power to the amplifier.

Optically coupled isolation amplifiers can principally be realized using only a single LED and photodiode combination. While useful for a wide range of digital applications, this design has fundamental limitations as to its linearity and stability as a function of time and temperature. A matched photodiode design, as used in the Burr-Brown 3650/3652 isolation amplifier, overcomes these difficulties [Burr-Brown, 1994]. Operation of the amplifier requires an isolated power supply to drive the input stages. Transformer coupled low leakage current isolated dc/dc converters are commonly used for this purpose. In some particular applications, especially in cases where the signal is transmitted over a longer distance by fiber optics, for example, ECG amplifiers used for gated magnetic resonance imaging, batteries are used to power the amplifier. Fiber optic coupling in isolation amplifiers is another option that offers the advantage of higher flexibility in the placement of parts on the amplifier board.

Biopotential amplifiers have to provide sufficient protection from electrical shock to both user and patient. Electrical-safety codes and standards specify the minimum safety requirements for the equipment, especially the maximum leakage currents for chassis and patient leads, and the power distribution system [Webster, 1992; AAMI, 1993].

Special attention to patient safety is required in situations where biopotential amplifiers are connected to personal computers which are more and more often used to process and store physiological signals and data. Due to the design of the power supplies used in standard PCs permitting high leakage currents — an inadequate situation for a medical environment — there is a potential risk involved even when the patient is isolated from the PC through an isolation amplifier stage or optical signal transmission from the amplifier to the computer. This holds especially in those cases where, due to the proximity of the PC to the patient, an operator might touch patient and computer at the same time, or the patient might touch the computer. It is required that a special power supply with sufficient limitation of leakage currents is used in the computer, or that an additional, medical grade isolation transformer is used to provide the necessary isolation between power outlet and PC.

52.2.3 Surge Protection

The isolation amplifiers described in the preceding paragraph are primarily used for the protection of the patient from electric shock. Voltage surges between electrodes as they occur during the application of a defibrillator or electrosurgical instrumentation also present a risk to the biopotential amplifier. Biopotential amplifiers should be protected against serious damage to the electronic circuits. This is also part of the patient safety since defective input stages could otherwise apply dangerous current levels to the patient. To achieve this protection, voltage limiting devices are connected between each measuring electrode and electric ground. Ideally, these devices do not represent a shunt impedance and thus do not lower the input impedance of the preamplifier as long as the input voltage remains in a range considered safe for the equipment. They appear as an open circuit. As soon as the voltage drop across the device reaches a critical value V_b, the impedance of the device changes sharply and current passes through it to such an extent that the voltage cannot exceed V_b due to the voltage drop across the series resistor R as indicated in Figure 52.8.

Devices used for amplifier protection are diodes, Zener diodes, and gas-discharge tubes. Parallel silicon diodes limit the voltage to approximately 600 mV. The transition from nonconducting to conducting state is not very sharp, and signal distortion begins at about 300 mV which can be within the range of

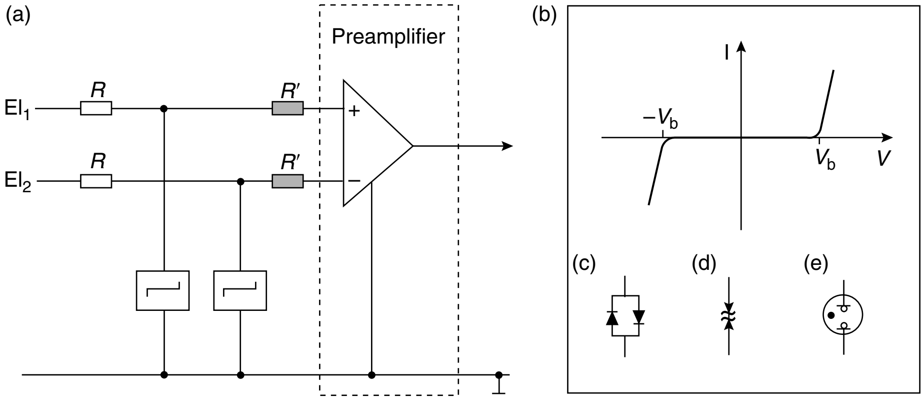


FIGURE 52.8 Protection of the amplifier input against high-voltage transients. The connection diagram for voltage-limiting elements is shown in panel (a) with two optional resistors R′ at the input. A typical current–voltage characteristic is shown in panel (b). Voltage-limiting elements shown are the anti-parallel connection of diodes (c), anti-parallel connection of Zener diodes (d), and gas-discharge tubes (e).

input voltages depending on the electrodes used. The breakdown voltage can be increased by connecting several diodes in series. Higher breakdown voltages are achieved by Zener diodes connected back to back. One of the diodes will be biased in the forward direction and the other in the reverse direction. The breakdown voltage in the forward direction is approximately 600 mV, but the breakdown voltage in the reverse direction is higher, generally in the range of 3 to 20 V, with a sharper voltage–current characteristic than the diode circuit.

A preferred voltage-limiting device for biopotential amplifiers is the *gas-discharge tube*. Due to its extremely high impedance in the nonconducting state, this device appears as an open circuit until it reaches its breakdown voltage. At the breakdown voltage which is in the range of 50 to 90 V, the tube switches to the conducting state and maintains a voltage that is usually several volts less than the breakdown voltage. Though the voltage maintained by the gas-discharge tube is still too high for some amplifiers, it is low enough to allow the input current to be easily limited to a safe value by simple circuit elements such as resistors like the resistors R′ indicated in Figure 52.8a. Preferred gas discharge tubes for biomedical applications are miniature neon lamps which are very inexpensive and have a symmetric characteristic.

52.2.4 Input Guarding

The common mode input impedance and thus the CMRR of an amplifier can be greatly increased by guarding the input circuit. The common mode signal can be obtained by two averaging resistors connected between the outputs of the two input op-amps of an instrumentation amplifier as shown in Figure 52.9. The buffered common-mode signal at the output of op-amp 4 can be used as guard voltage to reduce the effects of cable capacitance and leakage.

In many modern biopotential amplifiers, the reference electrode is not grounded. Instead, it is connected to the output of an amplifier for the common mode voltage, op-amp 3 in Figure 52.10, which works as an inverting amplifier. The inverted common mode voltage is fed back to the reference electrode. This negative feedback reduces the common-mode voltage to a low value [Webster, 1992]. Electrocardiographs based on this principle are called driven-right-leg systems replacing the right leg ground electrode of ordinary electrocardiographs by an actively driven electrode.

52.2.5 Dynamic Range and Recovery

With an increase of either the common mode or differential input voltage there will be a point where the amplifier will overload and the output voltage will no longer be representative for the input voltage. Similarly, with a decrease of the input voltage, there will be a point where the noise components of the output voltage cover the output signal to a degree that a measurement of the desired biopotential is no

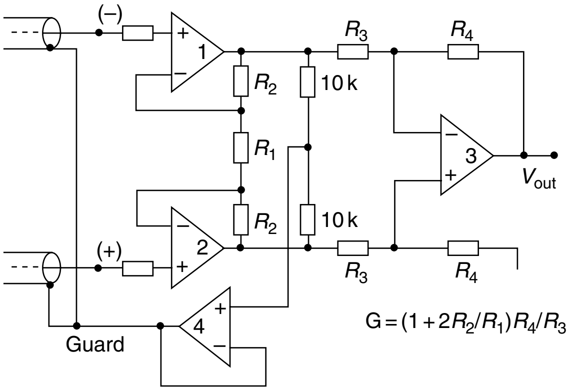


FIGURE 52.9 Instrumentation amplifier providing input guarding.

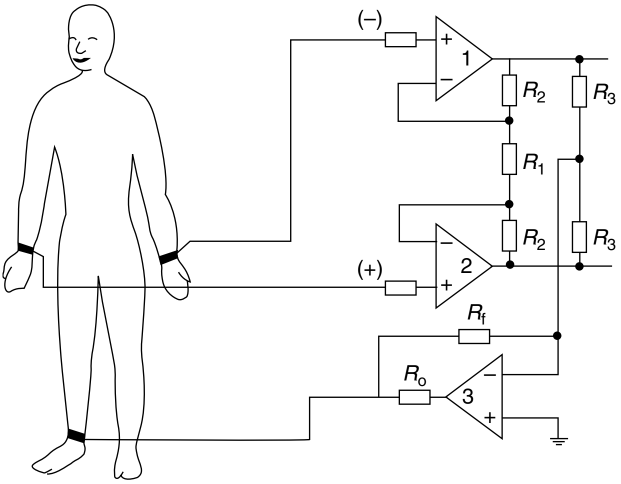


FIGURE 52.10 Driven-right-leg circuit reducing common-mode interference.

longer possible. The dynamic range of the amplifier, that is, the range between the smallest and largest possible input signal to be measured, has to cover the whole amplitude range of the physiological signal of interest. The required dynamic range of biopotential amplifiers can be quite large. In an application like fetal monitoring for example, two signals are recorded simultaneously from the electrodes which are quite different in their amplitudes: the fetal and the maternal ECG. While the maternal ECG shows an amplitude of up to 10 mV, the fetal ECG often does not reach more than 1 μV. Assuming that the fetal ECG is separated from the composite signal and fed to an analog/digital converter for digital signal processing with a resolution of 10 bit (signed integer), the smallest voltage to be safely measured with the biopotential amplifier is 1/512 μV or about 2 nV vs. 10 mV for the largest signal, or even up to 300 mV in the presence of an electrode offset potential. This translates to a dynamic range of 134 dB for the signals alone and 164 dB if the electrode potential is included in the consideration. Though most applications are less demanding, even such extreme requirements can be realized through careful design of the biopotential amplifer and the use of adequate components. The penalty for using less expensive amplifiers with diminished performance would be a potentially severe loss of information.

Transients appearing at the input of the biopotential amplifier, like voltage peaks from a cardiac pacemaker or a defibrillator, can drive the amplifier into saturation. An important characteristic for the amplifier is the time it takes to recover from such overloads. The recovery time depends on the characteristics of the transient, like amplitude and duration, the specific design of the amplifier, like bandwidth, and the components used. Typical biopotential amplifiers may take several seconds to recover from severe overload. The recovery time can be reduced by disconnecting the amplifier inputs at the discovery of a transient using an electronic switch.

52.2.6 Passive Isolation Amplifiers

Increasingly, biopotentials have to be measured within implanted devices and need to be transmitted to an external monitor or controller. Such applications include cardiac pacemakers transmitting the intracardiac ECG and functional electrical stimulation where, for example, action potentials measured at one eyelid serve to stimulate the other lid to restore the physiological function of a damaged lid at least to some degree. In these applications, the power consumption of the implanted biopotential amplifier limits the lifespan of the implanted device. The usual solution to this problem is an inductive transmission of power into the implanted device that serves to recharge an implanted battery. In applications where the size of the implant is of concern, it is desirable to eliminate the need for the battery and the related circuitry by using a quasi passive biopotential amplifer, that is, an amplifier that does not need a power supply.

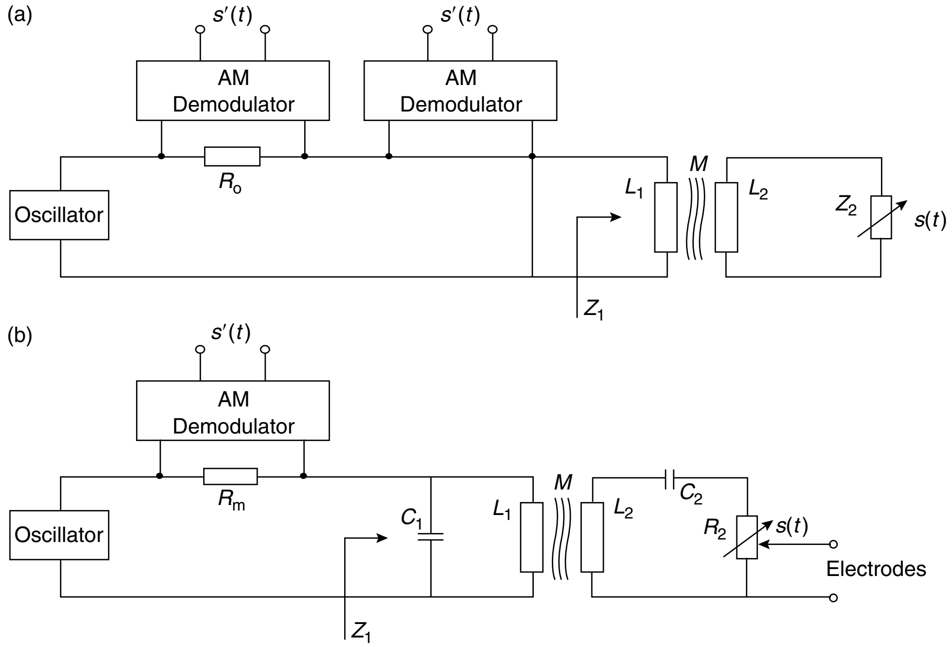


FIGURE 52.11 The *passive* isolation amplifier can be operated without the need for an isolated power supply. The biological source provides the power to modulate the load impedance of an inductive transformer. As an easy realization shown in panel (b), a FET can be directly connected to two electrodes. The source-drain resistance changes as a linear function of the biopotential which is then reflected by the input impedance of the transformer.

The function of passive telemetric amplifiers for biopotentials is based on the ability of the biological source to drive a low power device such as a FET and the sensing of the biopotentials through inductive or acoustic coupling of the implanted and external devices [Nagel et al., 1982]. In an inductive system, a FET serves as a load to an implanted secondary LC-circuit which is stimulated inductively by an extracorporal oscillator (Figure 52.11). Depending on the special realization of the system, the biopotential is available in the external circuit from either an amplitude or frequency-modulated carrier-signal. The input impedance of the inductive transmitter as a function of the secondary load impedance Z_2 is given by:

$$Z_1 = j\omega L_1 + \frac{(\omega M)^2}{Z_2 + j\omega L_2} \tag{52.6}$$

In an amplitude-modulated system, the resistive part of the input-impedance Z_1 must change as a linear function of the biopotential. The signal is obtained as the envelope of the carrier signal, measured across a resistor R_m. A frequency-modulated system is realized when the frequency of the signal generator is determined at least in part by the impedance Z_1 of the inductive transmitter. In both cases, the signal-dependent changes of the secondary impedance Z_2 can be achieved by a junction-FET. Using the field effect transistor as a variable load resistance changing its resistance in proportion to the source-gate voltage which is determined by the electrodes of this two-electrode amplifier, the power supplied by the biological source is sufficient to drive the amplifier. The input impedance can be in the range of 10^{10} Ω.

Optimal transmission characteristics are achieved with AM systems. Different combinations of external and implanted resonance circuits are possible to realize in an AM system, but primary parallel with secondary serial resonance yields the best characteristics. In this case, the input impedance is given by:

$$Z_1 = \frac{1}{j\omega C_1} + \left(\frac{L_1}{M}\right)^2 \cdot R_2 \tag{52.7}$$

The transmission factor $(L_1/M)^2$ is optimal since the secondary inductivity, that is, the implanted inductivity, can be small, only the external inductivity determines the transmission factor and the mutual inductivity should be small, a fact that favors the loose coupling that is inherent to two coils separated by skin and tissue. There are, of course, limits to M which cannot be seen from Equation 52.7. In a similar fashion, two piezoelectric crystals can be employed to provide the coupling between input and output.

This 2-lead isolation amplifier design is not limited to telemetric applications. It can also be used in all other applications where its main advantage lies in its simplicity and the resulting substantial cost savings as compared to other isolation amplifiers which require additional amplifier stages and an additional isolated power supply.

52.2.7 Digital Electronics

The ever increasing density of integrated digital circuits together with their extremely low power consumption permits digitizing and preprocessing of signals already on the isolated patient-side of the amplifiers, thus improving signal quality and eliminating the problems normally related to the isolation barrier, especially those concerning isolation voltage interferences and long-term stability of the isolation amplifiers. Digital signal transmission to a remote monitoring unit, a computer system, or computer network can be achieved without any risk of picking up transmission line interferences, especially when implemented with fiberoptical cables.

Digital techniques also offer an easy means of controlling the front-end of the amplifier. Gain factors can be easily adapted, and changes of the electrode potential resulting from electrode polarization or from interferences which might drive the differential amplifier into saturation can easily be detected and compensated.

52.3 Summary

Biopotential amplifiers are a crucial component in many medical and biological measurements, and largely determine the quality and information content of the measured signals. The extremely wide range of necessary specifications with regard to bandwidth, sensitivity, dynamic range, gain, CMRR, and patient safety leaves only little room for the application of general purpose biopotential amplifiers, and mostly requires the use of special purpose amplifiers.

Defining Terms

Common Mode Rejection Ratio (CMRR): The ratio between the amplitude of a common mode signal and the amplitude of a differential signal that would produce the same output amplitude or as the ratio of the differential gain over the common-mode gain: CMRR $= G_D/G_{CM}$. Expressed in decibels, the common mode rejection is 20 $\log_{10}$ CMRR. The common mode rejection is a function of frequency and source-impedance unbalance.

Isolation Mode Rejection Ratio (IMRR): The ratio between the isolation voltage, V_{ISO}, and the amplitude of the isolation signal appearing at the output of the isolation amplifier, or as isolation voltage divided by output voltage V_{OUT} in the absence of differential and common mode signal: IMRR $= V_{ISO}/V_{OUT}$.

Operational Amplifier (op-amp): A very high gain dc-coupled differential amplifier with single-ended output, high voltage gain, high input impedance, and low output impedance. Due to its high open-loop gain, the characteristics of an op-amp circuit only depend on its feedback network. Therefore, the integrated circuit op-amp is an extremely convenient tool for the realization of linear amplifier circuits [Horowitz and Hill, 1980].

References

AAMI, 1993. *AAMI Standards and Recommended Practices, Biomedical Equipment*, Vol. 2, 4th ed. AAMI, Arlington, VA.

Burr-Brown, 1994. *Burr-Brown Integrated Circuits Data Book, Linear Products*, Burr-Brown Corp., Tucson, AZ.

Horowitz, P. and Hill, W., 1980. *The Art of Electronics*, Cambridge University Press, Cambridge, UK.

Hutten, H. (Hrsg) 1992. *Biomedizinische Technik*, Springer-Verlag, Berlin.

Nagel, J., Ostgen, M., and Schaldach, M., 1982. *Telemetriesystem*, German Patent Application, P 3233240.8-15.

Pallás-Areny, R. and Webster, J.G., 1990. Composite Instrumentation Amplifier for Biopotentials. *Annals of Biomedical Engineering* 18, 251–262.

Strong, P., 1970. *Biophysical Measurements*, Tektronix, Inc., Beaverton, OR.

Webster, J.G., Ed., 1992. *Medical Instrumentation, Application and Design*, 2nd ed. Houghton Mifflin Company, Boston, MA.

Song, Y., Ozdamar, O., and Lu, C.C., 1998. Pasteless Electrode/Amplifier System for Auditory Brainstem Response (ABR) Recording. *Annals of Biomedical Engineering* 26, S-103.

Further Information

Detailed information on the realization of amplifiers for biomedical instrumentation and the availability of commercial products can be found in the references and in the Data Books and Application Notes of various manufacturers of integrated circuit amplifiers like Burr-Brown, Analog Devices, and Precision Monolithics Inc. as well as manufacturers of laboratory equipment like Gould and Grass.

53

Bioelectric Impedance Measurements

Robert Patterson
The University of Minnesota

Bioelectric tissue impedance measurements to determine or infer biological information have a long history dating back to before the turn of the century. The start of modern clinical applications of bioelectric impedance (BEI) measurements can be attributed in large part to the reports by Nyboer [1970]. BEI measurements are commonly used in **apnea** monitoring, especially for infants, and in the detection of venous **thrombus.** Many papers report the use of the BEI technique for peripheral blood flow, cardiac stroke volume, and body composition. Commercial equipment is available for these latter three applications, although the reliability, validity, and accuracy of these applications have been questioned and, therefore, have not received widespread acceptance in the medical community.

BEI measurements can be classified into two types. The first and most common application is in the study of the small pulsatile impedance changes associated with heart and respiratory action. The goal of this application is to give quantitative and qualitative information on the volume changes (**plethysmography**) in the lung, heart, peripheral arteries, and veins. The second application involves the determination of body characteristics such as total body fluid volume, inter and extra-cell volume, percent body fat, and cell and tissue viability. In this application, the total impedance is used and in some cases measured as a function of frequency, which is referred to as *impedance spectroscopy.*

53.1 Measurement Methods

Most single frequency BEI measurements are in the range of 50 to 100 kHz (at these frequencies no significant electrical shock hazard exists) using currents from 0.5 to 4 mA RMS. Currents at these levels

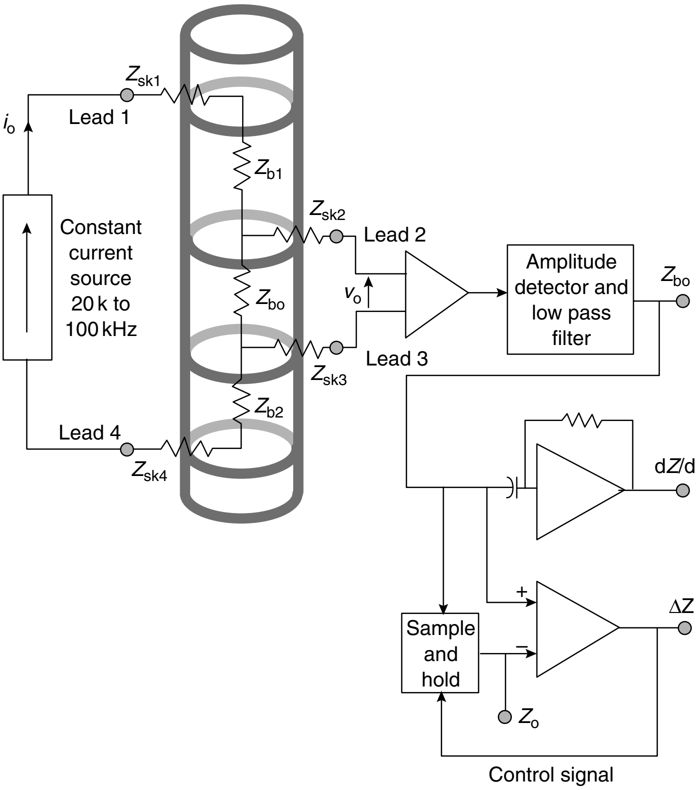


FIGURE 53.1 The four-electrode impedance measurement technique and the associated instrumentation.

are usually necessary to obtain a good signal-to-noise ratio when recording the small pulsatile changes that are in the range of 0.1 to 1% of the total impedance. The use of higher frequencies creates instrumentation design problems due to stray capacity.

BEI measurements in the 50–100 kHz range have typical skin impedance values 2–10 times the value of the underlying body tissue of interest depending on electrode area. In order to obtain BEI values that can be used to give quantitative biological information, the skin impedance contribution must be eliminated. This is accomplished by using the four electrode impedance measurement method shown in Figure 53.1, along with other signal processing blocks used in typical impedance plethysmographs.

Z_{bo} is the internal section of tissue we wish to measure. If we used two electrodes to make the measurement, we would include two skin impedances (i.e., Z_{sk1} and Z_{sk4}) and two internal tissue impedances (i.e., Z_{b1} and Z_{b2}) which would make it impossible to estimate an accurate value for Z_{bo}.

A constant current source supplies current, I_o, to the outside two electrodes 1 and 4. This current flows through the skin and body tissue independent of tissue and skin impedance values. The voltage V_o is measured across Z_{bo} with a voltage amplifier using electrodes 2 and 3. Assuming the output impedance of the current source is $\gg Z_{sk1} + Z_{b1} + Z_{bo} + Z_{b2} + Z_{sk4}$ and the input impedance of the voltage amplifier is $\gg Z_{sk2} + Z_{bo} + Z_{sk3}$, then

$$Z_{bo} = Z_o + \Delta Z, \qquad Z_o = V_o/I_o, \quad \text{and} \quad \Delta Z = \Delta V_o/I_o \tag{53.1}$$

where Z_o is the non-time-varying portion of the impedance and ΔZ is the impedance change typically-associated with the pulsation of blood in the region of measurement.

The output from the voltage pick-up amplifier (Figure 53.1) is connected to the amplitude detector and low pass filter which removes the high frequency carrier signal, which results in an output voltage proportional to $Z_{bo} \cdot Z_{bo}$ has a large steady part which is proportional to the magnitude of the tissue impedance (Z_o) and a small (0.1 to 1%) part, ΔZ, that represents the change due to respiratory or cardiac activity. In order to obtain a signal representing ΔZ, Z_o must be removed from Z_{bo} and the signal amplified. This can be accomplished by capacity coupling or by subtracting a constant that represents Z_o. The latter is usually done because many applications require near dc response. The output of the ΔZ amplifier will be a waveform oscillating around zero volts. The output from the ΔZ amplifier controls the sample and hold circuit. When the ΔZ output exceeds a given value, usually plus or minus a few tenths of an ohm, the sample and hold circuit updates its value of Z_o. The output from the sample and hold circuit is subtracted from Z_{bo} by the ΔZ amplifier. The derivative of Z_{bo} is frequently obtained in instruments intended for cardiac use.

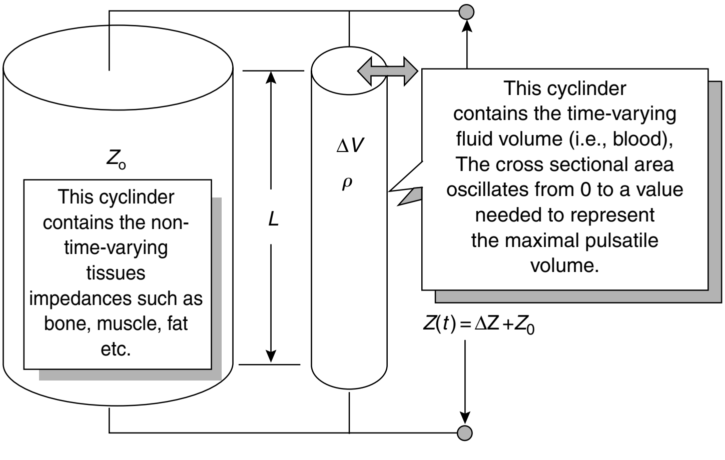
53.2 Modeling and Formula Development

To relate the ΔZ obtained on the thorax or peripheral limbs to the pulsatile blood volume change, the parallel column model, first described by Nyboer [1970], is frequently used (Figure 53.2). The model consists of a conducting volume with impedance Z_o in parallel with a time-varying column with resistivity ρ, length L, and a time-varying cross-sectional area which oscillates from zero to a finite value. At the time in the cardiac cycle when the pulsatile volume is at a minimum, all of the conducting tissues and fluids are represented by the volume labeled Z_o. This volume can be a heterogeneous mixture of all of the non-time-varying tissues such as fat, bone, muscle, etc. in the region under measurement. The only information needed about this volume is its impedance Z_o and that it is electrically in parallel with the small time varying column. During the cardiac cycle, the volume change in the right column starts with a zero cross-sectional area and increases in area until its volume equals the blood volume change. If the impedance of this volume is much greater than Z_o, then the following relation holds:

$$\Delta V = \rho(L^2/Z_o^2)\Delta Z \tag{53.2}$$

where

ΔV = the pulsatile volume change with resistivity ρ
ρ = the resistivity of the pulsatile volume in Ω-cm (typically the resistivity of blood)
L = the length of the cylinder
Z_o = the impedance measured when the pulsatile volume is at a minimum
ΔZ = the magnitude of the pulsatile impedance change.



Z_o
This cyclinder contains the non-time-varying tissues impedances such as bone, muscle, fat etc.
L
ΔV
ρ
This cyclinder contains the time-varying fluid volume (i.e., blood), The cross sectional area oscillates from 0 to a value needed to represent the maximal pulsatile volume.
$Z(t) = \Delta Z + Z_0$

FIGURE 53.2 Parallel column model.

The resistivity of blood, ρ in Ω-cm, is a function of hematocrit (H) expressed as a percentage and can be calculated as $\rho = 67.919 \exp(0.0247H)$ [Mohapatra et al., 1977]. The typical value used for blood is 150 Ω-cm.

53.3 Respiration Monitoring and Apnea Detection

If the BEI is measured across the thorax, a variation of approximately 1 to 2 Ω/l of lung volume change is observed, which increases with inspiration. The most common position of the electrodes for respiratory measurements is on each side of the thorax along the midaxillary line. The largest signal is generally obtained at the level of the xiphsternal joint although a more linear signal is obtained higher up near the axilla [Geddes and Baker, 1989].

The problems encountered with the quantitative use of BEI for respiration volume are movement artifacts and the change in the response depending on whether diaphragmatic or intercostal muscles are used. For most applications, the most serious problem is body movement and positional changes artifacts which can cause impedance changes significantly larger than the change caused by respiration.

The determination of apnea or whether respiration has stopped [Neuman, 1988] in infants is one of the most widely used applications of BEI. For convenience and due to the lack of space on the thorax of infants, only two electrodes are used. These are placed at the mid-thoracic level along the midaxillary line and are also used to obtain the ECG. No effort is usually made to quantitate the volume change. Filtering is used to reduce movement artifacts and automatic gain controls and adaptive threshold detection is used in the breath detection circuits. Due to movement artifacts, the normal breath detection rate in infants is not highly reliable. When respiration stops, body movement ceases which eliminates the movement artifacts and then apnea can be detected. Ventation detection problems can occur if the airway is obstructed and the infant makes inspiratory movement efforts or cardiac-induced impedance changes are interpreted as a respiratory signal. Figure 53.3 shows a typical impedance measurement during an apneic period.

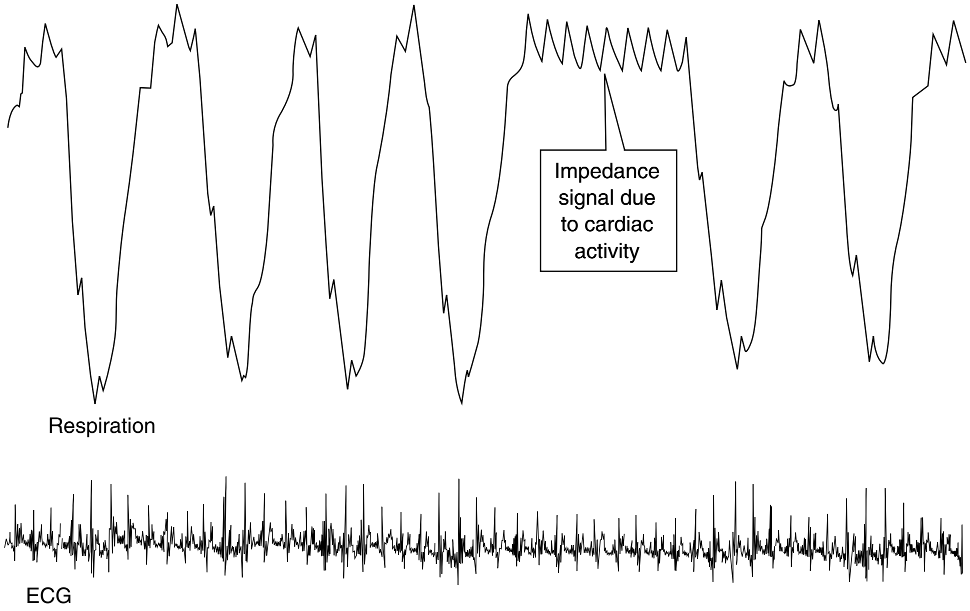


FIGURE 53.3 Example of BEI respiration signal and ECG.

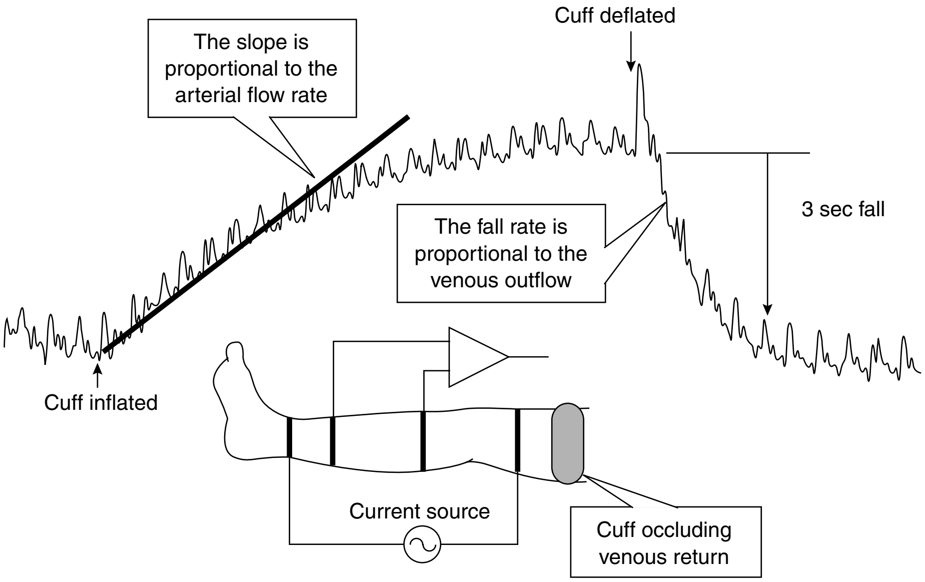

The slope is proportional to the arterial flow rate
Cuff deflated
3 sec fall
The fall rate is proportional to the venous outflow
Cuff inflated
Current source
Cuff occluding venous return

FIGURE 53.4 The measurement of arterial inflow and venous outflow.

53.4 Peripheral Blood Flow

BEI measurements are made on limbs to determine arterial blood flow into the limb or for the detection of venous **thrombosis**. In both applications, an occluding cuff is inflated above venous pressure to prevent outflow for a short period of time.

Figure 53.4 shows the typical electrode arrangement on the leg and the position of the occluding cuff. The cuff is rapidly inflated to 40 to 50 mmHg, which prevents venous outflow without significantly changing arterial inflow. The arterial inflow causes an increase in the volume of the limb. The slope of the initial impedance change as determined by the first three or four beats is used to measure the arterial flow rate. Equation 53.2 is used to calculate the volume change from the impedance change. The flow (the slope of the line in Figure 53.4) is determined by dividing the volume change by the time over which the impedance change was measured. The volume change that occurs after the impedance change reaches a plateau is a measure of the **compliance** of the venous system.

After the volume change of the leg has stabilized, the cuff is quickly deflated which results in an exponential decrease in volume. If a thrombosis exists in the veins, the time constant of the outflow lengthens. The initial slope, the time constant, and percentage change at 3 sec after cuff deflation have been used to quantitate the measurement. The percentage change at 3 sec has been reported to show the best agreement with **venograms**. The determination of deep venous thrombus is frequently made by combining the maximal volume change with the outflow rate. The agreement with a venogram is 94% for the detection of deep venous thrombus proximal to the knee [Anderson, 1988].

53.5 Cardiac Measurements

The measurements of chest impedance changes due to cardiac activity have been reported starting in 1930s. One of the most popular techniques, first reported by Patterson et al. [1964], for quantitative measurements uses band electrodes around the ends of the thorax as shown in Figure 53.5. Each heart beat causes a pulsatile decrease in impedance of 0.1 to 0.2 Ω (decreasing ΔZ and negative $\mathrm{d}Z/\mathrm{d}t$ are shown in an upward direction). The empirical formula for stroke volume based on this method follows

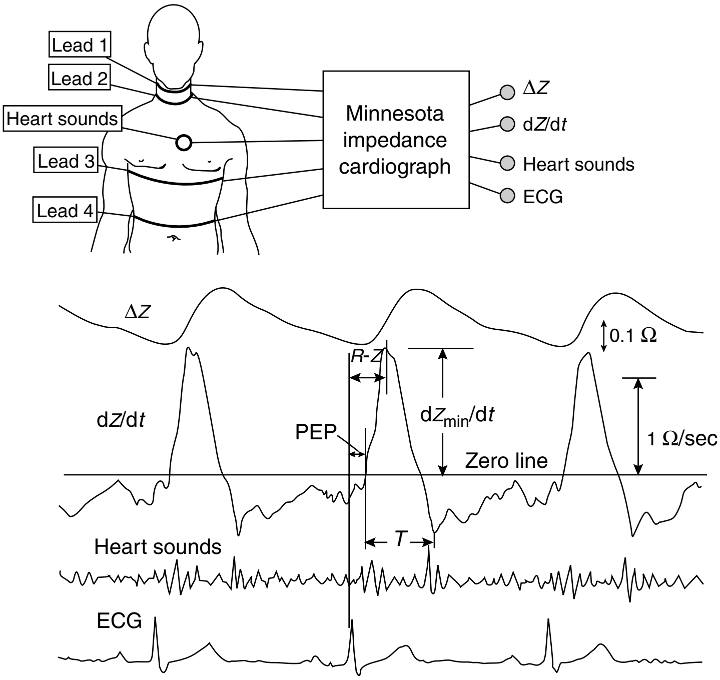


FIGURE 53.5 Impedance cardiographic waveforms.

from Equation 53.2:

$$\Delta V = \rho(L^2/Z_o^2)T\,\mathrm{d}Z_{\min}/\mathrm{d}t \tag{53.3}$$

where

ΔV = cardiac stroke volume (ml)
ρ = resistivity of blood (Ω-cm)
L = distance between the inner band electrodes (cm)
Z_o = base impedance (Ω)
$\mathrm{d}Z_{\min}/\mathrm{d}t$ = the magnitude of the largest negative derivative of the impedance change occurring during systole (Ω/sec)
T = systolic ejection time (sec)

Many studies have been conducted comparing the impedance technique with other standard methods of measuring stroke volume and cardiac output. In general, the correlation coefficient in subjects without valve problems or heart failure and with a stable circulatory system is 0.8 to 0.9. In patients with a failing circulatory system or valvular or other cardiovascular problems, the correlation coefficient may be less than 0.6 [Patterson, 1989].

Experimental physiological studies and computer modeling show that multiple sources contribute to the impedance signal. The anatomical regions that make significant contributions are the aorta, lungs, and atria. Recent studies have reported that the blood resistivity change with flow, pulsatile changes in the neck region, and the movement of the heart significantly contribute to the signal [Patterson et al., 1991; Wang and Patterson, 1995]. It appears that a number of different sources of the thoracic impedance change combine in a fortuitous manner to allow for a reasonable correlation between BEI measured stroke volume and other accepted techniques. However, in patients with cardiac problems where the contributions of the different sources may vary, the stroke volume calculated from BEI measurements may have significant error.

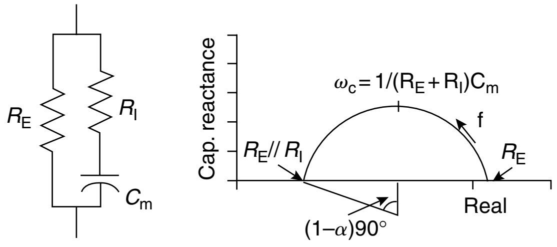
53.6 Body Composition (Single Frequency Measurement)

The percentage of body fat has been an important parameter in sports medicine, physical conditioning, weight loss programs, and predicting optimum body weight. To determine body composition, the body is configured as two parallel cylinders similar to the model described earlier. One cylinder is represented by fat and the other as fat-free body tissue. Since the resistivity of fat is much larger than muscle and other body fluids, the volume determined from the total body impedance measurement is assumed to represent the fat-free body volume. Studies have been conducted that calculate the fat-free body mass by determining the volume of the fat-free tissue cylinder using the impedance measured between the right hand and right foot, and using the subject's height as a measure of the cylinder's length with an empirical constant used to replace ρ [Lukaski et al., 1985]. Knowing the total weight and assuming body density factors, the percentage of body fat can be calculated as the difference between total weight and the weight of the fat-free tissue. Many studies have reported the correlation of the percentage of body fat calculated from BEI with other accepted standard techniques from approximately 0.88 to 0.98 [Schoeller and Kushner, 1989]. The physical model used for the equation development is a poor approximation to the actual body because it assumes a uniform cross-sectional area for the body between the hand the foot. Patterson [1989] pointed out the main problem with the technique: the measured impedance depends mostly on the characteristics of the arms and legs and not of the trunk. Therefore, determination of body fat with this method may often be inaccurate.

53.7 Impedance Spectroscopy

By measuring BEI over a range of frequencies (typically between 1 kHz and 1 MHz), the material properties of the tissues can be determined [Ackmann and Seitz, 1984]. Figure 53.6 shows the typical complex plane plot of the real and imaginary part of impedance and the model used to fit the data. R_E represents the extra-cellular space, R_I the intra-cellular space, and C_m the cell membrane. The parameter α is proportional to the angle of the suppression of the semicircle. It exists to account for the distribution of time constants in the tissue. At low frequencies, the current flows in the extra-cellular space and at high frequencies the current is capacitively passed across the cell membrane while the extra and intra cell spaces are in parallel.

Using these model parameters, studies have shown positive results in determining intra and extra-cellular fluid volumes [Kanai et al., 1987], body fat [De Lorenzo et al., 1997], tissue ischemica [Cinca et al., 1997] and cancerous tissues [Jossient, 1998].



$$Z = \left(\frac{R_E}{R_E + R_I}\right)\left(R_I + \frac{R_E}{1 + (j\omega C_m(R_E + R_I))^{1-\alpha}}\right)$$

FIGURE 53.6 Typical impedance spectroscopy data, model, and the equation used to fit the data.

53.8 Summary

Electrical impedance instrumentation is not relatively costly, which has encouraged its possible application in many different areas. The impedance measurement is influenced by many different factors including geometry, tissue conductivity, and blood flow. Because of this complexity, it is difficult to reliably measure an isolated physiological parameter, which has been the principle factor limiting its use. The applications that are widely used in clinical medicine are apnea monitoring and the detection of venous thrombosis. The other applications described above will need more study before becoming a reliable and useful measurement.

Defining Terms

Apnea: A suspension of respiration.

Compliance: The volume change divided by the pressure change. The higher the compliance, the more easily the vessel will expand as pressure increases.

Plethysmography: The measurement of the volume change of an organ or body part.

Thrombosis: The formation of a thrombus.

Thrombus: A clot of blood formed within the heart or vessel.

Venogram: An x-ray image of the veins using an injected radiopaque contrast material.

References

Ackmann, J.J. and Seitz, M.A. (1984). Methods of complex impedance measurements in biologic tissue. *Crit. Rev. Biomed. Eng.* 11: 281–311.

Anderson, F.A. Jr. (1988). Impedance plethysmography, in J.G. Webster (Ed.) *Encyclopedia of Medical Devices and Instrumentation,* Wiley, New York, pp. 1632–1643.

Cinca, J., Warran, M., Carreno, A., Tresanchez, M., Armadans, L., Gomez, P., and Soler-Soler, J. (1997). Changes in myocardial electrical impedance induced by coronary artery occlusion in pigs with and without preconditioning. *Circulation* 96: 3079–3086.

De Lorenzo, A., Andreoi, A., Matthie, J., and Withers, P. (1997). Predicting body cell mass with bioimpedance by using theoretical methods: a technological review. *J. Appl. Physiol.* 82: 1542–1558.

Geddes, L.A. and Baker, L.E. (1989). *Principles of Applied Biomedical Instrumentation — Third Edition*, Wiley, New York, pp. 569–572.

Jossient, J. (1998). The impedivity of freshly excised human breast tissue. *Physiol. Meas.* 19: 61–75.

Kanai, H., Haeno, M., and Sakamoto, K. (1987). Electrical measurement of fluid distribution in human legs and arms. *Med. Prog. Technol.* 12: 159–170.

Lukaski, H.C., Johnson, P.E., Bolonchuk, W.W., and Lykken, G.I. (1985). Assessment of fat-free mass using bioelectric impedance measurements of the human body. *Am. J. Clin. Nutr.* 41: 810–817.

Mohapatra, S.N., Costeloe, K.L., and Hill, D.W. (1977). Blood resistivity and its implications for the calculations of cardiac output by the thoracic electrical impedance technique. *Intensive Care Med.* 3: 63.

Neuman, M.R. (1988). Neonatal monitoring, in J.G. Webster (Ed.) *Encyclopedia of Medical Devices and Instrumentation*, Wiley, New York, pp. 2020–2034.

Nyboer, J. (1970). *Electrical Impedance Plethysmography*, 2nd ed., Charles C. Thomas, Springfield, IL.

Patterson, R.P. (1989). Fundamentals of impedance cardiography, *IEEE Eng. Med. Biol. Mag.* 8: 35–38.

Patterson, R.P. (1989). Body fluid determinations using multiple impedance measurements, *IEEE Eng. Med. Biol. Mag.* 8: 16–18.

Patterson, R., Kubicek, W.G., Kinnen, E., Witsoe, D., and Noren, G. (1964). Development of an electrical impedance plethysmography system to monitor cardiac output. *Proc of the First Ann. Rocky Mountain Bioengineering Symposium*, pp. 56–71.

Patterson, R.P., Wang, L., and Raza, S.B. (1991). Impedance cardiography using band and regional electrodes in supine, sitting, and during exercise. *IEEE Trans. BME* 38: 393–400.

Schoeller, D.A. and Kushner, R.F. (1989). Determination of body fluids by the impedance technique. *IEEE Eng. Med. Biol. Mag.* 8: 19–21.

Wang, L. and Patterson, R.P. (1995). Multiple sources of the impedance cardiogram based on 3D finite difference human thorax models. *IEEE Trans. Biomed. Eng.* 42: 393–400.

Further Information

The book by Nyboer, *Electrical Impedance Plethysmography* contains useful background information. *The Encyclopedia of Medical Devices and Instrumentation* edited by J.G. Webster and *Principles of Applied Biomedical Instrumentation* by L.A. Geddes and L.E. Baker give a more in-depth description of many applications and describe some usual measurements.

54

Implantable Cardiac Pacemakers

Michael Forde
Pat Ridgely
Medtronic, Inc.

The practical use of an implantable device for delivering a controlled, rhythmic electric stimulus to maintain the heartbeat is relatively recent: cardiac pacemakers have been in clinical use only slightly more than 30 years. Although devices have gotten steadily smaller over this period (from 250 g in 1960 to 25 g today), the technological evolution goes far beyond size alone. Early devices provided only single-chamber, asynchronous, nonprogrammable pacing coupled with questionable reliability and longevity. Today, advanced electronics afford dual-chamber multi*programmability*, diagnostic functions, rate response, data collection, and exceptional reliability, and lithium-iodine power sources extend longevity to upward of 10 years. Continual advances in a number of clinical, scientific, and engineering disciplines have so expanded the use of pacing that it now provides cost-effective benefits to an estimated 350,000 patients worldwide each year.

The modern pacing system is comprised of three distinct components: pulse generator, lead, and programmer (Figure 54.1). The pulse generator houses the battery and the circuitry which generates the stimulus and senses electrical activity. The lead is an insulated wire that carries the stimulus from the generator to the heart and relays intrinsic cardiac signals back to the generator. The programmer is a telemetry device used to provide two-way communications between the generator and the clinician. It can alter the therapy delivered by the pacemaker and retrieve diagnostic data that are essential for optimally titrating that therapy. Ultimately, the therapeutic success of the pacing prescription rests on the clinician's choice of an appropriate system, use of sound implant technique, and programming focused on patient outcomes.

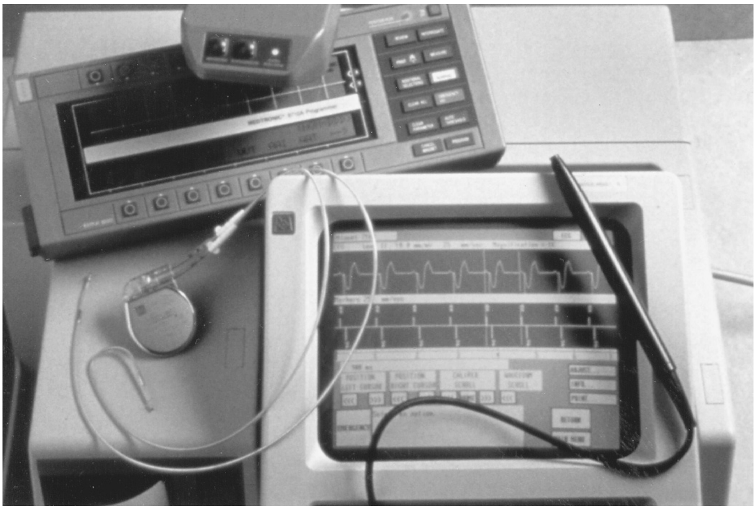


FIGURE 54.1 The pacing systems comprise a programmer, pulse generator, and lead. There are two programmers pictured above; one is portable, and the other is an office-based unit.

This chapter discusses in further detail the components of the modern pacing system and the significant evolution that has occurred since its inception. Our focus is on system design and operations, but we also briefly overview issues critical to successful clinical performance.

54.1 Indications

The decision to implant a permanent pacemaker for bradyarrhythmias usually is based on the major goals of symptom relief (at rest and with physical activity), restoration of functional capacity and quality of life, and reduced mortality. As with other healthcare technologies, appropriate use of pacing is the intent of indications guidelines established by Medicare and other third-party payors.

In 1984 and again in 1991, a joint commission of the American College of Cardiology and the American Heart Association established guidelines for pacemaker implantation (Committee on Pacemaker Implantation, 1991). In general, pacing is indicated when there is a dramatic slowing of the heart rate or a failure in the connection between the atria and ventricles resulting in decreased cardiac output manifested by such symptoms as syncope, light-headedness, fatigue, and exercise intolerance. Failure of impulse formation and/or conduction is the overriding theme of all pacemaker indications. There are four categories of pacing indications:

1. Heart block (e.g., complete heart block, symptomatic 2° AV block)
2. Sick sinus syndrome (e.g., symptomatic bradycardia, sinus arrest, sinus exit block)
3. Myocardial infarction (e.g., conduction disturbance related to the site of infarction)
4. Hypersensitive carotid sinus syndrome (e.g., recurrent syncope)

Within each of these four categories the ACC/AHA provided criteria for classifying a condition as group I (pacing is considered necessary), group II (pacing may be necessary), or group III (pacing is considered inappropriate).

New indications for cardiac pacing are being evaluated under the jurisdiction of the Food and Drug Administration. For example, **hypertrophic obstructive cardiomyopathy** (HOCM) is one of these new potential indications, with researchers looking at dual-chamber pacing as a means of reducing left ventricular outflow obstruction. Though efforts in these areas are ongoing and expanding, for now they remain unapproved as standard indications for pacing.

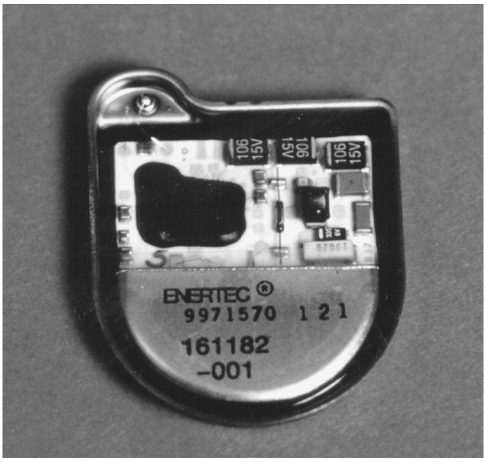


FIGURE 54.2 Internal view of pulse generator.

54.2 Pulse Generators

The pulse generator contains a power source, output circuit, sensing circuit, and a timing circuit (Figure 54.2). A telemetry coil is used to send and receive information between the generator and the programmer. **Rate-adaptive** pulse generators include the sensor components along with the circuit to process the information measured by the sensor.

Modern pacemakers use **CMOS circuit** technology. One to 2 kilobytes of read-only memory (ROM) are used to direct the output and sensing circuits; 16 to 512 bytes of random-access memory (RAM) are used to store diagnostic data. Some manufacturers offer fully RAM-based pulse generators, providing greater storage of diagnostic data and the flexibility for changing feature sets after implantation.

All components of the pulse generator are housed in a *hermetically* sealed titanium case with a connector block that accepts the lead(s). Because pacing leads are available with a variety of different connector sites and configurations, the pulse generator is available with an equal variety of connectors. The outer casing is laser-etched with the manufacturer, name, type (e.g., single- versus dual-chamber), model number, serial number, and the lead connection diagram for each identification. Once implanted, it may be necessary to use an x-ray to reveal the identity of the generator. Some manufacturers use radiopaque symbols and ID codes for this purpose, whereas others give their generators characteristic shapes.

54.2.1 Sensing Circuit

Pulse generators have two basic functions, pacing and sensing. Sensing refers to the recognition of an appropriate signal by the pulse generator. This signal is the intrinsic cardiac depolarization from the chamber or chambers in which the leads are placed. It is imperative for the sensing circuit to discriminate between these intracardiac signals and unwanted electrical interference such as far-field cardiac events, diastolic potentials, skeletal muscle contraction, and pacing stimuli. An intracardiac electrogram (Figure 54.3) shows the waveform as seen by the pacemaker; it is typically quite different from the corresponding event as shown on the surface ECG.

Sensing (and pacing) is accomplished with one of two configurations, bipolar and unipolar. In bipolar, the anode and cathode are close together, with the anode at the tip of the lead and the cathode a ring electrode about 2 cm proximal to the tip. In unipolar, the anode and cathode may be 5 to 10 cm apart. The anode is at the lead tip and the cathode is the pulse generator itself (usually located in the pectoral region).

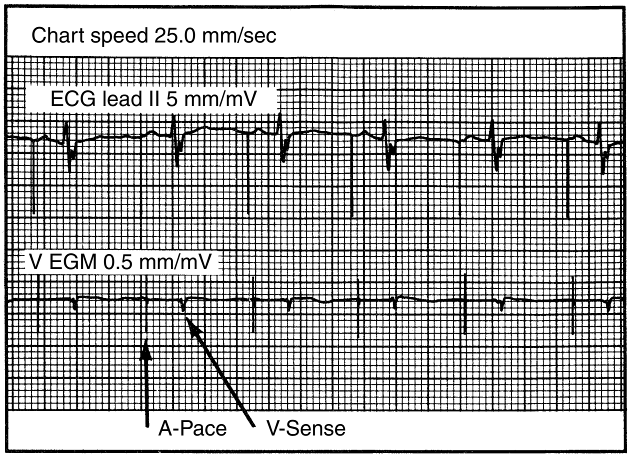


FIGURE 54.3 The surface ECG (ECG LEAD II) represents the sum total of the electrical potentials of all depolarizing tissue. The intracardiac electrogram (V EGM) shows only the potentials measured between the lead electrodes. This allows the evaluation of signals that may be hidden within the surface ECG.

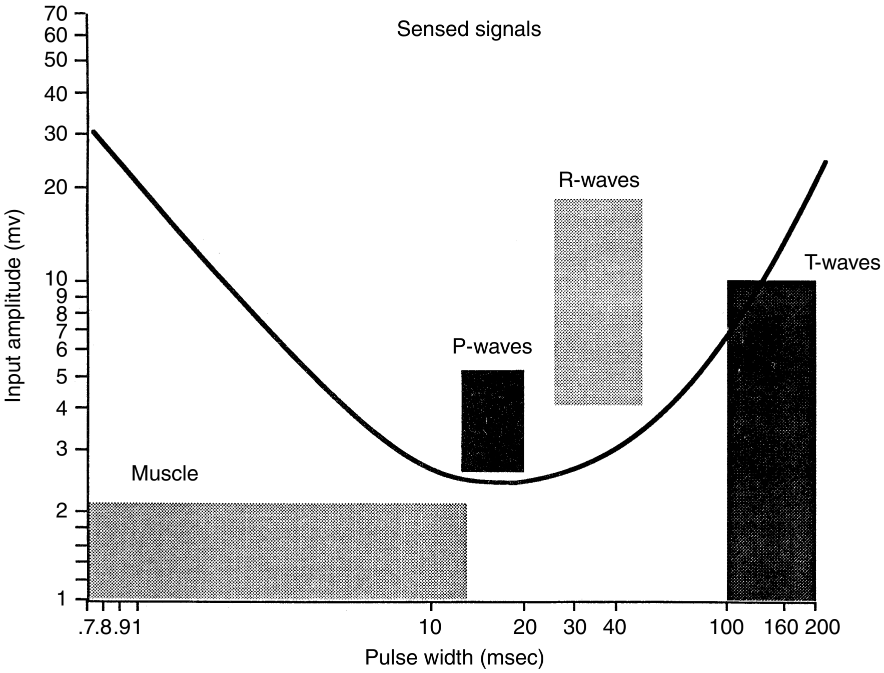


FIGURE 54.4 This is a conceptual depiction of the bandpass filter demonstrating the typical filtering of unwanted signals by discriminating between those with slew rates that are too low and/or too high.

In general, bipolar and unipolar sensing configurations have equal performance. A drawback of the unipolar approach is the increased possibility of sensing noncardiac signals: the large electrode separation may, for example, sense myopotentials from skeletal muscle movement, leading to inappropriate inhibition of pacing. Many newer pacemakers can be programmed to sense or pace in either configuration.

Once the electrogram enters the sensing circuit, it is scrutinized by a bandpass filter (Figure 54.4). The frequency of an R-wave is 10–30 Hz. The center frequency of most sensing amplifiers is 30 Hz. *T*-waves are slower, broad signals that are composed of lower frequencies (approximately 5 Hz or less). Far-field signals are also lower-frequency signals, whereas skeletal muscle falls in the range of 10–200 Hz.

At the implant, the voltage amplitude of the *R*-wave (and the *P*-wave, in the case of dual-chamber pacing) is measured to ensure the availability on an adequate signal. *R*-wave amplitudes are typically 5–25 mV, and *P*-wave amplitudes are 2–6 mV. The signals passing through the sense amplifier are compared to an adjustable reference voltage called the **sensitivity**. Any signal below the reference voltage is not sensed, and those above it are sensed. Higher-sensitivity settings (high-reference voltage) may lead to substandard sensing, and a lower reference voltage may result in oversensing. A minimum 2 : 1 safety margin should be maintained between the sensitivity setting and the amplitude of the intracardiac signal. The circuit is protected from extremely high voltages by a Zener diode.

The slope of the signal is also surveyed by the sensing circuit and is determined by the slew rate (the time rate of change in voltage). A slew rate that is too flat or too steep may be eliminated by the bandpass filter. On the average, the slew rate measured at implant should be between 0.75 and 2.50 V/sec.

The last line of defense in an effort to remove undesirable signals is to "blind" the circuit at specific times during the cardiac cycle. This is accomplished with blanking and refractory periods. Some of these periods are **programmable**. During the blanking period the sensing circuit is turned off, and during the refractory period the circuit can see the signal but does not initiate any of the basic timing intervals. Virtually all paced and sensed events begin concurrent blanking and refractory periods, typically ranging from 10 to 400 msec. These are especially helpful in dual-chamber pacemakers where there exists the potential for the pacing output of the atrial side to inhibit the ventricular pacing output, with dangerous consequences for patients in complete heart block.

Probably the most common question asked by the general public about pacing systems is the effect of electromagnetic interference (EMI) on their operation. EMI outside of the hospital is an infrequent problem, though patients are advised to avoid such sources of strong electromagnetic fields as arc welders, high-voltage generators, and radar antennae. Some clinicians suggest that patients avoid standing near antitheft devices used in retail stores. Airport screening devices are generally safe, though they may detect a pacemaker's metal case. Microwave ovens, ham radio equipment, video games, computers, and office equipment rarely interfere with the operation of modern pacemakers. A number of medical devices and procedures may on occasion do so, however; electrocautery, cardioversion and defibrillation, MRI, lithotripsy, diathermy, TENS units, and radiation therapy.

Pacemakers affected by interference typically respond with temporary loss of output or temporary reversion to asynchronous pacing (pacing at a fixed rate, with no inhibition from intrinsic cardiac events). The usual consequence for the patient is a return of the symptoms that originally led to the pacemaker implant.

54.2.2 Output Circuit

Pacing is the most significant drain on the pulse generator power source. Therefore, current drain must be minimized while maintaining an adequate safety margin between the **stimulation threshold** and the programmed output stimulus. Modern permanent pulse generators use constant voltage. The voltage remains at the programmed value while current fluctuates in relation to the source impedance.

Output energy is controlled by two programmable parameters, pulse amplitude and pulse duration. Pulse amplitudes range from 0.8 to 5 V and, in some generators, can be as high as 10 V (used for troubleshooting or for pediatric patients). Pulse duration can range from 0.05 to 1.5 msec. The prudent selection of these parameters will greatly influence the longevity of the pulse generator.

The output pulse is generated from the discharge of a capacitor charged by the battery. Most modern pulse generators contain a 2.8 V battery. The higher voltages are achieved using voltage multipliers (smaller capacitors used to charge the large capacitor). The voltage can be doubled by charging two smaller capacitors in parallel, with the discharge delivered to the output capacitor in series. Output pulses are emitted at a rate controlled by the timing circuit; output is commonly inhibited by sensed cardiac signals.

54.2.3 Timing Circuit

The timing circuit regulates such parameters as the pacing cycle length, refractory and blanking periods, pulse duration, and specific timing intervals between atrial and ventricular events. A crystal oscillator generating frequencies in the kHz range sends a signal to a digital timing and logic control circuit, which in turn operates internally generated clocks at divisions of the oscillatory frequency.

A rate-limiting circuit is incorporated into the timing circuit to prevent the pacing rate from exceeding an upper limit should a random component failure occur (an extremely rare event). This is also referred to as "runaway" protection and is typically 180 to 200 ppm.

54.2.4 Telemetry Circuit

Today's pulse generators are capable of both transmitting information from an RF antenna and receiving information with an RF decoder. This two-way communication occurs between the pulse generator and the programmer at approximately 300 Hz. Real-time telemetry is the term used to describe the ability of the pulse generator to provide information such as pulse amplitude, pulse duration, lead impedance, battery impedance, lead current, charge, and energy. The programmer, in turn, delivers coded messages to the pulse generator to alter any of the programmable features and to retrieve diagnostic data. Coding requirements reduce the likelihood of inappropriate programming alterations by environmental sources of radiofrequency and magnetic fields. It also prevents the improper use of programmers from other manufacturers.

54.2.5 Power Source

Over the years, a number of different battery technologies have been tried, including mercury-zinc, rechargeable silver-modified-mercuric-oxide-zinc, rechargeable nickel-cadmium, radioactive plutonium or promethium, and lithium with a variety of different cathodes. Lithium-cupric-sulfide and mercury-zinc batteries were associated with corrosion and early failure. Mercury-zinc produced hydrogen gas as a by-product of the battery reaction; the venting required made it impossible to hermetically seal the generator. This led to fluid infiltration followed by the risk of sudden failure.

The longevity of very early pulse generators was measured in hours. With the lithium-iodide technology now used, longevity has been reported as high as 15 years. The clinical desire to have a generator that is small and full-featured yet also long-lasting poses a formidable challenge to battery designers. One response by manufacturers has been to offer different models of generators, each offering a different balance between therapy, size, and longevity. Typical **battery capacity** is in the range of 0.8 to 3.0 amp-hours.

Many factors affect longevity, including pulse amplitude and duration, pacing rate, single- versus dual-chamber pacing, degree to which the patient uses the pacemaker, lead design, and static current drain from the sensing circuits. Improvements in lead design are often overlooked as a factor in improving longevity, but electrodes used in 1960 required a pulse generator output of 675 μJ for effective stimulation, whereas the electrodes of the 1990s need only 3 to 6 μJ.

Another important factor in battery design lies in the electrolyte that separates the anode and the cathode. The semisolid layer of lithium iodide that is used gradually thickens over the life of the cell, increasing the internal resistance of the battery. The voltage produced by lithium-iodine batteries is inversely related to this resistance and is linear from 2.8 V to approximately 2.4 V, representing about 90% of the usable battery life. It then declines exponentially to 1.8 V as the internal battery resistance increases from 10,000 to 40,000 Ω (Figure 54.5).

When the battery reaches between 2.0 and 2.4 V (depending on the manufacturer), certain functions of the pulse generator are altered so as to alert the clinician. These alterations are called the elective-replacement indicators (ERI). They vary from one pulse generator to another and include signature decreases in rate, a change to a specific pacing **mode**, pulse duration stretching, and the telemetered battery voltage. When the battery voltage reaches 1.8 V, the pulse generator may operate erratically or

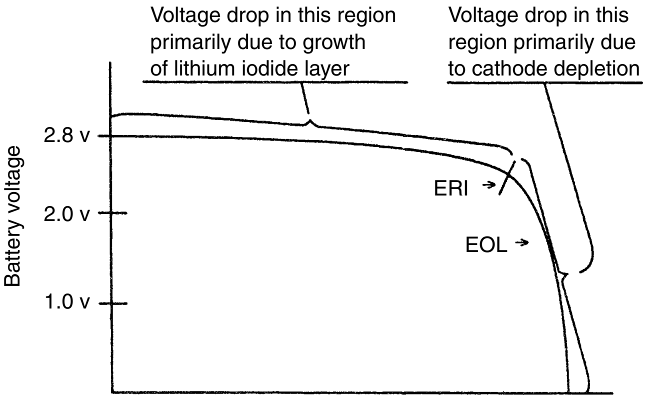


FIGURE 54.5 The initial decline in battery voltage is slow and then more rapid after the battery reaches the ERI voltage. An important aspect of battery design is the predictability of this decline so that timely generator replacement is anticipated.

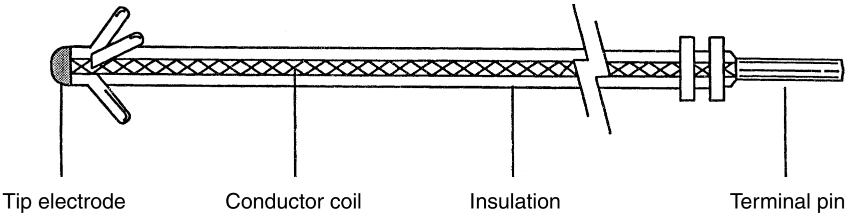


FIGURE 54.6 The four major lead components.

cease to function and is said to have reached "end of life." The time period between appearance of the ERI and end-of-life status averages about 3 to 4 months.

54.3 Leads

Implantable pacing leads must be designed not only for consistent performance within the hostile environment of the body but also for easy handling by the implanting physician. Every lead has four major components (Figure 54.6): the electrode, the conductor, the insulation, and the connector pin(s).

The electrode is located at the tip of the lead and is in direct contact with the myocardium. Bipolar leads have a tip electrode and a ring electrode (located about 2 cm proximal to the tip); unipolar leads have tip electrodes only. A small-radius electrode provides increased current density resulting in lower stimulation thresholds. The electrode also increases resistance at the electrode-myocardial interface, thus lowering the current drain further and improving battery longevity. The radius of most electrodes is 6 to 8 mm^2, though there are clinical trials underway using a "high-impedance" lead with a tip radius as low as 1.5 mm^2.

Small electrodes, however, historically have been associated with inferior sensing performance. Lead designers were able to achieve both good pacing and good sensing by creating porous-tip electrodes containing thousands of pores in the 20 to 100 μm range. The pores allow the ingrowth of tissue, resulting in the necessary increase in effective sensing area while maintaining a small pacing area. Some commonly used electrode materials include platinum-iridium. Elgiloy (an alloy of cobalt, iron, chromium, molybdenum, nickel, and manganese), platinum coated with platinized titanium, and vitreous or pyrolytic carbon coating a titanium or graphite core.

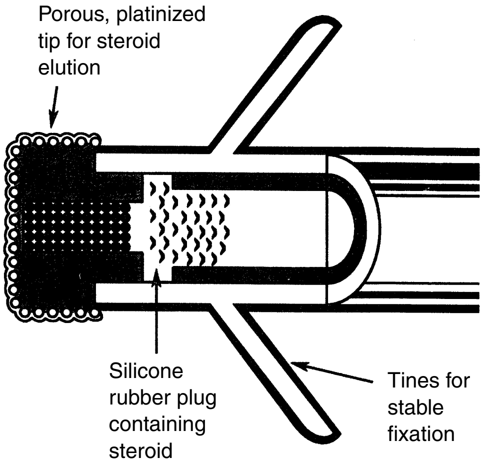

Porous, platinized tip for steroid elution
Silicone rubber plug containing steroid
Tines for stable fixation

FIGURE 54.7 The steroid elution electrode.

Another major breakthrough in lead design is the steroid-eluting electrode. About 1 mg of a corticosteroid (dexamethasone sodium phosphate) is contained in a silicone core that is surrounded by the electrode material (Figure 54.7). The "leaking" of the steroid into the myocardium occurs slowly over several years and reduces the inflammation that results from the lead placement. It also retards the growth of the fibrous sack that forms around the electrode which separates it from viable myocardium. As a result, the dramatic rise in acute thresholds that is seen with nonsteroid leads over the 8 to 16 weeks postimplant is nearly eliminated. This makes it possible to program a lower pacing output, further extending longevity.

Once a lead has been implanted, it must remain stable (or fixated). The fixation device is either active or passive. The active fixation leads incorporate corkscrew mechanisms, barbs, or hooks to attach themselves to the myocardium. The passive fixation leads are held into place with tines that become entangled into the netlike lining (trabeculae) of the heart. Passive leads generally have better acute pacing and sensing performance but are difficult to remove chronically. Active leads are easier to remove chronically and have the advantage of unlimited placement sites. Some implanters prefer to use active-fixation leads in the atrium and passive-fixation leads in the ventricle.

The conductor carries electric signals to the pulse generator and delivers the pacing pulses to the heart. It must be strong and flexible to withstand the repeated flexing stress placed on it by the beating heart. The early conductors were a single, straight wire that was vulnerable to fracturing. They have evolved into coiled (for increased flexibility) multifilar (to prevent complete failure with partial fractures) conductors. The conductor material is a nickel alloy called MP35N. Because of the need for two conductors, bipolar leads are usually larger in diameter than unipolar leads. Current bipolar leads have a coaxial design that has significantly reduced the diameter of bipolar leads.

Insulation materials (typically silicone and polyurethane) are used to isolate the conductor. Silicone has a longer history and the exclusive advantage of being repairable. Because of low tear strength, however, silicone leads tend to be thicker than polyurethane leads. Another relative disadvantage of silicone is its high coefficient of friction in blood, which makes it difficult for two leads to pass through the same vein. A coating applied to silicone leads during manufacturing has diminished this problem.

A variety of generator-lead connector configurations and adapters are available. Because incompatibility can result in disturbed (or even lost) pacing and sensing, an international standards (IS-1) has been developed in an attempt to minimize incompatibility.

Leads can be implanted epicardially and endocardially. *Epicardial* leads are placed on the outer surface of the heart and require the surgical exposure of a small portion of the heart. They are used when venous occlusion makes it impossible to pass a lead transvenously, when abdominal placement of the pulse generator is needed (as in the case of radiation therapy to the pectoral area), or in children (to allow for growth). *Endocardial* leads are more common and perform better in the long term. These leads are passed through the venous system and into the right side of the heart. The subclavian or cephalic veins in the pectoral region are common entry sites. Positioning is facilitated by a thin, firm wire stylet that passes through the central lumen of the lead, stiffening it. Fluoroscopy is used to visualize lead positioning and to confirm the desired location.

Manufacturers are very sensitive to the performance reliability of the leads. Steady improvements in materials, design, manufacturing, and implant technique have led to reliability well in excess of 99% over 3-year periods.

54.4 Programmers

Noninvasive reversible alteration of the functional parameters of the pacemaker is critical to ongoing clinical management. For a pacing system to remain effective throughout its lifetime, it must be able to adjust to the patient's changing needs. The programmer is the primary clinical tool for changing settings, for retrieving diagnostic data, and for conducting noninvasive tests.

The pacing rate for programmable pacemakers of the early 1960s was adjusted via a Keith needle manipulated percutaneously into a knob on the side of the pacemaker; rotating the needle changed the pacing rate. Through the late 1960s and early 1970s, magnetically attuned reed switches in the pulse generator made it possible to noninvasively change certain parameters such as rate, output, sensitivity, and polarity. The application of a magnet could alter the parameters which were usually limited to only one of two choices. It was not until the late 1970s, when radiofrequency energy was incorporated as the transmitter of information, that programmability began to realize its full potential. Radiofrequency transmission is faster, provides bidirectional telemetry, and decreases the possibility of unintended programming from inappropriate sources.

Most manufacturers today are moving away from a dedicated proprietary instrument and toward a PC-based design. The newer designs are generally more flexible, more intuitive to use, and more easily updated when new devices are released. Manufacturers and clinicians alike are becoming more sensitive to the role that time-efficient programming can play in the productivity of pacing clinics, which may provide follow-up for as many as 500 to 1000 patients a year.

54.5 System Operation

Pacemakers have gotten steadily more powerful over the last three decades, but at the cost of steadily greater complexity. Manufacturers have come to realize the challenge that this poses for busy clinicians and have responded with a variety of interpretive aids (Figure 54.8).

Much of the apparent complexity of the timing rules that determine pacemaker operation is due to a design goal of mimicking normal cardiac function without interfering with it. One example is the dual-chamber feature that provides sequential stimulation of the atrium before the ventricle.

Another example is rate response, designed for patients who lack the normal ability to increase their heart rate in response to a variety of physical conditions (e.g., exercise). Introduced in the mid-1980s, rate-responsive systems use some sort of sensor to measure the change in a physical variable correlated to heart rate. The sensor output is signal-processed and then used by the output circuit to specify a target pacing rate. The clinician controls the aggressiveness of the rate increase through a variety of parameters (including a choice of transfer function); pacemaker-resident diagnostics provide data helpful in titrating the rate-response therapy.

The most common sensor is the activity sensor, which uses piezoelectric materials to detect vibrations caused by body movement. Systems using a transthoracic-impedance sensor to estimate pulmonary **minute ventilation** are also commercially available. Numerous other sensors (e.g., stroke volume, blood temperature or pH, oxygen saturation, preejection interval, right ventricular pressure) are in various

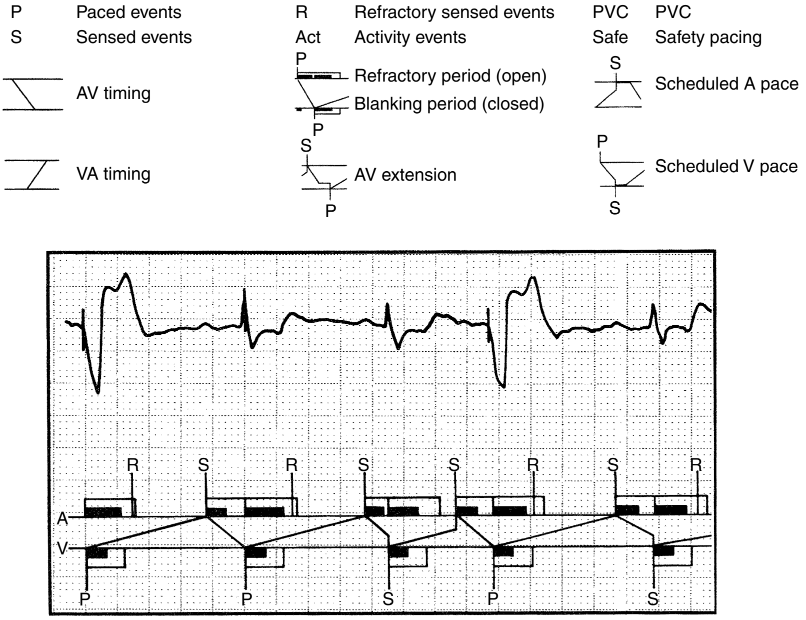


FIGURE 54.8 The Marker Channel Diagram is just one tool that makes interpretation of the ECG strip faster and more reliable for the clinician. It allows quick checking of the timing operations of the system.

stages of clinical research or have been market released outside the United States. Some of these systems are dual-sensor, combining the best features of each sensor in a single pacing system.

To make it easier to understand the gross-level system operation of modern pacemakers, a five-letter code has been developed by the North American Society of Pacing and Electrophysiology and the British Pacing and Electrophysiology Group [Bernstein et al., 1987]. The first letter indicates the chamber (or chambers) that are paced. The second letter reveals those chambers in which sensing takes place, and the third letter describes how the pacemaker will respond to a sensed event. The pacemaker will "inhibit" the pacing output when intrinsic activity is sensed or will "trigger" a pacing output based on a specific previously sensed event. For example, in DDD mode:

D: Pacing takes place in the atrium and the ventricle.
D: Sensing takes place in the atrium and the ventricle.
D: Both inhibition and triggering are the response to a sensed event. An atrial output is inhibited with an atrial-sensed event, whereas a ventricular output is inhibited with a ventricular-sensed event; a ventricular pacing output is triggered by an atrial-sensed event (assuming no ventricular event occurs during the A-V interval).

The fourth letter in the code is intended to reflect the degree of programmability of the pacemaker but is typically used to indicate that the device can provide rate response. For example, a DDDR device is one that is programmed to pace and sense in both chambers and is capable of sensor-driven rate variability. The fifth letter is reserved specifically for antitachycardia functions (Table 54.1).

Readers interested in the intriguing details of pacemaker timing operations are referred to the works listed at the end of this chapter.

54.6 Clinical Outcomes and Cost Implications

The demonstrable hemodynamic and symptomatic benefits provided by rate-responsive and dual-chamber pacing have led U.S. physicians to include at least one of these features in over three-fourths

TABLE 54.1 The NASPE/NPEG Code

Position	I	II	III	IV	V
Category	Chamber(s) paced	Chamber(s) sensed	Response to sensing	Programmability rate modulation	Antitachyarrhythmia function(s)
	O = None	O = None	O = None	O = None	O = None
	A = Atrium	A = Atrium	T = Triggered	P = Simple programmable	P = Packing
	V = Ventricle	V = Ventricle	I = Inhibited	M = Multiprogrammable	S = Shock
	D = Dual (A + V)	D = Dual (A + V)	D = Dual (T + I)	C = Communicating	D = Dual (P + S)
				R = Rate modulation	
Manufacturers' designation only	S = Single (A or V)	S = Single (A or V)			

Note: Positions I through III are used exclusively for antibradyarrhythmia function.
Source: From Bernstein A.D., et al., *PACE*, Vol. 10, July–Aug. 1987.

of implants in recent years. Also, new prospective data [Andersen et al., 1993] support a hypothesis investigated retrospectively since the mid-1980s: namely, that pacing the atrium in patients with sinus node dysfunction can dramatically reduce the incidence of such life-threatening complications as **congestive heart failure** and stroke associated with chronic **atrial fibrillation.** Preliminary analysis of the cost implications suggest that dual-chamber pacing is significantly cheaper to the U.S. healthcare system than is single-chamber pacing over the full course of therapy, despite the somewhat higher initial cost of implanting the dual-chamber system.

54.7 Conclusion

Permanent cardiac pacing is the beneficiary of three decades of advances in a variety of key technologies: biomaterials, electrical stimulation, sensing of bioelectrical events, power sources, microelectronics, transducers, signal analysis, and software development. These advances, informed and guided by a wealth of clinical experience acquired during that time, have made pacing a cost-effective cornerstone of cardiac arrhythmia management.

Defining Terms

Atrial fibrillation: An atrial arrhythmia resulting in chaotic current flow within the atria. The effective contraction of the atria is lost, allowing blood to pool and clot, leading to stroke if untreated.

Battery capacity: Given by the voltage and the current delivery. The voltage is a result of the battery chemistry, and current delivery (current × time) is measured in ampere hours and is related to battery size.

CMOS circuit: Abbreviation for complementary metallic oxide semiconductor, which is a form of semiconductor often used in pacemaker technology.

Congestive heart failure: The pathophysiologic state in which an abnormality of cardiac function is responsible for the failure of the heart to pump blood at a rate commensurate with the requirements of the body.

Endocardium: The inner lining of the heart.

Epicardium: The outer lining of the heart.

Hermeticity: The term, as used in the pacemaker industry, refers to a very low rate of helium gas leakage from the sealed pacemaker container. This reduces the chance of fluid intruding into the pacemaker generator and causing damage.

Hypertrophic obstructive cardiomyopathy: A disease of the myocardium characterized by thickening (hypertrophy) of the interventricular septum, resulting in the partial obstruction of blood from the left ventricle.

Minute ventilation: Respiratory rate × tidal volume (the amount of air taken in with each breath) = minute ventilation. This parameter is used as a biologic indicator for rate-adaptive pacing.

Mode: The type of pacemaker response to the patient's intrinsic heartbeat. The three commonly used modes are asynchronous, demand, and triggered.

Programmable: The ability to alter the pacemaker settings noninvasively. A variety of selections exist, each with its own designation.

Rate-adaptive: The ability to change the pacemaker stimulation interval caused by sensing a physiologic function other than the intrinsic atrial rhythm.

Sensitivity: A programmable setting that adjusts the reference voltage to which signals entering the sensing circuit are compared for filtering.

Stimulation threshold: The minimum output energy required to consistently "capture" (cause depolarization) of the heart.

References

Andersen H.R., Thuesen L., Bagger J.P. et al. (1993). Atrial versus ventricular pacing in sick sinus syndrome: a prospective randomized trial in 225 consecutive patients. *Eur. Heart. J.* 14: 252.

Bernstein A.D., Camm A.J., Fletcher R.D. et al. (1987). The NASPE/BPEG generic pacemaker code for antibradyarrhythmia and adaptive-rate pacing and antitachyarrhythmia devices. *PACE* 10: 794.

Committee on Pacemaker Implantation (1991). Guidelines for implantation of cardiac pacemakers and antiarrhythmic devices. *J. Am. Coll. Cardiol.* 18: 1.

Further Information

A good basic introduction to pacing from a clinical perspective is the third edition of *A Practical Guide to Cardiac Pacing* by H. Weston Moses, Joel Schneider, Brain Miller, and George Taylor (Little, Brown, 1991).

Cardiac Pacing (Blackwell Scientific, 1992), edited by Kenneth Ellenbogen, is an excellent intermediate treatment of pacing. The treatments of timing cycles and troubleshooting are especially good.

In-depth discussion of a wide range of pacing topics is provided by the third edition of *A Practice of Cardiac Pacing* by Seymour Furman, David Hayes, and David Holmes (Futura, 1993), and by *New Perspectives in Cardiac Pacing 3*, edited by Serge Barold and Jacques Mugica (Futura, 1993).

Detailed treatment of rate-responsive pacing is given in *Rate-Adaptive Cardiac Pacing: Single and Dual Chamber* by Chu-Pak Lau (Futura, 1993), and in *Rate-Adaptive Pacing*, edited by David Benditt (Blackwell Scientific, 1993).

The Foundations of Cardiac Pacing, Part I by Richard Sutton and Ivan Bourgeois (Futura, 1991) contains excellent illustrations of implantation techniques.

Readers seeking a historical perspective may wish to consult "Pacemakers, Pastmakers, and the Paced: An Informal History from A to Z," by Dwight Harken in the July/August 1991 issue of *Biomedical Instrumentation and Technology.*

PACE is the official journal of the North American Society of Pacing and Electrophysiology (NASPE) and of the International Cardiac Pacing and Electrophysiology Society. It is published monthly by Futura Publishing (135 Bedford Road, PO Box 418, Armonk, NY 10504 USA).

55

Noninvasive Arterial Blood Pressure and Mechanics

Gary Drzewiecki
Rutgers University

55.1 Introduction

The beginnings of noninvasive arterial pulse recording can be traced to the Renaissance. At that time, the Polish scientist, Strus (1555) had proposed that the arterial pulse possesses a waveform. Although instrumentation that he used was simple, he suggested that changes in the arterial pulse shape and strength may be related to disease conditions. Today, even though the technology is more advanced, noninvasive arterial blood pressure measurement still remains a challenge. Rigorous methods for extracting functional cardiovascular information from noninvasive pressure have been limited.

In this chapter, the most standard of noninvasive methods for arterial pressure measurements will be reviewed and future trends will be proposed. Two types of methods for noninvasive arterial pressure measurement may be defined; those that periodically sample and those that continuously record the pulse waveform. The sampling methods typically provide systolic and diastolic pressure and sometimes mean pressure. These values are collected over different heart beats during the course of 1 min. The continuous recording methods provide beat-to-beat measurements and often, the entire waveform. Some continuous methods only provide pulse pressure waveform and timing information.

The knowledge of systolic and diastolic pressure is fundamental for the evaluation of basic cardiovascular function and identifying disease. The choice of method depends on the type of study. For example, high blood pressure is a known precursor to many other forms of cardiovascular disease. A noninvasive method that samples blood pressure over the time course of months is usually adequate to study the progression of hypertension. The occlusive cuff-based methods fall into this category. These methods have been automated with recent instruments designed for ambulatory use [Graettinger et al., 1988]. Twenty-four to forty-eight hours ambulatory monitors have been applied to monitor the diurnal variation of a patient's blood pressure. This type of monitoring can alleviate the problem of "white coat hypertension," that is, the elevation of blood pressure associated with a visit to the physician's office [Pickering et al., 1988].

Short-term hemodynamic information obtained from the noninvasive arterial pulse waveform is a virtually untapped arena. While a great deal of knowledge has been gathered on the physics of the arterial pulse [Noordergraaf, 1978], it has been lacking in application because continuous pressure waveform monitors have not been available. A review of methods for continuous pulse monitoring that fill this gap will be provided.

Some applications for pulse monitoring can be found. In the recording time span of less than 1 min, the importance of the pressure waveform dominates, as well as beat-to-beat variations. This type of monitoring is critical in situations where blood pressure can alter quickly, such as due to trauma or anesthesia. Other applications for acute monitoring have been in aerospace, biofeedback, and lie detection. Moreover, pulse dynamics information becomes available such as wave reflection [Li, 1986]. Kelly et al. [1989] have shown that elevated systolic pressure can be attributed to pulse waveform changes due to a decrease in arterial compliance and increased wave reflection. Lastly, information on the cardiovascular control process can be obtained from pulse pressure variability [Omboni et al., 1993].

It has become increasingly popular to provide simultaneous recording of such variables as noninvasive blood pressure, oxygen saturation via pulse oximetry, body temperature, etc., in a single instrument. It is apparent that the advances in computer technology impinge on this practice, making it a clear trend. While this practice is likely to continue, the forefront of this approach will be those instruments that provide more than just a mere marriage of technology in a single design. It will be possible to extract functional information in addition to just pressure. As an example of this, a method for noninvasive measurement of the arterial pressure–lumen area curve will be provided.

55.2 Long-Term Sampling Methods

55.2.1 Vascular Unloading Principle

The vascular unloading principle is fundamental to all occlusive cuff-based methods of determining arterial blood pressure. It is performed by applying an external compression pressure or force to a limb such that it is transmitted to the underlying vessels. It is usually assumed that the external pressure and the tissue pressure (stress) are in equilibrium. The underlying vessels are then subjected to altered transmural pressure (internal minus external pressure) by varying the external pressure. It is further assumed [Marey, 1885] that the tension within the wall of the vessel is zero when transmural pressure is zero. Hence, the term vascular unloading originated.

Various techniques have been developed that attempt to detect vascular unloading. These generally rely on the fact that once a vessel is unloaded, further external pressure will cause it to collapse. In summary,

$$\text{If } P_a > P_c \Rightarrow \text{Lumen open}$$

or

$$\text{If } P_a > P_c \Rightarrow \text{Lumen closed}$$

where P_a is the arterial pressure and P_c is the cuff pressure. Most methods that employ the occlusive arm cuff rely on this principle and differ in the means of detecting whether the artery is open or closed. Briefly, some approaches are the skin flush, palpatory, Korotkoff (auscultatory), oscillometric, and ultrasound methods [Drzewiecki et al., 1987]. Of these, the methods of Korotkoff and oscillometry are in most common use and will be reviewed here.

The idea of using lumen opening and closure as an indication of blood pressure has survived since its introduction by Marey. This simple concept is complicated by the fact that lumen closure may not necessarily occur at zero transmural pressure. Instead, transmural pressure must be negative by 5 to 20 mmHg for complete closure. *In vitro* and *in vivo* human measurements have revealed that vessel buckling more closely approximates zero transmural pressure [Drzewiecki et al., 1997; Drzewiecki and Pilla, in press]. Buckling may be defined as the point at which the vessel switches from wall stretch to wall bending as a means of supporting its pressure load. The vessel is maximally compliant at this point and is approximately 25% open.

The validity of employing the buckling concept to blood pressure determination was tested. Sheth and Drzewiecki [1998] employed feedback control to regulate the pressure in a flexible diaphragm tonometer (see the section on *flexible diaphragm tonometer*). The buckling point was determined from the volume pulse. Thus, to detect vessel buckling, the derivative of the volume pulse with respect to mean pressure was computed. This was performed on a beat-to-beat basis. The feedback system was then used to adjust pressure such that this derivative was maximized, indicating the point of greatest instability. Thus, the underlying blood vessel was maintained in a constant state of buckling. According to the buckling theory, the transmural pressure should be nearly zero and tonometer pressure is equal to arterial pressure. A sample 2 min recording of noninvasive arterial pressure by this approach is shown (Figure 55.1). The method was capable of tracking the subject's blood pressure in response to a Valsalva maneuver in this same record. This test demonstrates the feasibility of employing buckling to measure blood pressure. This example also illustrates that beat-to-beat pressure can be obtained by this method without the necessity to occlude blood flow. This approach should be useful for pressure variability studies.

55.2.2 Occlusive Cuff Mechanics

The occlusive arm cuff has evolved more out of convenience than engineering design. As such, its mechanical properties are relatively unknown. The current version of the cuff is designed to encircle the upper arm. It consists of a flat rubber bladder connected to air supply tubing. The bladder is covered externally by cloth material with Velcro fasteners at either end for easy placement and removal. While the cuff encircles

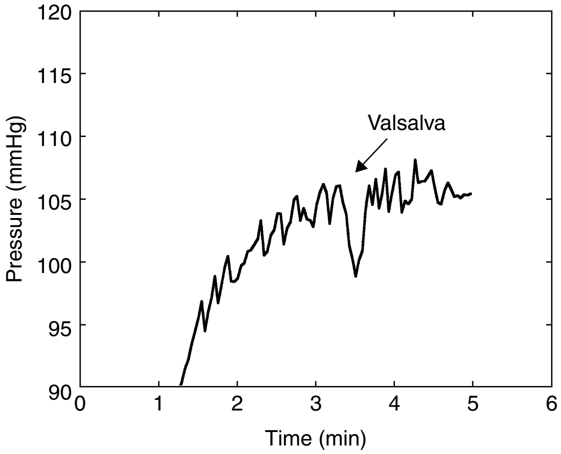


FIGURE 55.1 Application of arterial buckling to the continuous measurement of arterial pressure. The record shown tracks the subject's mean pressure in the brachial artery. The initial portion of the record illustrates the system as it locates the buckling point and the subject's blood pressure. The subject was directed to perform a brief Valsalva maneuver mid-way through the recording.

the entire arm, the bladder extends over approximately half the circumference. The bladder is pressurized with air derived from a hand pump, release valve, and manometer connected to the air supply tubing.

The cuff should accurately transmit pressure down to the tissue surrounding the brachial artery. A mechanical analysis revealed that the length of the cuff is required to be a specific fraction of the arm's circumference for pressure equilibrium [Alexander et al., 1977]. A narrow cuff resulted in the greatest error in pressure transmission and, thus, the greatest error in blood pressure determination. Geddes and Whistler [1978] experimentally examined the effect of cuff size on blood pressure accuracy for the Korotkoff method. Their measurements confirmed that a cuff-width-to-arm circumference ratio of 0.4 should be maintained. Cuff manufacturers, therefore, supply a range of cuff sizes appropriate for pediatric use up to large adults.

Another aspect of the cuff is its pressure response due to either internal air volume change or that of the limb. In this sense, the cuff can be thought of as a plethysmograph. It was examined by considering its pressure–volume characteristics. In an experiment, the cuff properties were isolated from that of the arm by applying it to a rigid cylinder of similar diameter [Drzewiecki et al., 1993]. Pressure–volume data were then obtained by injecting a known volume of air and noting the pressure. This was performed for a standard bladder cuff (13 cm width) over a typical range of blood pressure.

The cuff pressure–volume results for pressures less than 130 mmHg are nonlinear (Figure 55.2). For higher pressures, the data asymptotically approach a linear relationship. The cuff volume sensitivity, that

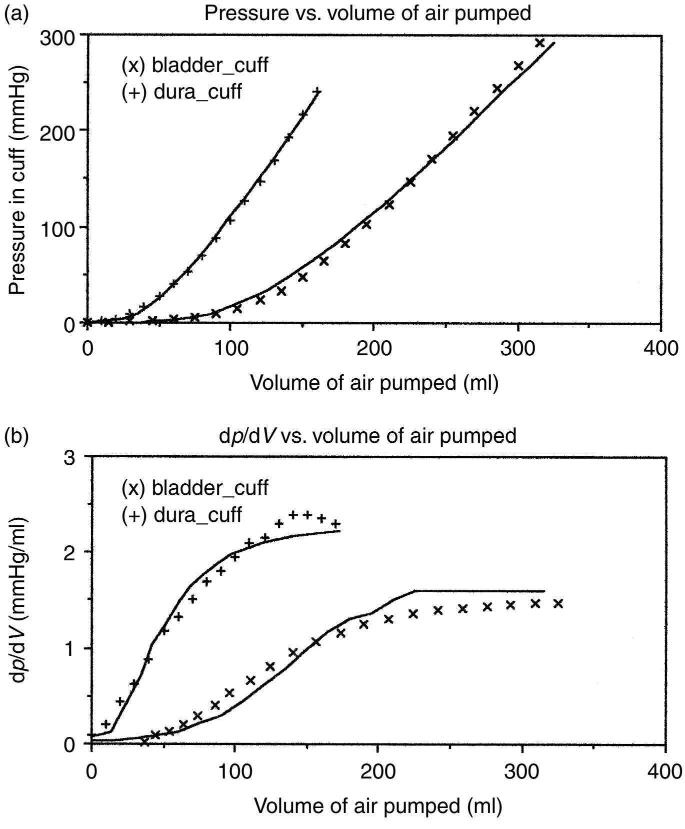


FIGURE 55.2 (a) Pressure–volume data obtained from two different occlusive arm cuffs. Inner surface of the cuff was fixed in this case to isolate cuff mechanics from that of the arm. (b) Derivative of the cuff pressure with respect to volume obtained from pressure–volume data of both cuffs. These curves indicate the pressure response of the cuff to volume change and are useful for plethysmography. Solid curves in both figures are the results of the occlusive cuff model.

is, its derivative with respect to volume, increased with cuff pressure (Figure 55.2). Above 130 mmHg, the cuff responded with nearly constant sensitivity.

A cuff mechanics theory was developed to explain the above cuff experiment [Drzewiecki et al., 1993]. Cuff mechanics was theorized to consist of two components. The first consists of the compressibility of air within the cuff. This was modeled using Boyle's gas law. The second component consists of elastic and geometric deformation of the cuff bladder. Cuff shape deformation proceeds at low pressures until the bladder reaches its final geometry, rendering a curved pressure–volume relationship. Then, elastic stretch of the rubber bladder takes over at high pressures, resulting in a nearly linear relationship. Solutions for this model are shown in comparison with the data in Figure 55.2 for two cuffs; a standard cuff and the Critikon Dura-cuf. This model was useful in linearizing the cuff for use as a plethysmograph and for application to oscillometry (below).

55.2.3 Method of Korotkoff

The auscultatory method or method of Korotkoff was introduced by the Russian army physician N. Korotkoff [1905]. In his experiments, Korotkoff discovered that sound emitted distally from a partially occluded limb. He realized that this sound was indicative of arterial flow and that together with the occlusive cuff could be used to determine blood pressure. The method, as employed today, utilizes a stethoscope placed distal to an arm cuff over the brachial artery at the antecubital fossa. The cuff is inflated to about 30 mmHg above systolic pressure and then allowed to deflate at a rate of 2 to 3 mm Hg/sec. With falling cuff pressure, sounds begin and slowly change their characteristics. The initial "tapping" sounds are referred to as Phase I Korotkoff sound and denote systolic pressure. The sounds increase in loudness during Phase II. The maximum intensity occurs in Phase III, where the tapping sound may be followed by a murmur due to turbulence. Finally, Phase IV Korotkoff sound is identified as muffled sound, and Phase V is the complete disappearance of sound. Phase IV is generally taken to indicate diastolic arterial pressure. But, Phase V has also been suggested to be a more accurate indication of diastolic pressure. This matter is a continued source of controversy. The long history of the Korotkoff sound provides much experimental information. For example, the frequency spectrum of sound, spatial variation along the arm, filtering effects, and timing are reviewed [Drzewiecki et al., 1989].

It is a long-held misconception that the origin of the Korotkoff sound is flow turbulence. Turbulence is thought to be induced by the narrowing of the brachial artery under an occlusive cuff as it is forced to collapse. There are arguments against this idea. First, the Korotkoff sounds do not sound like turbulence, that is, a murmur. Second, the Korotkoff sound can occur in low blood flow situations, while turbulence cannot. And, last, Doppler ultrasound indicates that peak flow occurs following the time occurrence of Korotkoff sound. An alternative theory suggests that the sound is due to nonlinear distortion of the brachial pulse, such that sound is introduced to the original pulse. This is shown to arise from flow limitation under the cuff in addition to curvilinear pressure–area relationship of the brachial artery [Drzewiecki et al., 1989]. Strong support for this theory comes from its ability to predict many of the Korotkoff sound's observable features.

The accuracy of the Korotkoff method is well known. London and London [1967] find that the Korotkoff method underestimates systolic pressure by 5 to 20 mmHg and overestimates diastolic pressure by 12 to 20 mmHg. However, certain subject groups, such as hypertensives or the elderly, can compound these errors [Spence et al., 1978]. In addition, it has been shown that the arm blood flow can alter the Korotkoff sound intensity and, thus, the accuracy of blood pressure measurement [Rabbany et al., 1993]. Disappearance of Korotkoff sound occurs early in Phase III for some subjects and is referred to as the auscultatory gap. This causes an erroneous indication of elevated diastolic pressure. The auscultatory gap error can be avoided by simply allowing cuff pressure to continue to fall, where the true Phase IV sounds return. This is particularly critical for automatic instruments to take into account. In spite of these errors, the method of Korotkoff is considered a documented noninvasive blood pressure standard by which other noninvasive methods may be evaluated [White et al., 1993].

The Korotkoff method is applicable to other vessels besides the brachial artery of the arm. For example, the temporal artery has been employed [Shenoy et al., 1993]. In this case, a pressure capsule is applied over the artery on the head in place of an occlusive cuff, to provide external pressure. This approach has been shown to be accurate and is applicable to aerospace. Pilots' cerebral vascular pressure often falls in high acceleration maneuvers, so that temporal artery pressure is a better indicator of this response.

55.2.4 Oscillometry

Oscillometric measurement of blood pressure predates the method of Korotkoff. The French physiologist Marey [1885] placed the arm within a compression chamber and observed that the chamber pressure fluctuated with the pulse. He also noted that the amplitude of pulsation varied with chamber pressure. Marey believed that the maximum pulsations or the onset of pulsations were associated with equality of blood pressure and chamber pressure. At that time, he was not certain what level of arterial pressure the maximum pulsations corresponded with. Recently, it has been demonstrated theoretically that the variation in cuff pressure pulsation is primarily due to the brachial artery buckling mechanics [Drzewiecki et al., 1994].

Today, oscillometry is performed using a standard arm cuff together with an in-line pressure sensor. Due to the requirement of a sensor, oscillometry is generally not performed manually, but, rather, with an automatic instrument. The recorded cuff pressure is high-pass-filtered above 1 Hz to observe the pulsatile oscillations as the cuff slowly deflates (Figure 55.3). It has been determined only recently that the maximum oscillations actually correspond with cuff pressure equal to mean arterial pressure (MAP) [Posey et al., 1969; Ramsey, 1979], confirming Marey's early idea. Systolic pressure is located at the point where the oscillations, O_s, are a fixed percentage of the maximum oscillations, O_m [Geddes et al., 1983]. In comparison with the intra-arterial pressure recordings, the systolic detection ratio is $O_s/O_m = 0.55$.

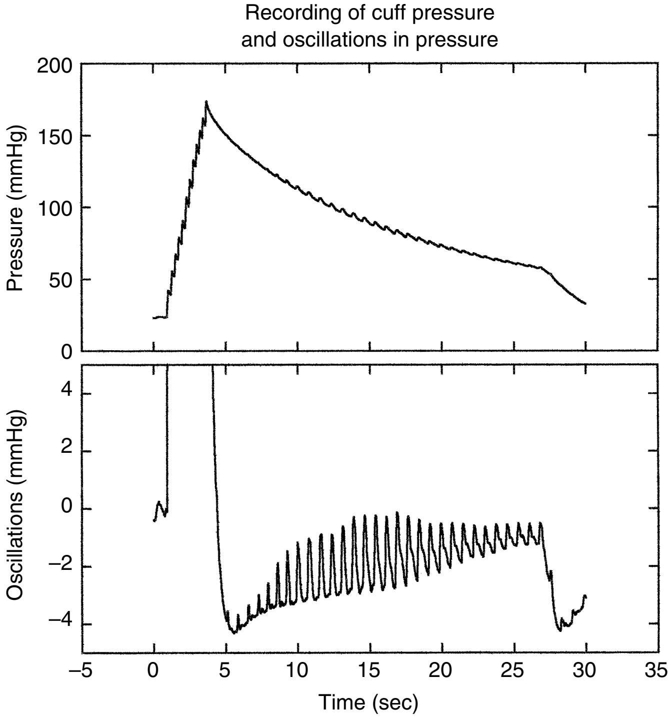


FIGURE 55.3 Sample recording of cuff pressure during oscillometric blood pressure measurement. Bottom panel shows oscillations in cuff pressure obtained by high pass filtering above 1/2 Hz.

Similarly, the diastolic pressure can be found as a fixed percentage of the maximum oscillations, as $O_d/O_m = 0.85$.

55.2.5 Derivative Oscillometry

The use of the oscillometric detection ratios to find systolic and diastolic pressure is an empirical approach. That is, the ratios are statistically valid over the population of subjects that form the total sample. This is a distinctly different approach than measuring blood pressure by detecting an event, such as the maximum in cuff pressure oscillations or the occurrence of Korotkoff sound. The event is constrained by the physics of arterial collapse under the cuff [Drzewiecki et al., 1994]. Therefore, it is more likely to be accurate under different conditions and, more importantly, for subjects outside of the sample population. In fact, the oscillometric determination of MAP is more accurate than systolic and diastolic pressures.

Drzewiecki et al. [1994] employed a model of oscillometry to evaluate the derivative of the oscillation amplitude curve (Figure 55.3) with respect to cuff pressure. When this derivative is plotted against cuff pressure, it was found that it reaches a maximum positive value. This occurred when cuff pressure equals diastolic pressure. Additionally, the minimum negative value was found to occur at systolic pressure. A measurement performed in our lab on a single subject illustrates the approach (Figure 55.4). The specific advantage offered is that the empirically based systolic and diastolic ratios are not necessary [Link, 1987]. This method may be referred to as derivative oscillometry.

Derivative oscillometry was evaluated experimentally in our lab. The values of systolic and diastolic pressures obtained by derivative oscillometry were compared with those obtained by the method of Korotkoff. Thirty recordings were obtained on normal subjects (Figure 55.5). The results indicated a high correlation of 0.93 between the two methods. Systolic mean error was determined to be 9% and diastolic mean error was −6%. Thus, derivative oscillometry was found to compare well with the method of Korotkoff in this preliminary evaluation. Before adopting derivative oscillometry, a more complete evaluation needs to be performed using a greater and more diverse subject population.

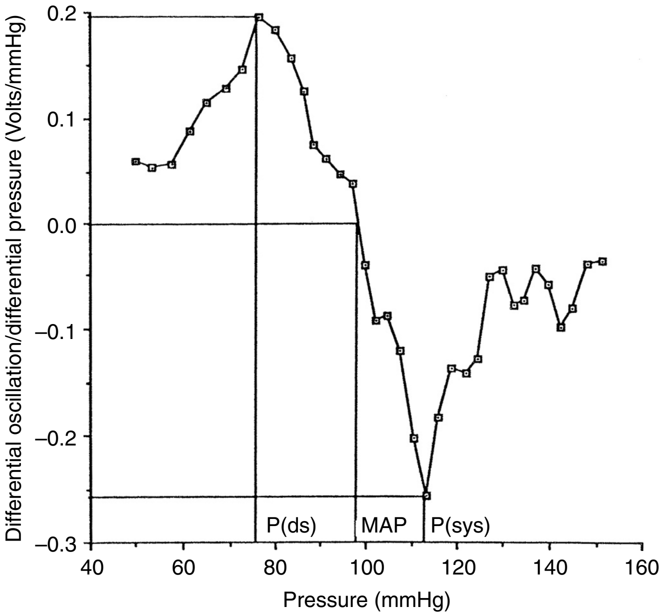


FIGURE 55.4 Method of derivative oscillometry. The derivative of cuff pressure oscillations data with respect to cuff pressure is shown from a single subject. The maximum and minimum values denote diastolic and systolic pressure, respectively. A zero derivative indicates MAP in this plot.

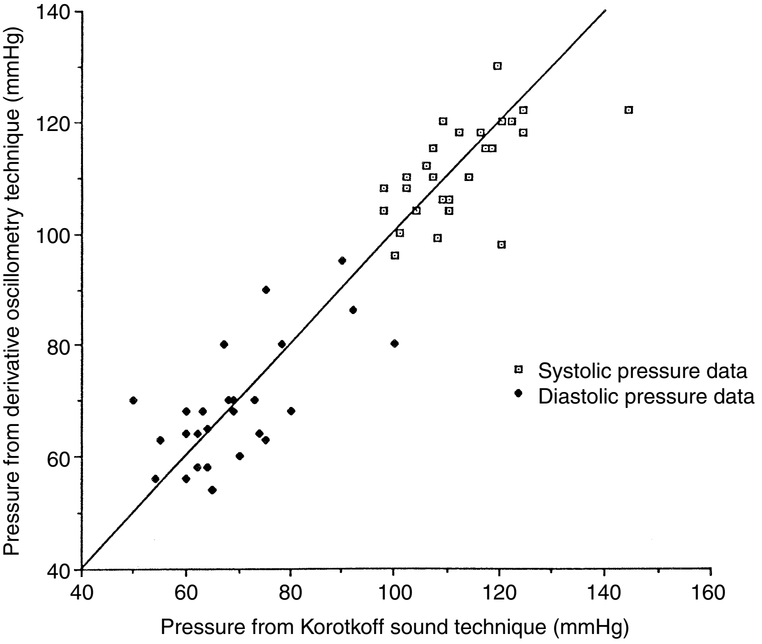


FIGURE 55.5 Experimental evaluation of derivative oscillometry using systolic and diastolic pressure from the method of Korotkoff as reference. The line indicates the result of linear regression to the data.

55.3 Pulse Dynamics Methods

55.3.1 R-Wave Time Interval Technique

One of the basic characteristics of pressure and flow under an occlusive cuff is that the pulse is apparently delayed with increasing cuff pressure. The R-wave of the ECG is often employed as a time reference. Arzbaecher et al. [1973] measured the time delay of the Korotkoff sound relative to the R-wave. They suggested that the curve obtained by plotting this data represents the rising portion of the arterial pulse waveform.

A Korotkoff sound model Drzewiecki et al., 1989 was employed to investigate the cuff delay effect. Time delay of the Korotkoff sound was computed relative to the proximal arterial pulse. Since the arterial pulse waveform was known in this calculation, the pressure-RK interval curve can be compared directly. The resemblance to a rising arterial pulse was apparent but some deviation was noted, particularly in the early portion of the RK interval curve. The model indicates that the pulse occurs earlier and with higher derivative than the true pulse waveform. In particular, the increased derivative that occurs at the foot of the pulse may mislead any study of wave propagation.

Recently, Sharir et al. [1993] performed comparisons of Doppler flow pulse delay and direct arterial pressure recordings. Their results confirm a consistent elevation in pulse compared with intra-arterial recording in the early portions of the wave. The average deviation was 10 mmHg in value. Hence, if accuracy is not important, this method can provide a reasonable estimate of blood pressure and its change. The commercial pulse watch has been a recent application of this approach.

55.3.2 Continuous Vascular Unloading

Penaz [1973] reasoned that if the cuff pressure could be continuously adjusted to equal the arterial pressure, the vessels would be in a constant state of vascular unloading. He employed mechanical feedback to continuously adjust the pressure in a finger chamber to apply this concept. The vascular volume was measured using photoplethysmography. When feedback was applied such that the vascular volume was held constant and at maximum pulsatile levels, the chamber pressure waveform was assumed to be equal to the arterial pressure.

Recent applications of the Penaz method have been developed by Wesseling et al. [1978] and Yamakoshi et al. [1980]. One instrument is commercially available as the FINAPRES [Ohmeda, Finapres, Englewood, CO]. These instruments have been evaluated in comparison with intra-arterial pressure recordings [Omboni et al., 1993]. Good waveform agreement has been obtained in comparison with intra-arterial measurements from the radial artery. But, it should be clear that the Penaz method employs finger pressure, which is a peripheral vascular location. This recording site is prone to pulse wave reflection effects and is therefore sensitive to the vascular flow resistance. It is anticipated that the technique would be affected by skin temperature, vasoactive drugs, and anesthetics. Moreover, mean pressure differences between the finger pulse and central aortic pressure should be expected.

55.3.3 Pulse Sensing

Pulse sensors attempt to measure the arterial pulse waveform from either the arterial wall deflection or force at the surface of the skin above a palpable vessel. Typically, these sensors are not directly calibrated in terms of pressure, but ideally respond proportionately to pressure. As such, they are primarily useful for dynamic information. While there are many designs available for this type of sensor, they generally fall into two categories. The first category is that of the volume sensor (Figure 55.6 [left]). This type of sensor relies on the adjacent tissues surrounding the vessel as a non-deflecting frame of reference as, for example, a photoplethysmograph. Skin deflections directly above the vessel are then measured relative to a reference frame to represent the arterial pressure. Several different transduction methods may be employed such as capacitive, resistive, inductive, optical, etc. Ideally, this type of sensor minimally restricts the motion of the skin so that contact force is zero. The drawback to pulse volume sensing is that the method responds to volume distention and *indirectly* to pressure. The nonlinear and viscoelastic nature of the vascular wall result in complex waveform alterations for this type of sensor that are difficult to correct in practice.

The second category is that of the pressure pulse sensor (Figure 55.6 [right]). This type of sensor measures stress due to arterial pressure transmitted though the skin above the pulse artery. The pressure pulse sensor requires that surface deflections are zero, as opposed to the volume sensor. Thus, the contact forces are proportionate to arterial pressure at the skin surface.

The differences in pulse waveforms that are provided by the above pulse recording techniques are clear in comparison with intra-arterial recordings [Van der Hoeven and Beneken, 1970]. In all cases, the pressure pulse method was found to provide superior waveform accuracy, free of the effects of vascular nonlinear viscoelasticity. Alternatively, the stiffness of the sensor can best characterize its pulse accuracy. High stiffness relative to the artery and surrounding tissue is required to best approximate the pressure pulse method.

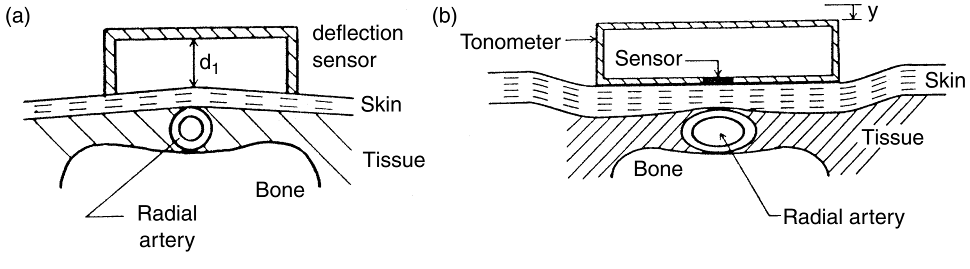


FIGURE 55.6 (a) Illustration of volume pulse method. (b) Illustration of pressure pulse method and arterial tonometry.

Arterial pulse recording is performed while the subject is stationary and refrains from moving the pulse location. But, it has become of interest to acquire ambulatory records. Without restraint, pulse recording is quite difficult due to motion artifact. For example, hand or finger motion can appear in recordings of the radial artery. A change in sensor positioning or acceleration can result in other types of artifact. The artifacts are often comparable in magnitude and frequency to the pulse, rendering simple filtering methods useless. Recently though, artifact cancellation techniques that employ sensor arrays have been applied with good success [Ciaccio et al., 1989].

55.3.4 Arterial Tonometry

Pulse sensing methods do not provide calibrated pressure. Arterial tonometry [Pressman and Newgard, 1963] is a pressure pulse method that can noninvasively record calibrated pressure in superficial arteries with sufficient bony support, such as the radial artery.

A tonometer is applied by first centering a contact stress sensor over the vessel. This is accomplished by repositioning the device until the largest pulse is detected. An array of sensors [Weaver et al., 1978] has been used to accomplish this electronically. Then, the tonometer is depressed towards the vessel. This leads to applanation of the vessel wall (Figure 55.6). If the vessel is not flattened sufficiently, the tonometer measures forces due to arterial wall tension and bending of the vessel. As depression is continued, the arterial wall is applanated further, but not so much as to occlude blood flow. At this intermediate position, wall tension becomes parallel to the tonometer sensing surface. Arterial pressure is then the remaining stress perpendicular to the surface and is measured by the sensor. This is termed the contact stress due to pressure. Ideally, the sensor should not measure skin shear (frictional) stresses. The contact stress is equal in magnitude to the arterial pressure when these conditions are achieved. The details of arterial tonometer calibration and design were analyzed by Drzewiecki et al. [1983, 1987]. In summary, tonometry requires that the contact stress sensor be flat, stiffer than the tissues, and small relative to the vessel diameter. Proper calibration can be attained either by monitoring the contact stress distribution (using a sensor array) or the maximum in measured pulse amplitude.

Recent research in tonometry has focused on miniaturization of semiconductor pressure sensor arrays [Weaver et al., 1978]. Alternatively, fiber optics have been employed by Drzewiecki [1985] and Moubarak et al. [1989], allowing extreme size reduction of the contact stress sensor. Commercial technology has been available (Jentow, Colin Electronics, Japan) and has been evaluated against intra-arterial records. Results indicate an average error of –5.6 mmHg for systolic pressure and –2.4 mmHg for diastole. Excellent pulse waveform quality is afforded by tonometry [Sato et al., 1993], making it a superior method for noninvasive pulse dynamics applications.

55.3.5 Flexible Diaphragm Tonometry

As an alternative to the high resolution tonometers under development, a new low resolution technology is introduced here. The basic advantage becomes one of cost, ease of positioning, and patient comfort, which is critical for long-term applications. In addition, while most tonometers have employed only the radial artery of the wrist for measurement, this technology is suitable for other superficial vessels and conforms to skin surface irregularities.

The flexible tonometer design applies the fact that tissue is incompressible in the short term [Bansal et al., 1994]. The concept is shown in Figure 55.7 using three tonometer volume compartments. These compartments are not physically separated and fluid can move between them. They are coupled to the skin and artery by means of the flexible diaphragm. When the arterial pressure exceeds that in the tonometer, the volume of the artery expands into Vb. Note also, that Va and Vc must increase to take up this expansion since water is incompressible. To restore the tonometer to a flat surface, the total volume of the tonometer is increased (Figure 55.7). In response, the tonometer pressure increases and the artery flattens. At this point, the volume in each compartment is equal, the diaphragm is flat, and the tonometer pressure is equal to arterial pressure. Thus, by maintaining the equilibrium of the relative volume compartments,

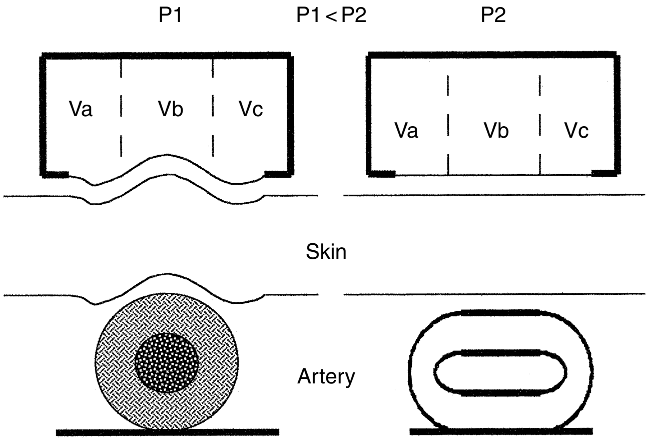


FIGURE 55.7 Concept of flexible diaphragm tonometry. P1 illustrates inappropriate level of pressure to provide arterial applanation tonometry. P2 indicates an increase in chamber volume so that the relative compartment volumes, V, are equal. In practice, relative compartment volumes are maintained constant via feedback control and applanation is continuous. Compartment pressure equals arterial pressure in P2.

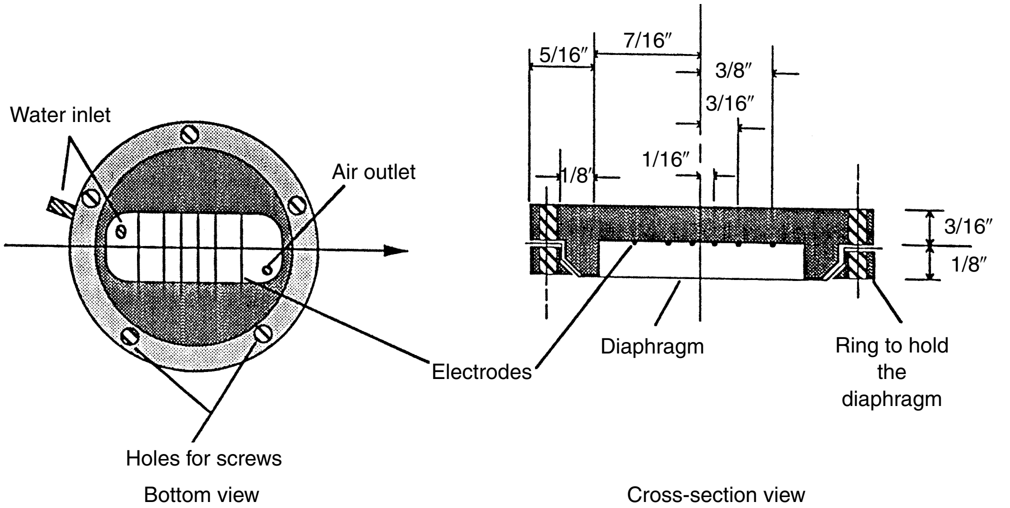


FIGURE 55.8 Design of a flexible diaphragm tonometer in accordance with the concept shown in Figure 55.7. Compartment volumes are obtained by means of impedance plethysmography. Electrode positions define the compartment boundaries.

applanation tonometry can be accomplished with a flexible diaphragm rather than a rigid one. In practice, instrumentation continuously adjusts the relative compartment volumes as arterial pressure changes.

A flexible diaphragm tonometer was machined from plexiglass (Figure 55.8). A rectangular channel was designed to contain saline. The front of the channel was sealed with a polyurethane sheet of 0.004 in. thickness. Two stainless steel electrodes were placed at each end of the channel. These were used to inject a current along the channel length. Near the center of the channel, four measuring electrodes were placed at equal spacing. Each pair of electrodes defined a volume compartment and the voltage across each pair was calibrated in terms of volume using impedance plethysmography. External to the tonometer, a saline-filled catheter was used to connect the channel to an electro-mechanical volume pump. The dynamics of the system were designed to possess appropriate frequency response.

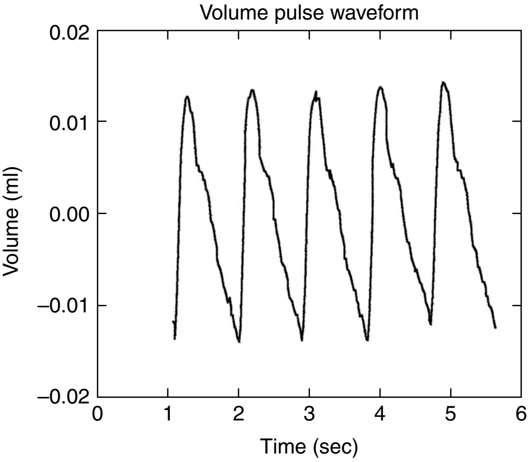


FIGURE 55.9 Sample of radial arterial volume recording from tonometer of Figure 55.8. Chamber pressure was fixed at 25 mmHg. Calibrated volume change is shown relative to the mean compartment volume.

Several beats of data were plotted from the volume pulse recordings of the flexible tonometer (Figure 55.9) for two adjacent compartments. The pulse volume is shown as a volume deviation from the mean compartment volume given a constant average tonometer pressure. The waveform shown illustrates the ability to provide calibrated noninvasive arterial volume information.

55.4 Noninvasive Arterial Mechanics

The flexible diaphragm tonometer and the occlusive cuff are capable of measuring volume in addition to pressure. Combining noninvasive measurements can extend the information provided by a single instrument, beyond that of blood pressure alone. Furthermore, functional information about the vasculature requires the simultaneous determination of pressure as well as geometry. In this section, the standard occlusive arm cuff will be employed as a plethysmograph to illustrate this concept.

The transmural pressure vs. lumen area (P–A) graph has been a convenient approach to analyze vessel function [Drzewiecki et al., 1997]. This curve can be applied to hemodynamics and portrays vessel function in a single graph. Unfortunately, its use has been limited to invasive studies or studies of isolated vessels where pressure can be experimentally controlled. The occlusive arm cuff offers a noninvasive solution to this problem by allowing the transmural pressure to be varied by altering the external pressure applied to a blood vessel. This approach has been used to noninvasively measure the transmural pressure vs. compliance relationship of the brachial artery [Mazhbich et al., 1983].

Drzewiecki and Pilla [1998] furthered the idea of noninvasively varying the transmural pressure by developing a cuff-based instrument that measures the P–A relationship. To perform the measurement, an occlusive arm cuff was pressurized by means of a diaphragm air pump. Cuff pressure was measured using a pressure sensor that provides pressure as an electronic signal. The pressure signal was input to a computer system via an analog-to-digital convertor for analysis. Two valves were used to control the flow of air. Initially, the valves were operated so that the pump inflated the cuff to well above the subject's systolic pressure. At that point, cuff air was recirculated through the pump. Under this condition, the mean cuff pressure remained constant, but the stroke volume of the pump resulted in cuff pressure oscillations at the frequency of the pump. Since the pump volume is a known quantity, it was used as a volume calibration source. The pump volume was divided by the cuff pressure oscillations due to the pump to yield the cuff compliance. With cuff compliance known, the cuff pressure arterial pulse was converted to a volume pulse by multiplying by cuff compliance. The pump was designed to operate at approximately 40 Hz, well

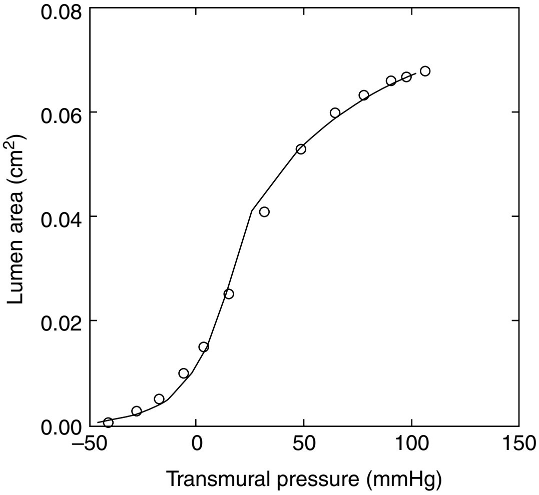


FIGURE 55.10 Arterial compliance obtained noninvasively by employing the occlusive arm cuff as a plethysmograph. Points are measured data. The solid curve is the model result obtained from Equation 55.1.

above the heart rate frequency components. This allowed the arterial pulse and pump pulse to be easily separated from the cuff pressure recording by using low pass and high pass digital filters with a cutoff frequency of 25 Hz.

The above procedure was repeated at every level of cuff pressure. The cuff pressure was released in steps of 5 to 10 mmHg until it was zero by briefly opening the release valve. The subject's pulse volume was divided by their pressure pulse (systolic minus diastolic pressures) to obtain the arterial compliance at every value of cuff pressure. The compliance per unit length was obtained by further dividing by the effective cuff length. The systolic, diastolic, and mean arterial pressures were evaluated by oscillometry and the Korotkoff method. Arterial compliance was then plotted for every value of transmural pressure (Figure 55.10). The compliance was then numerically integrated to obtain the corresponding brachial artery P–A curve. Since the vessel was collapsed for large negative transmural pressures, the initial constant of integration was chosen to be zero lumen area.

The arterial compliance curve possesses a maximum near zero transmural pressure. This point corresponds with the onset of vessel buckling. On either side of the buckling point, the vessel supports its pressure load differently. On the negative pressure side, the vessel partially collapses and undergoes wall bending. The compliance rapidly approaches zero during collapse. This occurs as the opposite walls of the vessel begin to contact each other and close the lumen. On the positive pressure side, the vessel supports its pressure load by wall stretch. The compliance slowly decreases with increasing pressure because of nonlinear wall elasticity. That is, the wall becomes stiffer with increasing stretch. The s-shaped lumen area curve is a consequence of these two different mechanisms. Preliminary studies of several subjects revealed that the shape of the noninvasive *P–A* curve is consistent for all subjects studied [Whitt and Drzewiecki, 1997; Drzewiecki and Pilla, 1998].

A mathematical model was developed that incorporates the fundamental physical and material properties that contribute to the collapsible *P–A* relationship [Drzewiecki et al., 1997; Drzewiecki and Pilla, 1998]. The basic form of the model is summarized by the following equation for transmural pressure:

$$P = -E((\lambda^{-1})^n - 1) + P_b + a(\mathrm{e}^{b(\lambda-1)} - 1) \qquad (55.1)$$

where a and b are arterial elastance constants for distension, E is vessel elastance during collapse, and n is a constant that determines the rate of change of pressure with respect to change in area during collapse. The quantity λ, was defined as the extension ratio and is evaluated from the lumen area divided by the lumen area at the buckling point, A/A_b. The first hyperbolic term is the pressure due to collapse and wall bending. The second term is the buckling pressure and is found when $A = A_b$. The third exponential term represents the pressure due to wall stretch. Some overlap in the contribution of each term may occur near the buckling pressure. Depending on the vessel type and material, Equation 55.1 can be improved by limiting the extension ratio and its inverse to unity.

Equation 55.1 was employed to analyze the brachial artery *P–A* data from each subject. The constants A_b and P_b were measured directly from the data as the point on the *P–A* curve that corresponds with maximum compliance or buckling. Their values were inserted into Equation 55.1 for each subject, leaving the remaining constants to be found by nonlinear least squares regression (Marquardt–Levenberg algorithm). The model was evaluated for the subject shown in Figure 55.10 and corresponds with the solid line curve. Similar results were obtained for all subjects studied ($N = 10$), with the mean error of estimate less than 3 mmHg. The above study suggests that no other physical properties need to be added to model vascular collapse. Thus, it can be considered a valid model for further studies of vascular properties and blood pressure determination.

While the other terms of the model were found to vary from subject to subject, the buckling pressure was discovered to be relatively constant (10 mmHg ± 11). Hence, buckling may be the important vascular feature that permits noninvasive blood pressure measurement to be feasible and may be the common thread that links all noninvasive methods. This idea was first examined in our theoretical examination of oscillometry above [Drzewiecki et al., 1994]. The successful use of buckling itself as a method to find arterial pressure was also described here [Sheth and Drzewiecki, 1998]. The phenomenon of buckling is a general effect that occurs independent of the technique used to find arterial pressure. It will also be independent of the quantitative differences in the *P–A* relationship of each specific subject.

55.5 Summary

Future research is open to noninvasive studies of how cardiovascular disease can alter blood vessel mechanics and the accuracy of blood pressure determination. The use of methods, such as occlusive cuff plethysmography together with blood pressure measurement and vascular mechanical modeling presented here, offers a means for noninvasive detection of the early stages of cardiovascular disease. Additionally, pulse dynamics and pressure variation offered by noninvasive pulse recording can provide new information about cardiovascular control.

Marey originally proposed that vessel closure is the important event that permits noninvasive blood pressure measurement. From the work presented here, this early concept is refocused to the instability of arterial buckling, or the *process* of closure, as the basic mechanism that enables noninvasive blood pressure determination.

Acknowledgments

The author wishes to express thanks to graduate students Vineet Bansal, Edward Ciaccio, Cindy Jacobs, James Pilla, Deepa Sheth, and Michael Whitt for their assistance with the experimental results presented here. The oscillometry studies were supported, in part, by Critikon, Inc., Tampa, FL and arterial tonometry was supported, in part, by IVAC Corp., San Diego, CA.

References

Alexander H., Cohen M., and Steinfeld L. (1977). Criteria in the choice of an occluding cuff for the indirect measurement of blood pressure. *Med. Biol. Eng. Comput.* 15: 2–10.

Arzbaecher R.C. and Novotney R.L. (1973). Noninvasive measurement of the arterial pressure contour in man. *Biblio. Cardiol.* 31: 63–69.

Bansal V., Drzewiecki G., and Butterfield R. (1994). Design of a flexible diaphragm tonometer. *Thirteenth S. Biomed. Eng. Conf.*, Washington, D.C., pp. 148–151.

Ciaccio E.J., Drzewiecki G.M., and Karam E. (1989). Algorithm for reduction of mechanical noise in arterial pulse recording with tonometry. *Proc. 15th Northeast Bioeng. Conf.*, Boston, pp. 161–162.

Drzewiecki G.M. (1985). The Origin of the Korotkoff Sound and Arterial Tonometry, Ph.D. dissertation. University of Pennsylvania, Philadelphia.

Drzewiecki G., Field S., Moubarak I., and Li J.K.-J. (1997). Vascular growth and collapsible pressure–area relationship. *Am. J. Physiol.* 273: H2030–H2043.

Drzewiecki G., Hood R., and Apple H. (1994). Theory of the oscillometric maximum and the systolic and diastolic detection ratios. *Ann. Biomed. Eng.* 22: 88–96.

Drzewiecki G.M., Karam E., Bansal V., Hood R., and Apple H. (1993). Mechanics of the occlusive arm cuff and its application as a volume sensor. *IEEE Trans. Biomed. Eng.* BME-40: 704–708.

Drzewiecki G.M., Melbin J., and Noordergraaf A. (1983). Arterial tonometry: review and analysis. *J. Biomech.* 16: 141–152.

Drzewiecki G.M., Melbin J., and Noordergraaf A. (1987). Noninvasive blood pressure recording and the genesis of Korotkoff sound. In *Handbook of Bioengineering*, S. Chien and R. Skalak. (eds.), pp. 8.1–8.36. New York, McGraw-Hill.

Drzewiecki G.M., Melbin J., and Noordergraaf A. (1989). The Korotkoff sound. *Ann. Biomed. Eng.* 17: 325–359.

Drzewiecki G. and Pilla J. (1998). Noninvasive measurement of the human brachial artery pressure–area relation in collapse and hypertension. *Ann. Biomed. Eng.* In press.

Graettinger W.F., Lipson J.L., Cheung D.G., and Weber M.A. (1988). Validation of portable noninvasive blood pressure monitoring devices: comparisons with intra-arterial and sphygmomanometer measurements. *Am. Heart J.* 116: 1155–1169.

Geddes L.A., Voelz M., Combs C., and Reiner D. (1983). Characterization of the oscillometric method for measuring indirect blood pressure. *Ann. Biomed. Eng.* 10: 271–280.

Geddes L.A. and Whistler S.J. (1978). The error in indirect blood pressure measurement with incorrect size of cuff. *Am. Heart J.* 96: 4–8.

Kelly R., Daley J., Avolio A., and O'Rourke M. (1989). Arterial dilation and reduced wave reflection — benefit of dilevalol in hypertension. *Hypertension* 14: 14–21.

Korotkoff N. (1905). On the subject of methods of determining blood pressure. *Bull. Imperial Mil. Med. Acad. (St. Petersburg)* 11: 365–367.

Li J.K.-J. (1986). Time domain resolution of forward and reflected waves in the aorta. *IEEE Trans. Biomed. Eng.* BME-33: 783–785.

Link W.T. (1987). Techniques for obtaining information associated with an individual's blood pressure including specifically a stat mode technique. US Patent #4,664,126.

London S.B. and London R.E. (1967). Comparison of indirect blood pressure measurements (Korotkoff) with simultaneous direct brachial artery pressure distal to cuff. *Adv. Intern. Med.* 13: 127–142.

Marey E.J. (1885). La Methode Graphique dans les Sciences Experimentales et Principalement en Physiologie et en Medicine. Masson, Paris.

Maurer A. and Noordergraaf A. (1976). Korotkoff sound filtering for automated three-phase measurement of blood pressure. *Am. Heart J.* 91: 584–591.

Mazhbich B.J. (1983). Noninvasive determination of elastic properties and diameter of human limb arteries. *Pflugers Arch.* 396: 254–259.

Moubarak I.F., Drzewiecki G.M., and Kedem J. (1989). Semi-invasive fiber — optic tonometer. *Proc. 15th Boston Northeast Bioeng. Conf.*, Boston, pp. 167–168.

Noordergraaf A. (1978). *Circulatory System Dynamics.* New York, Academic Press.

Omboni S., Parati G., Frattol A., Mutti E., Di Rienzo M., Castiglioni P., and Mancia G. (1993). Spectral and sequence analysis of finger blood pressure variability: comparison with analysis of intra-arterial recordings. *Hypertension* 22: 26–33.

Penaz J. (1973). Photoelectric measurement of blood pressure, volume, and flow in the finger. *Dig. 10th Intl. Conf. Med. Eng.*, Dresden, Germany, p. 104.

Pickering T., James G., Boddie C., Harshfield G., Blank S., and Laragh J. (1988). How common is white coat hypertension? *JAMA* 259: 225–228.

Posey J.A., Geddes L.A., Williams H., and Moore A.G. (1969). The meaning of the point of maximum oscillations in cuff pressure in the indirect measurement of blood pressure. Part 1. *Cardiovasc. Res. Cent. Bull.* 8: 15–25.

Pressman G.L. and Newgard P.M. (1963). A transducer for the continuous external measurement of arterial blood pressure. *IEEE Trans. Biomed. Eng.* BME-10: 73–81.

Rabbany S.Y., Drzewiecki G.M., and Noordergraaf A. (1993). Peripheral vascular effects on auscultatory blood pressure measurement. *J. Clin. Monitoring* 9: 9–17.

Ramsey III M. (1979). Noninvasive blood pressure determination of mean arterial pressure. *Med. Biol. Eng. Comput.* 17: 11–18.

Sato T., Nishinaga M., Kawamoto A., Ozawa T., and Takatsuji H. (1993). Accuracy of a continuous blood pressure monitor based on arterial tonometry. *Hypertension* 21: 866–874.

Sharir T., Marmor A., Ting C.-T., Chen J.-W., Liu C.-P., Chang M.-S., Yin F.C.P., and Kass D.A. (1993). Validation of a method for noninvasive measurement of central arterial pressure. *Hypertension* 21: 74–82.

Shenoy D., von Maltzahn W.W., and Buckley J.C. (1993). Noninvasive blood pressure measurement on the temporal artery using the auscultatory method. *Ann. Biomed. Eng.* 21: 351–360.

Sheth D. and Drzewiecki G. (1998). Using vessel buckling for continuous determination of arterial blood pressure. *Ann. Biomed. Eng.* 26: S–70.

Spence J.D., Sibbald W.J., and Cape R.D. (1978). Pseudohypertension in the elderly. *Clin. Sci. Mol. Med.* 55: 399s–402s.

Strus J. (1555). Sphygmicae artis jam mille ducentos annos peritae et desideratae, Libri V a Josephi Struthio Posnanience, medico recens conscripti, Basel. (As transcribed by A. Noordergraaf, Univ. of Pennsylvania, Philadelphia, PA).

Van der Hoeven G.M.A. and Beneken J.E.W. (1970). A reliable transducer for the recording of the arterial pulse wave. *Prog. Rep. 2. Inst. Med. Phys.* TNO, Utrecht.

Wesseling K.H., de Wit B., Snoeck B., Weber J.A.P., Hindman B.W., Nijland R., and Van der Hoeven G.M.A. (1978). An implementation of the Penaz method for measuring arterial blood pressure in the finger and the first results of an evaluation. *Prog. Rep. 6. Inst. Med. Phys.* TNO, Utrecht.

Weaver C.S., Eckerle J.S., Newgard P.M., Warnke C.T., Angell J.B., Terry S.C., and Robinson J. (1978). A study of noninvasive blood pressure measurement technique. In *Noninvasive Cardiovascular Measurements. Soc. Photo-Opt. Instr. Eng.* 167: 89.

Westerhof N., Bosman F., DeVries C.J., and Noordergraaf A. (1969). Analog studies of the human systemic arterial tree. *J. Biomech.* 2: 121–143.

White W.W., Berson A.S., Robbins C., Jamieson M.J., Prisant M., Roccella E., and Sheps S.G. (1993). National standard for measurement of resting and ambulatory blood pressures with automated sphygmomanometers. *Hypertension* 21: 504–509.

Whitt M. and Drzewiecki G. (1997). Repeatability of brachial artery area measurement. *Ann. Biomed. Eng.* 25: S-12.

Yamakoshi K., Shimazu H., and Togawa T. (1980). Indirect measurement of instantaneous arterial blood pressure in the human finger by the vascular unloading technique. *IEEE Trans. Biomed. Eng.* BME-27: 150.

56

Cardiac Output Measurement

Leslie A. Geddes
Purdue University

Cardiac output is the amount of blood pumped by the right or left ventricular per unit of time. It is expressed in liters per minute (L/min) and normalized by division by body surface area in square meters (m^2). The resulting quantity is called the cardiac index. Cardiac output is sometimes normalized to body weight, being expressed as mL/min per kilogram. A typical resting value for a wide variety of mammals is 70 mL/min per kg.

With exercise, cardiac output increases. In well-trained athletes, cardiac output can increase fivefold with maximum exercise. During exercise, heart rate increases, venous return increases, and the ejection fraction increases. Parenthetically, physically fit subjects have a low resting heart rate, and the time for the heart rate to return to the resting value after exercise is less than that for subjects who are not physically fit.

There are many direct and indirect (noninvasive) methods of measuring cardiac output. Of equal importance to the number that represents cardiac output is the left-ventricular ejection fraction (stroke volume divided by diastolic volume), which indicates the ability of the left ventricle to pump blood.

56.1 Indicator-Dilution Method

The principle underlying the indicator-dilution method is based on the upstream injection of a detectable indicator and on measuring the downstream concentration-time curve, which is called a *dilution curve.* The essential requirement is that the indicator mixes with all the blood flowing through the central mixing pool. Although the dilution curves in the outlet branches may be slightly different in shape, they all have the same area.

Figure 56.1a illustrates the injection of m g of indicator into an idealized flowing stream having the same velocity across the diameter of the tube. Figure 56.1b shows the dilution curve recorded downstream. Because of the flow-velocity profile, the cylinder of indicator and fluid becomes teardrop in shape, as shown in Figure 56.1c. The resulting dilution curve has a rapid rise and an exponential fall, as shown in

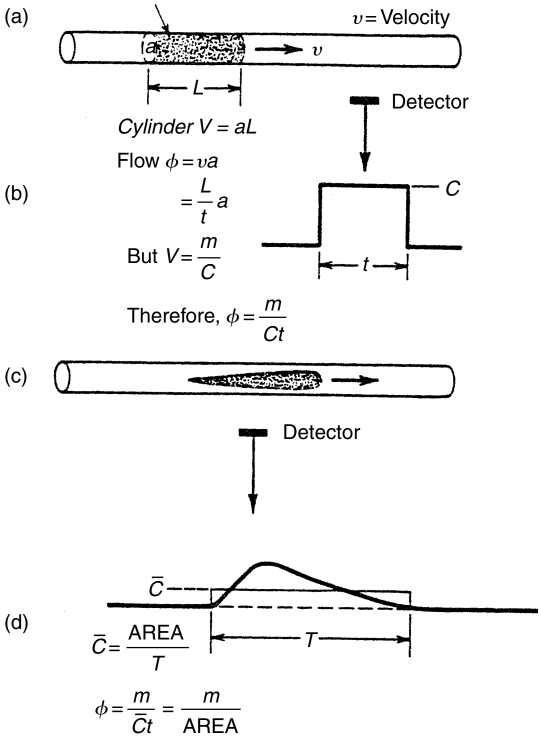


FIGURE 56.1 Genesis of the indicator-dilution curve.

Figure 56.1d. However, the area of the dilution curve is the same as that shown in Figure 56.1a. Derivation of the flow equation is shown in Figure 56.1, and the flow is simply the amount of indicator (m g) divided by the area of the dilution curve (g/mL × sec), which provides the flow in mL/sec.

56.1.1 Indicators

Before describing the various indicator-dilution methods, it is useful to recognize that there are two types of indicator, diffusible and nondiffusible. A diffusible indicator will leak out of the capillaries. A nondiffusible indicator is retained in the vascular system for a time that depends on the type of indicator. Whether cardiac output is overestimated with a diffusible indicator depends on the location of the injection and measuring sites. Table 56.1 lists many of the indictors that have been used for measuring cardiac output and the types of detectors used to obtain the dilution curve. It is obvious that the indicator selected must be detectable and not alter the flow being measured. Importantly, the indicator must be nontoxic and sterile.

When a diffusible indicator is injected into the right heart, the dilution curve can be detected in the pulmonary artery, and there is no loss of indicator because there is no capillary bed between these sites; therefore the cardiac output value will be accurate.

56.1.2 Thermal Dilution Method

Chilled 5% dextrose in water (D5W) or 0.9% NaCl can be used as indicators. The dilution curve represents a transient reduction in pulmonary artery blood temperature following injection of the indicator into the right atrium. Figure 56.2 illustrates the method and a typical thermodilution curve. Note that the indicator is really negative calories. The thermodilution method is based on heat exchange measured in calories, and the flow equation contains terms for the specific heat (C) and the specific gravity (S) of the indicator (i) and blood (b). The expression employed when a #7F thermistor-tipped catheter is used

TABLE 56.1 Indicators

Material	Detector	Retention data
Evans blue (T1824)	Photoelectric 640 μ	50% loss in 5 days
Indocyanine green	Photoelectric 800 μ	50% loss in 10 min
Coomassie blue	Photoelectric 585–600 μ	50% loss in 15–20 min
Saline (5%)	Conductivity cell	Diffusible[a]
Albumin 1^{131}	Radioactive	50% loss in 8 days
Na^{24}, K^{42}, D_2O, DHO	Radioactive	Diffusible[a]
Hot-cold solutions	Thermodetector	Diffusible[a]

[a] It is estimated that there is about 15% loss of diffusible indicators during the first pass through the lungs.

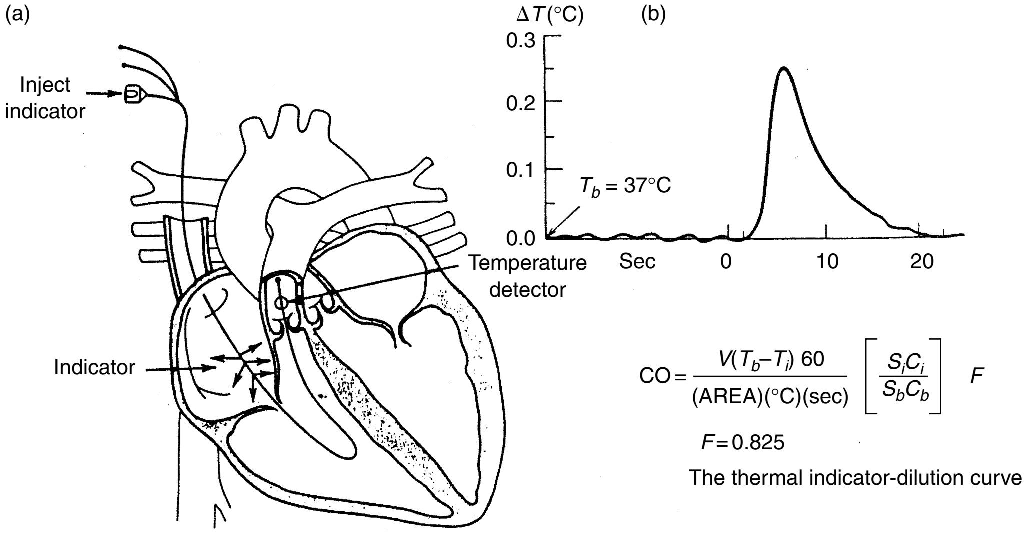

(a)
Inject indicator
Indicator
Temperature detector
(b)
ΔT(°C)
0.3
0.2
0.1
0.0
$T_b = 37°C$
Sec
0
10
20
$CO = \frac{V(T_b - T_i)\,60}{(AREA)(°C)(sec)} \left[\frac{S_iC_i}{S_bC_b}\right] F$
F = 0.825
The thermal indicator-dilution curve

FIGURE 56.2 The thermodilution method (a) and a typical dilution curve (b).

and chilled D5W is injected into the right atrium is as follows:

$$CO = \left[\frac{V(T_b - T_i)60}{A}\right]\left[\frac{S_iC_i}{S_bC_b}\right]F \tag{56.1}$$

where

V = Volume of indicator injected in mL
T_b = Temperature (average of pulmonary artery blood in (°C)
T_i = Temperature of the indicator (°C)
60 = Multiplier required to convert mL/sec into mL/min
A = Area under the dilution curve in (sec × °C)
S = Specific gravity of indicator (i) and blood (b)
C = Specific heat of indicator (i) and blood (b)
(S_iC_i/S_bC_b = 1.08 for 5% dextrose and blood of 40% packed-cell volume)
F = Empiric factor employed to correct for heat transfer through the injection catheter (for a #7F catheter, $F = 0.825$ [2]).

Entering these factors into the expression gives

$$\text{CO} = \frac{V(T_b - T_i)53.46}{A} \tag{56.2}$$

where CO = cardiac output in mL/min

$$53.46 = 60 \times 1.08 \times 0.825$$

To illustrate how a thermodilution curve is processed, cardiac output is calculated below using the dilution curve shown in Figure 56.2.

V = 5 ml of 5% dextrose in water
$T_b = 37°\text{C}$
$T_i = 0°\text{C}$
$A = 1.59°\text{C s}$
$\text{CO} = \dfrac{5(37 - 0)53.46}{1.59} = 6220$ mL/min

Although the thermodilution method is *the standard in clinical medicine*, it has a few disadvantages. Because of the heat loss through the catheter wall, several series 5-mL injections of indicator are needed to obtain a consistent value for cardiac output. If cardiac output is low, that is, the dilution curve is very broad, it is difficult to obtain an accurate value for cardiac output. There are respiratory-induced variations in PA blood temperature that confound the dilution curve when it is of low amplitude. Although room-temperature D5W can be used, chilled D5W provides a better dilution curve and a more reliable cardiac output value. Furthermore, it should be obvious that if the temperature of the indicator is the same as that of blood, there will be no dilution curve.

56.1.3 Indicator Recirculation

An ideal dilution curve shown in Figure 56.2 consists of a steep rise and an exponential decrease in indicator concentration. Algorithms that measure the dilution-curve area have no difficulty with such a curve. However, when cardiac output is low, the dilution curve is typically low in amplitude and very broad. Often the descending limb of the curve is obscured by recirculation of the indicator or by low-amplitude artifacts. Figure 56.3a is a dilution curve in which the descending limb is obscured by recirculation of the indicator. Obviously it is difficult to determine the practical end of the curve, which is often specified as the time when the indicator concentration has fallen to a chosen percentage (e.g., 1%) of the maximum amplitude ($C_{\max}$). Because the descending limb represents a good approximation of a decaying exponential curve (e^{-kt}), fitting the descending limb to an exponential allows reconstruction of the curve without a recirculation error, thereby providing a means for identifying the end for what is called the *first pass of the indicator*.

In Figure 56.3b, the amplitude of the descending limb of the curve in Figure 56.3a has been plotted on semilogarithmic paper, and the exponential part represents a straight line. When recirculation appears, the data points deviate from the straight line and therefore can be ignored, and the linear part (representing the exponential) can be extrapolated to the desired percentage of the maximum concentration, say 1% of $C_{\max}$. The data points representing the extrapolated part were replotted on Figure 56.3a to reveal the dilution curve undistorted by recirculation.

Commercially available indicator-dilution instruments employ digitization of the dilution curve. Often the data beyond about 30% of $C_{\max}$ are ignored, and the exponential is computed on digitally extrapolated data.

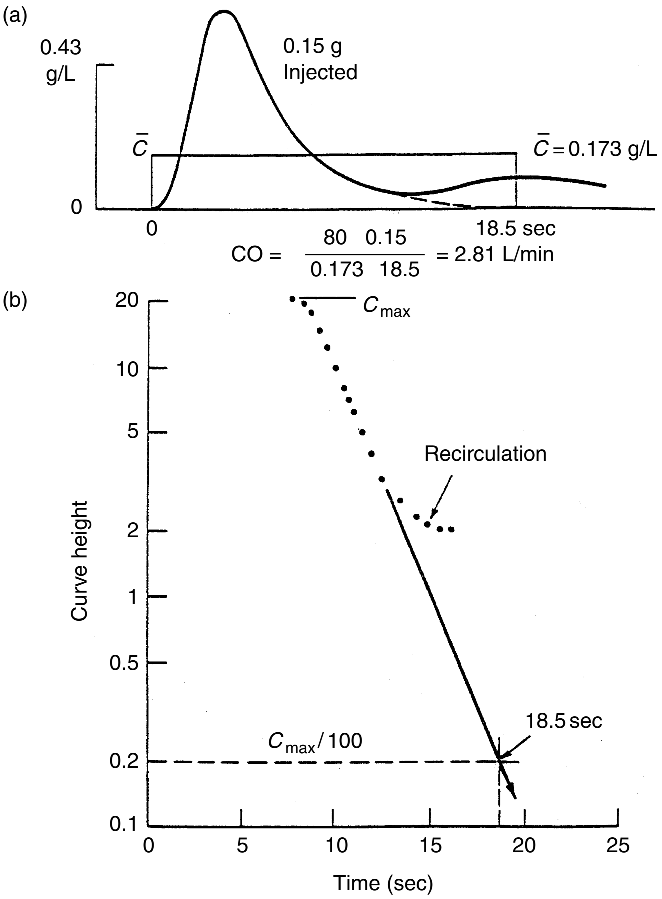


FIGURE 56.3 Dilution curve obscured by recirculation (a) and a semilogarithmic plot of the descending limb (b).

56.2 Fick Method

The Fick method *employs oxygen as the indicator* and the increase in oxygen content of venous blood as it passes through the lungs, along with the respiratory oxygen uptake, as the quantities that are needed to determine cardiac output (CO $=$ O_2 uptake/$A-VO_2$ difference). Oxygen uptake (mL/min) is measured at the airway, usually with an oxygen-filled spirometer containing a CO_2 absorber. The $A-VO_2$ difference is determined from the oxygen content (mL/100 mL blood) from any arterial sample and the oxygen content (mL/100 mL) of pulmonary arterial blood. The oxygen content of blood used to be difficult to measure. However, the new blood–gas analyzers that measure, pH, pO_2, pCO_2, hematocrit, and hemoglobin provide a value for O_2 content by computation using the oxygen-dissociation curve.

There is a slight technicality involved in determining the oxygen uptake because oxygen is consumed at body temperature but measured at room temperature in the spirometer. Consequently, the volume of O_2 consumed per minute displayed by the spirometer must be multiplied by a factor, F. Therefore the Fick equation is

$$\text{CO} = \frac{\text{O}_2 \text{ uptake/min}(F)}{A - V\text{O}_2 \text{ difference}} \tag{56.3}$$

Figure 56.4 is a spirogram showing a tidal volume riding on a sloping baseline that represents the resting expirating level (REL). The slope identifies the oxygen uptake at room temperature. In this subject,

Subject: 36-Year-old male
Height: 5′10½″
Weight: 175 lb″
Dody surface area: $A = W^{0.425} \quad H^{0.725} \quad 71.84$
where W = weight in kg
H = height in cm
A = area in cm^2
$A = 1.98\text{m}^2$

O_2 Uptake at 26°C: 400 mL/min
Barometer: 750 mmHg P_{H_2O} at 26°C: 25.2 mmHg

O_2 Uptake correction factor: F

$$F = \frac{273+37}{273+26} \quad \frac{750-25.2}{750-47}$$

$F = 1.069$

F from table 4.5 = 1.069

O_2 Uptake BTPS: 400 1.069 = 427.6
Blood O_2:

Arterial = 20 V%
Mixed venous = 15 V%

Cardiac output:

$$\text{CO} = \frac{427.6}{(20-15)/100} = 8552 \text{ mL/min}$$

Cardiac index:

$$\text{CI} = \frac{8.552}{1.98} = 4.32 \text{ L/min/m}^2$$

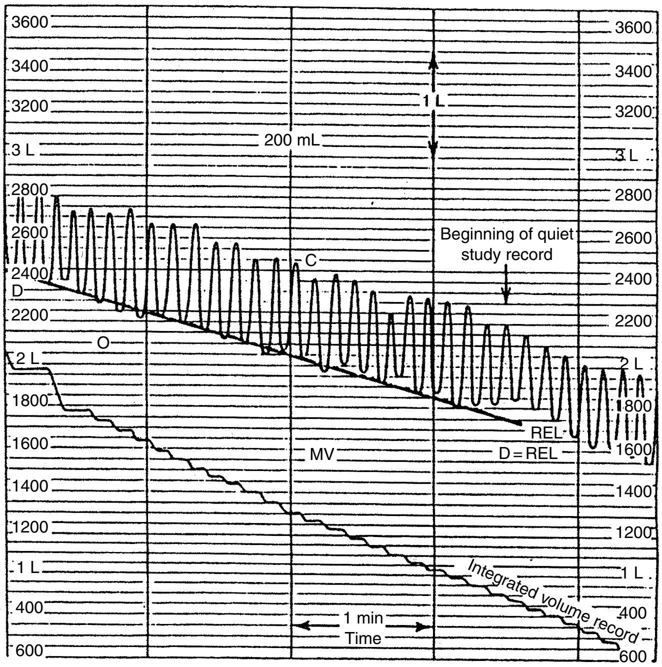


FIGURE 56.4 Measurement of oxygen uptake with a spirometer (*right*) and the method used to correct the measured volume (*left*).

the uncorrected oxygen consumption was 400 mL/min at 26°C in the spirometer. With a barometric pressure of 750 mmHg, the conversion factor F to correct this volume to body temperature (37°C) and saturated with water vapor is

$$F = \frac{273+37}{273+T_s} \times \frac{P_b - PH_2O}{P_b - 47} \tag{56.4}$$

where T_s is the spirometer temperature, P_b is the barometric pressure, and PH_2O at T_s is obtained from the water-vapor table (Table 56.2).

A sample calculation for the correction factor F is given in Figure 56.4, which reveals a value for F of 1.069. However, it is easier to use Table 56.3 to obtain the correction factor. For example, for a spirometer temperature of 26°C and a barometric pressure of 750 mmHg, $F = 1.0691$.

Note that the correction factor F in this case is only 6.9%. The error encountered by not including it may be less than the experimental error in making all other measurements.

The example selected shows that the $A - VO_2$ difference is 20–15 mL/100 mL blood and that the corrected O_2 uptake is 400×1.069; therefore the cardiac output is:

$$\text{CO} = \frac{400 \times 1.069}{(20-15)/100} = 8552 \text{ mL/min} \tag{56.5}$$

The Fick method does not require the addition of a fluid to the circulation and may have value in such a circumstance. However, its use requires stable conditions because an average oxygen uptake takes many minutes to obtain.

TABLE 56.2 Vapor Pressure of Water

Temp. °C	0.0	0.2	0.4	0.6	0.8
15	12.788	12.953	13.121	13.290	13.461
16	13.634	13.809	13.987	14.166	14.347
17	14.530	14.715	14.903	15.092	15.284
18	15.477	15.673	15.871	16.071	16.272
19	16.477	16.685	16.894	17.105	17.319
20	17.535	17.753	17.974	18.197	18.422
21	18.650	18.880	19.113	19.349	19.587
22	19.827	20.070	20.316	20.565	20.815
23	21.068	21.324	21.583	21.845	22.110
24	22.377	22.648	22.922	23.198	23.476
25	23.756	24.039	24.326	24.617	24.912
26	25.209	25.509	25.812	26.117	26.426
27	26.739	27.055	27.374	27.696	28.021
28	28.349	28.680	29.015	29.354	29.697
29	30.043	30.392	30.745	31.102	31.461
30	31.825	32.191	32.561	32.934	33.312
31	33.695	34.082	34.471	34.864	35.261
32	35.663	36.068	36.477	36.891	37.308
33	37.729	38.155	38.584	39.018	39.457
34	39.898	40.344	40.796	41.251	41.710
35	42.175	42.644	43.117	43.595	44.078
36	44.563	45.054	45.549	46.050	46.556
37	47.067	47.582	48.102	48.627	49.157
38	49.692	50.231	50.774	51.323	51.879
39	42.442	53.009	53.580	54.156	54.737
40	55.324	55.910	56.510	57.110	57.720
41	58.340	58.960	59.580	60.220	60.860

56.3 Ejection Fraction

The ejection fraction (EF) is one of the most convenient indicators of the ability of the left (or right) ventricle to pump the blood that is presented to it. Let v be the stroke volume (SV) and V be the end-diastolic volume (EDV); the ejection fraction is v/V or SV/EDV.

Measurement of ventricular diastolic and systolic volumes can be achieved radiographically, ultrasonically, and by the use of an indicator that is injected into the left ventricle where the indicator concentration is measured in the aorta on a beat-by-beat basis.

56.3.1 Indicator-Dilution Method for Ejection Fraction

Holt [1] described the method of injecting an indicator into the left ventricular during diastole and measuring the stepwise decrease in aortic concentration with successive beats (Figure 56.5). From this concentration-time record, end-diastolic volume, stroke volume, and ejection fraction can be calculated. No assumption need be made about the geometric shape of the ventricle. The following describes the theory of this fundamental method.

Let V be the end-diastolic ventricular volume. Inject m gm of indicator into this volume during diastole. The concentration (C_1) of indicator in the aorta for the first beat is m/V. By knowing the amount of indicator (m) injected and the calibration for the aortic detector, C_1 is established, and ventricular end-diastolic volume $V = m/C_1$.

After the first beat, the ventricle fills, and the amount of indicator left in the left ventricle is $m - mv/V$. The aortic concentration (C_2) for the second beat is therefore $m - mV/V = m(1 - v/V)$. Therefore

TABLE 56.3 Correction Factor F for Standardization of Collected Volume

°C/P_B	640	650	660	670	680	690	700	710	720	730	740	750	760	770	780
15	1.1388	1.1377	1.1367	1.1358	1.1348	1.1339	1.1330	1.1322	1.1314	1.1306	1.1298	1.1290	1.1283	1.1276	1.1269
16	1.1333	1.1323	1.1313	1.1304	1.1295	1.1286	1.1277	1.1269	1.1260	1.1253	1.1245	1.1238	1.1231	1.1224	1.1217
17	1.1277	1.1268	1.1266	1.1249	1.1240	1.1232	1.1224	1.1216	1.1208	1.1200	1.1193	1.1186	1.1179	1.1172	1.1165
18	1.1222	1.1212	1.1203	1.1194	1.1186	1.1178	1.1170	1.1162	1.1154	1.1147	1.1140	1.1133	1.1126	1.1120	1.1113
19	1.1165	1.1156	1.1147	1.1139	1.1131	1.1123	1.1115	1.1107	1.1100	1.1093	1.1086	1.1080	1.1073	1.1067	1.1061
20	1.1108	1.1099	1.1091	1.1083	1.1075	1.1067	1.1060	1.1052	1.1045	1.1039	1.1032	1.1026	1.1019	1.1094	1.1008
21	1.1056	1.1042	1.1034	1.1027	1.1019	1.1011	1.1004	1.0997	1.0990	1.0984	1.0978	1.0971	1.0965	1.0960	1.0954
22	1.0992	1.0984	1.0976	1.0969	1.0962	1.0964	1.0948	1.0941	1.0935	1.0929	1.0923	1.0917	1.0911	1.0905	1.0900
23	1.0932	1.0925	1.0918	1.0911	1.0904	1.0897	1.0891	1.0884	1.0878	1.0872	1.0867	1.0861	1.0856	1.0850	1.0845
24	1.0873	1.0866	1.0859	1.0852	1.0846	1.0839	1.0833	1.0827	1.0822	1.0816	1.0810	1.0805	1.0800	1.0795	1.0790
25	1.0812	1.0806	1.0799	1.0793	1.0787	1.0781	1.0775	1.0769	1.0764	1.0758	1.0753	1.0748	1.0744	1.0739	1.0734
26	1.0751	1.0710	1.0738	1.0732	1.0727	1.0721	1.0716	1.0710	1.0705	1.0700	1.0696	1.0691	1.0686	1.0682	1.0678
27	1.0688	1.0682	1.0677	1.0671	1.0666	1.0661	1.0656	1.0651	1.0640	1.0641	1.0637	1.0633	1.0629	1.0624	1.0621
28	1.0625	1.0619	1.0614	1.0609	1.0604	1.0599	1.0595	1.0591	1.0586	1.0582	1.0578	1.0574	1.0570	1.0566	1.0563
29	1.0560	1.0555	1.0550	1.0546	1.0548	1.0537	1.0533	1.0529	1.0525	1.0521	1.0518	1.0514	1.0519	1.0507	1.0504
30	1.0494	1.0496	1.0486	1.0482	1.0478	1.0474	1.0470	1.0467	1.0463	1.0460	1.0450	1.0453	1.0450	1.0447	1.0444

Source: From Kovach J.C., Paulos P., and Arabadjis C. 1955. *J. Thorac. Surg.* **29**: 552.
$V_s = FV_c$, where V_s is the standardized condition and V_c is the collected condition:

$$V = \frac{1 + 37/273}{1 + t°\mathrm{C}/273} \times \frac{P_B - \mathrm{PH_2O}}{P_B - 47} V_\mathrm{c} = FV_\mathrm{c}$$

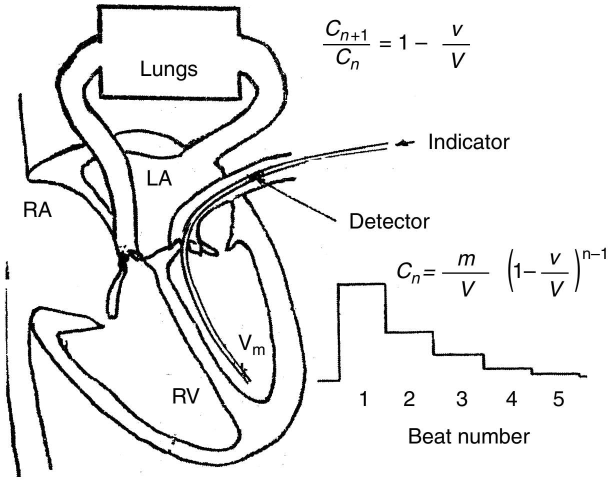

$\frac{C_{n+1}}{C_n} = 1 - \frac{v}{V}$
Lungs
Indicator
LA
RA
Detector
$C_n = \frac{m}{V}\left(1 - \frac{v}{V}\right)^{n-1}$
V_m
RV
1 2 3 4 5
Beat number

FIGURE 56.5 The saline method of measuring ejection fraction, involving injection of *m* g of NaCl into the left ventricle and detecting the aortic concentration (C) on a beat-by-beat basis.

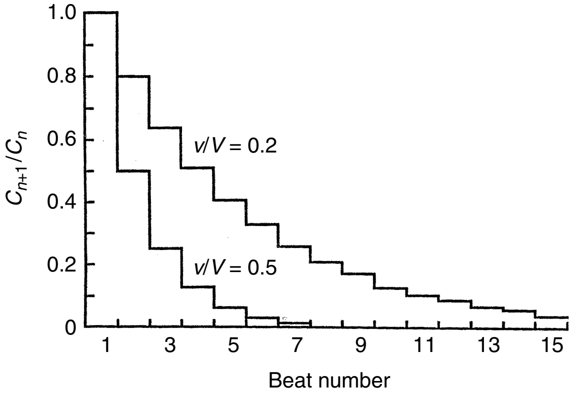


FIGURE 56.6 Stepwise decrease in indicator concentration (C) vs. beat number for ejection fraction (v/V) of 0.5 and 0.2.

the aortic concentration (C_2) for the second beat is

$$C_2 = \frac{m}{V}\left[1 - \frac{v}{V}\right] \tag{56.6}$$

By continuing the process, it is easily shown that the aortic concentration (C_n) for the nth beat is

$$C_n = \frac{m}{V}\left[1 - \frac{v}{V}\right]^{n-1} \tag{56.7}$$

Figure 56.6 illustrates the stepwise decrease in aortic concentration for ejection fractions (v/V) of 0.2 and 0.5, that is, 20% and 50%.

It is possible to determine the ejection fraction from the concentration ratio for two successive beats. For example,

$$C_n = \frac{m}{V}\left[1 - \frac{v}{V}\right]^{n-1} \tag{56.8}$$

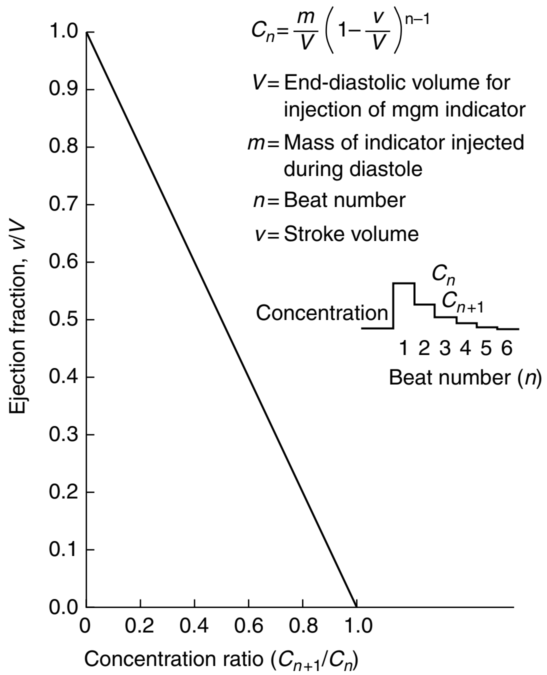


FIGURE 56.7 Ejection fraction (v/V) vs. the ratio of concentrations for successive beats (C_{n+1}/C_n).

$$C_n + 1 = \frac{m}{V}\left[1 - \frac{v}{V}\right]^n \tag{56.9}$$

$$\frac{C_{n+1}}{C_n} = 1 - \frac{v}{V} \tag{56.10}$$

from which

$$\frac{v}{V} = 1 - \frac{C_{n+1}}{C_n} \tag{56.11}$$

where v/V is the ejection fraction and C_{n+1}/C_n is the concentration ratio for two successive beats, for example, C_2/C_1 or C_3/C_2. Figure 56.7 illustrates the relationship between the ejection fraction v/V and the ratio of C_{n+1}/C_n. Observe that the detector need not be calibrated as long as there is a linear relationship between detector output and indicator concentration in the operating range.

References

[1] Holt, J.P. (1956). Estimation of the residual volume of the ventricle of the dog heart by two indicator-dilution techniques. *Circ. Res.* 4: 181.

[2] Weissel, R.D., Berger, R.L., and Hechtman, H.B. (1975). Measurement of cardiac output by thermodilution. *N. Engl. J. Med.* 292: 682.

57

External Defibrillators

Willis A. Tacker
Purdue University

Defibrillators are devices used to supply a strong electric shock (often referred to as a *countershock*) to a patient in an effort to convert excessively fast and ineffective heart rhythm disorders to slower rhythms that allow the heart to pump more blood. External defibrillators have been in common use for many decades for emergency treatment of life-threatening cardiac rhythms as well as for elective treatment of less threatening rapid rhythms. Figure 57.1 shows an external defibrillator.

Cardiac arrest occurs in more than 500,000 people annually in the United States, and more than 70% of the out-of-hospitals are due to cardiac arrhythmia treatable with defibrillators. The most serious arrhythmia treated by a defibrillator is ventricular fibrillation. Without rapid treatment using a defibrillator, ventricular fibrillation causes complete loss of cardiac function and death within minutes. Atrial fibrillation and the more organized rhythms of atrial flutter and ventricular tachycardia can be treated on a less emergent basis. Although they do not cause immediate death, their shortening of the interval between contractions can impair filling of the heart chambers and thus decrease cardiac output. Conventionally, treatment of ventricular fibrillation is called *defibrillation,* whereas treatment of the other tachycardias is called *cardioversion.*

57.1 Mechanism of Fibrillation

Fibrillation is chaotic electric excitation of the myocardium and results in loss of coordinated mechanical contraction characteristic of normal heart beats. Description of mechanisms leading to, and maintaining, fibrillation and other rhythm disorders are reviewed elsewhere [1] and are beyond the scope of this chapter. In summary, however, these rhythm disorders are commonly held to be a result of reentrant excitation pathways within the heart. The underlying abnormality that leads to the mechanism is the combination of conduction block of cardiac excitation plus rapidly recurring depolarization of the membranes of the cardiac cells. This leads to rapid repetitive propagation of a single excitation wave or of multiple

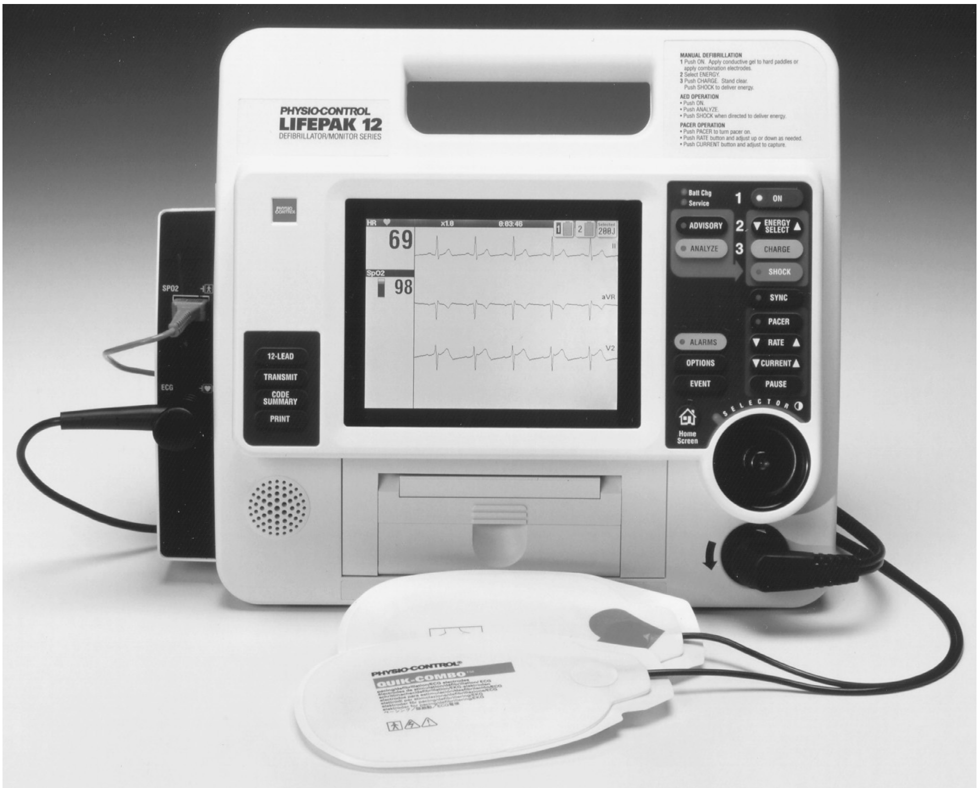


FIGURE 57.1 Photograph of a trans-chest defibrillator. (Provided by Physio-Control Corporation. With permission.)

excitatory waves throughout the heart. If the waves are multiple, the rhythm may degrade into total loss of synchronization of cardiac fiber contraction. Without synchronized contraction, the chamber affected will not contract, and this is fatal in the case of ventricular fibrillation. The most common cause of these conditions, and therefore of these rhythm disorders, is cardiac ischemia or infarction as a complication of atherosclerosis. Additional relatively common causes include other cardiac disorders, drug toxicity, electrolyte imbalances in the blood, hypothermia, and electric shocks (especially from alternating current).

57.2 Mechanism of Defibrillation

The corrective measure is to extinguish the rapidly occurring waves of excitation by simultaneously depolarizing most of the cardiac cells with a strong electric shock. The cells then can simultaneously repolarize themselves, and thus they will be back in phase with each other.

Despite years of intensive research, there is still no single theory for the mechanism of defibrillation that explains all the phenomena observed. However, it is generally held that the defibrillating shock must be adequately strong and have adequate duration to affect most of the heart cells. In general, longer duration shocks require less current than shorter duration shocks. This relationship is called the strength–duration relationship and is demonstrated by the curve shown in Figure 57.2. Shocks of strength and duration above and to the right of the current curve (or above the energy curve) have adequate strength to defibrillate, whereas shocks below and to the left do not. From the exponentially decaying current curve an energy curve can also be determined (also shown in Figure 57.2), which is high at very short durations due to

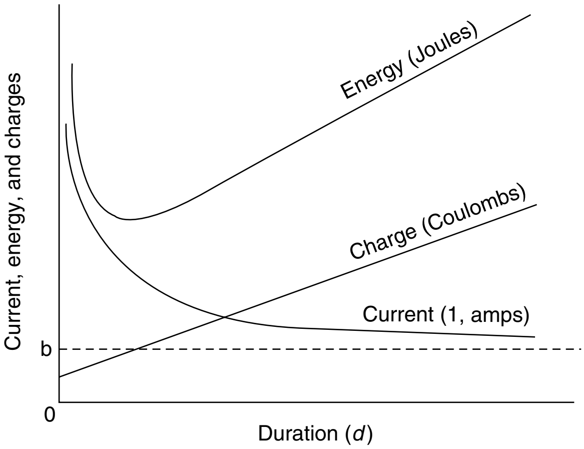


FIGURE 57.2 Strength–duration curves for current, energy, and charge. Adequate current shocks are above and to the right of the current curve. (Modified from Tacker W.A. and Geddes L.A. 1980. *Electrical Defibrillation,* Boca Raton, FL, CRC Press. With permission.)

high current requirements at short durations, but which is also high at longer durations due to additional energy being delivered as the pulse duration is lengthened at nearly constant current. Thus, for most electrical waveforms there is a minimum energy for defibrillation at approximate pulse durations of 3 to 8 msec. A strength–duration charge curve can also be determined as shown in Figure 57.2, which demonstrates that the minimum charge for defibrillation occurs at the shortest pulse duration tested. Very-short-duration pulses are not used, however, since the high current and voltage required is damaging to the myocardium. It is also important to note that excessively strong or long shocks may cause immediate refibrillation, thus failing to restore the heart function.

In practice, for a shock applied to electrodes on the skin surface of the patient's chest, durations are on the order of 3 to 10 msec and have an intensity of a few thousand volts and tens of amperes. The energy delivered to the subject by these shocks is selectable by the operator and is on the order of 50 to 360 J for most defibrillators. The exact shock intensity required at a given duration of electric pulse depends on several variables, including the intrinsic characteristics of the patient (such as the underlying disease problem or presence of certain drugs and the length of time the arrhythmia has been present), the techniques for electrode application, and the particular rhythm disorder being treated (more organized rhythms require less energy than disorganized rhythms).

57.3 Clinical Defibrillators

Defibrillator design has resulted from medical and physiologic research and advances in hardware technology. It is estimated that for each minute that elapses between onset of ventricular fibrillation and the first shock application, survival to leave hospital decreases by about 10%. The importance of rapid response led to development of portable, battery-operated defibrillators and more recently to automatic external defibrillators (AEDs) that enable emergency responders to defibrillate with minimal training.

All clinical defibrillators used today store energy in capacitors. Desirable capacitor specifications include small size, light weight, and capability to sustain several thousands of volts and many charge-discharge cycles. Energy storage capacitors account for at least one pound and usually several pounds of defibrillator weight. Energy stored by the capacitor is calculated from

$$W_s = \frac{1}{2}CE^2 \tag{57.1}$$

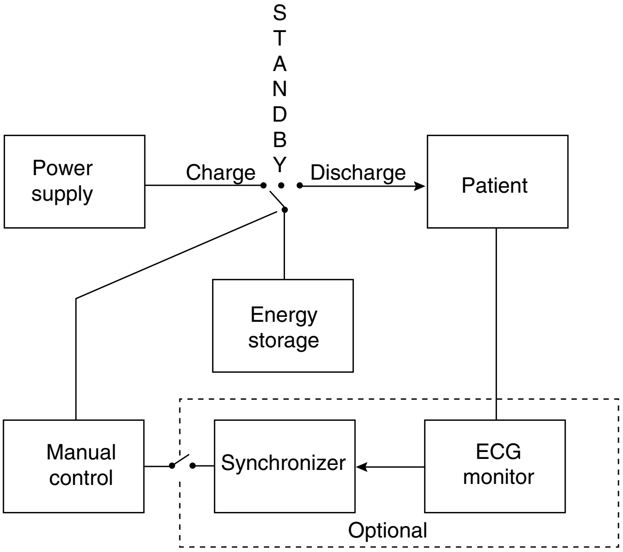


FIGURE 57.3 Block diagram of a typical defibrillator. (From Feinberg B. 1980. *Handbook Series in Clinical Laboratory Science,* Vol. 2, Boca Raton, FL, CRC Press. With permission.)

where W_s = stored energy in joules, C = capacitance in farads, and E = voltage applied to the capacitor. Delivered energy is expressed as

$$W_d = W_s \times \left(\frac{R}{R_i + R}\right) \tag{57.2}$$

where W_d = delivered energy, W_s = stored energy, R = subject resistance, and R_i = device resistance.

Figure 57.3 shows a block diagram for defibrillators. Most have a built-in monitor and synchronizer (dashed lines in Figure 57.3). Built-in monitoring speeds up diagnosis of potentially fatal arrhythmias, especially when the ECG is monitored through the same electrodes that are used to apply the defibrillating shock. The great preponderance of defibrillators for trans-chest defibrillation deliver shocks with either a damped sinusoidal waveform produced by discharge of an RCL circuit or a truncated exponential decay waveform (sometimes called trapezoidal). Basic components of exemplary circuits for damped sine waveform and trapezoidal waveform defibrillators are shown in Figure 57.4 and Figure 57.5. The shape of the waveforms generated by RCL defibrillators depend on the resistance of the patient as well as the energy storage capacitance and resistance and inductance of the inductor. When discharged into a 50-Ω load (to stimulate the patient's resistance), these defibrillators produce either a critically damped sine waveform or a slightly underdamped sine waveform (i.e., having a slight reversal of waveform polarity following the main waveform) into the 50-Ω load.

The exact waveform can be determined by application of Kirkchoff's voltage law to the circuit

$$L\frac{\mathrm{d}i}{\mathrm{d}t} + (R_i + R)i + \frac{1}{C}\int i\,\mathrm{d}t = 0 \tag{57.3}$$

where L = inductance in H, i = instantaneous current in amperes, t = time in seconds, R_i = device resistance, R = subject resistance, and C = capacitance. From this, the second-order differential equation describes the RCL defibrillator.

$$L\frac{\mathrm{d}^2 i}{\mathrm{d}t^2} + (R_i + R)\frac{\mathrm{d}i}{\mathrm{d}t} + \frac{1}{C}i = 0 \tag{57.4}$$

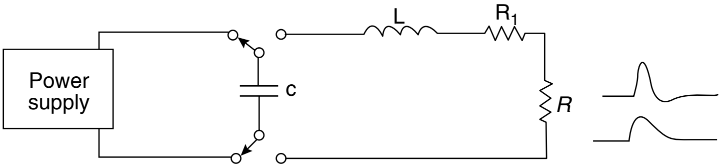


FIGURE 57.4 Resister–capacitor–inductor defibrillator. The patient is represented by *R*. (Modified from Feinberg B. 1980. *Handbook Series in Clinical Laboratory Science,* Vol. 2, Boca Raton, FL, CRC Press. With permission.)

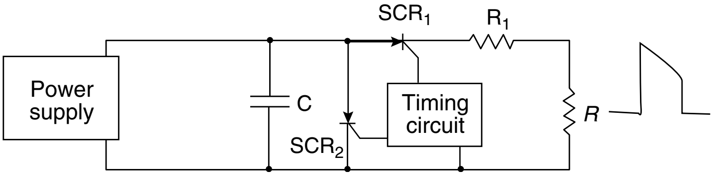


FIGURE 57.5 Trapezoidal wave defibrillator. The patient is represented by *R*. (Modified from Feinberg B. 1980. *Handbook Series in Clinical Laboratory Science,* Vol. 2, Boca Raton, FL, CRC Press. With permission.)

Trapezoidal waveform (actually, these are truncated exponential decay waveform) defibrillators are also used clinically. The circuit diagram in Figure 57.4 is exemplary of one design for producing such a waveform. Delivered energy calculation for this waveform is expressed as

$$W_\mathrm{d} = 0.5 I_\mathrm{i}^2 R \left[\frac{d}{\log_e \left(\frac{I_\mathrm{i}}{I_\mathrm{f}} \right)} \right] \left[1 - \left(\frac{I_\mathrm{f}}{I_\mathrm{i}} \right)^2 \right] \qquad (57.5)$$

where W_d = delivered energy, I_i = initial current in amperes, I_f = final current, R = resistance of the patient, and d = pulse duration in seconds. Both RCL and trapezoidal waveforms defibrillate effectively. Implantable defibrillators now use alternative waveforms such as a biphasic exponential decay waveform, in which the polarity of the electrodes is reversed part way through the shock. Use of the biphasic waveform has reduced the shock intensity required for implantable defibrillators but has not yet been extended to trans-chest use except on an experimental basis.

RCL defibrillators are the most widely available. They store up to about 440 J and deliver up to about 360 J into a patient with 50-Ω impedance. Several selectable energy intensities are available, typically from 5 to 360 J, so that pediatric patients, very small patients, or patients with easily converted arrhythmias can be treated with low-intensity shocks. The pulse duration ranges from 3 to 6 msec. Because the resistance (*R*) varies between patients (25 to 150 Ω) and is part of the RCL discharge circuit, the duration and damping of the pulse also varies; increasing patient impedance lengthens and dampens the pulse. Figure 57.6 shows waveforms from RCL defibrillators with critically damped and with underdamped pulses.

57.4 Electrodes

Electrodes for external defibrillation are metal and from 70 to 100 cm^2 in surface area. They must be coupled to the skin with an electrically conductive material to achieve low impedance across the electrode-patient interface. There are two types of electrodes: hand-held (to which a conductive liquid or solid gel is applied) and adhesive, for which an adhesive conducting material holds the electrode in place. Hand-held electrodes are reusable and are pressed against the patient's chest by the operator during shock delivery. Adhesive electrodes are disposable and are applied to the chest before the shock delivery and left in place for reuse if subsequent shocks are needed. Electrodes are usually applied with both electrodes on the

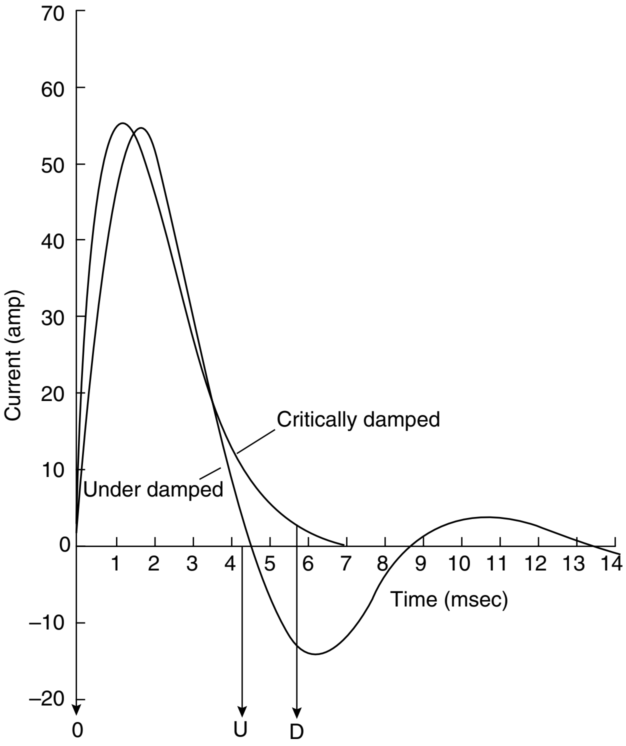


FIGURE 57.6 The damped sine wave. The interval *O–D* represents a duration for the critically and overdamped sine waves. By time *D*, more than 99% of the energy has been delivered. *O–U* is taken as the duration for an underdamped sine wave. (Modified from Tacker W.A. and Geddes L.A. 1980. *Electrical Defibrillation,* Boca Raton, FL, CRC Press. With permission.)

anterior chest as shown in Figure 57.7 or in anterior-to-posterior (front-to-back) position, as shown in Figure 57.8.

57.5 Synchronization

Most defibrillators for trans-chest use have the feature of synchronization, which is an electronic sensing and triggering mechanism for application of the shock during the QRS complex of the ECG. This is required when treating arrhythmias other than ventricular fibrillation, because inadvertent application of a shock during the *T* wave of the ECG often produces ventricular fibrillation. Selection by the operator of the synchronized mode of defibrillator operation will cause the defibrillator to automatically sense the QRS complex and apply the shock during the QRS complex. Furthermore, on the ECG display, the timing of the shock on the QRS is graphically displayed so the operator can be certain that the shock will not fall during the *T* wave (see Figure 57.9).

57.6 Automatic External Defibrillators

Automatic external defibrillators (AEDs) are defibrillators that automatically or semiautomatically recognize and treat rapid arrhythmias, usually under emergency conditions. Their operation requires less training than operation of manual defibrillators because the operator need not know which ECG waveforms indicate rhythms requiring a shock. The operator applies adhesive electrodes from the AED to the

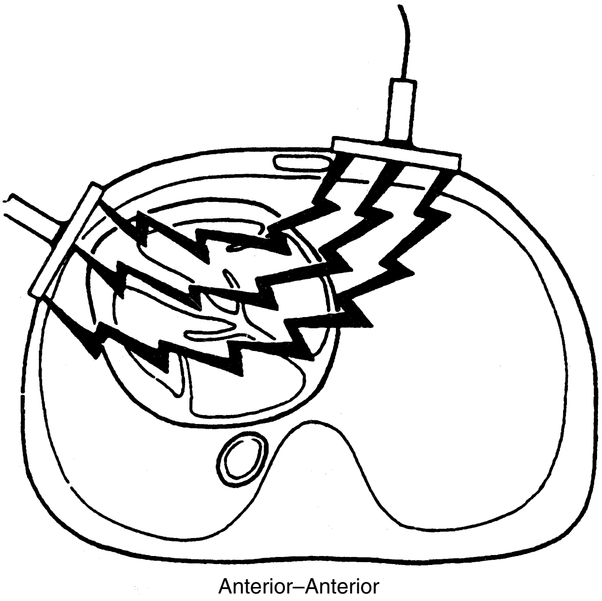


FIGURE 57.7 Cross-sectional view of the chest showing position for standard anterior wall (precordial) electrode placement. Lines of presumed current flow are shown between the electrodes on the skin surface. (Modified from Tacker W.A. [ed]. 1994. *Defibrillation of the Heart: ICDs, AEDs and Manual,* St. Louis, Mosby-Year Book. With permission.)

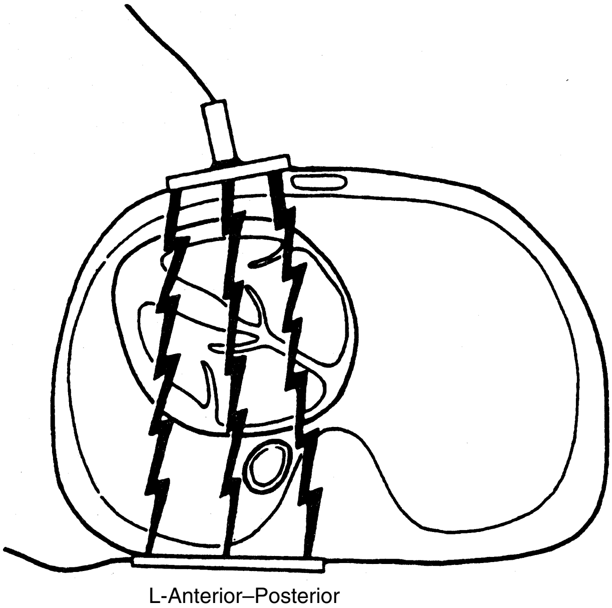


FIGURE 57.8 Cross-sectional view of the chest showing position for front-to-back electrode placement. Lines of presumed current flow are shown between the electrodes on the skin surface. (Modified from Tacker W.A. [ed]. 1994. *Defibrillation of the Heart: ICDs, AEDs and Manual,* St. Louis, Mosby-Year Book. With permission.)

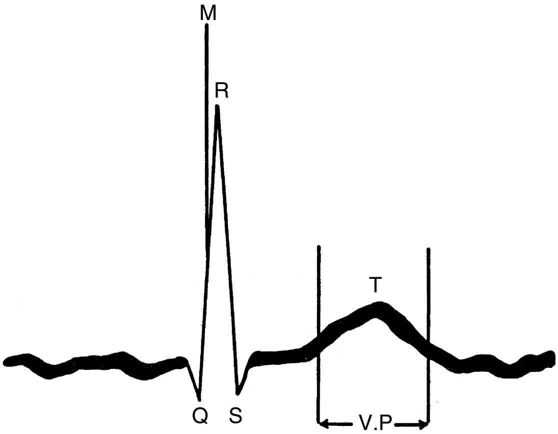


FIGURE 57.9 Timing mark (*M*) as shown on a synchronized defibrillator monitor. The *M* designates when a shock will be applied in the cardiac cycle. The *T* wave must be avoided, since a shock during the vulnerable period (V.P.) may fibrillate the ventricles. This tracing shows atrial fibrillation as identified by the irregular wavy baseline of the ECG. (Modified from Feinberg B. 1980. *Handbook Series in Clinical Laboratory Science,* Vol. 2, Boca Raton, FL, CRC Press. With permission.)

patient and turns on the AED, which monitors the ECG and determines by built-in signal processing whether or not and when to shock the patient. In a completely automatic mode, the AED does not have a manual control as shown in Figure 57.3 but instead has an automatic control. In semiautomatic mode, the operator must confirm the shock advisory from the AED to deliver the shock. AEDs have substantial potential for improving the chances of survival from cardiac arrest because they enable emergency personnel, who typically reach the patient before paramedics do, to deliver defibrillating shocks. Furthermore, the reduced training requirements make feasible the operation of AEDs in the home by a family member of a patient at high risk of ventricular fibrillation.

57.7 Defibrillator Safety

Defibrillators are potentially dangerous devices because of their high electrical output characteristics. The danger to the patient of unsynchronized shocks has already been presented, as has the synchronization design to prevent inadvertent precipitation of fibrillation by a cardioversion shock applied during the *T* wave.

There are other safety issues. Improper technique may result in accidental shocking of the operator or other personnel in the vicinity, if someone is in contact with the electric discharge pathway. This may occur if the operator is careless in holding the discharge electrodes or if someone is in contact with the patient or with a metal bed occupied by the subject when the shock is applied. Proper training and technique is necessary to avoid this risk.

Another safety issue is that of producing damage to the patient by application of excessively strong or excessively numerous shocks. Although cardiac damage has been reported after high-intensity and repetitive shocks to experimental animals and human patients, it is generally held that significant cardiac damage is unlikely if proper clinical procedures and guidelines are followed.

Failure of a defibrillator to operate correctly may also be considered a safety issue, since inability of a defibrillator to deliver a shock in the absence of a replacement unit means loss of the opportunity to resuscitate the patient. A recent review of defibrillator failures found that operator errors, inadequate defibrillator care and maintenance, and, to a lesser extent, component failure accounted for the majority of defibrillator failures [7].

References

[1] Tacker, W.A. Jr. (ed.). (1994). *Defibrillation of the Heart: ICDs, AEDs, and Manual.* St. Louis, Mosby-Year Book.

[2] Tacker, W.A. Jr. and Geddes, L.A. (1980). *Electrical Defibrillation.* Boca Raton, FL, CRC Press.

[3] Emergency Cardiac Care Committees, American Heart Association (1992). Guidelines for cardiopulmonary resuscitation and emergency cardiac care. *JAMA* 268: 2199.

[4] American National Standard ANSI/AAMI DF2 (1989). (second edition, revision of ANSI/AAMI DF2-1981) Safety and performance standard: cardiac defibrillator devices.

[5] Canadian National Standard CAN/CSA C22.2 No. 601.2.4-M90 (1990). Medical electrical equipment, part 2: particular requirements for the safety of cardiac defibrillators and cardiac defibrillator/monitors.

[6] International Standard IEC 601-2-4 (1983). Medical electrical equipment, part 2: particular requirements for the safety of cardiac defibrillators and cardiac defibrillator/monitors.

[7] Cummins, R.O., Chesemore, K., White, R.D., and the Defibrillator Working Group (1990). Defibrillator failures: causes of problems and recommendations for improvement. *JAMA* 264: 1019.

Further Information

Detailed presentation of material on defibrillator waveforms, algorithms for ECG analysis, and automatic defibrillation using AEDs, electrodes, design, clinical use, effects of drugs on shock strength required to defibrillate, damage due to defibrillator shocks, and use of defibrillators during open-thorax surgical procedures or trans-esophageal defibrillation are beyond the scope of this chapter. Also, the historical aspects of defibrillation are not presented here. For more information, the reader is referred to the publications at the end of this chapter [1–3]. For detailed description of specific defibrillators with comparisons of features, the reader is referred to articles from *Health Devices*, a monthly publication of ECRI, 5200 Butler Pike, Plymouth Meeting, Pa USA. For American, Canadian, and European defibrillator standards, the reader is referred to published standards [3–6] and Charbonnier's discussion of standards [1].

58 Implantable Defibrillators

Edwin G. Duffin
Medtronic, Inc.

The implantable cardioverter defibrillator (ICD) is a therapeutic device that can detect ventricular tachycardia or fibrillation and automatically deliver high-voltage (750 V) shocks that will restore normal sinus rhythm. Advanced versions also provide low-voltage (5 to 10 V) pacing stimuli for painless termination of ventricular tachycardia and for management of bradyarrhythmias. The proven efficacy of the automatic implantable defibrillator has placed it in the mainstream of therapies for the prevention of sudden arrhythmic cardiac death.

The implantable defibrillator has evolved significantly since first appearing in 1980. The newest devices can be implanted in the patient's pectoral region and use electrodes that can be inserted transvenously, eliminating the traumatic thoracotomy required for placement of the earlier epicardial electrode systems. Transvenous systems provide rapid, minimally invasive implants with high assurance of success and greater patient comfort. Advanced arrhythmia detection algorithms offer a high degree of sensitivity with reasonable specificity, and extensive monitoring is provided to document performance and to facilitate appropriate programming of arrhythmia detection and therapy parameters. Generator longevity can now exceed 4 years, and the cost of providing this therapy is declining.

58.1 Pulse Generators

The implantable defibrillator consists of a primary battery, high-voltage capacitor bank, and sensing and control circuitry housed in a hermetically sealed titanium case. Commercially available devices weigh between 197 and 237 g and range in volume from 113 to 145 cm^3. Clinical trials are in progress on devices with volumes ranging from 178 to 60 cm^3 and weights between 275 and 104 g. Further size reductions will be achieved with the introduction of improved capacitor and integrated circuit technologies and lead systems offering lower pacing and defibrillation thresholds. Progress should parallel that made with antibradycardia pacemakers that have evolved from 250-g, nonprogrammable, VOO units with 600-μJ pacing outputs to 26-g, multiprogrammable, DDDR units with dual 25-μJ outputs.

Implantable defibrillator circuitry must include an amplifier, to allow detection of the millivolt-range cardiac electrogram signals; noninvasively programmable processing and control functions, to evaluate the sensed cardiac activity and to direct generation and delivery of the therapeutic energy; high-voltage switching capability; dc–dc conversion functions to step up the low battery voltages; random access memories, to store appropriate patient and device data; and radiofrequency telemetry systems, to allow communication to and from the implanted device. Monolithic integrated circuits on hybridized substrates have made it possible to accomplish these diverse functions in a commercially acceptable and highly reliable form.

Defibrillators must convert battery voltages of approximately 6.5 to the 600–750 V needed to defibrillate the heart. Since the conversion process cannot directly supply this high voltage at current strengths needed for defibrillation, charge is accumulated in relatively large (Ý85–120 μF effective capacitance) aluminum electrolytic capacitors that account for 20 to 30% of the volume of a typical defibrillator. These capacitors must be charged periodically to prevent their dielectric from deteriorating. If this is not done, the capacitors become electrically leaky, yielding excessively long charge times and delay of therapy. Early defibrillators required that the patient return to the clinic periodically to have the capacitors reformed, whereas newer devices do this automatically at preset or programmable times. Improved capacitor technology, perhaps ceramic or thin-film, will eventually offer higher storage densities, greater shape variability for denser component packaging, and freedom from the need to waste battery capacity performing periodic reforming charges. Packaging density has already improved from 0.03 J/cm^3 for devices such as the early cardioverter to 0.43 J/cm^3 with some investigational ICDs. Capacitors that allow conformal shaping could readily increase this density to more than 0.6 J/cm^3.

Power sources used in defibrillators must have sufficient capacity to provide 50–400 full energy charges (Ý34 J) and 3 to 5 years of bradycardia pacing and background circuit operation. They must have a very low internal resistance in order to supply the relatively high currents needed to charge the defibrillation capacitors in 5–15 S. This generally requires that the batteries have large surface area electrodes and use chemistries that exhibit higher rates of internal discharge than those seen with the lithium iodide batteries used in pacemakers. The most commonly used defibrillator battery chemistry is lithium silver vanadium oxide.

58.2 Electrode Systems ("Leads")

Early implantable defibrillators utilized patch electrodes (typically a titanium mesh electrode) placed on the surface of the heart, requiring entry through the chest (Figure 58.1). This procedure is associated with approximately 3 to 4% perioperative mortality, significant hospitalization time and complications, patient discomfort, and high costs. Although subcostal, subxiphoid, and thoracoscopic techniques can minimize the surgical procedure, the ultimate solution has been development of fully transvenous lead systems with acceptable defibrillation thresholds.

Currently available transvenous leads are constructed much like pacemaker leads, using polyurethane or silicone insulation and platinum–iridium electrode materials. Acceptable thresholds are obtained in 67 to 95% of patients, with mean defibrillation thresholds ranging from 10.9 to 18.1 J. These lead systems use a combination of two or more electrodes located in the right ventricular apex, the superior vena cava, the coronary sinus, and sometimes, a subcutaneous patch electrode is placed in the chest region. These leads offer advantages beyond the avoidance of major surgery. They are easier to remove should there be infections or a need for lead system revision. The pacing thresholds of current transvenous defibrillation electrodes are typically 0.96±0.39 V, and the electrogram amplitudes are on the order of 16.4±6.4 mV. The eventual application of steroid-eluting materials in the leads should provide increased pacing efficiency

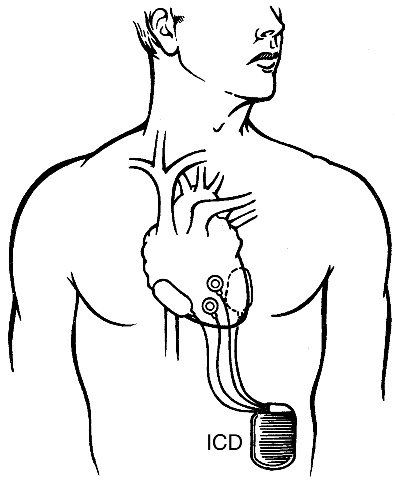


FIGURE 58.1 Epicardial ICD systems typically use two or three large defibrillating patch electrodes placed on the epicardium of the left and right ventricles and a pair of myocardial electrodes for detection and pacing. The generator is usually placed in the abdomen. (Copyright Medtronic, Inc. With permission.)

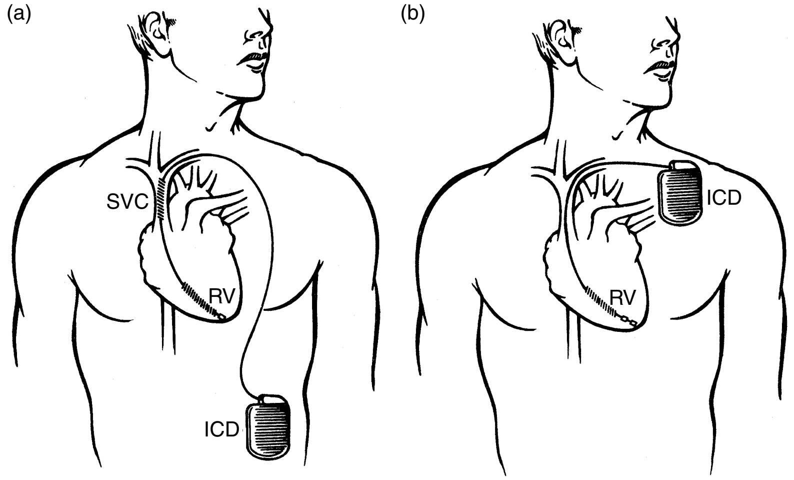


FIGURE 58.2 The latest transvenous fibrillation systems employ a single catheter placed in the right ventricular apex. In panel (a) a single transvenous catheter provides defibrillation electrodes in the superior vena cava and in the right ventricle. This catheter provides a single pace/sense electrode which is used in conjunction with the right ventricular high-voltage defibrillation electrode for arrhythmia detection and antibradycardia/antitachycardia pacing (a configuration that is sometimes referred to as *integrated bipolar*). With pulse generators small enough to be placed in the pectoral region, defibrillation can be achieved by delivering energy between the generator housing and one high-voltage electrode in the right ventricle (analogous to unipolar pacing) as is shown in panel (b). This catheter provided bipolar pace/sense electrodes for arrhythmia detection and antibradycardia/antitachycardia pacing. (Copyright Medtronic, Inc. With permission.)

with transvenous lead systems, thereby reducing the current drain associated with pacing and extending pulse generator longevity.

Lead systems are being refined to simplify the implant procedures. One approach is the use of a single catheter having a single right ventricular low-voltage electrode for pacing and detection, and a pair of high-voltage defibrillation electrodes spaced for replacement in the right ventricle and in the superior vena cava (Figure 58.2a). A more recent approach parallels that used for unipolar pacemakers.

A single right-ventricular catheter having bipolar pace/sense electrodes and one right ventricular high-voltage electrode is used in conjunction with a defibrillator housing that serves as the second high-voltage electrode (Figure 58.2b). Mean biphasic pulse defibrillation thresholds with the generator-electrode placed in the patient's left pectoral region are reported to be 9.8 ± 6.6 J ($n = 102$). This approach appears to be practicable only with generators suitable for pectoral placement, but such devices will become increasingly available.

58.3 Arrhythmia Detection

Most defibrillator detection algorithms rely primarily on heart rate to indicate the presence of a treatable rhythm. Additional refinements sometimes include simple morphology assessments, as with the probability density function, and analysis of rhythm stability and rate of change in rate.

The *probability density function* evaluates the percentage of time that the filtered ventricular electrogram spends in a window centered on the baseline. The rate-of-change-in-rate or *onset* evaluation discriminates sinus tachycardia from ventricular tachycardia on the basis of the typically gradual acceleration of sinus rhythms vs. the relatively abrupt acceleration of many pathologic tachycardias. The *rate stability* function is designed to bar detection of tachyarrhythmias as long as the variation in ventricular rate exceeds a physician-programmed tolerance, thereby reducing the likelihood of inappropriate therapy delivery in response to atrial fibrillation. This concept appears to be one of the more successful detection algorithm enhancements.

Because these additions to the detection algorithm reduce sensitivity, some defibrillator designs offer a supplementary detection mode that will trigger therapy in response to any elevated ventricular rate of prolonged duration. These *extended-high-rate* algorithms bypass all or portions of the normal detection screening, resulting in low specificity for rhythms with prolonged elevated rates such as exercise-induced sinus tachycardia. Consequently, use of such algorithms generally increases the incidence of inappropriate therapies.

Improvements in arrhythmia detection specificity are desirable, but they must not decrease the excellent sensitivity offered by current algorithms. The anticipated introduction of defibrillators incorporating dual-chamber pacemaker capability will certainly help in this quest, since it will then be possible to use atrial electrograms in the rhythm classification process. It would also be desirable to have a means of evaluating the patient's hemodynamic tolerance of the rhythm, so that the more comfortable pacing sequences could be used as long as the patient was not syncopal yet branch quickly to a definitive shock should the patient begin to lose consciousness.

Although various enhanced detection processes have been proposed, many have not been tested clinically, in some cases because sufficient processing power was not available in implantable systems, and in some cases because sensor technology was not yet ready for chronic implantation. Advances in technology may eventually make some of these very elegant proposals practicable. Examples of proposed detection enhancements include extended analyses of cardiac event timing (PR and RR stability, AV interval variation, temporal distribution of atrial electrogram intervals and of ventricular electrogram intervals, timing differences and/or coherency of multiple ventricular electrograms, ventricular response to a provocative atrial extrastimuli), electrogram waveform analyses (paced depolarization integral, morphology analyses of right ventricular or atrial electrograms), analyses of hemodynamic parameters (right-ventricular pulsatile pressure, mean right atrial and mean right ventricular pressures, wedge coronary sinus pressure, static right ventricular pressure, right atrial pressure, right ventricular stroke volume, mixed venous oxygen saturation and mixed venous blood temperature, left ventricular impedance, intramyocardial pressure gradient, aortic and pulmonary artery flow), and detection of physical motion.

Because defibrillator designs are intentionally biased to overtreat in preference to the life-threatening consequences associated with failure to treat, there is some incidence of inappropriate therapy delivery. Unwarranted therapies are usually triggered by supraventricular tachyarrhythmias, especially atrial fibrillation, or sinus tachycardia associated with rates faster than the ventricular tachycardia detection rate threshold. Additional causes include nonsustained ventricular tachycardia, oversensing of T waves, double counting of R waves and pacing stimuli from brady pacemakers, and technical faults such as loose leadgenerator connections or lead fractures.

Despite the bias for high detection sensitivity, undersensing does occur. It has been shown to result from inappropriate detection algorithm programming, such as an excessively high tachycardia detection rate; inappropriate amplifier gain characteristics; and electrode designs that place the sensing terminals too close to the high-voltage electrodes with a consequent reduction in electrogram amplitude following shocks. Undersensing can also result in the induction of tachycardia should the amplifier gain control algorithm result in undersensing of sinus rhythms.

58.4 Arrhythmia Therapy

Pioneering implantable defibrillators were capable only of defibrillation shocks. Subsequently, synchronized cardioversion capability was added. Antibradycardia pacing had to be provided by implantation of a standard pacemaker in addition to the defibrillator, and, if antitachycardia pacing was prescribed, it was necessary to use an antitachycardia pacemaker. Several currently marketed implantable defibrillators offer integrated ventricular demand pacemaker function and tiered antiarrhythmia therapy (pacing/cardioversion/defibrillation). Various burst and ramp antitachycardia pacing algorithms are offered, and they all seem to offer comparably high success rates. These expanded therapeutic capabilities improve patient comfort by reducing the incidence of shocks in conscious patients, eliminate the problems and discomfort associated with implantation of multiple devices, and contribute to a greater degree of success, since the prescribed regimens can be carefully tailored to specific patient needs. Availability of devices with antitachy pacing capability significantly increases the acceptability of the implantable defibrillator for patients with ventricular tachycardia.

Human clinical trials have shown that biphasic defibrillation waveforms are more effective than monophasic waveforms, and newer devices now incorporate this characteristic. Speculative explanations for biphasic superiority include the large voltage change at the transition from the first to the second phase or hyperpolarization of tissue and reactivation of sodium channels during the initial phase, with resultant tissue conditioning that allows the second phase to more readily excite the myocardium.

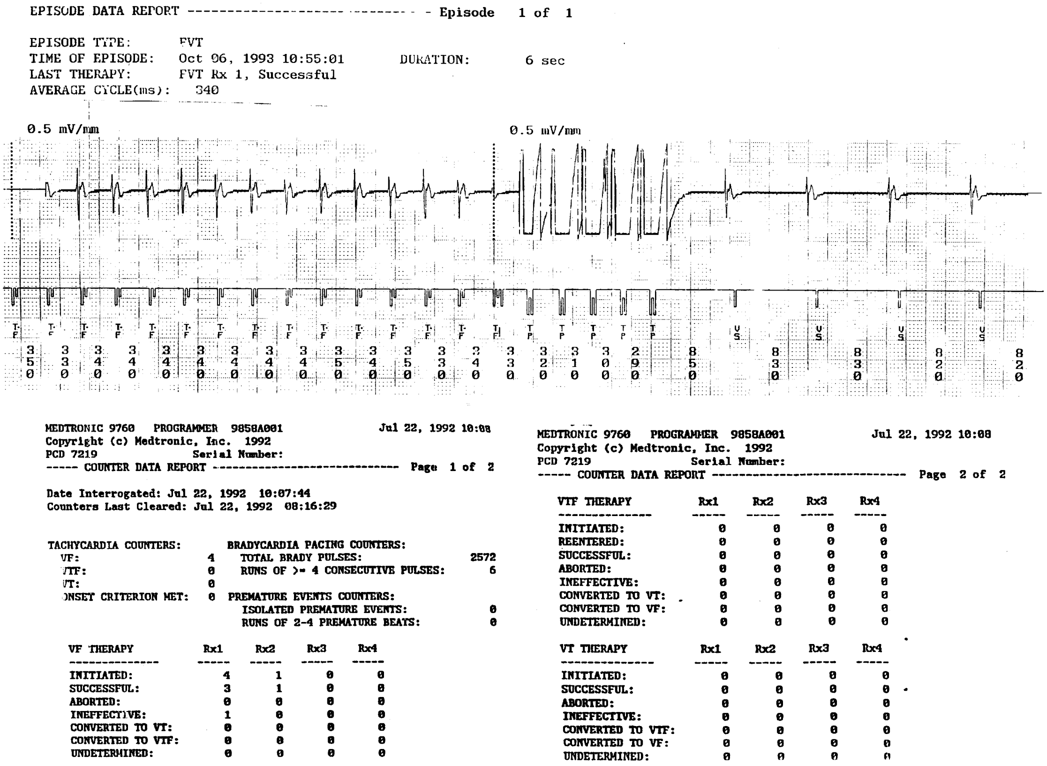
Antitachycardia pacing and cardioversion are not uniformly successful. There is some incidence of ventricular arrhythmia acceleration with antitachycardia pacing and cardioversion, and it is also not unusual for cardioversion to induce atrial fibrillation that in turn triggers unwarranted therapies. An ideal therapeutic solution would be one capable of preventing the occurrence of tachycardia altogether. Prevention techniques have been investigated, among them the use of precisely timed subthreshold stimuli, simultaneous stimulation at multiple sites, and pacing with elevated energies at the site of the tachycardia, but none has yet proven practical.

The rudimentary VVI antibradycardia pacing provided by current defibrillators lacks rate responsiveness and atrial pacing capability. Consequently, some defibrillator patients require implantation of a separate dual-chamber pacemaker for hemodynamic support. It is inevitable that future generations of defibrillators will offer dual-chamber pacing capabilities.

Atrial fibrillation, occurring either as a consequence of defibrillator operation or as a natural progression in many defibrillator patients, is a major therapeutic challenge. It is certainly possible to adapt implantable defibrillator technology to treat atrial fibrillation, but the challenge is to do so without causing the patient undue discomfort. Biphasic waveform defibrillation of acutely induced atrial fibrillation has been demonstrated in humans with an 80% success rate at 0.4 J using *epicardial* electrodes. Stand-alone atrial defibrillators are in development, and, if they are successful, it is likely that this capability would be integrated into the mainstream ventricular defibrillators as well. However, most conscious patients find shocks above 0.5 J to be very unpleasant, and it remains to be demonstrated that a clinically acceptable energy level will be efficacious when applied with transvenous electrode systems to spontaneously occurring atrial fibrillation. Moreover, a stand-alone atrial defibrillator either must deliver an atrial shock with complete assurance of appropriate synchronization to ventricular activity or must restrict the therapeutic energy delivery to atrial structures in order to prevent inadvertent induction of a malignant ventricular arrhythmia.

58.5 Implantable Monitoring

Until recently, defibrillator data recording capabilities were quite limited, making it difficult to verify the adequacy of arrhythmia detection and therapy settings. The latest devices record electrograms and diagnostic channel data showing device behavior during multiple tachyarrhythmia episodes. These devices also include counters (number of events detected, success and failure of each programmed therapy, and so on) that present a broad, though less specific, overview of device behavior (Figure 58.3). Monitoring capability in some of the newest devices appears to be the equivalent of 32 Kbytes of random access memory, allowing electrogram waveform records of approximately 2-min duration, with some opportunity for later expansion by judicious selection of sampling rates and data compression techniques. Electrogram storage has proven useful for documenting false therapy delivery due to atrial fibrillation, lead fractures, and sinus



```
EPISODE DATA REPORT ---------------------------------- - Episode   1 of  1

EPISODE TYPE:        FVT
TIME OF EPISODE:     Oct 06, 1993 10:55:01      DURATION:      6 sec
LAST THERAPY:        FVT Rx 1, Successful
AVERAGE CYCLE(ms):    340
```

```
MEDTRONIC 9760   PROGRAMMER  9858A001              Jul 22, 1992 10:08
Copyright (c) Medtronic, Inc.  1992
PCD 7219                Serial Number:
----- COUNTER DATA REPORT ----------------------------- Page  1 of  2

Date Interrogated: Jul 22, 1992  10:07:44
Counters Last Cleared: Jul 22, 1992  08:16:29

TACHYCARDIA COUNTERS:        BRADYCARDIA PACING COUNTERS:
  VF:                     4    TOTAL BRADY PULSES:                 2572
  VTF:                    0    RUNS OF >= 4 CONSECUTIVE PULSES:       6
  VT:                     0
  ONSET CRITERION MET:    0  PREMATURE EVENTS COUNTERS:
                               ISOLATED PREMATURE EVENTS:             0
                               RUNS OF 2-4 PREMATURE BEATS:           0

   VF THERAPY           Rx1     Rx2     Rx3     Rx4
   --------------      -----   -----   -----   -----
   INITIATED:              4       1       0       0
   SUCCESSFUL:             3       1       0       0
   ABORTED:                0       0       0       0
   INEFFECTIVE:            1       0       0       0
   CONVERTED TO VT:        0       0       0       0
   CONVERTED TO VTF:       0       0       0       0
   UNDETERMINED:           0       0       0       0
```

```
MEDTRONIC 9760   PROGRAMMER  9858A001              Jul 22, 1992 10:08
Copyright (c) Medtronic, Inc.  1992
PCD 7219                Serial Number:
----- COUNTER DATA REPORT ----------------------------- Page  2 of  2

   VTF THERAPY          Rx1     Rx2     Rx3     Rx4
   --------------      -----   -----   -----   -----
   INITIATED:              0       0       0       0
   REENTERED:              0       0       0       0
   SUCCESSFUL:             0       0       0       0
   ABORTED:                0       0       0       0
   INEFFECTIVE:            0       0       0       0
   CONVERTED TO VT:        0       0       0       0
   CONVERTED TO VF:        0       0       0       0
   UNDETERMINED:           0       0       0       0

   VT THERAPY           Rx1     Rx2     Rx3     Rx4
   --------------      -----   -----   -----   -----
   INITIATED:              0       0       0       0
   SUCCESSFUL:             0       0       0       0
   ABORTED:                0       0       0       0
   INEFFECTIVE:            0       0       0       0
   CONVERTED TO VTF:       0       0       0       0
   CONVERTED TO VF:        0       0       0       0
   UNDETERMINED:           0       0       0       0
```

FIGURE 58.3 Typical data recorded by an implantable defibrillator include stored intracardiac electrograms with annotated markers indicating cardiac intervals, paced and sensed events, and device classification of events (TF = fast tachycardia; TP = antitachy pacing stimulus; VS = sensed nontachy ventricular event). In the example, five rapid pacing pulses convert a ventricular tachycardia with a cycle length of 340 msec into sinus rhythm with a cycle length of 830 msec. In the lower portion of the figure is an example of the summary data collected by the ICD, showing detailed counts of the performance of the various therapies (Rx) for ventricular tachycardia (VT), fast ventricular (VTF), and ventricular (VF). (Copyright Medtronic, Inc. With permission.)

tachycardia, determining the triggers of arrhythmias; documenting rhythm accelerations in response to therapies; and demonstrating appropriate device behavior when treating asymptomatic rhythms.

Electrograms provide useful information by themselves, yet they cannot indicate how the device interpreted cardiac activity. Increasingly, electrogram records are being supplemented with event markers that indicate how the device is responding on a beat-by-beat basis. These records can include measurements of the sensed and paced intervals, indication as to the specific detection zone an event falls in, indication of charge initiation, and other device performance data.

58.6 Follow-Up

Defibrillator patients and their devices require careful follow-up. In one study of 241 ICD patients with epicardial lead systems, 53% of the patients experienced one or more complications during an average exposure of 24 months. These complications included infection requiring device removal in 5%, postoperative respiratory complications in 11%, postoperative bleeding and/or thrombosis in 4%, lead system migration or disruption in 8%, and documented inappropriate therapy delivery, most commonly due to atrial fibrillation, in 22%. A shorter study of 80 patients with transvenous defibrillator systems reported no postoperative pulmonary complications, transient nerve injury (1%), asymptomatic subclavian vein occlusion (2.5%), pericardial effusion (1%), subcutaneous patch pocket hematoma (5%), pulse generator pocket infection (1%), lead fracture (1%), and lead system dislodgement (10%). During a mean follow-up period of 11 months, 7.5% of the patients in this series experienced inappropriate therapy delivery, half for atrial fibrillation and the rest for sinus tachycardia.

Although routine follow-up can be accomplished in the clinic, detection and analysis of transient events depends on the recording capabilities available in the devices or on the use of various external monitoring equipment.

58.7 Economics

The annual cost of ICD therapy is dropping as a consequence of better longevity and simpler implantation techniques. Early generators that lacked programmability, antibradycardia pacing capability, and event recording had 62% survival at 18 months and 2% at 30 months. Some recent programmable designs that include VVI pacing capability and considerable event storage exhibit 96.8% survival at 48 months. It has been estimated that an increase in generator longevity from 2 to 5 years would lower the cost per life-year saved by 55% in a hypothetical patient population with a 3-year sudden mortality of 28%. More efficient energy conversion circuits and finer line-width integrated circuit technology with smaller, more highly integrated circuits and reduced current drains will yield longer-lasting defibrillators while continuing the evolution to smaller volumes.

Cost of the implantation procedure is clearly declining as transvenous lead systems become commonplace. Total hospitalization duration, complication rates, and use of costly hospital operating rooms and intensive care facilities all are reduced, providing significant financial benefits. One study reported requiring half the intensive care unit time and a reduction in total hospitalization from 26 to 15 days when comparing transvenous to epicardial approaches. Another center reported a mean hospitalization stay of 6 days for patients receiving transvenous defibrillation systems.

Increasing sophistication of the implantable defibrillators paradoxically contributes to cost efficacy. Incorporation of single-chamber brady pacing capability eliminates the cost of a separate pacemaker and lead for those patients who need one. Eventually even dual-chamber pacing capability will be available. Programmable detection and therapy features obviate the need for device replacement that was required when fixed parameter devices proved to be inappropriately specified or too inflexible to adapt to a patient's physiologic changes.

Significant cost savings may be obtained by better patient selection criteria and processes, obviating the need for extensive hospitalization and costly electrophysiologic studies prior to device implantation in some patient groups. One frequently discussed issue is the prophylactic role that implantable defibrillators will or should play. Unless a means is found to build far less expensive devices that can be placed with minimal time and facilities, the life-saving yield for prophylactic defibrillators will have to be high if they are to be cost-effective. This remains an open issue.

58.8 Conclusion

The implantable defibrillator is now an established and powerful therapeutic tool. The transition to pectoral implants with biphasic waveforms and efficient yet simple transvenous lead systems is simplifying the implant procedure and drastically reducing the number of unpleasant VF inductions required to demonstrate adequate system performance These advances are making the implantable defibrillator easier to use, less costly, and more acceptable to patients and their physicians.

Acknowledgments

Portions of this text are derived from Duffin E.G. and Barold S.S. 1994. Implantable cardioverter-defibrillators: an overview and future directions, Chapter 28 of I. Singer (ed), Implantable Cardioverter-Defibrillator, and are used with permission of Futura Publishing Company, Inc.

References

Josephson, M. and Wellens H. (eds). (1992). *Tachycardias: Mechanisms and Management.* Mount Kisco, NY, Futura Publishing.

Kappenberger, L. and Lindemans, F. (eds). (1992). *Practical Aspects of Staged Therapy Defibrillators.* Mount Kisco, NY, Futura Publishing.

Singer, I. (ed.). (1994). *Implantable Cardioverter-Defibrillator.* Mount Kisco, NY, Futura Publishing.

Tacker, W. (ed.). (1994). *Defibrillation of the Heart: ICD's, AED's, and Manual.* St. Louis, Mosby.

Memorial issue on implantable defibrillators honoring Michel Mirowski. *PACE*, 14: 865.

59

Implantable Stimulators for Neuromuscular Control

Primoz Strojnik
P. Hunter Peckham
Case Western Reserve University

59.1 Functional Electrical Stimulation

Implantable stimulators for neuromuscular control are the technologically most advanced versions of functional electrical stimulators. Their function is to generate contraction of muscles, which cannot be controlled volitionally because of the damage or dysfunction in the neural paths of the central nervous system (CNS). Their operation is based on the electrical nature of conducting information within nerve fibers, from the neuron cell body (soma), along the axon, where a travelling action potential is the carrier of excitation. While the action potential is naturally generated chemically in the head of the axon, it may also be generated artificially by depolarizing the neuron membrane with an electrical pulse. A train of electrical impulses with certain amplitude, width, and repetition rate, applied to a muscle innervating nerve (a motor neuron) will cause the muscle to contract, very much like in natural excitation. Similarly, a train of electrical pulses applied to the muscular tissue close to the motor point will cause muscle contraction by stimulating the muscle through the neural structures at the motor point.

59.2 Technology for Delivering Stimulation Pulses to Excitable Tissue

A practical system used to stimulate a nerve consists of three components (1) a *pulse generator* to generate a train of pulses capable of depolarizing the nerve, (2) a **lead wire**, the function of which is to deliver the pulses to the stimulation site, and (3) an *electrode*, which delivers the stimulation pulses to the excitable tissue in a safe and efficient manner.

In terms of location of the above three components of an electrical stimulator, stimulation technology can be described in the following terms:

Surface or transcutaneous stimulation, where all three components are outside the body and the electrodes are placed on the skin above or near the motor point of the muscle to be stimulated. This method has been used extensively in medical rehabilitation of nerve and muscle. Therapeutically, it has been used to prevent atrophy of paralyzed muscles, to condition paralyzed muscles before the application of functional stimulation, and to generally increase the muscle bulk. As a functional tool, it has been used in rehabilitation of plegic and paretic patients. Surface systems for functional stimulation have been developed to correct drop-foot condition in hemiplegic individuals [Liberson, 1961], for hand control [Rebersek, 1973], and for standing and stepping in individuals with **paralysis** of the lower extremities [Kralj and Bajd, 1989]. This fundamental technology was commercialized by Sigmedics, Inc. [Graupe, 1998]. The inability of surface stimulation to reliably excite the underlying tissue in a repeatable manner and to selectively stimulate deep muscles has limited the clinical applicability of surface stimulation.

Percutaneous stimulation employs electrodes which are positioned inside the body close to the structures to be stimulated. Their lead wires permanently penetrate the skin to be connected to the external pulse generator. State of the art embodiments of percutaneous electrodes utilize a small-diameter insulated stainless steel lead that is passed through the skin. The electrode structure is formed by removal of the insulation from the lead and subsequent modification to ensure stability within the tissue. This modification includes forming barbs or similar anchoring mechanisms. The percutaneous electrode is implanted using a hypodermic needle as a trochar for introduction. As the needle is withdrawn, the anchor at the electrode tip is engaged into the surrounding tissue and remains in the tissue. A connector at the skin surface, next to the skin penetration point, joins the percutaneous electrode lead to the hardwired external stimulator. The penetration site has to be maintained and care must be taken to avoid physical damage of the lead wires. In the past, this technology has helped develop the existing implantable systems, and it may be used for short and long term, albeit not permanent, stimulation applications [Marsolais, 1986; Memberg, 1993].

The term *implantable stimulation* refers to stimulation systems in which all three components, pulse generator, lead wires, and electrodes, are permanently surgically implanted into the body and the skin is solidly closed after the implantation procedure. Any interaction between the implantable part and the outside world is performed using telemetry principles in a contact-less fashion. This chapter is focused on implantable neuromuscular stimulators, which will be discussed in more detail.

59.3 Stimulation Parameters

In functional **electrical stimulation**, the typical stimulation waveform is a train of rectangular pulses. This shape is used because of its effectiveness as well as relative ease of generation. All three parameters of a stimulation train, that is, frequency, amplitude, and pulse-width, have effect on muscle contraction. Generally, the stimulation frequency is kept as low as possible, to prevent muscle fatigue and to conserve stimulation energy. The determining factor is the muscle fusion frequency at which a smooth muscle response is obtained. This frequency varies; however, it can be as low as 12 to 14 Hz and as high as 50 Hz. In most cases, the stimulation frequency is kept constant for a certain application. This is true both for surface as well as implanted electrodes.

With surface electrodes, the common way of modulating muscle force is by varying the stimulation pulse amplitude at a constant frequency and pulse width. The stimulation amplitudes may be as low as 25 V at 200 μsec for the stimulation of the peroneal nerve and as high as 120 V or more at 300 μsec for activation of large muscles such as the gluteus maximus.

In implantable stimulators and electrodes, the stimulation parameters greatly depend on the implantation site. When the electrodes are positioned on or around the target nerve, the stimulation amplitudes are on the order of a few milliamperes or less. Electrodes positioned on the muscle surface (epimysial electrodes) or in the muscle itself (intramuscular electrodes), employ up to ten times higher amplitudes. For muscle force control, implantable stimulators rely either on pulse-width modulation or amplitude modulation. For example, in upper extremity applications, the current amplitude is usually a fixed paramter set to 16 or 20 mA, while the muscle force is modulated with pulse-widths within 0 to 200 μsec.

59.4 Implantable Neuromuscular Stimulators

Implantable stimulation systems use an encapsulated pulse generator that is surgically implanted and has subcutaneous leads that terminate at electrodes on or near the desired nerves. In low power consumption applications such as the cardiac pacemaker, a primary battery power source is included in the pulse generator case. When the battery is close to depletion, the pulse generator has to be surgically replaced.

Most implantable systems for neuromuscular application consist of an external and an implanted component. Between the two, an inductive radio-frequency link is established, consisting of two tightly coupled resonant coils. The link allows transmission of power and information, through the skin, from the external device to the implanted pulse generator. In more advanced systems, a back-telemetry link is also established, allowing transmission of data outwards, from the implanted to the external component.

Ideally, implantable stimulators for neuromuscular control would be stand alone, totally implanted devices with an internal power source and integrated sensors detecting desired movements from the motor cortex and delivering stimulation sequences to appropriate muscles, thus bypassing the neural damage. At the present developmental stage, they still need a control source and an external controller to provide power and stimulation information. The control source may be either operator driven, controlled by the user, or triggered by an event such as the heel-strike phase of the gait cycle. Figure 59.1 depicts a neuromuscular prosthesis developed at the Case Western Reserve University (CWRU) and Cleveland Veterans Affairs Medical Center for the restoration of hand functions using an implantable neuromuscular stimulator. In this application, the patient uses the shoulder motion to control opening and closing of the hand.

The internal electronic structure of an implantable neuromuscular stimulator is shown in Figure 59.2. It consists of receiving and data retrieval circuits, power supply, data processing circuits, and output stages.

59.4.1 Receiving Circuit

The stimulator's receiving circuit is an LC circuit tuned to the resonating frequency of the external transmitter, followed by a rectifier. Its task is to provide the raw DC power from the received **rf** signal and at the same time allow extraction of stimulation information embedded in the rf carrier. There are various encoding schemes allowing simultaneous transmission of power and information into an implantable electronic device. They include amplitude and frequency modulation with different modulation indexes as well as different versions of digital encoding such as Manchester encoding where the information is

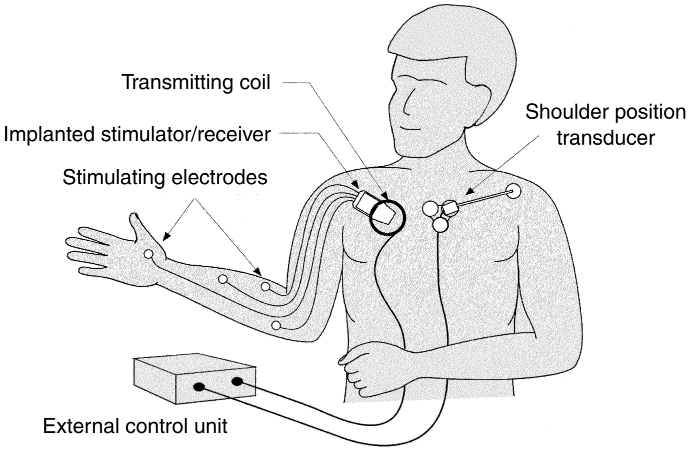


FIGURE 59.1 Implanted FES hand grasp system.

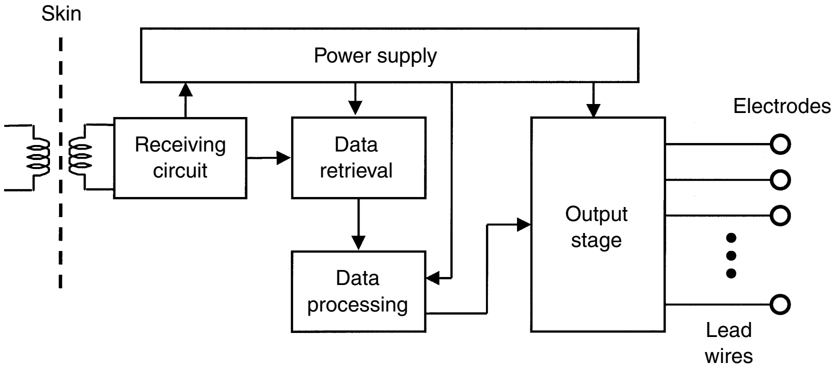


FIGURE 59.2 Block diagram of an implantable neuromuscular stimulator.

hidden in a logic value transition position rather than the logic value itself. Synchronous and asynchronous clock signals may be extracted from the modulated carrier to drive the implant's logic circuits.

The use of radiofrequency transmission for medical devices is regulated and in most countries limited to certain frequencies and radiation powers. (In the United States, the use of the rf space is regulated by the Federal Communication Commission [FCC].) Limited rf transmission powers as well as conservation of power in battery operated external controllers dictate high coupling efficiencies between the transmitting and receiving antennas. Optimal coupling parameters cannot be uniformly defined; they depend on application particularities and design strategies.

59.4.2 Power Supply

The amount of power delivered into an implanted electronic package depends on the coupling between the transmitting and the receiving coil. The coupling is dependent on the distance as well as the alignment between the coils. The power supply circuits must compensate for the variations in distance for different users as well as for the alignment variations due to skin movements and consequent changes in relative coil-to-coil position during daily usage. The power dissipated on power supply circuits must not raise the overall implant case temperature.

In implantable stimulators that require stimulation voltages in excess of the electronics power supply voltages (20 to 30 V), the stimulation voltage can be provided directly through the receiving coil. In that case, voltage regulators must be used to provide the electronics supply voltage (usually 5 V), which heavily taxes the external power transmitter and increases the implant internal power dissipation.

59.4.3 Data Retrieval

Data retrieval technique depends on the data-encoding scheme and is closely related to power supply circuits and implant power consumption. Most commonly, amplitude modulation is used to encode the in-going data stream. As the high quality factor of resonant LC circuits increases the efficiency of power transmission, it also effectively reduces the transmission bandwidth and therefore the transmission data rate. Also, high quality circuits are difficult to amplitude modulate since they tend to continue oscillating even with power removed. This has to be taken into account when designing the communication link in particular for the start-up situation when the implanted device does not use the power for stimulation and therefore loads the transmitter side less heavily, resulting in narrower and higher resonant curves. The load on the receiving coil may also affect the low pass filtering of the received rf signal.

Modulation index (m) or depth of modulation affects the overall energy transfer into the implant. At a given rf signal amplitude, less energy is transferred into the implanted device when 100% modulation is used ($m = 1$) as compared to 10% modulation ($m = 0.053$). However, retrieval of 100% modulated signal is much easier than retrieval of a 10% modulated signal.

59.4.4 Data Processing

Once the information signal has been satisfactorily retrieved and reconstructed into logic voltage levels, it is ready for logic processing. For synchronous data processing a clock signal is required. It can be generated locally within the implant device, reconstructed from the incoming data stream, or can be derived from the rf carrier. A crystal has to be used with a local oscillator to assure stable clock frequency. Local oscillator allows for asynchronous data transmission. Synchronous transmission is best achieved using Manchester data encoding. Decoding of Manchester encoded data recovers the original clock signal, which was used during data encoding. Another method is using the downscaled rf carrier signal as the clock source. In this case, the information signal has to be synchronized with the rf carrier. Of course, 100% modulation scheme cannot be used with carrier-based clock signal. Complex command structure used in multichannel stimulators requires intensive data decoding and processing and consequently extensive electronic circuitry. Custom-made, application specific circuits (ASIC) are commonly used to minimize the space requirements and optimize the circuit performance.

59.4.5 Output Stage

The output stage forms stimulation pulses and defines their electrical characteristics. Even though a mere rectangular pulse can depolarize a nervous membrane, such pulses are not used in clinical practice due to their noxious effect on the tissue and **stimulating electrodes**. These effects can be significantly reduced by charge balanced stimulating pulses where the cathodic stimulation pulse is followed by an anodic pulse containing the same electrical charge, which reverses the electrochemical effects of the cathodic pulse. Charge balanced waveforms can be assured by capacitive coupling between the pulse generator and stimulation electrodes. Charge balanced stimulation pulses include symmetrical and asymmetrical waveforms with anodic phase immediately following the cathodic pulse or being delayed by a short, 20 to 60 μsec interval.

The output stages of most implantable neuromuscular stimulators have constant current characteristics, meaning that the output current is independent on the electrode or tissue impedance. Practically, the constant current characteristics ensure that the same current flows through the excitable tissues regardless of the changes that may occur on the electrode-tissue interface, such as the growth of fibrous tissue around the electrodes. Constant current output stage can deliver constant current only within the supply voltage — compliance voltage. In neuromuscular stimulation, with the electrode impedance being on the order of 1 kΩ, and the stimulating currents in the order of 20 mA, the compliance voltage must be above

20 V. Considering the voltage drops and losses across electronic components, the compliance voltage of the output stage may have to be as high as 33 V.

The stimulus may be applied through either monopolar or bipolar electrodes. The monopolar electrode is one in which a single active electrode is placed near the excitable nerve and the return electrode is placed remotely, generally at the implantable unit itself. Bipolar electrodes are placed at the stimulation site, thus limiting the current paths to the area between the electrodes. Generally, in monopolar stimulation the active electrode is much smaller than the return electrode, while bipolar electrodes are the same size.

59.5 Packaging of Implantable Electronics

Electronic circuits must be protected from the harsh environment of the human body. The packaging of implantable electronics uses various materials, including polymers, metals, and ceramics. The encapsulation method depends somewhat on the electronic circuit technology. Older devices may still use discrete components in a classical form, such as leaded transistors and resistors. The newer designs, depending on the sophistication of the implanted device, may employ application-specific integrated circuits (ASICs) and thick film hybrid circuitry for their implementation. Such circuits place considerable requirements for hermeticity and protection on the implanted circuit packaging.

Epoxy encapsulation was the original choice of designers of implantable neuromuscular stimulators. It has been successfully used with relatively simple circuits using discrete, low impedance components. With epoxy encapsulation, the receiving coil is placed around the circuitry to be "potted" in a mold, which gives the implant the final shape. Additionally, the epoxy body is coated with silicone rubber that improves the **biocompatibility** of the package. Polymers do not provide an impermeable barrier and therefore cannot be used for encapsulation of high density, high impedance electronic circuits. The moisture ingress ultimately will reach the electronic components, and surface ions can allow electric shorting and degradation of leakage-sensitive circuitry and subsequent failure.

Hermetic packaging provides the implant electronic circuitry with a long-term protection from the ingress of body fluids. Materials that provide hermetic barriers are metals, ceramics, and glasses. Metallic packaging generally uses a titanium capsule machined from a solid piece of metal or deep-drawn from a piece of sheet metal. Electrical signals, such as power and stimulation, enter and exit the package through hermetic feedthroughs, which are hermetically welded onto the package walls. The **feedthrough** assembly utilizes a ceramic or glass insulator to allow one or more wires to exit the package without contact with the package itself. During the assembly procedures, the electronic circuitry is placed in the package and connected internally to the feedthroughs, and the package is then welded closed. Tungsten Inert Gas (TIG), electron beam, or laser welding equipment is used for the final closure. Assuming integrity of all components, hermeticity with this package is ensured. This integrity can be checked by detecting gas leakage from the capsule. Metallic packaging requires that the receiving coil be placed outside the package to avoid significant loss of rf signal or power, thus requiring additional space within the body to accommodate the volume of the entire implant. Generally, the hermetic package and the receiving antenna are jointly imbedded in an epoxy encapsulant, which provides electric isolation for the metallic antenna and stabilizes the entire implant assembly. Figure 59.3 shows such an implantable stimulator designed and made by the CWRU/Veterans Administration Program. The hermetic package is open, displaying the electronic **hybrid circuit**. More recently, alumina-based ceramic packages have been developed that allow hermetic sealing of the electronic circuitry together with enclosure of the receiving coil [Strojnik, 1994]. This is possible due to the rf transparency of ceramics. The impact of this type of enclosure is still not fully investigated. The advantage of this approach is that the volume of the implant can be reduced, thus minimizing the biologic response, which is a function of volume. Yet, an unexplored issue of this packaging method is the effect of powerful electromagnetic fields on the implant circuits, lacking the protection of the metal enclosure. This is a particular concern with high gain (EMG, ENG, or EKG sensing) amplifiers, which in the future may be included in the implant package as part of back-telemetry circuits. Physical strength of ceramic packages and their resistance to impact will also require future investigation.

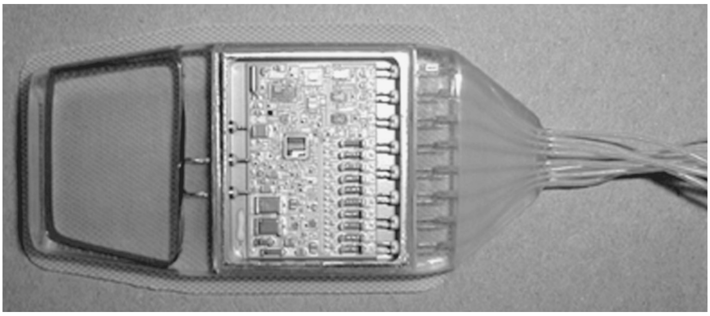


FIGURE 59.3 Photograph of a multichannel implantable stimulator telemeter. Hybrid circuit in titanium package is shown exposed. Receiving coil (left) is imbedded in epoxy resin together with titanium case. Double feedthroughs are seen penetrating titanium capsule wall on the right.

59.6 Leads and Electrodes

Leads connect the pulse generator to the electrodes. They must be sufficiently flexible to move across the joints while at the same time sufficiently sturdy to last for the decades of the intended life of the device. They must also be stretchable to allow change of distance between the pulse generator and the electrodes, associated with body movements. Ability to flex and to stretch is achieved by coiling the lead conductor into a helix and inserting the helix into a small-diameter silicone tubing. This way, both flexing movements and stretching forces exerted on the lead are attenuated, while translated into torsion movements and forces exerted on the coiled conductor. Using multi-strand rather than solid conductors further enhances the longevity. Several individually insulated multi-strand conductors can be coiled together, thus forming a multiple conductor lead wire. Most lead configurations include a connector at some point between the implant and the terminal electrode, allowing for replacement of the implanted receiver or leads in the event of failure. The connectors used have been either single pin in-line connectors located somewhere along the lead length or a multiport/multilead connector at the implant itself. Materials used for lead wires are stainless steels, MP35N (Co, Cr, Ni alloy), and noble metals and their alloys.

Electrodes deliver electrical charge to the stimulated tissues. Those placed on the muscle surface are called epimysial, while those inserted into the muscles are called intramuscular. Nerve stimulating electrodes are called epineural when placed against the nerve, or cuff electrodes when they encircle the nerve. Nerve electrodes may embrace the nerve in a spiral manner individually, or in an array configuration. Some implantable stimulation systems merely use exposed lead-wire conductor sutured to the epineurium as the electrode. Generally, nerve electrodes require approximately one-tenth of the energy for muscle activation as compared to muscle electrodes. However, they require more extensive surgery and may be less selective, but the potential for neural damage is greater than, for example, nerve encircling electrodes.

Electrodes are made of corrosion resistant materials, such as noble metals (platinum or iridium) and their alloys. For example, a platinum–iridium alloy consisting of 10% iridium and 90% platinum is commonly used as an electrode material. Epimysial electrodes developed at CWRU use Ø4 mm Pt90Ir10 discs placed on Dacron reinforced silicone backing. CWRU intramuscular electrodes employ a stainless steel lead-wire with the distal end de-insulated and configured into an electrode tip. A small, umbrella-like anchoring barb is attached to it. With this arrangement, the diameter of the electrode tip does not differ much from the lead wire diameter and this electrode can be introduced into a deep muscle with a trochar-like insertion tool. Figure 59.4 shows enlarged views of these electrodes.

59.7 Safety Issues of Implantable Stimulators

The targeted lifetime of implantable stimulators for neuromuscular control is the lifetime of their users, which is measured in tens of years. Resistance to premature failure must be assured by manufacturing

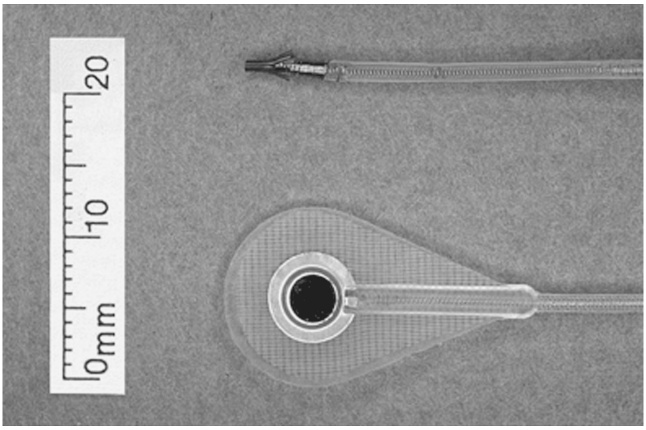


FIGURE 59.4 Implantable electrodes with attached lead wires. Intramuscular electrode (top) has stainless steel tip and anchoring barbs. Epimysial electrode has PtIr disk in the center and is backed by silicone-impregnated Dacron mesh.

processes and testing procedures. Appropriate materials must be selected that will withstand the working environment. Protection against mechanical and electrical hazards that may be encountered during the device lifetime must be incorporated in the design. Various procedures are followed and rigorous tests must be performed during and after its manufacturing to assure the quality and reliability of the device.

- *Manufacturing and testing*— Production of implantable electronic circuits and their encapsulation in many instances falls under the standards governing production and encapsulation of integrated circuits. To minimize the possibility of failure, the implantable electronic devices are manufactured in controlled clean-room environments, using high quality components and strictly defined manufacturing procedures. Finished devices are submitted to rigorous testing before being released for implantation. Also, many tests are carried out during the manufacturing process itself. To assure maximum reliability and product confidence, methods, tests, and procedures defined by military standards, such as MILSTD-883, are followed.
- *Bio-compatibility*— Since the implantable stimulators operate surgically implanted in living tissue, an important part of their design has to be dedicated to biocompatibility, that is, their ability to dwell in living tissue without disrupting the tissue in its functions, creating adverse tissue response, or changing its own properties due to the tissue environment. Elements of biocompatibility include tissue reaction to materials, shape, and size, as well as electrochemical reactions on stimulation electrodes. There are known biomaterials used in the making of implantable stimulators. They include stainless steels, titanium and tantalum, noble metals such as platinum and iridium, as well as implantable grades of selected epoxy and silicone-based materials.
- *Susceptibility to electromagnetic interference (EMI) and electrostatic discharge (ESD)* — Electromagnetic fields can disrupt the operation of electronic devices, which may be lethal in situations with life support systems, but they may also impose risk and danger to users of neuromuscular stimulators. Emissions of EMI may come from outside sources; however, the external control unit is also a source of electromagnetic radiation. Electrostatic discharge shocks are not uncommon during the dry winter season. These shocks may reach voltages as high as 15 kV and more. Sensitive electronic components can easily be damaged by these shocks unless protective design measures are taken. The electronic circuitry in implantable stimulators is generally protected by the metal case. However, the circuitry can be damaged through the feedthroughs either by handling or during the implantation procedure by the electrocautery equipment. ESD damage may happen even after implantation when long lead-wires are utilized. There are no standards directed specifically towards implantable electronic devices. The general standards put in place for

electromedical equipment by the International Electrotechnical Commission provide guidance. The specifications require survival after 3 kV and 8 kV ESD discharges on all conductive and nonconductive accessible parts, respectively.

59.8 Implantable Stimulators in Clinical Use

59.8.1 Peripheral Nerve Stimulators

- *Manipulation* — Control of complex functions for movement, such as hand control, requires the use of many channels of stimulation. At the Case Western Reserve University and Cleveland VAMC, an eight-channel stimulator has been developed for grasp and release [Smith, 1987]. This system uses eight channels of stimulation and a titanium-packaged, thick-film hybrid circuit as the pulse generator. The implant is distributed by the Neurocontrol Corporation (Cleveland, OH) under the name of Freehand®. It has been implanted in approximately 150 patients in the United States, Europe, Asia, and Australia. The implant is controlled by a dual-microprocessor external unit carried by the patient with an input control signal provided by the user's remaining volitional movement. Activation of the muscles provides two primary grasp patterns and allows the person to achieve functional performance that exceeds his or her capabilities without the use of the implanted system. This system received pre-market approval from the FDA in 1998.
- *Locomotion* — The first implantable stimulators were designed and implanted for the correction of the foot drop condition in hemiplegic patients. Medtronic's Neuromuscular Assist (NMA) device consisted of an rf receiver implanted in the inner thigh and connected to a cuff electrode embracing the peroneal nerve just beneath the head of fibula at the knee [McNeal, 1977; Waters 1984]. The Ljubljana peroneal implant had two versions [Vavken, 1976; Strojnik, 1987] with the common feature that the implant–rf receiver was small enough to be implanted next to the peroneal nerve in the fossa poplitea region. Epineural stimulating electrodes were an integral part of the implant. This feature and the comparatively small size make the Ljubljana implant a precursor of the microstimulators described in Section 59.9. Both NMA and the Ljubljana implants were triggered and synchronized with gait by a heel switch.

  The same implant used for hand control and developed by the CWRU has also been implanted in the lower extremity musculature to assist incomplete quadriplegics in standing and transfer operations [Triolo, 1996]. Since the design of the implant is completely transparent, it can generate any stimulation sequence requested by the external controller. For locomotion and transfer-related tasks, stimulation sequences are preprogrammed for individual users and activated by the user by means of pushbuttons. The implant (two in some applications) is surgically positioned in the lower abdominal region. Locomotion application uses the same electrodes as the manipulation system; however, the lead wires have to be somewhat longer.
- *Respiration* — Respiratory control systems involve a two-channel implantable stimulator with electrodes applied bilaterally to the phrenic nerve. Most of the devices in clinical use were developed by Avery Laboratories (Dobelle Institute) and employed discrete circuitry with epoxy encapsulation of the implant and a nerve cuff electrode. Approximately 1000 of these devices have been implanted in patients with respiratory disorders such as high-level tetraplegia [Glenn, 1986]. Activation of the phrenic nerve results in contraction of each hemidiaphragm in response to electrical stimulation. In order to minimize damage to the diaphragms during chronic use, alternation of the diaphragms has been employed, in which one hemidiaphragm will be activated for several hours followed by the second. A review of existing systems was given by Creasy et al. [1996]. Astrotech of Finland also recently introduced a phrenic stimulator. More recently, DiMarco [1997] has investigated use of CNS activation of a respiratory center to provide augmented breathing.
- *Urinary control* — Urinary control systems have been developed for persons with spinal cord injury. The most successful of these devices has been developed by Brindley [1982] and is manufactured by Finetech, Ltd. (England). The implanted receiver consists of three separate stimulator devices,

each with its own coil and circuitry, encapsulated within a single package. The sacral roots (S2, S3, and S4) are placed within a type of encircling electrode, and stimulation of the proper roots will generate contraction of both the bladder and the external sphincter. Cessation of stimulation results in faster relaxation of the external sphincter than of the bladder wall, which then results in voiding. Repeated trains of pulses applied in this manner will eliminate most urine, with only small residual amounts remaining. Approximately 1500 of these devices have been implanted around the world. This technology also has received FDA pre-market approval and is currently distributed by NeuroControl Corporation.

- *Scoliosis treatment* — Progressive lateral curvature of the adolescent vertebral column with simultaneous rotation is known as idiopathic scoliosis. Electrical stimulation applied to the convex side of the curvature has been used to stop or reduce its progression. Initially rf powered stimulators have been replaced by battery powered totally implanted devices [Bobechko, 1979; Herbert, 1989]. Stimulation is applied intermittently, stimulation amplitudes are under 10.5 V (510 Ω), and frequency and pulsewidth are within usual FES parameter values.

59.8.2 Stimulators of Central Nervous System

Some stimulation systems have electrodes implanted on the surface of the central nervous system or in its deep areas. They do not produce functional movements; however, they "modulate" a pathological motor brain behavior and by that stop unwanted motor activity or abnormality. Therefore, they can be regarded as stimulators for neuromuscular control.

- *Cerebellar stimulation* — Among the earliest stimulators from this category are cerebellar stimulators for control of reduction of effects of cerebral palsy in children. Electrodes are placed on the cerebellar surface with the leads penetrating cranium and dura. The pulse generator is located subcutaneously in the chest area and produces intermittent stimulation bursts. There are about 600 patients using these devices [Davis, 1997].
- *Vagal stimulation* — Intermittent stimulation of the vagus nerve with 30 sec on and five min off has been shown to reduce frequency of epileptic seizures. A pacemaker-like device, developed by Cyberonics, is implanted in the chest area with a bipolar helical electrode wrapped around the left vagus nerve in the neck. The stimulation sequence is programmed (most often parameter settings are 30 Hz, 500 μsec, 1.75 mA); however, patients have some control over the device using a hand-held magnet [Terry, 1991]. More than 3000 patients have been implanted with this device, which received the pre-marketing approval (PMA) from the FDA in 1997.
- *Deep brain stimulation* — Recently, in 1998, an implantable stimulation device (Activa by Medtronic) was approved by the FDA that can dramatically reduce uncontrollable tremor in patients with Parkinson's disease or essential tremor [Koller, 1997]. With this device, an electrode array is placed stereotactically into the ventral intermediate nucleus of thalamic region of the brain. Lead wires again connect the electrodes to a programmable pulse generator implanted in the chest area. Application of high frequency stimulation (130 Hz, 60 to 210 μsec, 0.25 to 2.75 V) can immediately suppress the patient's tremor.

59.9 Future of Implantable Electrical Stimulators

59.9.1 Distributed Stimulators

One of the major concerns with multichannel implantable neuromuscular stimulators is the multitude of leads that exit the pulse generator and their management during surgical implantation. Routing of multiple leads virtually increases the implant size and by that the burden that an implant imposes on the tissue. A solution to that may be distributed stimulation systems with a single outside controller and multiple single-channel implantable devices implanted throughout the structures to be stimulated.

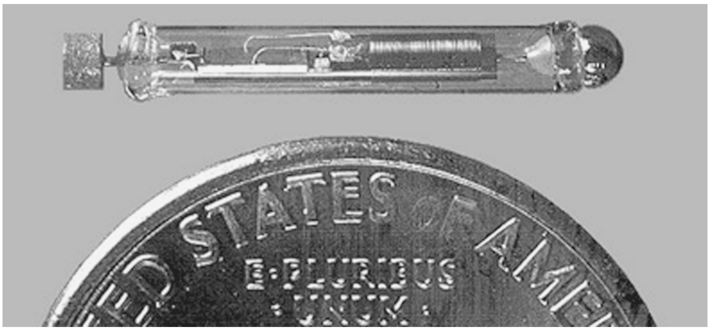


FIGURE 59.5 Microstimulator developed at A.E. Mann Foundation. Dimensions are roughly 2 × 16 mm. Electrodes at the ends are made of tantalum and iridium, respectively.

This concept has been pursued both by the Alfred E. Mann Foundation [Strojnik, 1992; Cameron, 1997] and the University of Michigan [Ziaie, 1997]. Micro-injectable stimulator modules have been developed that can be injected into the tissue, into a muscle, or close to a nerve through a lumen of a hypodermic needle. A single external coil can address and activate a number of these devices located within its field, on a pulse-to-pulse basis. A glass-encapsulated microstimulator developed at the AEMF is shown in Figure 59.5.

59.9.2 Sensing of Implantable Transducer-Generated and Physiological Signals

External command sources such as the shoulder-controlled joystick utilized by the Freehand® system impose additional constraints on the implantable stimulator users, since they have to be donned by an attendant. Permanently implanted control sources make neuro-prosthetic devices much more attractive and easier to use. An implantable joint angle transducer (IJAT) has been developed at the CWRU that consists of a magnet and an array of magnetic sensors implanted in the distal and the proximal end of a joint, respectively [Smith, 1998]. The sensor is connected to the implantable stimulator package, which provides the power and also transmits the sensor data to the external controller, using a back-telemetry link. Figure 59.6 shows a radiograph of the IJAT implanted in a patient's wrist. Myoelectric signals (MES) from muscles not affected by paralysis are another attractive control source for implantable neuromuscular stimulators. Amplified and bin-integrated EMG signal from uninvolved muscles, such as the sterno-cleido-mastoid muscle, has been shown to contain enough information to control an upper extremity neuroprosthesis [Scott, 1996]. EMG signal is being utilized by a multichannel stimulator-telemeter developed at the CWRU, containing 12 stimulator channels and 2 MES channels integrated into the same platform [Strojnik, 1998].

59.10 Summary

Implantable stimulators for neuromuscular control are an important tool in rehabilitation of paralyzed individuals with preserved neuro-muscular apparatus, as well as in the treatment of some neurological disorders that result in involuntary motor activity. Their impact on rehabilitation is still in its infancy; however, it is expected to increase with further progress in microelectronics technology, development of smaller and better sensors, and with improvements of advanced materials. Advancements in neuro-physiological science are also expected to bring forward wider utilization of possibilities offered by implantable neuromuscular stimulators.

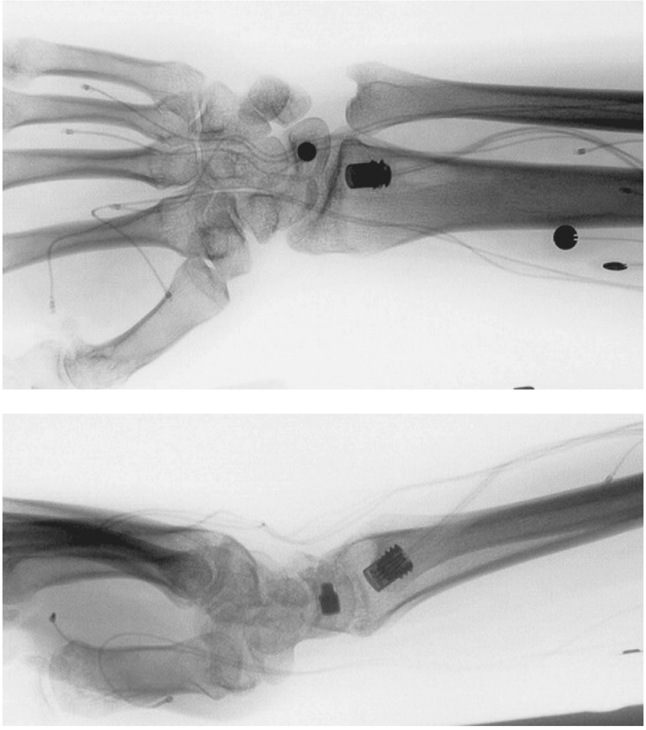


FIGURE 59.6 Radiograph of the joint angle transducer (IJAT) implanted in the wrist. The magnet is implanted in the lunate bone (top) while the magnetic sensor array is implanted in the radius. Leads going to the implant case can be seen as well as intramuscular and epimysial electrodes with their individual lead wires.

Defining Terms

Biocompatibility: Ability of a foreign object to coexist in a living tissue.

Electrical stimulation: Diagnostic, therapeutic, and rehabilitational method used to excite motor nerves with the aim of contracting the appropriate muscles and obtain limb movement.

EMG activity: Muscular electrical activity associated with muscle contraction and production of force.

Feedthrough: Device that allows passage of a conductor through a hermetic barrier.

Hybrid circuit: Electronic circuit combining miniature active and passive components on a single ceramic substrate.

Implantable stimulator: Biocompatible electronic stimulator designed for surgical implantation and operation in a living tissue.

Lead wire: Flexible and strong insulated conductor connecting pulse generator to stimulating electrodes.

Paralysis: Loss of power of voluntary movement in a muscle through injury to or disease to its nerve supply.

Stimulating electrode: Conductive device that transfers stimulating current to a living tissue. On its surface, the electric charge carriers change from electrons to ions or vice versa.

rf-radiofrequency: Pertaining to electromagnetic propagation of power and signal in frequencies above those used in electrical power distribution.

References

Bobechko, W.P., Herbert, M.A., and Friedman, H.G. 1979. Electrospinal instrumentation for scoliosis: current status. *Orthop. Clin. North. Am.* 10: 927.

Brindley, G.S., Polkey, C.E., and Rushton, D.N. 1982. Sacral anterior root stimulators for bladder control in paraplegia. *Paraplegia* 20: 365.

Cameron, T., Loeb, G.E., Peck, R.A., Schulman, J.H., Strojnik, P., and Troyk, P.R. 1997. Micromodular implants to provide electrical stimulation of paralyzed muscles and limbs. *IEEE Trans. Biomed. Eng.* 44: 781.

Creasey, G., Elefteriades, J., DiMarco, A., Talonen, P., Bijak, M., Girsch, W., and Kantor, C. 1996. Electrical stimulation to restore respiration. *J. Rehab. Res. Dev.* 33: 123.

Davis, R.: 1997. Cerebellar stimulation for movement disorders. In P.L. Gildenberg and R.R. Tasker (eds), *Textbook of Stereotactic and Functional Neurosurgery*, McGraw-Hill, New York.

DiMarco, A.F., Romaniuk, J.R., Kowalski, K.E., and Supinski, G.S. 1997. Efficacy of combined inspiratory intercostal and expiratory muscle pacing to maintain artificial ventilation. *Am. J. Respir. Crit. Care Med.* 156: 122.

Glenn, W.W., Phelps, M.L., Elefteriades, J.A., Dentz, B., and Hogan, J.F. 1986. Twenty years of experience in phrenic nerve stimulation to pace the diaphragm pacing. *Clin. Electrophysiol.* 9: 780.

Graupe, D. and Kohn, K.H. 1998. Functional neuromuscular stimulator for short-distance ambulation by certain thoracic-level spinal-cord-injured paraplegics. *Surg. Neurol.* 50: 202.

Herbert, M.A. and Bobechko, W.P. 1989. Scoliosis treatment in children using a programmable, totally implantable muscle stimulator (ESI). *IEEE Trans. Biomed. Eng.* 36: 801.

Koller, W., Pahwa, R., Busenbark, K., Hubble, J., Wilkinson, S., Lang, A., Tuite, P., Sime, E., Lazano, A., Hauser, R., Malapira, T., Smith, D., Tarsy, D., Miyawaki, E., Norregaard, T., Kormos, T., and Olanow, C.W. 1997. High-frequency unilateral thalamic stimulation in the treatment of essential and parkinsonian tremor. *Ann. Neurol.* 42: 292.

Kralj, A. and Bajd, T. 1989. *Functional Electrical Stimulation: Standing and Walking After Spinal Cord Injury*, CRC Press, Inc., Boca Raton, FL.

Liberson, W.T., Holmquest, H.J., Scot, D., and Dow, M. 1961. Functional electrotherapy: stimulation of the peroneal nerve synchronized with the swing phase of the gait of hemiplegic patients. *Arch. Phys. Med. Rehab.* 42: 101.

Marsolais, E.B. and Kobetic, R. 1986. Implantation techniques and experience with percutaneous intramuscular electrodes in lower extremities. *J. Rehab. Res. Dev.* 23: 1.

McNeal, D.R., Waters, R., and Reswick, J. 1977. Experience with implanted electrodes. *Neurosurgery* 1: 228.

Memberg, W., Peckham, P.H., Thorpe, G.B., Keith, M.W., and Kicher, T.P. 1993. An analysis of the reliability of percutaneous intramuscular electrodes in upper extremity FNS applications. *IEEE Trans. Biomed. Eng.* 1: 126.

Rebersek, S. and Vodovnik, L. 1973. Proportionally controlled functional electrical stimulation of hand. *Arch. Phys. Med. Rehab.* 54: 378.

Scott, T.R.D., Peckham, P.H., and Kilgore, K.L. 1996. Tri-state myoelectric control of bilateral upper extremity neuroprosthesies for tetraplegic individuals. *IEEE Trans. Rehab. Eng.* 2: 251.

Smith, B., Peckham, P.H., Keith, M.W., and Roscoe, D.D. 1987. An externally powered, multichannel, implantable stimulator for versatile control of paralyzed muscle. *IEEE Trans. Biomed. Eng.* 34: 499.

Smith, B., Tang, Johnson, M.W., Pourmehdi, S., Gazdik, M.M., Buckett, J.R., and Peckham, P.H. 1998. An externally powered, multichannel, implantable stimulator-telemeter for control of paralyzed muscle. *IEEE Trans. Biomed. Eng.* 45: 463.

Strojnik, P., Pourmehdi, S., and Peckham, P. 1998. Incorporating FES control sources into implanatable stimulators. *Proc. 6th Vienna International Workshop on Functional Electrostimulation*, Vienna, Austria.

Strojnik, P., Meadows, P., Schulman, J.H., and Whitmoyer, D. 1994. Modification of a cochlear stimulation system for FES applications. *Basic Appl. Myology.* BAM 4: 129.

Strojnik, P., Acimovic, R., Vavken, E., Simic, V., and Stanic, U. 1987. Treatment of drop foot using an implantable peroneal underknee stimulator. *Scand. J. Rehab. Med.* 19: 37.

Strojnik, P., Schulman, J., Loeb, G., and Troyk, P. 1992. Multichannel FES system with distributed microstimulators. *Proc. 14th Ann. Int. Conf. IEEE*, MBS, Paris, p. 1352.

Terry, R.S., Tarver, W.B., and Zabara, J. 1991. The implantable neurocybernetic prosthesis system. *Pacing Clin. Electrophysiol.* 14: 86.

Triolo, R.J., Bieri, C., Uhlir, J., Kobetic, R., Scheiner, A., and Marsolais, E.B. 1996. Implanted functional neuromuscular stimulation systems for individuals with cervical spinal cord injuries: clinical case reports. *Arch. Phys. Med. Rehabil.* 77: 1119.

Vavken, E. and Jeglic, A. 1976. Application of an implantable stimulator in the rehabilitation of paraplegic patients. *Int. Surg.* 61: 335–339.

Waters, R.L., McNeal, D.R., and Clifford, B. 1984. Correction of footdrop in stroke patients via surgically implanted peroneal nerve stimulator. *Acta Orthop. Belg.* 50: 285.

Ziaie, B., Nardin, M.D., Coghlan, A.R., and Najafi, K. 1997. A single-channel implantable microstimulator for functional neuromuscular stimulation. *IEEE Trans. Biomed. Eng.* 44: 909.

Further Information

Additional references on early work in FES which augment peer review publications can be found in Proceedings from Conferences in Dubrovnik and Vienna. These are the External Control of Human Extremities and the Vienna International Workshop on Electrostimulation, respectively.

60

Respiration

Leslie A. Geddes
Purdue University

60.1 Lung Volumes

The amount of air flowing into and out of the lungs with each breath is called the tidal volume (TV). In a typical adult this amounts to about 500 ml during quiet breathing. The respiratory system is capable of moving much more air than the tidal volume. Starting at the *resting expiratory level* (REL in Figure 60.1), it is possible to inhale a volume amounting to about seven times the tidal volume; this volume is called the *inspiratory capacity* (IC). A measure of the ability to inspire more than the tidal volume is the *inspiratory reserve volume* (IRV), which is also shown in Figure 60.1. Starting from REL, it is possible to forcibly exhale a volume amounting to about twice the tidal volume; this volume is called the *expiratory reserve volume* (ERV). However, even with the most forcible expiration, it is not possible to exhale all the air from the lungs; a *residual volume* (RV) about equal to the expiratory reserve volume remains. The sum of the expiratory reserve volume and the residual volume is designated the *functional residual capacity* (FRC). The volume of air exhaled from a maximum inspiration to a maximum expiration is called the *vital capacity* (VC). The *total lung capacity* (TLC) is the total air within the lungs, that is, that which can be moved in a vital-capacity maneuver plus the residual volume. All except the residual volume can be determined with a volume-measuring instrument such as a spirometer connected to the airway.

60.2 Pulmonary Function Tests

In addition to the static lung volumes just identified, there are several time-dependent volumes associated with the respiratory act. The *minute volume* (MV) is the volume of air per breath (tidal volume) multiplied by the respiratory rate (R), that is, $MV = (TV)\ R$. It is obvious that the same minute volume can be produced by rapid shallow or slow deep breathing. However, the effectiveness is not the same, because not all the respiratory air participates in gas exchange, there being a dead space volume. Therefore the alveolar ventilation is the important quantity which is defined as the tidal volume (TV) minus the dead space (DS) multiplied by the respiratory rate R, that is, alveolar ventilation $= (TV - DS)\ R$. In a normal adult subject, the dead space amounts to about 150 ml, or 2 ml/kg.

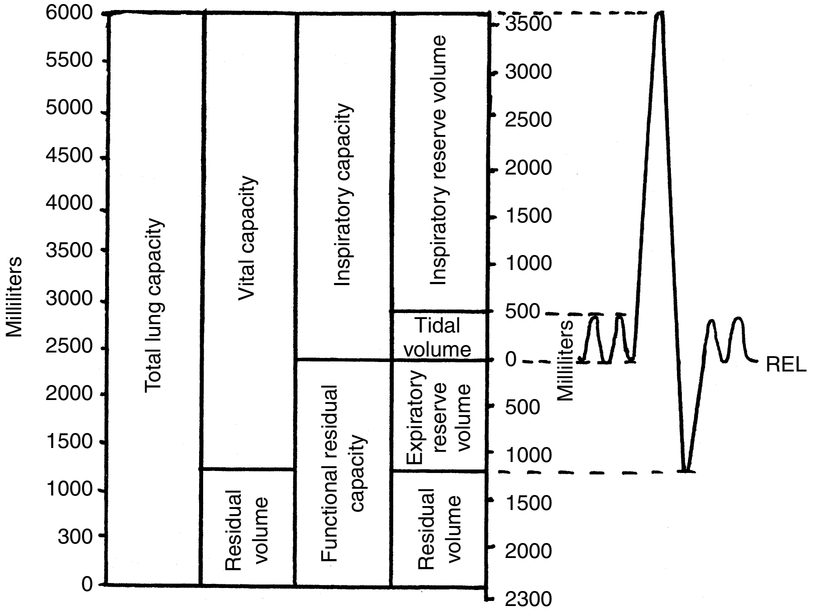


FIGURE 60.1 Lung volumes.

60.2.1 Dynamic Tests

Several timed respiratory volumes describe the ability of the respiratory system to move air. Among these are *forced vital capacity* (FVC), *forced expiratory volume* in t sec (FEV$_t$), the *maximum ventilatory volume* (MVV), which was previously designated the *maximum breathing capacity* (MBC), and the *peak flow* (PF). These quantities are measured with a spirometer without valves and CO_2 absorber or with a pneumotachograph coupled to an integrator.

60.2.1.1 Forced Vital Capacity

Forced vital capacity (FVC) is shown in Figure 60.2 and is measured by taking the maximum inspiration and forcing all of the inspired air out as rapidly as possible. Table 60.1 presents normal values for males and females.

60.2.1.2 Forced Expiratory Volume

Forced expiratory volume in t seconds (FEV$_t$) is shown in Figure 60.2, which identifies FEV$_{0.5}$ and FEV$_{1.0}$, and Table 60.1 presents normal values for FEV$_{1.0}$.

60.2.1.3 Maximum Voluntary Ventilation

Maximum voluntary ventilation (MVV) is the volume of air moved in 1 min when breathing as deeply and rapidly as possible. The test is performed for 20 sec and the volume scaled to a 1-min value; Table 60.1 presents normal values.

60.2.1.4 Peak Flow

Peak flow (PF) in l/min is the maximum flow velocity attainable during an FEV maneuver and represents the maximum slope of the expired volume–time curve (Figure 60.2); typical normal values are shown in Table 60.1.

60.2.1.5 The Water-Sealed Spirometer

The water-sealed spirometer was the traditional device used to measure the volume of air moved in respiration. The Latin word *spirare* means to breathe. The most popular type of spirometer consists of

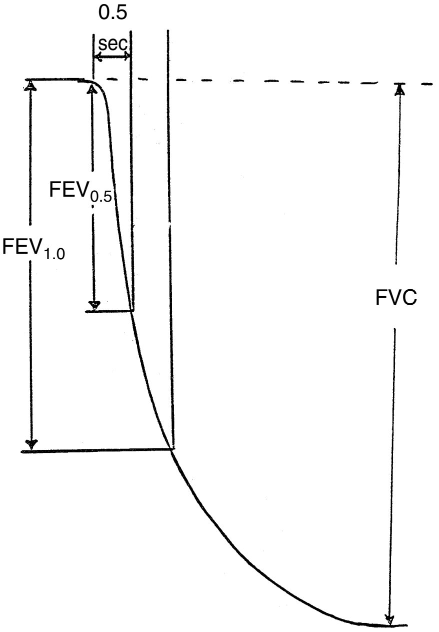


FIGURE 60.2 The measurement of timed forced expiratory volume (FEV$_t$) and forced vital capacity (FVC).

TABLE 60.1 Dynamic Volumes

Males
FVC (l) = 0.133H − 0.022A − 3.60(SEE = 0.58)[a]
FEV1 (l) = 0.094H − 0.028A − 1.59(SEE = 0.52)[a]
MVV (l/min) = 3.39H − 1.26A − 21.4(SEE = 29)[a]
PF (l/min) = (10.03 − 0.038A)H[b]
Females
FVC (l) = 0.111H − 0.015A − 3.16(SD = 0.42)[c]
FEV1 (l) = 0.068H − 0.023A − 0.92(SD = 0.37)[c]
MVV (l/min) = 2.05H − 0.57A − 5.5(SD = 10.7)[c]
PF (l/min) = (7.44 − 0.0183A)H[c]

H = height in inches, A = age in years, l = liters, l/min = liters per minute, SEE = standard error of estimate, SD = standard deviation

[a] Kory, Callahan, Boren, Syner (1961). *Am. J. Med.* 30: 243.

[b] Leiner, Abramowitz, Small, Stenby, Lewis (1963). *Am. Rev. Resp. Dis.* 88: 644.

[c] Lindall, Medina, Grismer (1967). *Am. Rev. Resp. Dis.* 95: 1061.

a hollow cylinder closed at one end, inverted and suspended in an annular space filled with water to provide an air-tight seal. Figure 60.3 illustrates the method of suspending the counterbalanced cylinder (bell), which is free to move up and down to accommodate the volume of air under it. Movement of the bell, which is proportional to volume, is usually recorded by an inking pen applied to a graphic record which is caused to move with a constant speed. Below the cylinder, in the space that accommodates the volume of air, are inlet and outlet breathing tubes. At the end of one or both of these tubes is a check valve designed to maintain a unidirectional flow of air through the spirometer. Outside the spirometer the two breathing tubes are brought to a Y tube which is connected to a mouthpiece. With a pinch clamp placed

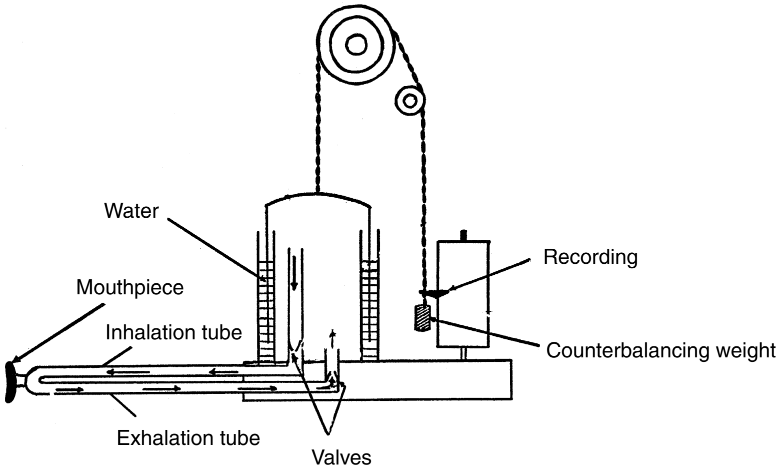


FIGURE 60.3 The simple spirometer.

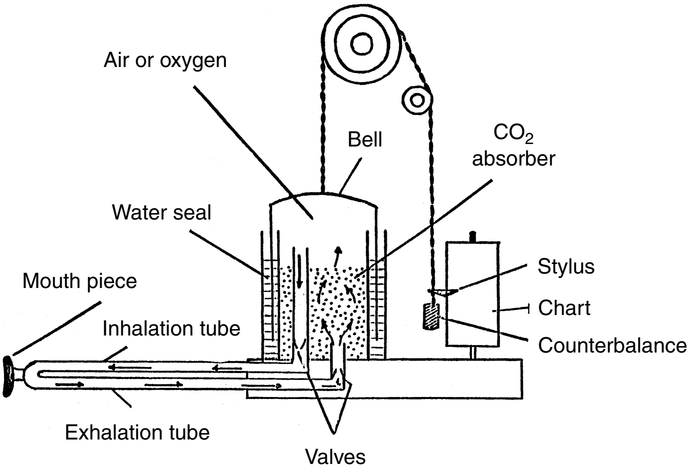


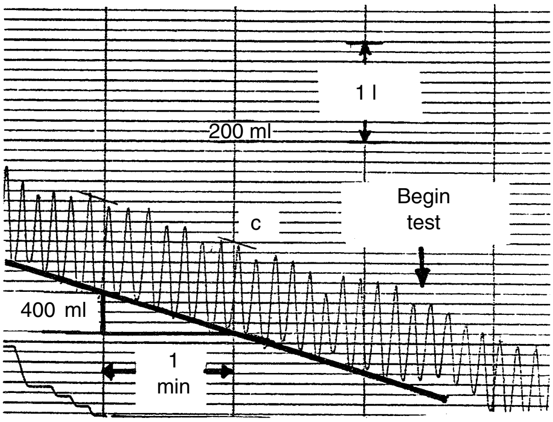
FIGURE 60.4 The spirometer with CO_2 absorber and a record of oxygen uptake (Figure 60.5).

on the nose, inspiration diminishes the volume of air under the bell, which descends, causing the stylus to rise on the graphic record. Expiration produces the reverse effect. Thus, starting with the spirometer half-filled, quiet respiration causes the bell to rise and fall. By knowing the "bell factor," the volume of air moved per centimeter excursion of the bell, the volume change can be quantitated. Although a variety of flowmeters are now used to measure respiratory volumes, the spirometer with a CO_2 absorber is ideally suited to measure oxygen uptake.

60.2.1.6 Oxygen Uptake

A second and very important use for the water-filled spirometer is measurement of oxygen used per unit of time, designated the *O_2 uptake*. This measurement is accomplished by incorporating a soda-lime, carbon-dioxide absorber into the spirometer as shown in Figure 60.4a. Soda-lime is a mixture of calcium hydroxide, sodium hydroxide, and silicates of sodium and calcium. The exhaled carbon dioxide combines with the soda-lime and becomes solid carbonates. A small amount of heat is liberated by this reaction.

Starting with a spirometer filled with oxygen and connected to a subject, respiration causes the bell to move up and down (indicating tidal volume) as shown in Figure 60.5. With continued respiration



$$V_{BTPS} = V_{MEAS} \quad F$$

$$F = \frac{273 + 37}{273 + T} \quad \frac{P_B - P_{H2O}}{P_B - 47}$$

$$V_{MEAS} = 400\,\text{ml @ } 26°\text{C} = T$$
$$(P_B = 750\text{ mmHg})$$

$$F = \frac{273 + 37}{273 + 26} \quad \frac{750 - 25.2^*}{750 - 47}$$
$$= 1.069$$

$$V_{BTPS} = 400 \quad 1.069$$
$$= 427.6\,\text{ml}$$

FIGURE 60.5 Oxygen consumption.

the baseline of the recording rises, reflecting disappearance of oxygen from under the bell. By measuring the slope of the baseline on the spirogram, the volume of oxygen consumed per minute can be determined. Figure 60.5 presents a typical example along with calculation.

60.2.1.7 The Dry Spirometer

The water-sealed spirometer was the most popular device for measuring the volumes of respiratory gases; however, it is not without its inconveniences. The presence of water causes corrosion of the metal parts. Maintenance is required to keep the device in good working order over prolonged periods. To eliminate these problems, manufacturers have developed dry spirometers. The most common type employs a collapsible rubber or plastic bellows, the expansion of which is recorded during breathing. The earlier rubber models had a decidedly undesirable characteristic which caused their abandonment. When the bellows was in its mid-position, the resistance to breathing was a minimum; when fully collapsed, it imposed a slight negative resistance; and when fully extended it imposed a slight positive resistance to breathing. Newer units with compliant plastic bellows minimize this defect.

60.2.2 The Pneumotachograph

The pneumotachograph is a device which is placed directly in the airway to measure the velocity of air flow. The volume per breath is therefore the integral of the velocity–time record during inspiration or expiration. Planimetric integration of the record, or electronic integration of the velocity–time signal, yields the tidal volume. Although tidal volume is perhaps more easily recorded with the spirometer, the dynamics of respiration are better displayed by the pneumotachograph, which offers less resistance to the air stream and exhibits a much shorter response time — so short in most instruments that cardiac impulses are often clearly identifiable in the velocity–time record.

If a specially designed resistor is placed in a tube in which the respiratory gases flow, a pressure drop will appear across it. Below the point of turbulent flow, the pressure drop is linearly related to air-flow velocity. The resistance may consist of a wire screen or a series of capillary tubes; Figure 60.6 illustrates both types. Detection and recording of this pressure differential constitutes a pneumotachogram; Figure 60.7

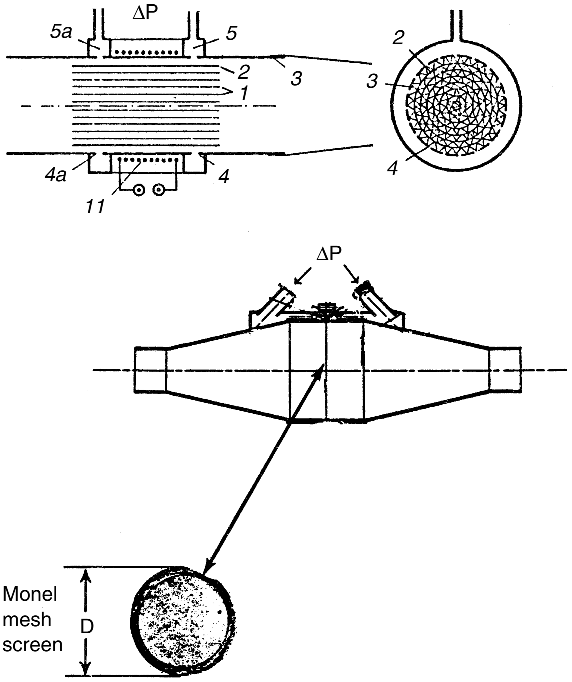


FIGURE 60.6 Pneumotachographs.

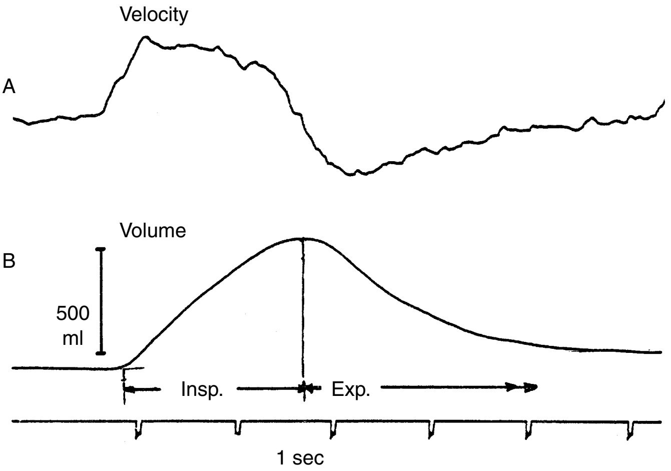


FIGURE 60.7 Velocity (A) and volume changes (B) during normal, quiet breathing; B is the integral of A.

presents a typical air–velocity record, along with the spirogram, which is the integral of the flow signal. The small-amplitude artifacts in the pneumotachogram are cardiac impulses.

For human application, linear flow rates up to 200 l/min should be recordable with fidelity. The resistance to breathing depends upon the flow rate, and it is difficult to establish an upper limit of tolerable resistance. Silverman and Whittenberger [1950] stated that a resistance of 6 mm H_2O is perceptible to human subjects. Many of the high-fidelity pneumotachographs offer 5 to 10 mm H_2O resistance at 100 and 200 l/min. It would appear that such resistances are acceptable in practice.

Response times of 15 to 40 msec seem to be currently in use. Fry and coworkers [1957] analyzed the dynamic characteristics of three types of commercially available, differential-pressure pneumotachographs which employed concentric cylinders, screen mesh, and parallel plates for the air resistors. Using a high-quality, differential-pressure transducer with each, they measured total flow resistance ranging from 5 to 15 cm H_2O. Frequency response curves taken on one model showed fairly uniform response to 40 Hz; the second model showed a slight increase in response at 50 Hz, and the third exhibited a slight drop in response at this frequency.

60.2.3 The Nitrogen-Washout Method for Measuring FRC

The *functional residual capacity* (FRC) and the *residual volume* (RV) are the only lung compartments that cannot be measured with a volume-measuring device. Measuring these requires use of the nitrogen analyzer and application of the dilution method.

Because nitrogen does not participate in respiration, it can be called a *diluent.* Inspired and expired air contain about 80% nitrogen. Between breaths, the FRC of the lungs contains the same concentration of nitrogen as in environmental air, that is, 80%. By causing a subject to inspire from a spirometer filled with 100% oxygen and to exhale into a second collecting spirometer, all the nitrogen in the FRC is replaced by oxygen, that is, the nitrogen is "washed out" into the second spirometer. Measurement of the concentration of nitrogen in the collecting spirometer, along with a knowledge of its volume, permits calculation of the amount of nitrogen originally in the functional residual capacity and hence allows calculation of the FRC, as now will be shown.

Figure 60.8 illustrates the arrangement of equipment for the nitrogen-washout test. Note that two check valves, (I, EX) are on both sides of the subject's breathing tube, and the nitrogen meter is connected to the mouthpiece. Valve V is used to switched the subject from breathing environmental air to the measuring system. The left-hand spirometer contains 100% oxygen, which is inhaled by the subject via valve I. Of course, a nose clip must be applied so that all the respired gases flow through the breathing tube connected

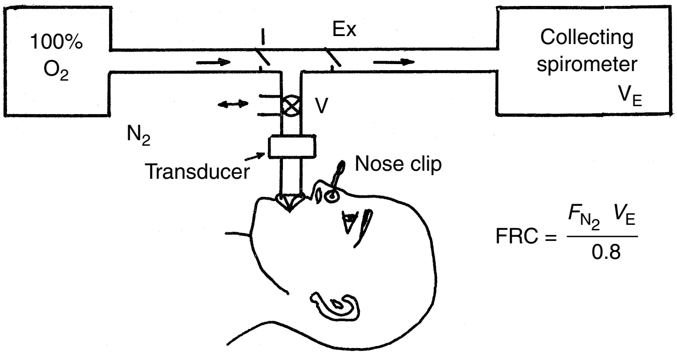


FIGURE 60.8 Arrangement of equipment for the nitrogen-washout technique. Valve V allows the subject to breathe room air until the test is started. The test is started by operating valve V at the end of a normal breath, that is, the subject starts breathing 100% O_2 through the inspiratory valve (I) and exhales the N_2 and O_2 mixture into a collecting spirometer via the expiratory valve EX.

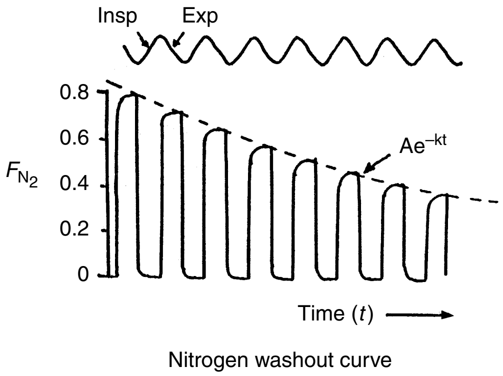


FIGURE 60.9 The nitrogen washout curve.

to the mouthpiece. It is in this tube that the sampling inlet for the nitrogen analyzer is located. Starting at the resting expiratory level, inhalation of pure oxygen causes the nitrogen analyzer to indicate zero. Expiration closes valve I and opens valve EX. The first expired breath contains nitrogen derived from the FRC (diluted by the oxygen which was inspired); the nitrogen analyzer indicates this percentage. The exhaled gases are collected in the right-hand spirometer. The collecting spirometer and all the interconnecting tubing was first flushed with oxygen to eliminate all nitrogen. This simple procedure eliminates the need for applying corrections and facilitates calculation of the FRC. With continued breathing, the nitrogen analyzer indicates less and less nitrogen because it is being washed out of the FRC and is replaced by oxygen. Figure 60.9 presents a typical record of the diminishing concentration of expired nitrogen throughout the test. In most laboratories, the test is continued until the concentration of nitrogen falls to about 1%. The nitrogen analyzer output permits identification of this concentration. In normal subjects, virtually all the nitrogen can be washed out of the FRC in about 5 min.

If the peaks on the nitrogen washout record are joined, a smooth exponential decay curve is obtained in normal subjects. A semilog of N_2 vs. time provides a straight line. In subjects with trapped air, or poorly ventilated alveoli, the nitrogen-washout curve consists of several exponentials as the multiple poorly ventilated regions give up their nitrogen. In such subjects, the time taken to wash out all the nitrogen usually exceeds 10 min. Thus, the nitrogen concentration–time curve provides useful diagnostic information on ventilation of the alveoli.

If it is assumed that all the collected (washed-out) nitrogen was uniformly distributed within the lungs, it is easy to calculate the FRC. If the environmental air contains 80% nitrogen, then the volume of nitrogen in the functional residual capacity is 0.8 (FRC). Because the volume of expired gas in the collecting spirometer is known, it is merely necessary to determine the concentration of nitrogen in this volume. To do so requires admitting some of this gas to the inlet valve of the nitrogen analyzer. Note that this concentration of nitrogen (F_{N_2}) exists in a volume which includes the volume of air expired (V_E) plus the original volume of oxygen in the collecting spirometer (V_0) at the start of the test and the volume of the tubing (V_t) leading from the expiratory collecting valve. It is therefore advisable to start with an empty collecting spirometer ($V_0 = 0$). Usually the tubing volume (V_t) is negligible with respect to the volume of expired gas collected in a typical washout test. In this situation the volume of nitrogen collected is $V_E\ F_{N_2}$, where F_{N_2} is the fraction of nitrogen within the collected gas. Thus, 0.80 (FRC) $= F_{N_2}$ (V_E). Therefore

$$\text{FRC} = \frac{F_{N_2} V_E}{0.80} \tag{60.1}$$

It is important to note that the value for FRC so obtained is at ambient temperature and pressure and is saturated with water vapor (ATPS). In respiratory studies, this value is converted to body temperature and saturated with water vapor (BTPS).

In the example shown in Figure 60.9, the washout to 1% took about 44 breaths. With a breathing rate of 12/min, the washout time was 220 sec. The volume collected (V_E) was 22 l and the concentration of nitrogen in this volume was 0.085 (F_{N_2}); therefore

$$\text{FRC} = \frac{0.085 \times 22{,}000}{0.80} = 2{,}337 \text{ ml} \qquad (60.2)$$

60.3 Physiologic Dead Space

The volume of ventilated lung that does not participate in gas exchange is the physiologic dead space (V_d). It is obvious that the physiologic dead space includes anatomic dead space, as well as the volume of any alveoli that are not perfused. In the lung, there are theoretically four types of alveoli, as shown in Figure 60.10. The normal alveolus (A) is both ventilated and perfused with blood. There are alveoli that are ventilated but not perfused (B); such alveoli contribute significantly to the physiologic dead space. There are alveoli that are not ventilated but perfused (C); such alveoli do not provide the exchange of respiratory gases. Finally, there are alveoli that are both poorly ventilated and poorly perfused (D); such alveoli contain high CO_2 and N_2 and low O_2. These alveoli are the last to expel their CO_2 and N_2 in washout tests.

Measurement of physiologic dead space is based on the assumption that there is almost complete equilibrium between alveolar pCO_2 and pulmonary capillary blood. Therefore, the arterial pCO_2 represents mean alveolar pCO_2 over many breaths when an arterial blood sample is drawn for analysis of pCO_2. The Bohr equation for physiologic dead space is

$$V_d = \left[\frac{paCO_2 - pECO_2}{paCO_2}\right] V_E \qquad (60.3)$$

In this expression, $paCO_2$ is the partial pressure in the arterial blood sample which is withdrawn slowly during the test; $pECO_2$ is the partial pressure of CO_2 in the volume of expired air; V_E is the volume of expired air per breath (tidal volume).

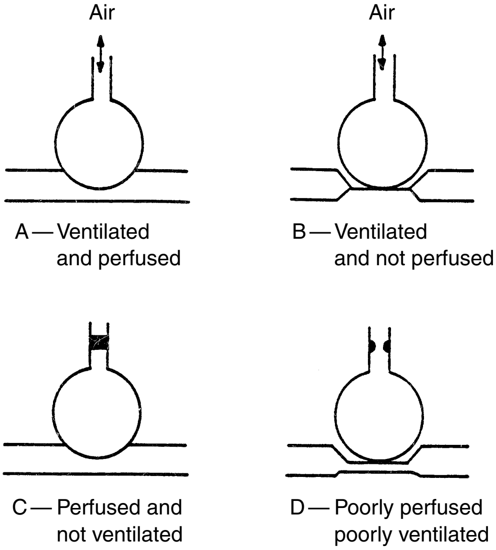

Air
Air
A — Ventilated and perfused
B — Ventilated and not perfused
C — Perfused and not ventilated
D — Poorly perfused poorly ventilated

FIGURE 60.10 The four types of alveoli.

In a typical test, the subject would breathe in room air and exhale into a collapsed (Douglas) bag. The test is continued for 3 min or more, and the number of breaths is counted in that period. An arterial blood sample is withdrawn during the collection period. The $\mathrm{pCO_2}$ in the expired gas is measured, and then the volume of expired gas is measured by causing it to flow into a spirometer or flowmeter by collapsing the collecting bag.

In a typical 3-min test, the collected volume is 33 l, and the $\mathrm{pCO_2}$ in the expired gas is 14.5 mmHg. During the test, the $\mathrm{pCO_2}$ in the arterial blood sample was 40 mmHg. The number of breaths was 60; therefore, the average tidal volume was $33{,}000/60 = 550$ ml. The physiologic dead space (V_d) is:

$$V_\mathrm{d} = \left[\frac{40 - 14.5}{40}\right] 550 = 350 \text{ ml} \tag{60.4}$$

It is obvious that an elevated physiological dead space indicates lung tissue that is not perfused with blood.

References

Fry D.I., Hyatt R.E., and McCall C.B. 1957. Evaluation of three types of respiratory flowmeters. *Appl. Physiol.* 10: 210.

Silverman L. and Whittenberger J. 1950. Clinical pneumotachograph. *Meth. Med. Res.* 2: 104.

61
Mechanical Ventilation

Khosrow Behbehani
The University of Texas at Arlington
The University of Texas Southwestern Medical Center

61.1 Introduction

This chapter presents an overview of the structure and function of mechanical ventilators. Mechanical ventilators, which are often also called respirators, are used to artificially ventilate the lungs of patients who are unable to naturally breathe from the atmosphere. In almost 100 years of development, many mechanical ventilators with different designs have been developed [Mushin et al., 1980; Philbeam, 1998]. The very early devices used bellows that were manually operated to inflate the lungs. Today's respirators employ an array of sophisticated components such as microprocessors, fast response servo valves, and precision transducers to perform the task of ventilating the lungs. The changes in the design of ventilators have come about as the result of improvements in engineering the ventilator components and the advent of new therapy modes by clinicians. A large variety of ventilators are now available for short-term treatment of acute respiratory problems as well as long-term therapy for chronic respiratory conditions.

It is reasonable to broadly classify today's ventilators into two groups. The first and indeed the largest group encompasses the intensive care respirators used primarily in hospitals to support patients following certain surgical procedures or assist patients with acute respiratory disorders. The second group includes less complicated machines that are primarily used at home to treat patients with chronic respiratory disorders.

The level of engineering design and sophistication for the intensive care ventilators is higher than the ventilators used for chronic treatment. However, many of the engineering concepts employed in designing

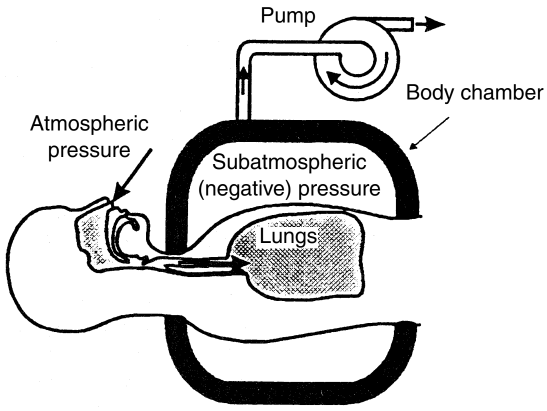


FIGURE 61.1 A simplified illustration of a negative-pressure ventilator.

intensive care ventilators can also be applied in the simpler chronic care units. Therefore, this presentation focuses on the design of intensive care ventilators; the terms respirator, mechanical ventilator, or ventilator that will be used from this point on refer to the intensive care unit respirators.

At the beginning, the designers of mechanical ventilators realized that the main task of a respirator was to ventilate the lungs in a manner as close to natural respiration as possible. Since natural inspiration is a result of negative pressure in the pleural cavity generated by distention of the diaphragm, designers initially developed ventilators that created the same effect. These ventilators are called *negative-pressure ventilators.* However, more modern ventilators use pressures greater than atmospheric pressures to ventilate the lungs; they are known as *positive-pressure ventilators.*

61.2 Negative-Pressure Ventilators

The principle of operation of a negative-pressure respirator is shown in Figure 61.1. In this design, the flow of air to the lungs is created by generating a negative pressure around the patient's thoracic cage. The negative pressure moves the thoracic walls outward expanding the intra-thoracic volume and dropping the pressure inside the lungs. The pressure gradient between the atmosphere and the lungs causes the flow of atmospheric air into the lungs. The inspiratory and expiratory phases of the respiration are controlled by cycling the pressure inside the body chamber between a sub-atmospheric level (inspiration) and the atmospheric level (exhalation). Flow of the breath out of the lungs during exhalation is caused by the recoil of thoracic muscles.

Although it may appear that the negative-pressure respirator incorporates the same principles as natural respiration, the engineering implementation of this concept has not been very successful. A major difficulty has been in the design of a chamber for creating negative pressure around the thoracic walls. One approach has been to make the chamber large enough to house the entire body with the exception of the head and neck. Using foam rubber around the patient's neck, one can seal the chamber and generate a negative pressure inside the chamber. This design configuration, commonly known as the iron lung, was tried back in the 1920s and proved to be deficient in several aspects. The main drawback was that the negative pressure generated inside the chamber was applied to the chest as well as the abdominal wall, thus creating a venous blood pool in the abdomen and reducing cardiac output.

More recent designs have tried to restrict the application of the negative pressure to the chest walls by designing a chamber that goes only around the chest. However, this has not been successful because obtaining a seal around the chest wall (Figure 61.1) is difficult.

Negative-pressure ventilators also made the patient less accessible for patient care and monitoring. Further, synchronization of the machine cycle with the patient's effort has been difficult and they are also

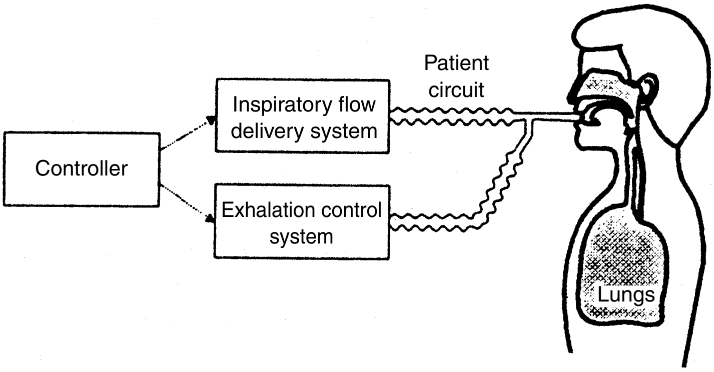


FIGURE 61.2 A simplified diagram of the functional blocks of a positive-pressure ventilator.

typically noisy and bulky [McPherson and Spearman, 1990]. These deficiencies of the negative-pressure ventilators have led to the development of the positive-pressure ventilators.

61.3 Positive-Pressure Ventilators

Positive-pressure ventilators generate the inspiratory flow by applying a positive pressure (greater than the atmospheric pressure) to the airways. Figure 61.2 shows a simplified block diagram of a positive-pressure ventilator. During inspiration, the inspiratory flow delivery system creates a positive pressure in the tubes connected to the patient airway, called **patient circuit**, and the exhalation control system closes a valve at the outlet of the tubing to the atmosphere. When the ventilator switches to exhalation, the inspiratory flow delivery system stops the positive pressure and the exhalation system opens the valve to allow the patient's exhaled breath to flow to the atmosphere. The use of a positive pressure gradient in creating the flow allows treatment of patients with high lung resistance and low compliance. As a result, positive-pressure ventilators have been very successful in treating a variety of breathing disorders and have become more popular than negative-pressure ventilators.

Positive-pressure ventilators have been employed to treat patients ranging from neonates to adults. Due to anatomical differences between various patient populations, the ventilators and their modes of treating infants are different than those for adults. Nonetheless, their fundamental design principles are similar and adult ventilators comprise a larger percentage of ventilators manufactured and used in clinics. Therefore, the emphasis here is on the description of adult positive-pressure ventilators. Also, the concepts presented will be illustrated using a microprocessor-based design example, as almost all modern ventilators use microprocessor instrumentation.

61.4 Ventilation Modes

Since the advent of respirators, clinicians have devised a variety of strategies to ventilate the lungs based on patient conditions. For instance, some patients need the respirator to completely take over the task of ventilating their lungs. In this case, the ventilator operates in **mandatory mode** and delivers mandatory breaths. On the other hand, some patients are able to initiate a breath and breathe on their own, but may need oxygen-enriched air flow or slightly elevated airway pressure. When a ventilator assists a patient who is capable of demanding a breath, the ventilator delivers spontaneous breaths and operates in **spontaneous mode**. In many cases, it is first necessary to treat the patient with mandatory ventilation and as the patient's condition improves spontaneous ventilation is introduced; it is used primarily to wean the patient from mandatory breathing.

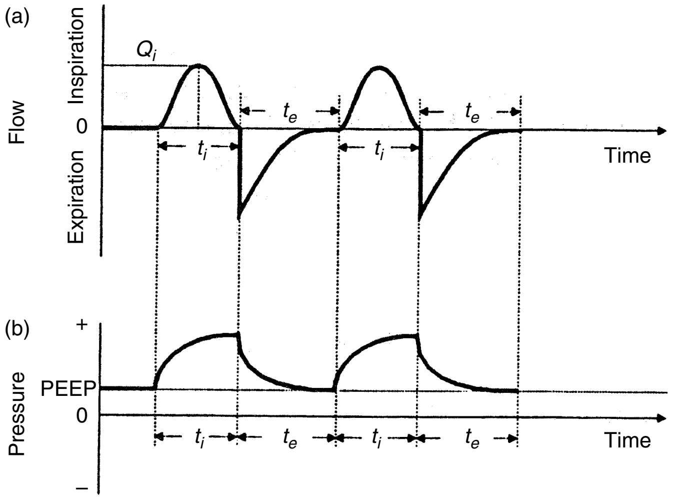


FIGURE 61.3 (a) Inspiratory flow for a controlled mandatory volume controlled ventilation breath, (b) airway pressure resulting from the breath delivery with a non-zero PEEP.

61.4.1 Mandatory Ventilation

Designers of adult ventilators have employed two rather distinct approaches for delivering mandatory breaths: **volume controlled ventilation** and **pressure controlled ventilation**. Volume controlled ventilation, which presently is more popular, refers to delivering a specified tidal volume to the patient during the inspiratory phase. Pressure controlled ventilation, however, refers to raising the airway pressure to a level, set by the therapist, during the inspiratory phase of each breath. Regardless of the type, a ventilator operating in mandatory mode must control all aspects of breathing such as tidal volume, respiration rate, inspiratory flow pattern, and oxygen concentration of the breath. This is often labeled as **controlled mandatory ventilation (CMV).**

Figure 61.3 shows the flow and pressure waveforms for a volume controlled ventilation (CMV). In this illustration, the inspiratory flow waveform is chosen to be a half sinewave. In Figure 61.3a, t_{i} is the inspiration duration, t_{e} is the exhalation period, and Q_{i} is the amplitude of inspiratory flow. The ventilator delivers a tidal volume equal to the area under the flow waveform in Figure 61.3a at regular intervals ($t_{\mathrm{i}} + t_{\mathrm{e}}$) set by the therapist. The resulting pressure waveform is shown in Figure 61.3b. It is noted that during volume controlled ventilation, the ventilator delivers the same volume irrespective of the patient's respiratory mechanics. However, the resulting pressure waveform such as the one shown in Figure 61.3b, will be different among patients. Of course, for safety purposes, the ventilator limits the maximum applied airway pressure according to the therapist's setting.

As can be seen in Figure 61.3b, the airway pressure at the end of exhalation may not end at atmospheric pressure (zero gauge). The **positive end expiratory pressure (PEEP)** is sometimes used to keep the alveoli from collapsing during expiration [Norwood, 1990]. In other cases, the expiration pressure is allowed to return to the atmospheric level.

Figure 61.4a shows a plot of the pressure and flow during a mandatory pressure controlled ventilation. In this case, the respirator raises and maintains the airway pressure at the desired level independent of patient airway compliance and resistance. The level of pressure during inspiration, P_{i}, is set by the therapist. While the ventilator maintains the same pressure trajectory for patients with different respiratory resistance and compliance, the resulting flow trajectory, shown in Figure 61.4b, will depend on the respiratory mechanics of each patient.

In the following, the presentation will focus on volume ventilators, as they are more common. Further, in a microprocessor-based ventilator, the mechanism for delivering mandatory volume and pressure

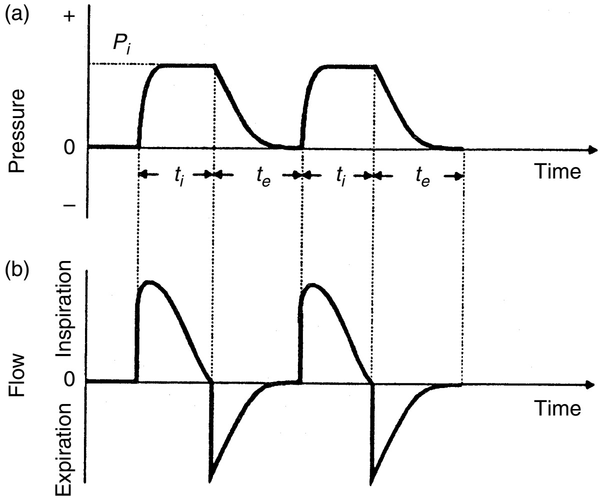


FIGURE 61.4 (a) Inspiratory pressure pattern for a controlled mandatory pressure controlled ventilation breath, (b) airway flow pattern resulting from the breath delivery. Note that PEEP is zero.

controlled ventilation have many similar main components. The primary difference lies in the control algorithms governing the delivery of breaths to the patient.

61.4.2 Spontaneous Ventilation

An important phase in providing respiratory therapy to a recovering pulmonary patient is weaning the patient from the respirator. As the patient recovers and gains the ability to breathe independently, the ventilator must allow the patient to initiate a breath and control the breath rate, flow rate, and the tidal volume. Ideally, when a respirator is functioning in the spontaneous mode, it should let the patient take breaths with the same ease as breathing from the atmosphere. This, however, is difficult to achieve because the respirator does not have an infinite gas supply or an instantaneous response. In practice, the patient generally has to exert more effort to breathe spontaneously on a respirator than from the atmosphere. However, patient effort is reduced as the ventilator response speed increases [McPherson, 1990]. Spontaneous ventilation is often used in conjunction with mandatory ventilation since the patient may still need breaths that are delivered entirely by the ventilator. Alternatively, when a patient can breathe completely on his own but needs oxygen-enriched breath or elevated airway pressure, spontaneous ventilation alone may be used.

As in the case of mandatory ventilation, several modes of spontaneous ventilation have been devised by therapists. Two of the most important and popular spontaneous breath delivery modes are described below.

61.4.2.1 Continuous Positive Airway Pressure (CPAP) in Spontaneous Mode

In this mode, the ventilator maintains a positive pressure at the airway as the patient attempts to inspire. Figure 61.5 illustrates a typical airway pressure waveform during CPAP breath delivery. The therapist sets the sensitivity level lower than PEEP. When the patient attempts to breathe, the pressure drops below the sensitivity level and the ventilator responds by supplying breathable gases to raise the pressure back to the PEEP level. Typically, the PEEP and sensitivity levels are selected such that the patient will be impelled to exert effort to breathe independently. As in the case of the mandatory mode, when the patient exhales the ventilator shuts off the flow of gas and opens the exhalation valve to allow the exhaled gases to flow into the atmosphere.

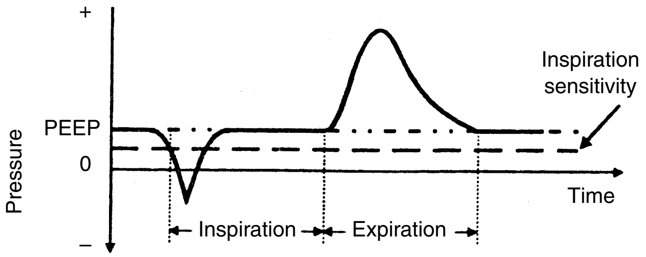


FIGURE 61.5 Airway pressure during a CPAP spontaneous breath delivery.

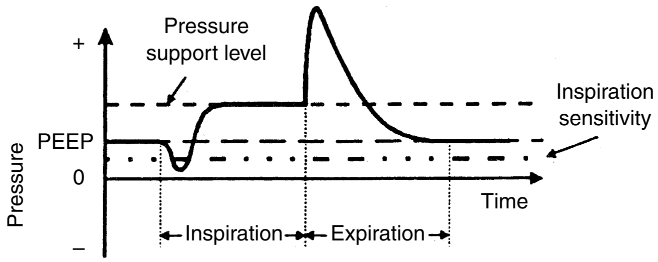


FIGURE 61.6 Airway pressure during a pressure support spontaneous breath delivery.

61.4.2.2 Pressure Support in Spontaneous Mode

This mode is similar to the CPAP mode with the exception that during the inspiration the ventilator attempts to maintain the patient airway pressure at a level above PEEP. In fact, CPAP may be considered a special case of **pressure support** ventilation in which the support level is fixed at the atmospheric level.

Figure 61.6 shows a typical airway pressure waveform during the delivery of a pressure support breath. In this mode, when the patient's airway pressure drops below the therapist-set sensitivity line, the ventilator inspiratory breath delivery system raises the airway pressure to the **pressure support level** (>PEEP), selected by the therapist. The ventilator stops the flow of breathable gases when the patient starts to exhale and controls the exhalation valve to achieve the set PEEP level.

61.5 Breath Delivery Control

Figure 61.7 shows a simplified block diagram for delivering mandatory or spontaneous ventilation. Compressed air and oxygen are normally stored in high pressure tanks (≅1400 kPa) that are attached to the inlets of the ventilator. In some ventilators, an air compressor is used in place of a compressed air tank. Manufacturers of mechanical respirators have designed a variety of blending and metering devices [McPherson, 1990]. The primary mission of the device is to enrich the inspiratory air flow with the proper level of oxygen and to deliver a tidal volume according to the therapist's specifications. With the introduction of microprocessors for control of metering devices, electromechanical valves have gained popularity [Puritan-Bennett, 1987]. In Figure 61.7, the air and oxygen valves are placed in closed feedback loops with the air and oxygen flow sensors. The microprocessor controls each the valves to deliver the desired inspiratory air and oxygen flows for mandatory and spontaneous ventilation. During inhalation, the exhalation valve is closed to direct all the delivered flows to the lungs. When exhalation starts, the microprocessor actuates the exhalation valve to achieve the desired PEEP level. The airway pressure sensor, shown on the right side of Figure 61.7, generates the feedback signal necessary for maintaining the desired PEEP (in both mandatory and spontaneous modes) and airway pressure support level during spontaneous breath delivery.

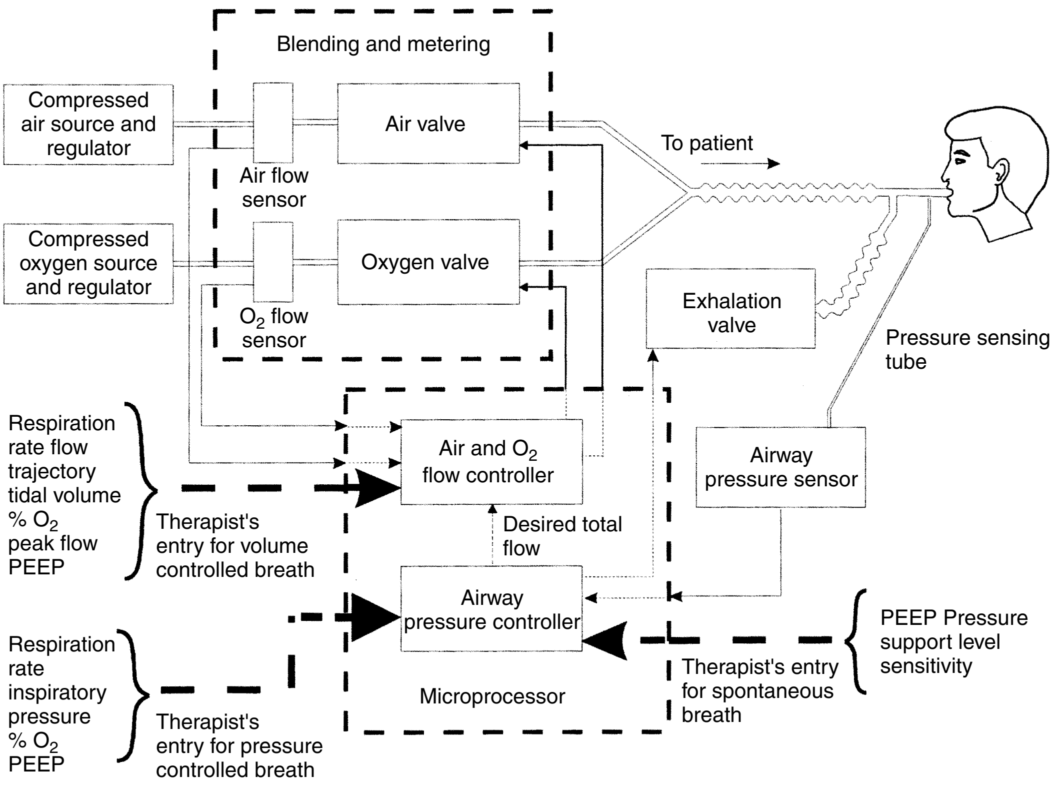

Blending and metering
Compressed air source and regulator
Air valve
Air flow sensor
Compressed oxygen source and regulator
Oxygen valve
O2 flow sensor
To patient
Exhalation valve
Pressure sensing tube
Respiration rate flow trajectory tidal volume % O2 peak flow PEEP
Therapist's entry for volume controlled breath
Air and O2 flow controller
Desired total flow
Airway pressure sensor
Airway pressure controller
PEEP Pressure support level sensitivity
Respiration rate inspiratory pressure % O2 PEEP
Therapist's entry for pressure controlled breath
Microprocessor
Therapist's entry for spontaneous breath

FIGURE 61.7 A simplified block diagram of a control structure for mandatory and spontaneous breath delivery.

61.5.1 Mandatory Volume Controlled Inspiratory Flow Delivery

In a microprocessor-controlled ventilator (Figure 61.7), the electronically actuated valves open from a closed position to allow the flow of blended gases to the patient. The control of flow through each valve depends on the therapist's specification for the mandatory breath. That is, the clinician must specify the following parameters for delivery of CMV breaths (1) respiration rate; (2) flow waveform; (3) tidal volume; (4) oxygen concentration (of the delivered breath); (5) peak flow; and (6) PEEP, as shown in the lower left side of Figure 61.7. It is noted that the PEEP selected by the therapist in the mandatory mode is only used for control of exhalation flow; that will be described in the following section. The microprocessor utilizes the first five of the above parameters to compute the total desired inspiratory flow trajectory. To illustrate this point, consider the delivery of a tidal volume using a half sinewave as shown in Figure 61.3. If the therapist selects a tidal volume of V_t (L), a respiration rate of n breaths per minute (bpm), the amplitude of the respirator flow, Q_i(L/s), then the total desired inspiratory flow, $Q_d(t)$, for a single breath, can be computed from the following equation:

$$Q_d(t) = \begin{cases} Q_i \sin \frac{\pi t}{t_i}, & 0 \le t < t_i \\ 0, & t_i < t \le t_e \end{cases} \tag{61.1}$$

where t_i signifies the duration of inspiration and is computed from the following relationship:

$$t_i = \frac{V_t}{2Q_i} \tag{61.2}$$

The duration of expiration in seconds is obtained from

$$t_e = \frac{60}{n} - t_i \tag{61.3}$$

The ratio of inspiratory to expiratory periods of a mandatory breath is often used for adjusting the respiration rate. This ratio is represented by **I : E (ratio)** and is computed as follows. First, the inspiratory and expiratory periods are normalized with respect to t_i. Hence, the normalized inspiratory period becomes unity and the normalized expiratory period is given by $R = t_e/t_i$. Then, the **I : E ratio** is simply expressed as 1 : R.

To obtain the desired oxygen concentration in the delivered breath, the microprocessor computes the discrete form of $Q_d(t)$ as $Q_d(k)$ where k signifies the kth sample interval. Then, the total desired flow, $Q_d(k)$, is partitioned using the following relationships:

$$Q_{da}(k) = \frac{(1 - m)Q_d(k)}{(1 - c)} \tag{61.4}$$

and

$$Q_{dx}(k) = \frac{(m - c)Q_d(k)}{(1 - c)} \tag{61.5}$$

where k signifies the sample interval, $Q_{da}(k)$ is the desired air flow (the subscript da stands for desired air), $Q_{dx}(k)$ is the desired oxygen flow (the subscript dx stands for desired oxygen), m is the desired oxygen concentration, and c is the oxygen concentration of the ventilator air supply.

A number of control design strategies may be appropriate for the control of the air and oxygen flow delivery valves. A simple controller is the proportional plus integral controller that can be readily implemented in a microprocessor. For example, the controller for the air valve has the following form:

$$I(k) = K_p E(k) + K_i A(k) \tag{61.6}$$

where $E(k)$ and $A(k)$ are given by

$$E(k) = Q_{da}(k) - Q_{sa}(k) \tag{61.7}$$

$$A(k) = A(k - 1) + E(k) \tag{61.8}$$

where $I(k)$ is the input (voltage or current) to the air valve at the kth sampling interval, $E(k)$ is the error in the delivered flow, $Q_{da}(k)$ is the desired air flow, $Q_{sa}(k)$ is the sensed or actual air flow (the subscript sa stands for sensed air flow), $A(k)$ is the integral (rectangular integration) part of the controller, and K_p and K_i are the controller proportionality constants. It is noted that the above equations are applicable to the control of either the air or oxygen valve. For control of the oxygen flow valve, $Q_{dx}(k)$ replaces $Q_{da}(k)$ and $Q_{sx}(k)$ replaces $Q_{sa}(k)$ where $Q_{sx}(k)$ represents the sensed oxygen flow (the subscript sx stands for sensed oxygen flow).

The control structure shown in Figure 61.7 provides the flexibility of quickly adjusting the percentage of oxygen in the enriched breath gases. That is, the controller can regulate both the total flow and the percent oxygen delivered to the patient. Since the internal volume of the flow control valve is usually small (<50 ml), the desired change in the oxygen concentration of the delivered flow can be achieved within one inspiratory period. In actual clinical applications, rapid change of percent oxygen from one breath to another is often desirable, as it reduces the waiting time for the delivery of the desired oxygen concentration. A design similar to the one shown in Figure 61.7 has been successfully implemented in a microprocessor-based ventilator [Behbehani, 1984] and is deployed in hospitals around the world.

61.5.2 Pressure Controlled Inspiratory Flow Delivery

The therapist entry for pressure-controlled ventilation is shown in Figure 61.7 (lower left-hand side). In contrast to the volume-controlled ventilation where $Q_\mathrm{d}(t)$ was computed directly from the operator's entry, the total desired flow is generated by a closed loop controller labeled as Airway Pressure Controller in Figure 61.7. This controller uses the therapist-selected inspiratory pressure, respiration rate, and the I:E ratio to compute the desired inspiratory pressure trajectory. The trajectory serves as the controller reference input. The controller then computes the flow necessary to make the actual airway pressure track the reference input. Assuming a proportional-plus-integral controller, the governing equations are

$$Q_\mathrm{d}(k) = C_p E_p(k) + C_\mathrm{i} A_p(k) \tag{61.9}$$

where Q_d is the computed desired flow, C_p and C_i are the proportionality constants, k represents the sample interval, and $E_p(k)$ and $A_p(k)$ are computed using the following equations:

$$E_p(k) = P_\mathrm{d}(k) - P_\mathrm{s}(k) \tag{61.10}$$

$$A_p(k) = A_p(k-1) + E_p(k) \tag{61.11}$$

where $E_p(k)$ is the difference between the desired pressure trajectory, $P_\mathrm{d}(k)$, and the sensed airway pressure, $P_\mathrm{s}(k)$, the parameter $A_p(k)$ represents the integral portion of the controller. Using Q_d from Equation 61.9, the control of air and O_2 valves is accomplished in the same manner as in the case of volume-controlled ventilation described earlier (Equation 61.4 through Equation 61.8).

61.5.3 Expiratory Pressure Control in Mandatory Mode

It is often desirable to keep the patient's lungs inflated at the end of expiration at a pressure greater than atmospheric level [Norwood, 1990]. That is, rather than allowing the lungs to deflate during the exhalation, the controller closes the exhalation valve when the airway pressure reaches the PEEP level. When expiration starts, the ventilator terminates flow to the lungs; hence, the regulation of the airway pressure is achieved by controlling the flow of patient exhaled gases through the exhalation valve.

In a microprocessor-based ventilator, an electronically actuated valve can be employed that has adequate dynamic response (≅20 msec rise time) to regulate PEEP. For this purpose, the pressure in the patient breath delivery circuit is measured using a pressure transducer (Figure 61.7). The microprocessor will initially open the exhalation valve completely to minimize resistance to expiratory flow. At the same time, it will sample the pressure transducer's output and start to close the exhalation valve as the pressure begins to approach the desired PEEP level. Since the patient's exhaled flow is the only source of pressure, if the airway pressure drops below PEEP, it cannot be brought back up until the next inspiratory period. Hence, an overrun (i.e., a drop to below PEEP) in the closed-loop control of PEEP cannot be tolerated.

61.5.4 Spontaneous Breath Delivery Control

The small diameter (≅5 mm) pressure sensing tube, shown on the right side of Figure 61.7, pneumatically transmits the pneumatic pressure signal from the patient airway to a pressure transducer placed in the ventilator. The output of the pressure transducer is amplified, filtered, and then sampled by the microprocessor. The controller receives the therapist's inputs regarding the spontaneous breath characteristics such as the PEEP, sensitivity, and oxygen concentration, as shown on the lower right-hand side of Figure 61.7. The desired airway pressure is computed from the therapist entries of PEEP, pressure support level, and sensitivity. The multiple-loop control structure shown in Figure 61.7 is used to deliver a CPAP or a pressure support breath. The sensed proximal airway pressure is compared with the desired airway pressure. The airway pressure controller computes the total inspiratory flow level required to raise the airway pressure to the desired level. This flow level serves as the reference input or total desired flow for the flow control loop. Hence, in general, the desired total flow trajectory for the spontaneous breath delivery may be different for each inspiratory cycle. If the operator has specified oxygen concentration greater than 21.6% (the atmospheric air oxygen concentration of the ventilator air supply), the controller will partition the total required flow into the air and oxygen flow rates using Equation 61.4 and Equation 61.5. The flow controller then uses the feedback signals from air and oxygen flow sensors and actuates the air and oxygen valves to deliver the desired flows.

For a microprocessor-based ventilator, the control algorithm for regulating the airway pressure can also be a proportional plus integral controller [Behbehani, 1984; Behbehani and Watanabe, 1986]). In this case, the governing equations are identical to Equation 61.9 through Equation 61.11.

If a non-zero PEEP level is specified, the same control strategy as the one described for mandatory breath delivery can be used to achieve the desired PEEP.

61.6 Summary

Today's mechanical ventilators can be broadly classified into negative-pressure and positive-pressure ventilators. Negative-pressure ventilators do not offer the flexibility and convenience that positive-pressure ventilators provide; hence, they have not been very popular in clinical use. Positive-pressure ventilators have been quite successful in treating patients with pulmonary disorders. These ventilators operate in either mandatory or spontaneous mode. When delivering mandatory breaths, the ventilator controls all parameters of the breath such as tidal volume, inspiratory flow waveform, respiration rate, and oxygen content of the breath. Mandatory breaths are normally delivered to the patients that are incapable of breathing on their own. In contrast, spontaneous breath delivery refers to the case where the ventilator responds to the patient's effort to breathe independently. Therefore, the patient can control the volume and the rate of the respiration. The therapist selects the oxygen content and the pressure at which the breath is delivered. Spontaneous breath delivery is typically used for patients who are on their way to full recovery, but are not completely ready to breathe from the atmosphere without mechanical assistance.

Defining Terms

Continuous positive airway pressure (CPAP): A spontaneous ventilation mode in which the ventilator maintains a constant positive pressure, near or below PEEP level, in the patient's airway while the patient breathes at will.

I : E ratio: The ratio of normalized inspiratory interval to normalized expiratory interval of a mandatory breath. Both intervals are normalized with respect to the inspiratory period. Hence, the normalized inspiratory period is always unity.

Mandatory mode: A mode of mechanically ventilating the lungs where the ventilator controls all breath delivery parameters such as tidal volume, respiration rate, flow waveform, etc.

Patient circuit: A set of tubes connecting the patient airway to the outlet of a respirator.

Positive end expiratory pressure (PEEP): A therapist-selected pressure level for the patient airway at the end of expiration in either mandatory or spontaneous breathing.

Pressure controlled ventilation: A mandatory mode of ventilation where during the inspiration phase of each breath, a constant pressure is applied to the patient's airway independent of the patient's airway resistance and/or compliance respiratory mechanics.

Pressure support: A spontaneous breath delivery mode during which the ventilator applies a positive pressure greater than PEEP to the patient's airway during inspiration.

Pressure support level: Refers to the pressure level, above PEEP, that the ventilator maintains during the spontaneous inspiration.

Spontaneous mode: A ventilation mode in which the patient initiates and breathes from the ventilator-supplied gas at will.

Volume controlled ventilation: A mandatory mode of ventilation where the volume of each breath is set by the therapist and the ventilator delivers that volume to the patient independent of the patient's airway resistance and/or compliance respiratory mechanics.

References

Behbehani, K. (1984). PLM-Implementation of a Multiple Closed-Loop Control Strategy for a Microprocessor-Controlled Respirator. *Proc. ACC Conf.*, pp. 574–576.

Behbehani, K. and Watanabe, N.T. (1986). A New Application of Digital Computer Simulation in the Design of a Microprocessor-Based Respirator. *Summer Simulation Conf.*, pp. 415–420.

McPherson, S.P. and Spearman, C.B. (1990). *Respiratory Therapy Equipment,* 4th ed., C.V. Mosby Co., St. Louis, MO.

Norwood, S. (1990). Physiological Principles of Conventional Mechanical Ventilation. In *Clinical Application of Ventilatory Support,* Kirby, R.R., Banner, M.J., and Downs, J.B., Eds., Churchill Livingstone, New York, pp. 145–172.

Pilbeam, S.P. (1998). *Mechanical Ventilation: Physiological and Clinical Applications,* 3rd ed., Mosby, St. Louis, MO.

Puritan-Bennett 7200 Ventilator System Series, "Ventilator, Options and Accessories," Part No. 22300A, Carlsbad, CA, September 1990.

62
Essentials of Anesthesia Delivery

A. William Paulsen
Emory University

The intent of this chapter is to provide an introduction to the practice of anesthesiology and to the technology currently employed. Limitations on the length of this work and the enormous size of the topic require that this chapter rely on other elements within this Handbook and other texts cited as general references for many of the details that inquisitive minds desire and deserve.

The practice of anesthesia includes more than just providing relief from pain. In fact, pain relief can be considered a secondary facet of the specialty. In actuality, the modern concept of the safe and efficacious delivery of anesthesia requires consideration of three fundamental tenets, which are ordered here by relative importance:

1. Maintenance of vital organ function
2. Relief of pain
3. Maintenance of the "internal milieu"

The first, maintenance of vital organ function, is concerned with preventing damage to cells and organ systems that could result from inadequate supply of oxygen and other nutrients. The delivery of blood and cellular substrates is often referred to as perfusion of the cells or tissues. During the delivery of an anesthetic, the patient's "vital signs" are monitored in an attempt to prevent inadequate tissue perfusion. However, the surgery itself, the patient's existing pathophysiology, drugs given for the relief of pain, or even the management of blood pressure may compromise tissue perfusion. Why is adequate perfusion of tissues a higher priority than providing relief of pain for which anesthesia is named? A rather obvious extreme example is that without cerebral perfusion, or perfusion of the spinal cord, delivery of an anesthetic is not necessary. Damage to other organ systems may result in a range of complications from delaying the patient's recovery to diminishing their quality of life to premature death.

In other words, the primary purpose of anesthesia care is to maintain adequate delivery of required substrates to each organ and cell, which will hopefully preserve cellular function. The second principle of anesthesia is to relieve the pain caused by surgery. Chronic pain and suffering caused by many disease states is now managed by a relatively new sub-specialty within anesthesia, called Pain Management.

The third principle of anesthesia is the maintenance of the internal environment of the body, for example, the regulation of electrolytes (sodium, potassium, chloride, magnesium, calcium, etc.), acid-base balance, and a host of supporting functions on which cellular function and organ system communications rest.

The person delivering anesthesia may be an Anesthesiologist (physician specializing in anesthesiology), an Anesthesiology Physician Assistant (a person trained in a medical school at the masters level to administer anesthesia as a member of the care team lead by an Anesthesiologist), or a nurse anesthetist (a nurse with Intensive Care Unit experience that has additional training in anesthesia provided by advanced practice nursing programs). There are three major categories of anesthesia provided to patients (1) general anesthesia; (2) conduction anesthesia; and (3) monitored anesthesia care. General anesthesia typically includes the intravenous injection of anesthetic drugs that render the patient unconscious and paralyze their skeletal muscles. Immediately following drug administration a plastic tube is inserted into the trachea and the patient is connected to an electropneumatic system to maintain ventilation of the lungs. A liquid anesthetic agent is vaporized and administered by inhalation, sometimes along with nitrous oxide, to maintain anesthesia for the surgical procedure. Often, other intravenous agents are used in conjunction with the inhalation agents to provide what is called a balanced anesthetic.

Conduction anesthesia refers to blocking the conduction of pain and possibly motor nerve impulses travelling along specific nerves or the spinal cord. Common forms of conduction anesthesia include spinal and epidural anesthesia, as well as specific nerve blocks, for example, axillary nerve blocks. In order to achieve a successful conduction anesthetic, local anesthetic agents such as lidocaine, are injected into the proximity of specific nerves to block the conduction of electrical impulses. In addition, sedation may be provided intravenously to keep the patient comfortable while he/she is lying still for the surgery.

Monitored anesthesia care refers to monitoring the patient's vital signs while administering sedatives and analgesics to keep the patient comfortable, and treating complications related to the surgical procedure. Typically, the surgeon administers topical or local anesthetics to alleviate the pain.

In order to provide the range of support required, from the paralyzed mechanically ventilated patient to the patient receiving monitored anesthesia care, a versatile anesthesia delivery system must be available to the anesthesia care team. Today's anesthesia delivery system is composed of six major elements:

1. The primary and secondary sources of gases (O_2, air, N_2O, vacuum, gas scavenging, and possibly CO_2 and helium)
2. The gas blending and vaporization system
3. The breathing circuit (including methods for manual and mechanical ventilation)
4. The excess gas scavenging system that minimizes potential pollution of the operating room by anesthetic gases
5. Instruments and equipment to monitor the function of the anesthesia delivery system
6. Patient monitoring instrumentation and equipment

The traditional anesthesia machine incorporated elements 1, 2, 3, and more recently 4. The evolution to the anesthesia delivery system adds elements 5 and 6. In the text that follows, references to the "anesthesia machine" refer to the basic gas delivery system and breathing circuit as contrasted with the "anesthesia delivery system" which includes the basic "anesthesia machine" and all monitoring instrumentation.

62.1 Gases Used During Anesthesia and Their Sources

Most inhaled anesthetic agents are liquids that are vaporized in a device within the anesthesia delivery system. The vaporized agents are then blended with other breathing gases before flowing into the breathing circuit and being administered to the patient. The most commonly administered form of anesthesia is called a balanced general anesthetic, and is a combination of inhalation agent plus intravenous analgesic drugs. Intravenous drugs often require electromechanical devices to administer an appropriately controlled flow of drug to the patient.

Gases needed for the delivery of anesthesia are generally limited to oxygen (O_2), air, nitrous oxide (N_2O), and possibly helium (He) and carbon dioxide (CO_2). Vacuum and gas scavenging lines are also required. There needs to be secondary sources of these gases in the event of primary failure or questionable contamination. Typically, primary sources are those supplied from a hospital distribution system at 345 kPa (50 psig) through gas columns or wall outlets. The secondary sources of gas are cylinders hung on yokes on the anesthesia delivery system.

62.1.1 Oxygen

Oxygen provides an essential metabolic substrate for all human cells, but it is not without dangerous side effects. Prolonged exposure to high concentrations of oxygen may result in toxic effects within the lungs that decrease diffusion of gas into and out of the blood, and the return to breathing air following prolonged exposure to elevated O_2 may result in a debilitating explosive blood vessel growth in infants. Oxygen is usually supplied to the hospital in liquid form (boiling point of −183°C), stored in cryogenic tanks, and supplied to the hospital piping system as a gas. The efficiency of liquid storage is obvious since 1 l of liquid becomes 860 l of gas at standard temperature and pressure. The secondary source of oxygen within an anesthesia delivery system is usually one or more E cylinders filled with gaseous oxygen at a pressure of 15.2 MPa (2200 psig).

62.1.2 Air (78% N_2, 21% O_2, 0.9% Ar, 0.1% Other Gases)

The primary use of air during anesthesia is as a diluent to decrease the inspired oxygen concentration. The typical primary source of medical air (there is an important distinction between "air" and "medical air" related to the quality and the requirements for periodic testing) is a special compressor that avoids hydrocarbon based lubricants for purposes of medical air purity. Dryers are employed to rid the compressed air of water prior to distribution throughout the hospital. Medical facilities with limited need for medical air may use banks of H cylinders of dry medical air. A secondary source of air may be available on the anesthesia machine as an E cylinder containing dry gas at 15.2 MPa.

62.1.3 Nitrous Oxide

Nitrous oxide is a colorless, odorless, and non-irritating gas that does not support human life. Breathing more than 85% N_2O may be fatal. N_2O is not an anesthetic (except under hyperbaric conditions), rather it is an analgesic and an amnestic. There are many reasons for administering N_2O during the course of an anesthetic including: enhancing the speed of induction and emergence from anesthesia; decreasing the concentration requirements of potent inhalation anesthetics (i.e., halothane, isoflurane, etc.); and as an essential adjunct to narcotic analgesics. N_2O is supplied to anesthetizing locations from banks of H cylinders that are filled with 90% liquid at a pressure of 5.1 MPa (745 psig). Secondary supplies are available on the anesthesia machine in the form of E cylinders, again containing 90% liquid. Continual exposure to low levels of N_2O in the workplace has been implicated in a number of medical problems including spontaneous abortion, infertility, birth defects, cancer, liver and kidney disease, and others. Although there is no conclusive evidence to support most of these implications, there is a recognized need to scavenge all waste anesthetic gases and periodically sample N_2O levels in the workplace to maintain the lowest possible levels consistent with reasonable risk to the operating room personnel and cost to the institution [Dorsch and Dorsch, 1998]. Another gas with analgesic properties similar to N_2O is xenon, but its use is experimental, and its cost is prohibitive at this time.

62.1.4 Carbon Dioxide

Carbon dioxide is colorless and odorless, but very irritating to breathe in higher concentrations. CO_2 is a byproduct of human cellular metabolism and is not a life-sustaining gas. CO_2 influences many physiologic processes either directly or through the action of hydrogen ions by the reaction $CO_2 + H_2O \leftrightarrow H_2CO_3 \leftrightarrow H^+ + HCO_3^-$. Although not very common in the U.S. today, in the past CO_2 was administered during anesthesia to stimulate respiration that was depressed by anesthetic agents and to cause increased blood flow in otherwise compromised vasculature during some surgical procedures. Like N_2O, CO_2 is supplied as a liquid in H cylinders for distribution in pipeline systems or as a liquid in E cylinders that are located on the anesthesia machine.

62.1.5 Helium

Helium is a colorless, odorless, and non-irritating gas that will not support life. The primary use of helium in anesthesia is to enhance gas flow through small orifices as in asthma, airway trauma, or tracheal stenosis. The viscosity of helium is not different from other anesthetic gases (refer to Table 62.1) and is therefore of no benefit when airway flow is laminar. However, in the event that ventilation must be performed through abnormally narrow orifices or tubes which create turbulent flow conditions, helium is the preferred carrier gas. Resistance to turbulent flow is proportional to the density rather than viscosity of the gas and helium is an order of magnitude less dense than other gases. A secondary advantage of helium is that it has a large specific heat relative to other anesthetic gases and therefore can carry the heat from laser surgery out of the airway more effectively than air, oxygen, or nitrous oxide.

TABLE 62.1 Physical Properties of Gases Used During Anesthesia

Gas	Molecular wt.	Density (g/l)	Viscosity (cp)	Specific heat (KJ/Kg°C)
Oxygen	31.999	1.326	0.0203	0.917
Nitrogen	28.013	1.161	0.0175	1.040
Air	28.975	1.200	0.0181	1.010
Nitrous oxide	44.013	1.836	0.0144	0.839
Carbon dioxide	44.01	1.835	0.0148	0.850
Helium	4.003	0.1657	0.0194	5.190

TABLE 62.2 Physical Properties of Currently Available Volatile Anesthetic Agents

Agent generic name	Boiling point (°C at 760 mmHg)	Vapor pressure (mmHg at 20°C)	Liquid density (g/ml)	MAC[a] (%)
Halothane	50.2	243	1.86	0.75
Enflurane	56.5	175	1.517	1.68
Isoflurane	48.5	238	1.496	1.15
Desflurane	23.5	664	1.45	6.0
Sevoflurane	58.5	160	1.51	2.0

[a] Minimum alveolar concentration is the percentage of the agent required to provide surgical anesthesia to 50% of the population in terms of a cummulative dose response curve. The lower the MAC, the more potent the agent.

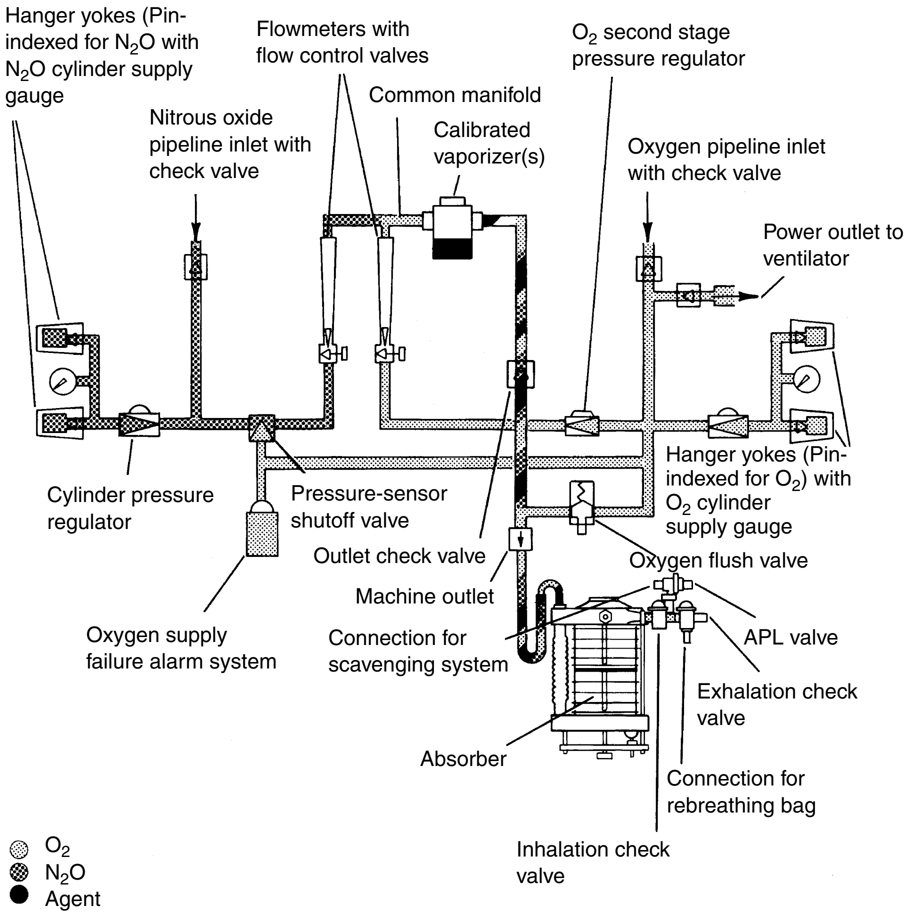

Hanger yokes (Pin-indexed for N_2O with N_2O cylinder supply gauge
Nitrous oxide pipeline inlet with check valve
Flowmeters with flow control valves
Common manifold
Calibrated vaporizer(s)
O_2 second stage pressure regulator
Oxygen pipeline inlet with check valve
Power outlet to ventilator
Hanger yokes (Pin-indexed for O_2) with O_2 cylinder supply gauge
Oxygen flush valve
APL valve
Exhalation check valve
Connection for rebreathing bag
Inhalation check valve
Absorber
Connection for scavenging system
Machine outlet
Outlet check valve
Pressure-sensor shutoff valve
Cylinder pressure regulator
Oxygen supply failure alarm system
O_2
N_2O
Agent

FIGURE 62.1 Schematic diagram of gas piping within a simple two-gas (oxygen and nitrous oxide) anesthesia machine.

62.2 Gas Blending and Vaporization System

The basic anesthesia machine utilizes primary low pressure gas sources of 345 kPa (50 psig) available from wall or ceiling column outlets, and secondary high pressure gas sources located on the machine as pictured schematically in Figure 62.1. Tracing the path of oxygen in the machine demonstrates that oxygen comes from either the low pressure source, or from the 15.2 MPa (2200 psig) high pressure yokes via cylinder pressure regulators and then branches to service several other functions. First and foremost, the second stage pressure regulator drops the O_2 pressure to approximately 110 kPa (16 psig) before it enters the needle valve and the rotameter type flowmeter. From the flowmeter O_2 mixes with gases from other flowmeters and passes through a calibrated agent vaporizer where specific inhalation anesthetic agents are vaporized and added to the breathing gas mixture. Oxygen is also used to supply a reservoir canister that sounds a reed alarm in the event that the oxygen pressure drops below 172 kPa (25 psig). When the oxygen pressure drops to 172 kPa or lower, then the nitrous oxide pressure sensor shutoff valve closes and N_2O is prevented from entering its needle valve and flowmeter and is therefore eliminated from the breathing gas mixture. In fact, all machines built in the U.S. have pressure sensor shutoff valves installed in the lines to every flowmeter, except oxygen, to prevent the delivery of a hypoxic gas mixture in the event of an oxygen pressure failure. Oxygen may also be delivered to the common gas outlet or machine outlet via a momentary normally closed flush valve that typically provides a flow of 65 to 80 l of O_2 per min directly

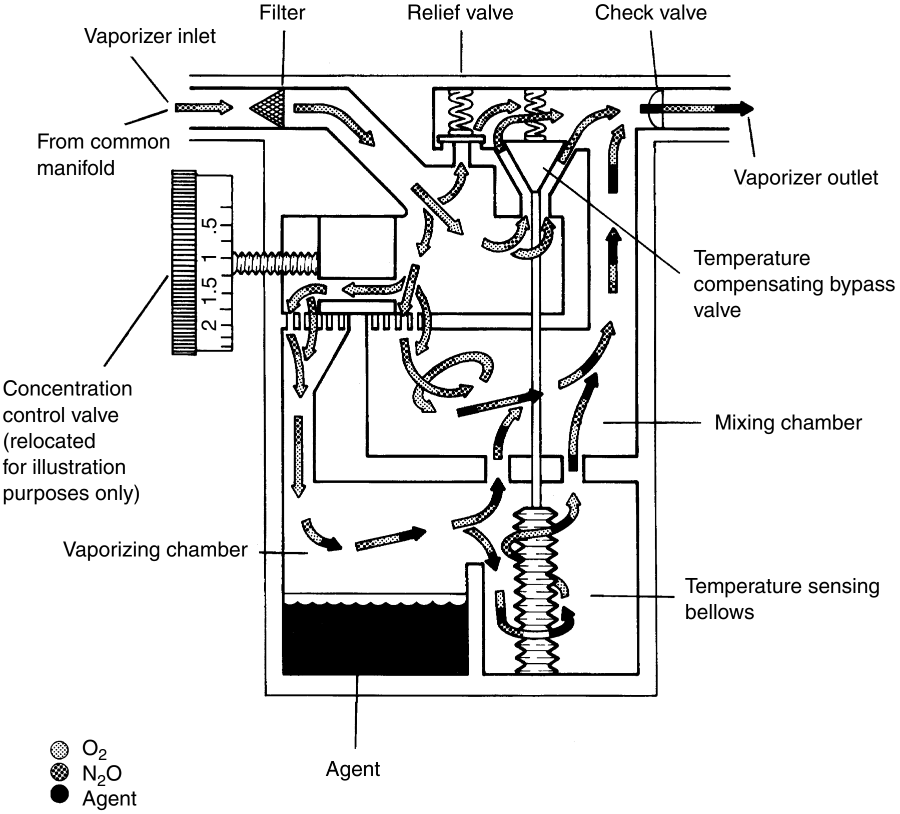


FIGURE 62.2 Schematic diagram of a calibrated in-line vaporizer that uses the flow-over technique for adding anesthetic vapor to the breathing gas mixture.

into the breathing circuit. Newer machines are required to have a safety system for limiting the minimum concentration of oxygen that can be delivered to the patient to 25%. The flow paths for nitrous oxide and other gases are much simpler in the sense that after coming from the high pressure regulator or the low pressure hospital source, gas is immediately presented to the pressure sensor shutoff valve from where it travels to its specific needle valve and flowmeter to join the common gas line and enter the breathing circuit.

Currently all anesthesia machines manufactured in the United States use only calibrated flow-through vaporizers, meaning that all of the gases from the various flowmeters are mixed in the manifold prior to entering the vaporizer. Any given vaporizer has a calibrated control knob that, once set to the desired concentration for a specific agent, will deliver that concentration to the patient. Some form of interlock system must be provided such that only one vaporizer may be activated at any given time. Figure 62.2 schematically illustrates the operation of a purely mechanical vaporizer with temperature compensation. This simple flow-over design permits a fraction of the total gas flow to pass into the vaporizing chamber where it becomes saturated with vapor before being added back to the total gas flow. Mathematically this is approximated by:

$$F_A = \frac{Q_{VC} * P_A}{P_B * (Q_{VC} + Q_G) - P_A * Q_G}$$

where F_A is the fractional concentration of agent at the outlet of the vaporizer, Q_G is the total flow of gas entering the vaporizer, Q_{VC} is the amount of Q_G that is diverted into the vaporization chamber, P_A is the vapor pressure of the agent, and P_B is the barometric pressure.

From Figure 62.2, the temperature compensator would decrease Q_{VC} as temperature increased because vapor pressure is proportional to temperature. The concentration accuracy over a range of clinically expected gas flows and temperatures is approximately ± 15%. Since vaporization is an endothermic process, anesthetic vaporizers must have sufficient thermal mass and conductivity to permit the vaporization process to proceed independent of the rate at which the agent is being used.

62.3 Breathing Circuits

The concept behind an effective breathing circuit is to provide an adequate volume of a controlled concentration of gas to the patient during inspiration, and to carry the exhaled gases away from the patient during exhalation. There are several forms of breathing circuits which can be classified into two basic types (1) open circuit, meaning no rebreathing of any gases and no CO_2 absorber present; and (2) closed circuit, indicating presence of CO_2 absorber and some rebreathing of other gases. Figure 62.3 illustrates the Lack modification of a Mapleson open circuit breathing system. There are no valves and no CO_2 absorber. There is a great potential for the patient to rebreath their own exhaled gases unless the fresh gas inflow is two to three times the patient's minute volume. Figure 62.4 illustrates the most popular form of breathing circuit, the circle system, with oxygen monitor, circle pressure gage, volume monitor

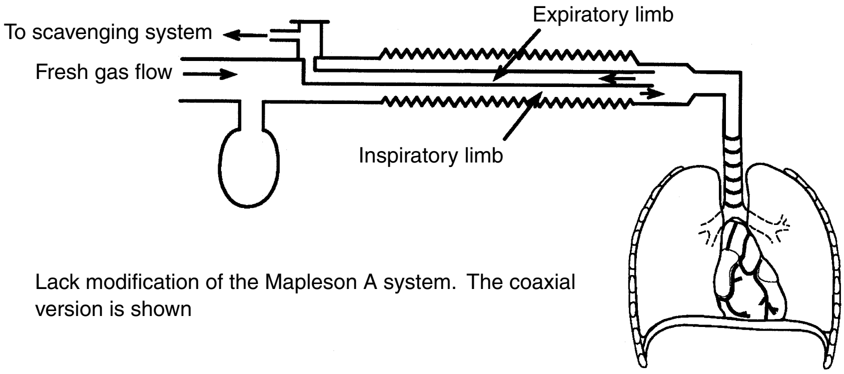

To scavenging system
Fresh gas flow
Expiratory limb
Inspiratory limb
Lack modification of the Mapleson A system. The coaxial version is shown

FIGURE 62.3 An example of an open circuit breathing system that does not use unidirectional flow valves or contain a carbon dioxide absorbent.

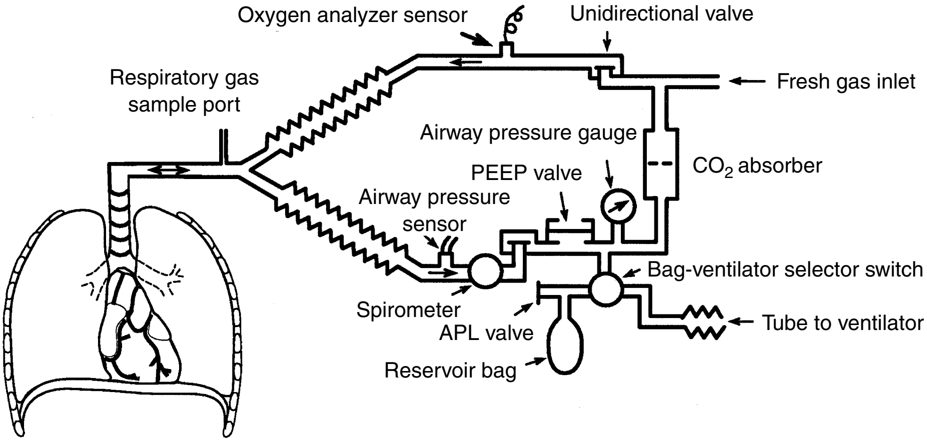


FIGURE 62.4 A diagram of a closed circuit circle breathing system with unidirectional valves, inspired oxygen sensor, pressure sensor, and CO_2 absorber.

(spirometer), and airway pressure sensor. The circle is a closed system, or semi-closed when the fresh gas inflow exceeds the patient's requirements. Excess gas evolves into the scavenging device, and some of the exhaled gas is rebreathed after having the CO_2 removed. The inspiratory and expiratory valves in the circle system guarantee that gas flows to the patient from the inspiratory limb and away from the patient through the exhalation limb. In the event of a failure of either or both of these valves, the patient will rebreath exhaled gas that contains CO_2, which is a potentially dangerous situation.

There are two forms of mechanical ventilation used during anesthesia (1) volume ventilation, where the volume of gas delivered to the patient remains constant regardless of the pressure that is required; and (2) pressure ventilation, where the ventilator provides whatever volume to the patient that is required to produce some desired pressure in the breathing circuit. Volume ventilation is the most popular since the volume delivered remains theoretically constant despite changes in lung compliance. Pressure ventilation is useful when compliance losses in the breathing circuit are high relative to the volume delivered to the lungs.

Humidification is an important adjunct to the breathing circuit because it maintains the integrity of the cilia that line the airways and promote the removal of mucus and particulate matter from the lungs. Humidification of dry breathing gases can be accomplished by simple passive heat and moisture exchangers inserted into the breathing circuit at the level of the endotracheal tube connectors, or by elegant dual servo electronic humidifiers that heat a reservoir filled with water and also heat a wire in the gas delivery tube to prevent rain-out of the water before it reaches the patient. Electronic safety measures must be included in these active devices due to the potential for burning the patient and the fire hazard.

62.4 Gas Scavenging Systems

The purpose of scavenging exhaled and excess anesthetic agents is to reduce or eliminate the potential hazard to employees who work in the environment where anesthetics are administered, including operating rooms, obstetrical areas, special procedures areas, physician's offices, dentist's offices, and veterinarian's surgical suites. Typically more gas is administered to the breathing circuit than is required by the patient, resulting in the necessity to remove excess gas from the circuit. The scavenging system must be capable of collecting gas from all components of the breathing circuit, including adjustable pressure level valves, ventilators, and sample withdrawal type gas monitors, without altering characteristics of the circuit such as pressure or gas flow to the patient. There are two broad types of scavenging systems as illustrated in Figure 62.5: the open interface is a simple design that requires a large physical space for the reservoir volume, and the closed interface with an expandable reservoir bag and which must include relief valves for handling the cases of no scavenged flow and great excess of scavenged flow.

Trace gas analysis must be performed to guarantee the efficacy of the scavenging system. The National Institutes of Occupational Safety and Health (NIOSH) recommends that trace levels of nitrous oxide be maintained at or below 25 parts per million (ppm) time weighted average and that halogenated anesthetic agents remain below 2 ppm.

62.5 Monitoring the Function of the Anesthesia Delivery System

The anesthesia machine can produce a single or combination of catastrophic events, any one of which could be fatal to the patient:

1. Delivery of a hypoxic gas mixture to the patient
2. The inability to adequately ventilate the lungs by not producing positive pressure in the patient's lungs, by not delivering an adequate volume of gas to the lungs, or by improper breathing circuit connections that permit the patient's lungs to receive only rebreathed gases
3. The delivery of an overdose of an inhalational anesthetic agent

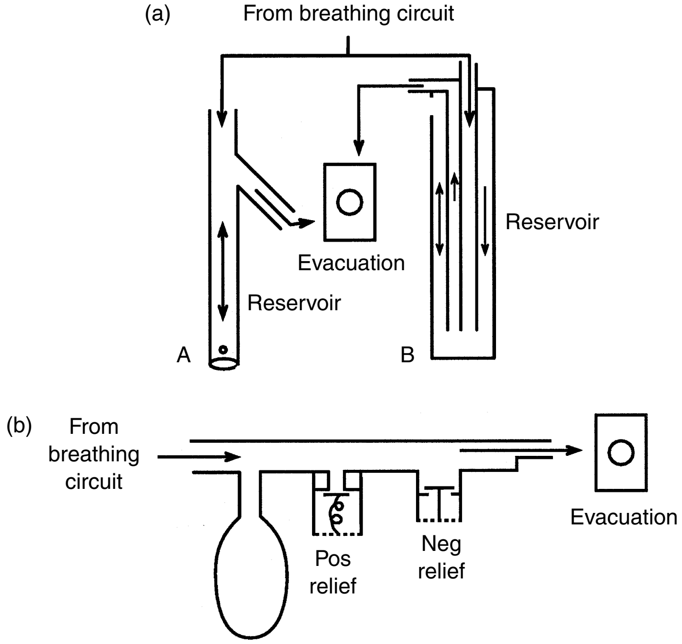


FIGURE 62.5 Examples of (a) open and (b) closed gas scavenger interfaces. The closed interface requires relief valves in the event of scavenging flow failure.

The necessary monitoring equipment to guarantee proper function of the anesthesia delivery system includes at least:

- Inspired Oxygen Concentration monitor with absolute low level alarm of 19%
- Airway Pressure Monitor with alarms for:
    1. Low pressure indicative of inadequate breathing volume and possible leaks
    2. Sustained elevated pressures that could compromise cardiovascular function
    3. High pressures that could cause pulmonary barotrauma
    4. Subatmospheric pressure that could cause collapse of the lungs
- Exhaled Gas Volume Monitor
- Carbon Dioxide Monitor (capnography)
- Inspired and Exhaled Concentration of anesthetic agents by any of the following:
    1. Mass spectrometer
    2. Raman spectrometer
    3. Infrared or other optical spectrometer

A mass spectrometer is a very useful cost-effective device since it alone can provide capnography, inspired and exhaled concentrations of all anesthetic agents, plus all breathing gases simultaneously (O_2, N_2, CO_2, N_2O, Ar, He, halothane, enflurane, isoflurane, desflurane, and suprane). The mass spectrometer is unique in that it may be tuned to monitor an assortment of exhaled gases while the patient is asleep, including (1) ketones for detection of diabetic ketoacidosis; (2) ethanol or other marker in the irrigation solution during transurethral resection of the prostate for early detection of the TURP syndrome, which results in a severe dilution of blood electrolytes; and (3) pentanes during the evolution of a heart attack, to mention a few.

Sound monitoring principles require (1) earliest possible detection of untoward events (before they result in physiologic derangements); and (2) specificity that results in rapid identification and resolution of the problem. An extremely useful rule to always consider is "*never monitor the anesthesia delivery system* *performance through the patient's physiologic responses.*" That is, never intentionally use a device like a pulse oximeter to detect a breathing circuit disconnection since the warning is very late and there is no specific information provided that leads to rapid resolution of the problem.

62.6 Monitoring the Patient

The anesthetist's responsibilities to the patient include: providing relief from pain and preserving all existing normal cellular function of all organ systems. Currently the latter obligation is fulfilled by monitoring essential physiologic parameters and correcting any substantial derangements that occur before they are translated into permanent cellular damage. The inadequacy of current monitoring methods can be appreciated by realizing that most monitoring modalities only indicate damage after an insult has occurred, at which point the hope is that it is reversible or that further damage can be prevented.

Standards for basic intraoperative monitoring of patients undergoing anesthesia, that were developed and adopted by the American Society of Anesthesiologists, became effective in 1990. Standard I concerns the responsibilities of anesthesia personnel, while Standard II requires that the patient's oxygenation, ventilation, circulation, and temperature be evaluated continually during all anesthetics. The following list indicates the instrumentation typically available during the administration of anesthetics.

Electrocardiogram
Pulse oximetry
Urine output
Cardiac output
Electroencephalogram (EEG)
Evoked potentials
Non-invasive or invasive blood pressure
Temperature
Nerve stimulators
Mixed venous oxygen saturation
Transesophageal echo cardiography (TEE)
Coagulation status
Blood gases and electrolytes (pO_2, pCO_2, pH, BE, Na^+, K^+, Cl^-, Ca^{++}, and glucose)
Mass spectrometry, Raman spectrometry, or infrared breathing gas analysis

62.6.1 Control of Patient Temperature

Anesthesia alters the thresholds for temperature regulation and the patient becomes unable to maintain normal body temperature. As the patient's temperature falls even a few degrees toward room temperature, several physiologic derangements occur (1) drug action is prolonged; (2) blood coagulation is impaired; and (3) post-operative infection rate increases. On the positive side, cerebral protection from inadequate perfusion is enhanced by just a few degrees of cooling. Proper monitoring of core body temperature and forced hot air warming of the patient is essential.

62.6.2 Monitoring the Depth of Anesthesia

There are two very unpleasant experiences that patients may have while undergoing an inadequate anesthetic (1) the patient is paralyzed and unable to communicate their state of discomfort, and they are feeling the pain of surgery and are aware of their surroundings; (2) the patient may be paralyzed, unable to communicate, and is aware of their surroundings, but is not feeling any pain. The ability to monitor the depth of anesthesia would provide a safeguard against these unpleasant experiences. However, despite numerous instruments and approaches to the problem it remains elusive. Brain stem auditory evoked responses have come the closest to depth of anesthesia monitoring, but it is difficult to perform, is expensive, and is not possible to perform during many types of surgery. A promising new technology, called bi-spectral index (BIS monitoring) is purported to measure the level of patient awareness through multivariate analysis of a single channel of the EEG.

62.6.3 Anesthesia Computer-Aided Record Keeping

Conceptually, every anesthetist desires an automated anesthesia record keeping system. Anesthesia care can be improved through the feedback provided by correct record keeping, but today's systems have an enormous overhead associated with their use when compared to standard paper record keeping. No doubt that automated anesthesia record keeping reduces the drudgery of routine recording of vital signs, but to enter drugs and drips and their dosages, fluids administered, urine output, blood loss, and other data requires much more time and machine interaction than the current paper system. Despite attempts to use every input/output device ever produced by the computer industry from keyboards to bar codes to voice and handwriting recognition, no solution has been found that meets wide acceptance. Tenants of a successful system must include:

1. The concept of a user transparent system, which is ideally defined as requiring no communication between the computer and the clinician (far beyond the concept of user friendly), and therefore that is intuitively obvious to use even to the most casual users
2. Recognition of the fact that educational institutions have very different requirements from private practice institutions
3. Real time hard copy of the record produced at the site of anesthetic administration that permits real time editing and notation
4. Ability to interface with a great variety of patient and anesthesia delivery system monitors from various suppliers
5. Ability to interface with a large number of hospital information systems
6. Inexpensive to purchase and maintain

62.6.4 Alarms

Vigilance is the key to effective risk management, but maintaining a vigilant state is not easy. The practice of anesthesia has been described as moments of shear terror connected by times of intense boredom. Alarms can play a significant role in redirecting one's attention during the boredom to the most important event regarding patient safety, but only if false alarms can be eliminated, alarms can be prioritized, and all alarms concerning anesthetic management can be displayed in a single clearly visible location.

62.6.5 Ergonomics

The study of ergonomics attempts to improve performance by optimizing the relationship between people and their work environment. Ergonomics has been defined as a discipline which investigates and applies information about human requirements, characteristics, abilities, and limitations to the design, development, and testing of equipment, systems, and jobs [Loeb, 1993]. This field of study is only in its infancy and examples of poor ergonomic design abound in the anesthesia workplace.

62.6.6 Simulation in Anesthesia

Complete patient simulators are hands-on realistic simulators that interface with physiologic monitoring equipment to simulate patient responses to equipment malfunctions, operator errors, and drug therapies. There are also crisis management simulators. Complex patient simulators, which are analogous to flight simulators, are currently being marketed for training anesthesia personnel. The intended use for these complex simulators is currently being debated in the sense that training people to respond in a preprogrammed way to a given event may not be adequate training.

62.6.7 Reliability

The design of an anesthesia delivery system is unlike the design of most other medical devices because it is a life support system. As such, its core elements deserve all of the considerations of the latest fail-safe technologies. Too often in today's quest to apply microprocessor technology to everything, tradeoffs are made among reliability, cost, and engineering elegance. The most widely accepted anesthesia machine designs continue to be based upon simple ultra-reliable mechanical systems with an absolute minimum of catastrophic failure modes. The replacement of needle valves and rotameters, for example, with microprocessor controlled electromechanical valves can only introduce new catastrophic failure modes. However, the inclusion of microprocessors can enhance the safety of anesthesia delivery if they are implemented without adding catastrophic failure modes.

Further Information

Blitt, C.D. and Hines, R.L., Eds. (1995). *Monitoring in Anesthesia and Critical Care Medicine*, 3rd ed. Churchill Livingstone, New York.

Dorsch, J.A. and Dorsch, S.E. (1998). *Understanding Anesthesia Equipment*, 4th ed. Williams and Wilkins, Baltimore, MD.

Ehrenwerth, J. and Eisenkraft, J.B. (1993). *Anesthesia Equipment: Principles and Applications.* Mosby, St. Louis, MO.

Gravenstein N. and Kirby, R.R., Eds. (1996). *Complications in Anesthesiology,* 2nd ed. Lippincott-Raven, Philadelphia, PA.

Loeb, R. (1993). Ergonomics of the anesthesia workplace. *STA Interface* 4: 18.

Miller, R.D. Ed. (1999). *Anesthesia*, 5th ed. Churchill Livingstone, New York.

Miller, R.D. Ed. (1998). *Atlas of Anesthesia.* Churchill Livingstone, New York.

Saidman, L.J. and Smith, N.T., Eds. (1993). *Monitoring in Anesthesia*, 3rd ed. Butterworth-Heinemann, Stoneham, MA.

63

Electrosurgical Devices

Jeffrey L. Eggleston
Valleylab, Inc.

Wolf W. von Maltzahn
Whitaker Foundation

An electrosurgical unit (ESU) passes high-frequency electric currents through biologic tissues to achieve specific surgical effects such as cutting, **coagulation**, or **desiccation**. Although it is not completely understood how electrosurgery works, it has been used since the 1920s to cut tissue effectively while at the same time controlling the amount of bleeding. Cutting is achieved primarily with a continuous sinusoidal waveform, whereas coagulation is achieved primarily with a series of sinusoidal wave packets. The surgeon selects either one of these waveforms or a blend of them to suit the surgical needs. An electrosurgical unit can be operated in two modes, the monopolar mode and the bipolar mode. The most noticeable difference between these two modes is the method in which the electric current enters and leaves the tissue. In the monopolar mode, the current flows from a small **active electrode** into the surgical site, spreads through the body, and returns to a large **dispersive electrode** on the skin. The high current density in the vicinity of the active electrode achieves tissue cutting or coagulation, whereas the low current density under the dispersive electrode causes no tissue damage. In the bipolar mode, the current flows only through the tissue held between two forceps electrodes. The monopolar mode is used for both cutting and coagulation. The bipolar mode is used primarily for coagulation.

This chapter begins with the theory of operation for electrosurgical units, outlines various modes of operation, and gives basic design details for electronic circuits and electrodes. It then describes how improper application of electrosurgical units can lead to hazardous situations for both the operator and the patient and how such hazardous situations can be avoided or reduced through proper monitoring methods. Finally, the chapter gives an update on current and future developments and applications.

63.1 Theory of Operation

In principle, electrosurgery is based on the rapid heating of tissue. To better understand the thermodynamic events during electrosurgery, it helps to know the general effects of heat on biologic tissue. Consider a tissue volume that experiences a temperature increase from normal body temperature to 45°C within a few seconds. Although the cells in this tissue volume show neither microscopic nor macroscopic changes, some cytochemical changes do in fact occur. However, these changes are reversible, and the cells return to their normal function when the temperature returns to normal values. Above 45°C, irreversible changes take place that inhibit normal cell functions and lead to cell death. First, between 45°C and 60°C, the proteins in the cell lose their quaternary configuration and solidify into a glutinous substance that resembles the white of a hard-boiled egg. This process, termed *coagulation*, is accompanied by tissue blanching. Further increasing the temperature up to 100°C leads to tissue drying; that is, the aqueous cell contents evaporate. This process is called *desiccation*. If the temperature is increased beyond 100°C, the solid contents of the tissue reduce to carbon, a process referred to as *carbonization*. Tissue damage depends not only on temperature, however, but also on the length of exposure to heat. Thus, the overall temperature-induced tissue damage is an integrative effect between temperature and time that is expressed mathematically by the Arrhenius relationship, where an exponential function of temperature is integrated over time [1].

In the monopolar mode, the active electrode either touches the tissue directly or is held a few millimeters above the tissue. When the electrode is held above the tissue, the electric current bridges the air gap by creating an electric discharge arc. A visible arc forms when the electric field strength exceeds 1 kV/mm in the gap and disappears when the field strength drops below a certain threshold level.

When the active electrode touches the tissue and the current flows directly from the electrode into the tissue without forming an arc, the rise in tissue temperature follows the bioheat equation

$$T - T_o = \frac{1}{\sigma \rho c} J^2 t \tag{63.1}$$

where T and T_o are the final and initial temperatures (K), σ is the electrical conductivity (S/m), ρ is the tissue density (kg/m^3), c is the specific heat of the tissue (Jkg^{-1}K^{-1}), J is the current density (A/m^2), and t is the duration of heat applications [1]. The bioheat equation is valid for short application times where secondary effects such as heat transfer to surrounding tissues, blood perfusion, and metabolic heat can be neglected. According to Equation 63.1, the surgeon has primarily three means of controlling the cutting or coagulation effect during electrosurgery: the contact area between active electrode and tissue, the electrical current density, and the activation time. In most commercially available electrosurgical generators, the output variable that can be adjusted is power. This power setting, in conjunction with the output power vs. tissue impedance characteristics of the generator, allow the surgeon some control over current. Table 63.1 lists typical output power and mode settings for various surgical procedures. Table 63.2 lists some typical impedance ranges seen during use of an ESU in surgery. The values are shown as ranges because the impedance increases as the tissue dries out, and at the same time, the output power of the ESU decreases. The surgeon may control current density by selection of the active electrode type and size.

63.2 Monopolar Mode

A continuous sinusoidal waveform cuts tissue with very little hemostasis. This waveform is simply called *cut* or *pure cut*. During each positive and negative swing of the sinusoidal waveform, a new discharge arc forms and disappears at essentially the same tissue location. The electric current concentrates at this tissue location, causing a sudden increase in temperature due to resistive heating. The rapid rise in temperature then vaporizes intracellular fluids, increases cell pressure, and ruptures the cell membrane, thereby parting the tissue. This chain of events is confined to the vicinity of the arc, because from there the electric current

TABLE 63.1 Typical ESU Power Settings for Various Surgical Procedures

Power-level range	Procedures
Low power	
<30 W cut	Neurosurgery
<30 W coag	Dermatology
	Plastic surgery
	Oral surgery
	Laparoscopic sterilization
	Vasectomy
Medium power	
30–150 W cut	General surgery
30–70 W coag	Laparotomies
	Head and neck surgery (ENT)
	Major orthopedic surgery
	Major vascular surgery
	Routine thoracic surgery
	Polypectomy
High power	
>150 W cut	Transurethral resection procedures (TURPs)
>70 W coag	Thoracotomies
	Ablative cancer surgery
	Mastectomies

Note: Ranges assume the use of a standard blade electrode. Use of a needle electrode, or other small current-concentrating electrode, allows lower settings to be used; users are urged to use the lowest setting that provides the desired clinical results.

TABLE 63.2 Typical Impedance Ranges Seen During Use of an ESU in Surgery

Cut mode application	Impedance range (Ω)
Prostate tissue	400–1700
Oral cavity	1000–2000
Liver tissue	
Muscle tissue	
Gall bladder	1500–2400
Skin tissue	1700–2500
Bowel tissue	2500–3000
Periosteum	
Mesentery	3000–4200
Omentum	
Adipose tissue	3500–4500
Scar tissue	
Adhesions	
Coag Mode Application	
Contact coagulation to stop bleeding	100–1000

spreads to a much larger tissue volume, and the current density is no longer high enough to cause resistive heating damage. Typical output values for ESUs, in cut and other modes, are shown in Table 63.3.

Experimental observations have shown that more hemostasis is achieved when cutting with an interrupted sinusoidal waveform or amplitude modulated continuous waveform. These waveforms are typically

TABLE 63.3 Typical Output Characteristics of ESUs

	Output voltage range open circuit, $V_{\text{peak-peak}}$, V	Output power range, W	Frequency, kHz	Crest factor ($V_{\text{Peak}}/V_{\text{rms}}$)	Duty cycle %
Monopolar modes					
Cut	200–5000	1–400	300–1750	1.4–2.1	100
Blend	1500–5800	1–300	300–1750	2.1–6.0	25–80
Desiccate	400–6500	1–200	240–800	3.5–6.0	50–100
Fulgurate/spray	6000–12000	1–200	300–800	6.0–20.0	10–70
Bipolar mode					
Coagulate/desiccate	200–1000	1–70	300–1050	1.6–12.0	25–100

called *blend* or *blended cut.* Some ESUs offer a choice of blend waveforms to allow the surgeon to select the degree of hemostasis desired.

When a continuous or interrupted waveform is used in contact with the tissue and the output voltage current density is too low to sustain arcing, desiccation of the tissue will occur. Some ESUs have a distinct mode for this purpose called *desiccation* or *contact coagulation.*

In noncontact coagulation, the duty cycle of an interrupted waveform and the crest factor (ratio of peak voltage to rms voltage) influence the degree of hemostasis. While a continuous waveform reestablishes the arc at essentially the same tissue location concentrating the heat there, an interrupted waveform causes the arc to reestablish itself at different tissue locations. The arc seems to dance from one location to the other raising the temperature of the top tissue layer to coagulation levels. These waveforms are called *fulguration* or *spray.* Since the current inside the tissue spreads very quickly from the point where the arc strikes, the heat concentrates in the top layer, primarily desiccating tissue and causing some carbonization. During surgery, a surgeon can easily choose between cutting, coagulation, or a combination of the two by activating a switch on the grip of the active electrode or by use of a footswitch.

63.3 Bipolar Mode

The bipolar mode concentrates the current flow between the two electrodes, requiring considerably less power for achieving the same coagulation effect than the monopolar mode. For example, consider coagulating a small blood vessel with 3-mm external diameter and 2-mm internal diameter, a tissue resistivity of 360 Ωcm, a contract area of 2×4 mm^2, and a distance between the forceps tips of 1 mm. The tissue resistance between the forceps is 450 Ω as calculated from $R = \rho L/A$, where ρ is the resistivity, L is the distance between the forceps, and A is the contact area. Assuming a typical current density of 200 mA/cm^2, then a small current of 16 mA, a voltage of 7.2 V, and a power level of 0.12 W suffice to coagulate this small blood vessel. In contrast, during monopolar coagulation, current levels of 200 mA and power levels of 100 W or more are not uncommon to achieve the same surgical effect. The temperature increase in the vessel tissue follows the bioheat equation, Equation 63.1. If the specific heat of the vessel tissue is 4.2 Jkg^{-1}K^{-1} and the tissue density is 1 g/cm^3, then the temperature of the tissue between the forceps increases from 37 to 57°C in 5.83 sec. When the active electrode touches the tissue, less tissue damage occurs during coagulation, because the charring and carbonization that accompanies **fulguration** is avoided.

63.4 ESU Design

Modern ESUs contain building blocks that are also found in other medical devices, such as microprocessors, power supplies, enclosures, cables, indicators, displays, and alarms. The main building blocks unique to ESUs are control input switches, the high-frequency power amplifier, and the safety monitor. The first two will be discussed briefly here, and the latter will be discussed later.

Control input switches include front panel controls, footswitch controls, and handswitch controls. In order to make operating an ESU more uniform between models and manufacturers, and to reduce the possibility of operator error, the ANSI/AAMI HF-18 standard [2] makes specific recommendations concerning the physical construction and location of these switches and prescribes mechanical and electrical performance standards. For instance, front panel controls need to have their function identified by a permanent label and their output indicated on alphanumeric displays or on graduated scales; the pedals of foot switches need to be labeled and respond to a specified activation force; and if the active electrode handle incorporates two finger switches, their position has to correspond to a specific function. Additional recommendations can be found in Reference 2.

Four basic high-frequency power amplifiers are in use currently; the somewhat dated vacuum tube/spark gap configuration, the parallel connection of a bank if bipolar power transistors, the hybrid connection of parallel bipolar power transistors cascaded with metal oxide silicon field effect transistors (MOSFETs), and the bridge connection of MOSFETs. Each has unique properties and represents a stage in the evolution of ESUs.

In a vacuum tube/spark gap device, a tuned-plate, tuned-grid vacuum tube oscillator is used to generate a continuous waveform for use in cutting. This signal is introduced to the patient by an adjustable isolation transformer. To generate a waveform for fulguration, the power supply voltage is elevated by a step-up transformer to about 1600 V rms which then connects to a series of spark gaps. The voltage across the spark gaps is capacitively coupled to the primary of an isolation transformer. The RLC circuit created by this arrangement generates a high crest factor, damped sinusoidal, interrupted waveform. One can adjust the output power and characteristics by changing the turns ratio or tap on the primary and/or secondary side of the isolation transformer, or by changing the spark gap distance.

In those devices that use a parallel bank of bipolar power transistors, the transistors are arranged in a Class A configuration. The bases, collectors, and emitters are all connected in parallel, and the collective base node is driven through a current-limiting resistor. A feedback RC network between the base node and the collector node stabilizes the circuit. The collectors are usually fused individually before the common node connects them to one side of the primary of the step-up transformer. The other side of the primary is connected to the high-voltage power supply. A capacitor and resistor in parallel to the primary create a resonance tank circuit that generates the output waveform at a specific frequency. Additional elements may be switched in and out of the primary parallel RLC to alter the output power and waveform for various electrosurgical modes. Small-value resistors between the emitters and ground improve the current sharing between transistors. This configuration sometimes requires the use of matched sets of high-voltage power transistors.

A similar arrangement exists in amplifiers using parallel bipolar transistors cascaded with a power MOSFET. This arrangement is called a *hybrid cascode amplifier*. In this type of amplifier, the collectors of a group if bipolar transistors are connected, via protection diodes, to one side of the primary of the step-up output transformer. The other side of the primary is connected to the high-voltage power supply. The emitters of two or three bipolar transistors are connected, via current limiting resistors, to the drain of an enhancement mode MOSFET. The source of the MOSFET is connected to ground, and the gate of the MOSFET is connected to a voltage-snubbing network driven by a fixed amplitude pulse created by a high-speed MOS driver circuit. The bases of the bipolar transistors are connected, via current control RC networks, to a common variable base voltage source. Each collector and base is separately fused. In cut modes, the gate drive pulse is a fixed frequency, and the base voltage is varied according to the power setting. In the coagulation modes, the base voltage is fixed and the width of the pulses driving the MOSFET is varied. This changes the conduction time of the amplifier and controls the amount of energy imparted to the output transformer and its load. In the coagulation modes and in high-power cut modes, the bipolar power transistors are saturated, and the voltage across the bipolar/MOSFET combination is low. This translates to high efficiency and low power dissipation.

The most common high-frequency power amplifier in use is a bridge connection of MOSFETs. In this configuration, the drains of a series of power MOSFETs are connected, via protection diodes, to one side of the primary of the step-up output transformer. The drain protection diodes protect the MOSFETs against the negative voltage swings of the transformer primary. The other side of the transformer primary is connected to the high-voltage power supply. The sources of the MOSFETs are connected to ground. The gate of each MOSFET has a resistor connected to ground and one to its driver circuitry. The resistor to ground speeds up the discharge of the gate capacitance when the MOSFET is turned on while the gate series resistor eliminates turn-off oscillations. Various combinations of capacitors and/or LC networks can be switched across the primary of the step-up output transformer to obtain different waveforms. In the cut mode, the output power is controlled by varying the high-voltage power supply voltage. In the coagulation mode, the output power is controlled by varying the on time of the gate drive pulse.

63.5 Active Electrodes

The monopolar active electrode is typically a small flat blade with symmetric leading and trailing edges that is embedded at the tip of an insulated handle. The edges of the blade are shaped to easily initiate discharge arcs and to help the surgeon manipulate the incision; the edges cannot mechanically cut tissue. Since the surgeon holds the handle like a pencil, it is often referred to as the "pencil." Many pencils contain in their handle one or more switches to control the electrosurgical waveform, primarily to switch between cutting and coagulation. Other active electrodes include needle electrodes, loop electrodes, and ball electrodes. Needle electrodes are used for coagulating small tissue volumes like in neurosurgery or plastic surgery. Loop electrodes are used to resect nodular structures such as polyps or to excise tissue samples for pathologic analysis. An example would be the LLETZ procedure where the transition zone of the cervix is excised. Electrosurgery at the tip of an endoscope or laparoscope requires yet another set of active electrodes and specialized training of the surgeon.

63.6 Dispersive Electrodes

The main purpose of the dispersive electrode is to return the high-frequency current to the electrosurgical unit without causing harm to the patient. This is usually achieved by attaching a large electrode to the patient's skin away from the surgical site. The large electrode area and a small contact impedance reduce the current density to levels where tissue heating is minimal. Since the ability of a dispersive electrode to avoid tissue heating and burns is of primary importance, dispersive electrodes are often characterized by their *heating factor*. The heating factor describes the energy dissipated under the dispersive electrode per Ω of impedance and is equal to $I^2 t$, where I is the rms current and t is the time of exposure. During surgery a typical value for the heating factor is 3 $A^2 s$, but factors of up to 9 $A^2 s$ may occur during some procedures [3].

Two types of dispersive electrodes are in common use today, the resistive type and the capacitive type. In disposable form, both electrodes have a similar structure and appearance. A thin, rectangular metallic foil has an insulating layer on the outside, connects to a gel-like material on the inside, and may be surrounded by an adhesive foam. In the resistive type, the gel-like material is made of an adhesive conductive gel, whereas in the capacitive type, the gel is an adhesive dielectric nonconductive gel. The adhesive foam and adhesive gel layer ensure that both electrodes maintain good skin contact to the patient, even if the electrode gets stressed mechanically from pulls on the electrode cable. Both types have specific advantages and disadvantages. Electrode failures and subsequent patient injury can be attributed mostly to improper application, electrode dislodgment, and electrode defects rather than to electrode design.

63.7 ESU Hazards

Improper use of electrosurgery may expose both the patient and the surgical staff to a number of hazards. By far the most frequent hazards are electric shock and undesired burns. Less frequent are undesired neuromuscular stimulation, interference with pacemakers or other devices, electrochemical effects from direct currents, implant heating, and gas explosions [1,4].

Current returns to the ESU through the dispersive electrode. If the contact area of the dispersive electrode is large and the current exposure time short, then the skin temperature under the electrode does not rise above 45°C, which has been shown to be the maximum safe temperature [5]. However, to include a safety margin, the skin temperature should not rise more than 6°C above the normal surface temperature of 29 to 33°C. The current density at any point under the dispersive electrode has to be significantly below the recognized burn threshold of 100 mA/cm^2 for 10 sec.

To avoid electric shock and burns, the American National Standard for Electrosurgical Devices [2] requires that "any electrosurgical generator that provides for a dispersive electrode and that has a rated output power of greater than 50 W shall have at least one patient circuit safety monitor." The most common safety monitors are the contact quality monitor for the dispersive electrode and the patient circuit monitor. A contact quality monitor consists of a circuit to measure the impedance between the two sides of a split dispersive electrode and the skin. A small high-frequency current flows from one section of the dispersive electrode through the skin to the second section of the dispersive electrode. If the impedance between these two sections exceeds a certain threshold, or changes by a certain percentage, an audible alarm sounds, and the ESU output is disabled.

Patient circuit monitors range from simple to complex. The simple ones monitor electrode cable integrity while the complex ones detect any abnormal condition that could result in electrosurgical current flowing in other than normal pathways. Although the output isolation transformer present in most modern ESUs usually provides adequate patient protection, some potentially hazardous conditions may still arise. If a conductor to the dispersive electrode is broken, undesired arcing between the broken conductor ends may occur, causing fire in the operating room and serious patient injury. Abnormal current pathways may also arise from capacitive coupling between cables, the patient, operators, enclosures, beds, or any other conductive surface or from direct connections to other electrodes connected to the patient. The patient circuit monitoring device should be operated from an isolated power source having a maximum voltage of 12 V rms. The most common device is a cable continuity monitor. Unlike the contact quality monitor, this monitor only checks the continuity of the cable between the ESU and the dispersive electrode and sounds an alarm if the resistance in that conductor is greater than 1 kΩ. Another implementation of a patient circuit monitor measures the voltage between the dispersive electrode connection and ground. A third implementation functions similarly to a ground fault circuit interrupter (GFCI) in that the current in the wire to the active electrode and the current in the wire to the dispersive electrode are measured and compared with each other. If the difference between these currents is greater than a preset threshold, the alarm sounds and the ESU is disconnected.

There are other sources of undesired burns. Active electrodes get hot when they are used. After use, the active electrode should be placed in a protective holster, if available, or on a suitable surface to isolate it from the patient and surgical staff. The correct placement of an active electrode will also prevent the patient and/or surgeon from being burned if an inadvertent activation of the ESU occurs (e.g., someone accidentally stepping on a foot pedal). Some surgeons use a practice called *buzzing the hemostat* in which a small bleeding vessel is grasped with a clamp or hemostat and the active electrode touched to the clamp while activating. Because of the high voltages involved and the stray capacitance to ground, the surgeon's glove may be compromised. If the surgical staff cannot be convinced to eliminate the practice of buzzing hemostats, the probability of burns can be reduced by use of a cut waveform instead of a coagulation waveform (lower voltage), by maximizing contact between the surgeon's hand and the clamp, and by not activating until the active electrode is firmly touching the clamp.

Although it is commonly assumed that neuromuscular stimulation ceases or is insignificant at frequencies above 10 kHz, such stimulation has been observed in anesthetized patients undergoing certain electrosurgical procedures. This undesirable side effect of electrosurgery is generally attributed to nonlinear events during the electric arcing between the active electrode and tissue. These events rectify the high-frequency current leading to both dc and low-frequency current components. These current components can reach magnitudes that stimulate nerve and muscle cells. To minimize the probability of unwanted neuromuscular stimulation, most ESUs incorporate in their output circuit a high-pass filter that suppresses dc and low-frequency current components.

The use of electrosurgery means the presence of electric discharge arcs. This presents a potential fire hazard in an operating room where oxygen and flammable gases may be present. These flammable gases may be introduced by the surgical staff (anesthetics or flammable cleaning solutions), or may be generated within the patients themselves (bowel gases). The use of disposable paper drapes and dry surgical gauze also provides a flammable material that may be ignited by sparking or by contact with a hot active electrode. Therefore, prevention of fires and explosions depends primarily on the prudence and judgment of the ESU operator.

63.8 Recent Developments

Electrosurgery is being enhanced by the addition of a controlled column of argon gas in the path between the active electrode and the tissue. The flow of argon gas assists in clearing the surgical site of fluid and improves visibility. When used in the coagulation mode, the argon gas is turned into a plasma allowing tissue damage and smoke to be reduced, and producing a thinner, more flexible eschar. When used with the cut mode, lower power levels may be used.

Many manufacturers have begun to include sophisticated computer-based systems in their ESUs that not only simplify the use of the device but also increase the safety of patient and operator [6]. For instance, in a so-called soft coagulation mode, a special circuit continuously monitors the current between the active electrode and the tissue and turns the ESU output on only after the active electrode has contacted the tissue. Furthermore, the ESU output is turned off automatically, once the current has reached a certain threshold level that is typical for coagulated and desiccated tissue. This feature is also used in a bipolar mode termed *autobipolar*. Not only does this feature prevent arcing at the beginning of the procedure, but it also keeps the tissue from being heated beyond 70°C. Some devices offer a so-called power-peak-system that delivers a very short power peak at the beginning of electrosurgical cutting to start the cutting arc. Other modern devices use continuous monitoring of current and voltage levels to make automatic power adjustments in order to provide for a smooth cutting action from the beginning of the incision to its end. Some manufacturers are developing waveforms and instruments designed to achieve specific clinical results such as bipolar cutting tissue lesioning, and vessel sealing. With the growth and popularity of laparoscopic procedures, additional electrosurgical instruments and waveforms tailored to this surgical specialty should also be expected.

Increased computing power, more sophisticated evaluation of voltage and current waveforms, and the addition of miniaturized sensors will continue to make ESUs more user-friendly and safer.

Defining Terms

Active electrode: Electrode used for achieving desired surgical effect.

Coagulation: Solidification of proteins accompanied by tissue whitening.

Desiccation: Drying of tissue due to the evaporation of intracellular fluids.

Dispersive electrode: Return electrode at which no electrosurgical effect is intended.

Fulguration: Random discharge of sparks between active electrode and tissue surface in order to achieve coagulation and/or desiccation.

Spray: Another term for *fulguration*. Sometimes this waveform has a higher crest factor than that used for fulguration.

References

[1] Pearce John A. 1986. *Electrosurgery*, New York, John Wiley.

[2] American National Standard for Electrosurgical Devices. 1994. HF18, American National Standards Institute.

[3] Gerhard Glen C. 1988. Electrosurgical unit. In J.G. Webster (Ed.), *Encyclopedia of Medical Devices and Instrumentation*, Vol. 2, pp. 1180–1203, New York, John Wiley.

[4] Gendron Francis G. 1988. *Unexplained Patient Burns: Investigating Latrogenic Injuries*, Brea, CA, Quest Publishing.

[5] Pearce J.A., Geddes L.A., and Van Vleet J.F. et al. 1983. Skin burns from electrosurgical current. *Med. Instrum.* 17: 225.

[6] Haag R. and Cuschieri A. 1993. Recent advances in high-frequency electrosurgery: development of automated systems. *J. R. Coll. Surg. Ednb.* 38: 354.

[7] LaCourse J.R., Miller W.T. III, and Vogt M. et al. 1985. Effect of high frequency current on nerve and muscle tissue. *IEEE Trans. Biomed. Eng.* 32: 83.

Further Information

American National Standards Institute, 1988. International Standard, Medical Electrical Equipment, Part 1: General Requirements for Safety, IEC 601-1, 2nd ed., New York.

American National Standards Institute, 1991. International Standard, Medical Electrical Equipment, Part 2: Particular Requirements for the Safety of High Frequency Surgical Equipment, IEC 601-2-2, 2nd ed., New York.

National Fire Protection Association, 1993. Standard for Health Care Facilities, NFPA 99.

64

Biomedical Lasers

Millard M. Judy
Baylor Research Institute

Approximately 20 years ago the CO_2 laser was introduced into surgical practice as a tool to photothermally ablate, and thus to incise and to debulk, soft tissues. Subsequently, three important factors have led to the expanding biomedical use of laser technology, particularly in surgery. These factors are (1) the increasing understanding of the wave-length selective interaction and associated effects of **ultraviolet-infrared (UV–IR) radiation** with biologic tissues, including those of acute damage and long-term healing, (2) the rapidly increasing availability of lasers emitting (essentially monochromatically) at those wavelengths that are strongly absorbed by molecular species within tissues, and (3) the availability of both optical fiber and lens technologies as well as of endoscopic technologies for delivery of the laser radiation to the often remote internal treatment site. Fusion of these factors has led to the development of currently available biomedical laser systems.

This chapter briefly reviews the current status of each of these three factors. In doing so, each of the following topics will be briefly discussed:

1. The physics of the interaction and the associated effects (including clinical efforts) of UV–IR radiation on biologic tissues
2. The fundamental principles that underlie the operations and construction of all lasers

3. The physical properties of the optical delivery systems used with the different biomedical lasers for delivery of the laser beam to the treatment site
4. The essential physical features of those biomedical lasers currently in routine use ranging over a number of clinical specialties, and brief descriptions of their use
5. The biomedical uses of other lasers used surgically in limited scale or which are currently being researched for applications in surgical and diagnostic procedures and the photosensitized inactivation of cancer tumors

In this review, effort is made in the text and in the last section to provide a number of key references and sources of information for each topic that will enable the reader's more in-depth pursuit.

64.1 Interaction and Effects of UV–IR Laser Radiation on Biologic Tissues

Electromagnetic radiation in the UV–IR spectral range propagates within biologic tissues until it is either scattered or absorbed.

64.1.1 Scattering in Biologic Tissue

Scattering in matter occurs only at the boundaries between regions having different optical refractive indices and is a process in which the energy of the radiation is conserved [Van de Hulst, 1957]. Since biologic tissue is structurally inhomogeneous at the microscopic scale, for example, both subcellular and cellular dimensions, and at the macroscopic scale, for example, cellular assembly (tissue) dimensions, and predominantly contains water, proteins, and lipids, all different chemical species, it is generally regarded as a scatterer of UV–IR radiation. The general result of scattering is deviation of the direction of propagation of radiation. The deviation is strongest when wavelength and scatterer are comparable in dimension (Mie scattering) and when wavelength greatly exceeds particle size (Rayleigh scattering) [Van de Hulst, 1957]. This dimensional relationship results in the deeper penetration into biologic tissues of those longer wavelengths which are not absorbed appreciably by pigments in the tissues. This results in the relative transparency of nonpigmented tissues over the visible and near-IR wavelength ranges.

64.1.2 Absorption in Biologic Tissue

Absorption of UV–IR radiation in matter arises from the wavelength-dependent resonant absorption of radiation by molecular electrons of optically absorbing molecular species [Grossweiner, 1989]. Because of the chemical inhomogeneity of biologic tissues, the degree of absorption of incident radiation strongly depends upon its wavelength. The most prevalent or concentrated UV–IR absorbing molecular species in biologic tissues are listed in Table 64.1 along with associated high-absorbance wavelengths. These species include the peptide bonds; the phenylalanine, tyrosine, and tryptophan residues of proteins, all of which absorb in the UV range; oxy- and deoxyhemoglobin of blood which absorb in the visible to near-IR range; melanin, which absorbs throughout the UV to near-IR range, which decreasing absorption occurring with increasing wavelength; and water, which absorbs maximally in the mid-IR range [Hale and Querry, 1973; Miller and Veitch, 1993; White et al., 1968]. Biomedical lasers and their emitted radiation wavelength values also are tabulated also in Table 64.1. The correlation between the wavelengths of clinically useful lasers and wavelength regions of absorption by constituents of biological tissues is evident. Additionally, exogenous light-absorbing chemical species may be intentionally present in tissues. These include:

1. Photosensitizers, such as porphyrins, which upon excitation with UV-visible light initiate photochemical reactions which are cytotoxic to the cells of the tissue, for example, a cancer which concentrates the photosensitizer relative to surrounding tissues [Spikes, 1989]

TABLE 64.1 UV–IR-Radiation-Absorbing Constituent of Biological Tissues and Biomedical Laser Wavelengths

		Optical absorption			
Constituent	Tissue type	Wavelength[a] (nm)	Relative[b] strength	Laser type	Wavelength (nm)
Proteins	All				
Peptide bond		<220 (r)	+++++++	ArF	193
Amino acid					
Residues					
Tryptophan		220–290 (r)	+		
Tyrosine		220–290 (r)	+		
Phenylalanine		220–2650 (r)	+		
Pigments					
Oxyhemoglobin	Blood	414 (p).	+++	Ar ion	488–514.5
	vascular tissues	537 (p).	++	frequency	532
		575 (p).	++	doubled	
		970 (p).	+	Nd:YAG	
		(690–1100) (r)		Diode	810
				Nd:YAG	1064
Deoxyhemoglobin	Blood	431 (p)	+++	Dye	400–700
	vascular tissues	554 (p)	++	Nd:YAG	1064
Melanin	Skin	220–1000 (r)	++++	Ruby	693
Water	All	2.1 (p)	+++	Ho:YAG	2100
		3.02 (p)	+++++++	Er:YAG	2940
		>2.94 (r)	++++	CO_2	10,640

[a] (p): Peak absorption wavelength; (r): wavelength range.

[b] The number of + signs qualitatively ranks the magnitude of the optical absorbtion.

2. Dyes such as indocyanine green which, when dispersed in a concentrate fibrin protein gel can be used to localize 810 nm *GaAlAs* diode laser radiation and the associated heating to achieve localized thermal denaturation and bonding of collagen to effect joining or welding of tissue [Bass et al., 1992; Oz et al., 1989]
3. Tattoo pigments including graphite (black) and black, blue, green, and red organic dyes [Fitzpatrick, 1994; McGillis et al., 1994]

64.2 Penetration and Effects of UV–IR Laser Radiation into Biologic Tissue

Both scattering and absorption processes affect the variations of the intensity of radiation with propagation into tissues. In the absence of scattering, absorption results in an exponential decrease of radiation intensity described simply by Beers law [Grossweiner, 1989]. With appreciable scattering present, the decrease in incident intensity from the surface is no longer monotonic. A maximum in local internal intensity is found to be present due to efficient back-scattering, which adds to the intensity of the incoming beam as shown, for example, by Miller and Veitch [1993] for visible light penetrating into the skin and by Rastegar et al. [1992] for 1.064 μm *Nd:YAG* laser radiation penetrating into the prostate gland. Thus, the relative contributions of absorption and scattering of incident laser radiation will stipulate the depth in a tissue at which the resulting tissue effects will be present. Since the absorbed energy can be released in a number of different ways including thermal vibrations, fluorescence, and resonant electronic energy transfer according to the identity of the absorber, the effects on tissue are in general different. Energy release from both hemoglobin and melanin pigments and from water is by molecular vibrations resulting in a local temperature rise. Sufficient continued energy absorption and release can result in local temperature increases which, as energy input increases, result in protein denaturation (41 to 65°C), water evaporation and boiling (up to ≃300°C under confining pressure of tissue), thermolysis of proteins, generation of gaseous decomposition products and of carbonaceous residue or char (≥300°C). The generation of residual char is minimized by sufficiently rapid energy input to support rapid gasification reactions. The clinical effect of this chain of thermal events is tissue ablation. Much smaller values of energy input result in coagulation of tissues due to protein denaturation.

Energy release from excited exogenous photosensitizing dyes is via formation of free-radical species or energy exchange with itinerant dissolved molecular oxygen [Spikes, 1989]. Subsequent chemical reactions following free-radical formation or formation of an activated or more reactive form of molecular oxygen following energy exchange can be toxic to cells with takeup of the photosensitizer.

Energy release following absorption of **visible (VIS) radiation** by fluorescent molecular species, either endogenous to tissue or exogenous, is predominantly by emission of longer wavelength radiation [Lakowicz, 1983]. Endogenous fluorescent species include tryptophan, tyrosine, phenylalanine, flavins, and metal-free porphyrins. Comparison of measured values of the intensity of fluorescence emission from hyperplastic (transformed precancerous) cervical cells to cancerous cervical cells with normal cervical epithelial cells shows a strong potential for diagnostic use in the automated diagnosis and staging of cervical cancer [Mahadevan et al., 1993].

64.3 Effects of Mid-IR Laser Radiation

Because of the very large absorption by water of radiation with wavelength in the IR range ≥2.0 μm, the radiation of *Ho:YAG, Er: YAG, and CO$_2$* lasers is absorbed within a very short distance of the tissue surface, and scattering is essentially unimportant. Using published values of the water absorption coefficient [Hale and Querry, 1973] and assuming an 80% water content and that the decrease in intensity is exponential with distance, the depth in the "average" soft tissue at which the intensity has decreased to 10% of the incident value (the optical penetration depth) is estimated to be 619, 13, and 170 μm, respectively, for Ho:YAG, Er:YAG, and CO$_2$ laser radiation. Thus, the absorption of radiation from these laser sources and thermalization of this energy results essentially in the formation of a surface heat source. With sufficient energy input, tissue ablation through water boiling and tissue thermolysis occur at the surface. Penetration of heat to underlying tissues is by diffusion alone; thus, the depth of coagulation of tissue below the surface region of ablation is limited by competition between thermal diffusion and the rate of descent of the heated surface impacted by laser radiation during ablation of tissue. Because of this competition, coagulation depths obtained in soft biologic tissues with use of mid-IR laser radiation are typically ≤205 to 500 μm, and the ability to achieve sealing of blood vessels leading to hemostatic ("bloodless") surgery is limited [Judy et al., 1992; Schroder et al., 1987].

64.4 Effects of Near-IR Laser Radiation

The 810-nm and 1064-μm radiation, respectively, of the GaAlAs diode laser and Nd:YAG laser penetrate more deeply into biologic tissues than the radiation of longer-wavelength IR lasers. Thus, the resulting thermal effects arise from absorption at greater depth within tissues, and the depths of coagulation and degree of hemostasis achieved with these lasers tend to be greater than with the longer-wavelength IR lasers. For example, the optical penetration depths (10% incident intensity) for 810-nm and 1.024-μm radiation are estimated to be 4.6 and ≃8.6 mm respectively in canine prostate tissue [Rastegar et al., 1992]. Energy deposition of 3600 J from each laser onto the urethral surface of the canine prostate results in maximum coagulation depths of 8 and 12 mm respectively using diode and Nd:YAG lasers [Motamedi et al., 1993]. Depths of optical penetration and coagulation in porcine liver, a more vascular tissue than prostate gland, of 2.8 and ≃9.6 mm, respectively, were obtained with a Nd:YAG laser beam, and of 7 and 12 mm respectively with an 810-nm diode laser beam [Rastegar et al., 1992]. The smaller penetration depth obtained with 810-nm diode radiation in liver than in prostate gland reflects the effect of greater vascularity (blood content) on near-IR propagation.

64.5 Effects of Visible-Range Laser Radiation

Blood and vascular tissues very efficiently absorb radiation in the visible wavelength range due to the strong absorption of hemoglobin. This absorption underlies, for example, the use of:

1. The argon ion laser (488 to 514.5 nm) in the localized heating and thermal coagulation of the vascular choroid layer and adjacent retina, resulting in the anchoring of the retina in treatment of retinal detachment [Katoh and Peyman, 1988].
2. The argon ion laser (488 to 514.5 nm), frequency-doubled Nd:YAG laser (532 nm), and dye laser radiation (585 nm) in the coagulative treatment of cutaneous vascular lesions such as port wine stains [Mordon et al., 1993].
3. The argon ion (488 to 514.5 nm) and frequency-doubled Nd:YAG lasers (532 nm) in the ablation of pelvic endometrial lesions which contain brown iron-containing pigments Keye et al., 1983].

Because of the large absorption by hemoglobin and iron-containing pigments, the incident laser radiation is essentially absorbed at the surface of the blood vessel or lesion, and the resulting thermal effects are essentially local [Miller and Veitch, 1993].

64.6 Effects of UV Laser Radiation

Whereas exposure of tissue to IR and visible-light-range laser energy result in removal of tissue by thermal ablation, exposure to *argon fluoride (ArF)* laser radiation of 193-nm wavelength results predominantly in ablation of tissue initiated by a photochemical process [Garrison and Srinivasan, 1985]. This ablation arises from repulsive forces between like-charged regions of ionized protein molecules that result from ejection of molecular electrons following UV photon absorption [Garrison and Srinivasan, 1985]. Because the ionization and repulsive processes are extremely efficient, little of the incident laser energy escapes as thermal vibrational energy, and the extent of thermal coagulation damage adjacent to the site of incidence is very limited [Garrison and Srinivasan, 1985]. This feature and the ability to tune very finely the fluence emitted by the ArF laser so that micrometer depths of tissue can be removed have led to ongoing clinical trials to investigate the efficiency of the use of the ArF laser to selectively remove tissue from the surface of the human cornea for correction of short-sighted vision to eliminate the need for corrective eyewear [Van Saarloos and Constable, 1993].

64.7 Effects of Continuous and Pulsed IR–Visible Laser Radiation and Associated Temperature Rise

Heating following absorption of IR-visible laser radiation arises from molecular vibration during loss of the excitation energy and initially is manifested locally within the exposed region of tissue. If incidence of the laser energy is maintained for a sufficiently long time, the temperature within adjacent regions of biologic tissue increases due to heat diffusion. The mean squared distance $\langle X^2 \rangle$ over which appreciable heat diffusion and temperature rise occur during exposure time t can be described in terms of the thermal diffusion time τ by the equation:

$$\langle X^2 \rangle = \tau t \tag{64.1}$$

where τ is defined as the ratio of the thermal conductivity to the product of the heat capacity and density. For soft biologic tissues τ is approximately 1×10^3 cm^2 sec^{-1} [Meijering et al., 1993]. Thus, with continued energy input, the distance over which thermal diffusion and temperature rise occurs increases. Conversely, with use of pulsed radiation, the distance of heat diffusion can be made very small; for example, with exposure to a 1-μsec pulse, the mean thermal diffusion distance is found to be approximately 0.3 μm, or about 3 to 10% of a biologic cell diameter. If the laser radiation is strongly absorbed and the ablation of tissues is efficient, then little energy diffuses away from the site of incidence, and lateral thermally induced coagulation of tissue can be minimized with pulses of short duration. The effect of limiting lateral thermal damage is desirable in the cutting of cornea [Hibst et al., 1992] and sclera of the eye [Hill et al., 1993], and joint cartilage [Maes and Sherk, 1994], all of which are avascular (or nearly so, with cartilage), and the hemostasis arising from lateral tissue coagulation is not required.

64.8 General Description and Operation of Lasers

Lasers emit a beam of intense electromagnetic radiation that is essentially monochromatic or contains at most a few nearly monochromatic wavelengths and is typically only weakly divergent and easily focused into external optical systems. These attributes of laser radiation depend on the key phenomenon which underlies laser operation, that of light amplification by stimulated emission of radiation, which in turn gives rise to the acronym *LASER*.

In practice, a laser is generally a generator of radiation. The generator is constructed by housing a light-emitting medium within a cavity defined by mirrors which provide feedback of emitted radiation through the medium. With sustained excitation of the ionic or molecular species of the medium to give a large density of excited energy states, the spontaneous and attendant stimulated emission of radiation from these states by photons of identical wavelength (a lossless process), which is amplified by feedback due to photon reflection by the cavity mirrors, leads to the generation of a very large photon density within the cavity. With one cavity mirror being partially transmissive, say 0.1 to 1%, a fraction of the cavity energy is emitted as an intense beam. With suitable selection of a laser medium, cavity geometry, and peak wavelengths of mirror reflection, the beam is also essentially monochromatic and very nearly collimated.

Identity of the lasing molecular species or laser medium fixes the output wavelength of the laser. Laser media range from gases within a tubular cavity, organic dye molecules dissolved in a flowing inert liquid carrier and heat sink, to impurity-doped transparent crystalline rods (solid state lasers) and semiconducting diode junctions [Lengyel, 1971]. The different physical properties of these media in part determine the methods used to excite them into lasing states.

Gas-filled, or gas lasers are typically excited by dc or rf electric current. The current either ionizes and excites the lasing gas, for example, argon, to give the electronically excited and lasing Ar+ ion, or ionizes a gaseous species in a mixture also containing the lasing species, for example, N_2, which by efficient energy transfer excites the lasing molecular vibrational states of the CO_2 molecule.

Dye lasers and so-called solid-state lasers are typically excited by intense light from either another laser or from a flash lamp. The excitation light wavelength range is selected to ensure efficient excitation at the absorption wavelength of the lasing species. Both excitation and output can be continuous, or the use of a pulsed flashlamp or pulsed exciting laser to pump a solid-state or dye laser gives pulsed output with high peak power and short pulse duration of 1 μsec to 1 msec. Repeated excitation gives a train of pulses. Additionally, pulses of higher peak power and shorter duration of approximately 10 nsec can be obtained from solid lasers by intracavity Q-switching [Lengyel, 1971]. In this method, the density of excited states is transiently greatly increased by impeding the path between the totally reflecting and partially transmitting mirror of the cavity interrupting the stimulated emission process. Upon rapid removal of the impeding device (a beam-interrupting or -deflecting device), stimulated emission of the very large population of excited lasing states leads to emission of an intense laser pulse. The process can give single pulses or can be repeated to give a pulse train with repetition frequencies typically ranging from 1 Hz to 1 kHz.

Gallium aluminum (GaAlAs) lasers are, as are all semiconducting diode lasers, excited by electrical current which creates excited hole-electron pairs in the vicinity of the diode junction. Those carrier pairs are the lasing species which emit spontaneously and with photon stimulation. The beam emerges parallel to the function with the plane of the function forming the cavity and thin-layer surface mirrors providing reflection. Use of continuous or pulsed excitation current results in continuous or pulsed output.

64.9 Biomedical Laser Beam Delivery Systems

Beam delivery systems for biomedical lasers guide the laser beam from the output mirror to the site of action on tissue. Beam powers of up to 100 W are transmitted routinely. All biomedical lasers incorporate a coaxial aiming beam, typically from a HeNe laser (632.8 nm) to illuminate the site of incidence on tissue.

Usually, the systems incorporate two different beam-guiding methods, either (1) a flexible fused silica (SiO_2) optical fiber or light guide, generally available currently for laser beam wavelengths between $\simeq$400 nm and $\simeq$2.1 μm, where $\mathrm{SiO_2}$ is essentially transparent and (2) an articulated arm having beam-guiding mirrors for wavelengths greater than circa 2.1 μm (e.g., $\mathrm{CO_2}$ lasers), for the Er:YAG and for pulsed lasers having peak power outputs capable of causing damage to optical fiber surfaces due to ionization by the intense electric field (e.g., pulsed ruby). The arm comprises straight tubular sections articulated together with high-quality power-handling dielectric mirrors at each articulation junction to guide the beam through each of the sections. Fused silica optical fibers usually are limited to a length of 1 to 3 m and to wavelengths in the visible-to-low midrange IR ($<$2.1 μm), because longer wavelengths of IR radiation are absorbed by water impurities ($<$2.9 μm) and by the $\mathrm{SiO_2}$ lattice itself (wavelengths $>$5 μm), as described by Levi [1980].

Since the flexibility, small diameter, and small mechanical inertia of optical fibers allow their use in either flexible or rigid endoscopes and offer significantly less inertia to hand movement, fibers for use at longer IR wavelengths are desired by clinicians. Currently, researchers are evaluating optical fiber materials transparent to longer IR wavelengths. Material systems showing promise are fused $\mathrm{Al_2O_3}$ fibers in short lengths for use with near-3-μm radiation of the Er:YAG laser and *Ag halide* fibers in short lengths for use with the $\mathrm{CO_2}$ laser emitting at 10.6 μm [Merberg, 1993]. A flexible hollow Teflon waveguide 1.6 mm in diameter having a thin metal film overlain by a dielectric layer has been reported recently to transmit 10.6 μm $\mathrm{CO_2}$ radiation with attenuation of 1.3 and 1.65 dB/m for straight and bent (5-mm radius, 90-degree bend) sections, respectively [Gannot et al., 1994].

64.9.1 Optical Fiber Transmission Characteristics

Guiding of the emitted laser beam along the optical fiber, typically of uniform circular cross-section, is due to total internal reflection of the radiation at the interface between the wall of the optical fiber core and the cladding material having refractive index n_1 less than that of the core n_2 [Levi, 1980]. Total internal reflection occurs for any angle of incidence θ of the propagating beam with the wall of the fiber core such that $\theta > \theta_c$ where

$$\sin\theta_c = \left(\frac{n_1}{n_2}\right) \tag{64.2}$$

or in terms of the complementary angle α_c

$$\cos\alpha_c = \left(\frac{n_1}{n_2}\right) \tag{64.3}$$

For a focused input beam with apical angle α_m incident upon the flat face of the fiber as shown in Figure 64.1, total internal reflection and beam guidance within the fiber core will occur [Levi, 1980] for

$$\mathrm{NA} = \sin(\alpha_m/2) \le \left[n_2^2 - n_1^2\right]^{0.5} \tag{64.4}$$

where NA is the numerical aperture of the fiber.

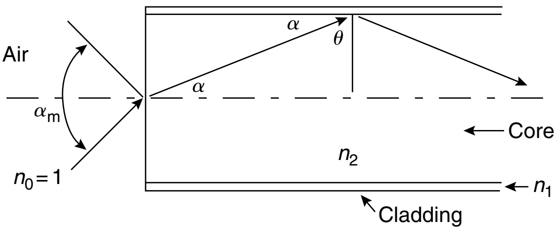


FIGURE 64.1 Critical reflection and propagation within an optical fiber.

This relationship ensures that the critical angle of incidence of the interface is not exceeded and that total internal reflection occurs [Levi, 1980]. Typical values of NA for fused SiO_2 fibers with polymer cladding are in the range of 0.36 to 0.40. The typical values of $\alpha_m = 14$ degrees used to insert the beam of the biomedical laser into the fiber is much smaller than those values ($\simeq$21 to 23 degrees) corresponding to typical NA values. The maximum value of the propagation angle α typically used in biomedical laser systems is $\simeq$4.8 degrees.

Leakage of radiation at the core–cladding interface of the fused SiO_2 fiber is negligible, typically being 0.3 dB/m at 400 nm and 0.01 dB/m at 1.064 μm. Bends along the fiber length always decrease the angle of the incidence at the core–cladding interface. Bends do not give appreciable losses for values of the bending radius sufficiently large that the angle of incidence θ of the propagating beam in the bent core does not becomes less than θ_c at the core–cladding interface [Levi, 1980]. The relationship given by Levi [1980] between the bending radius r_b, the fiber core radius r_o, the ratio (n_2/n_1) of fiber core to cladding refractive indices, and the propagation angle α in Figure 64.1 which ensures that the beam does not escape is

$$\frac{n_1}{n_2} > \frac{1-\rho}{1+\rho}\cos\alpha \tag{64.5}$$

where $\rho = (r_o/r_b)$. The inequality will hold for all $\alpha \leq \alpha_c$ provided that

$$\frac{n_1}{n_2} \leq \frac{1-\rho}{1+\rho} \tag{64.6}$$

Thus, the critical bending radius r_{bc} is the value of r_b such that Equation 64.6 is an equality. Use of Equation 64.6 predicts that bends with radii $\geq$12, 18, and 30 mm, respectively, will not result in appreciable beam leakage from fibers having 400-, 600-, and 1000-μm diameter cores, which are typical in biomedical use. Thus, use of fibers in flexible endoscopes usually does not compromise beam guidance.

Because the integrity of the core–cladding interface is critical to beam guiding, the clad fiber is encased typically in a tough but flexible protective fluoropolymer buffer coat.

64.9.2 Mirrored Articulated Arm Characteristics

Typically two or three relatively long tubular sections or arms of 50 to 80 cm length make up the portion of the articulated arm that extends from the laser output fixturing to the handpiece, endoscope, or operating microscope stage used to position the laser beam onto the tissue proper. Mirrors placed at the articulation of the arms and within the articulated handpiece, laparoscope, or operating microscope stage maintain the centration of the trajectory of the laser beam along the length of the delivery system. Dielectric multilayer mirrors [Levi, 1980] are routinely used in articulated devices. Their low high reflectivity $\leq$99.9 +% and power-handling capabilities ensure efficient power transmission down the arm. Mirrors in articulated devices typically are held in kinetically adjustable mounts for rapid stable alignment to maintain beam concentration.

64.9.3 Optics for Beam Shaping on Tissues

Since the rate of heating on tissue, and hence rates of ablation and coagulation, depends directly on energy input per unit volume of tissue, selection of ablation and coagulation rates of various tissues is achieved through control of the energy density (J/cm^2 or W sec/cm^2) of the laser beam. This parameter is readily achieved through use of optical elements such as discrete focusing lenses placed in the handpiece or rigid endoscope which control the spot size upon the tissue surface or by affixing a so-called contact tip to the end of an optical fiber. These are conical or spherical in shape with diameters ranging from 300 to 1200 μm and with very short focal lengths. The tip is placed in contact with the tissue and generates a submillimeter-sized focal spot in tissue very near the interface between the tip and tissue. One advantage of using the contact tip over a focused beam is that ablation proceeds with small lateral depth of attendant coagulation [Judy et al., 1993a]. This is because the energy of the tightly focused beam causes tissue thermolysis essentially at the tip surface and because the resulting tissue products strongly absorb the beam resulting in energy deposition and ablation essentially at the tip surface. This contrasts with the radiation penetrating deeply into tissue before thermolysis which occurs with a less tightly focused beam from a free lens or fiber. An additional advantage with the use of contact tips in the perception of the surgeon is that the kinesthetics of moving a contact tip along a resisting tissue surface more closely mimics the "touch" encountered in moving a scalpel across the tissue surface.

Recently a class of optical fiber tips has been developed which laterally directs the beam energy from a silica fiber [Judy et al., 1993b]. These tips, either a gold reflective micromirror or an angled refractive prism, offer a lateral angle of deviation ranging from 35 to 105 degrees from the optical fiber axis (undeviated beam direction). The beam reflected from a plane micromirror is unfocused and circular in cross-section, whereas the beam from a concave mirror and refractive devices is typically elliptical in shape, fused with distal diverging rays. Fibers with these terminations are currently finding rapidly expanding, large-scale application in coagulation (with 1.064-μm Nd:YAG laser radiation) of excess tissue lining the urethra in treatment of benign prostatic hypertrophy [Costello et al., 1992]. The capability for lateral beam direction may offer additional utility of these terminated fibers in other clinical specialties.

64.9.4 Features of Routinely Used Biomedical Lasers

Currently four lasers are in routine large-scale clinical biomedical use to ablate, dissect, and to coagulate soft tissue. Two, the carbon dioxide (CO_2) and argon ion (Ar-ion) lasers, are gas-filled lasers. The other two employ solid-state lasing media. One is the Neodymium–yttrium–aluminum–garnet (Nd:YAG) laser, commonly referred to as a solid-state laser, and the other is the gallium–aluminum arsenide (GaAlAs) semiconductor diode laser. Salient features of the operating characteristics and biomedical applications of those lasers are listed in Table 64.2 to Table 64.5. The operational descriptions are typical of the lasers currently available commercially and do not represent the product of any single manufacturer.

64.9.5 Other Biomedical Lasers

Some important biomedical lasers have smaller-scale use or currently are being researched for biomedical application. The following four lasers have more limited scales of surgical use:

The Ho:YAG (Holmium:YAG) laser, emitting pulses of 2.1 μm wavelength and up to 4 J in energy, used in soft tissue ablation in arthroscopic (joint) surgery (FDA approved).

The Q-switched Ruby ($Cr{:}Al_20_3$) laser, emitting pulses of 694-nm wavelength and up to 2 J in energy is used in dermatology to disperse black, blue, and green tattoo pigments and melanin in pigmented lesions (not melanoma) for subsequent removal by phagocytosis by macrophages (FDA approved).

The flashlamp pumped pulsed dye laser emitting 1- to 2-J pulses at either 577- or 585-nm wavelength (near the 537–577 absorption region of blood) is used for treatment of cutaneous vascular lesions and melanin pigmented lesions except melanoma. Use of pulsed radiation helps to localize the thermal damage to within the lesions to obtain low damage of adjacent tissue.

The following lasers are being investigated for clinical uses.

1. The Er:YAG laser, emitting at 2.94 μm near the major water absorption peak (OH stretch), is currently being investigated for ablation of tooth enamel and dentin (Li et al., 1992)
2. Dye lasers emitting at 630 to 690 nm are being investigated for application as light sources for exciting dihematoporphyrin ether or benzoporphyrin derivatives in investigation of the efficacy of these photosensitives in the treatment of esophageal, bronchial, and bladder carcinomas for the FDA approved process

TABLE 64.2 Operating Characteristics of Principal Biomedical Lasers

Characteristics	Ar ion laser	CO_2 laser
Cavity medium	Argon gas, 133 Pa	l0% CO_2 10% Ne, 80% He; 1330 Pa
Lasing species	Ar+ ion	CO_2 molecule
Excitation	Electric discharge, continuous	Electric discharge, continuous, pulsed
Electric input	208 V_{AC}, 60 A	110 V_{AC}, 15 A
Wall plug efficiency	$\simeq$0.06%	$\simeq$10%
characteristics	Nd:YAG laser	GaAlAs diode laser
Cavity medium	Nd-dopted YAG	n-p junction, GaAlAs diode
Lasing species	Nd3t in YAG lattice	Hole-electron pairs at diode junction
Excitation	Flashlamp, continuous, pulsed	Electric current, continuous pulsed
Electric input	208/240 V_{AC}, 30 A continuous 110 V_{AC}, 10 A pulsed	110 V_{AC}, 15A
Wall plug efficiency	$\simeq$1%	$\simeq$ 23%

TABLE 64.3 Output Beam Characteristics of Ar-Ion and CO_2 Biomedical Lasers

Output characteristics	Argon laser	CO_2 laser
Output power	2–8 W, continuous	1–100 W, continuous
Wavelength (sec)	Multiple lines (454.6–528.7 nm), 488, 514.5 dominant	10.6 μm
Electromagnetic wave propagation mode	TEM_{∞}	TEM_{∞}
Beam guidance, shaping	Fused silica optical fiber with contact tip or flat-ended for beam emission, lensed handpiece. Slit lamp with ocular lens	Flexible articulated arm with mirrors; lensed handpiece or mirrored microscope platen

TABLE 64.4 Output Beam Characteristics of Nd:YAG and GaAlAs Diode Biomedical Lasers

Output characteristics	Nd:YAG lasers	GaAlAs diode laser
Output power	1–100 W continuous at 1.064 millimicron 1–36 W continuous at 532 nm (frequency doubled with KTP)	1–25 W continuous
Wavelength(sec)	1.064 μm/532 nm	810 nm
Electromagnetic wave propagation modes	Mixed modes	Mixed modes
Beam guidance and shaping	Fused SiO_2 optical fiber with contact tip directing mirrored or refracture tip	Fused SiO_2 optical fiber with contact tip or laterally directing mirrored or refracture tip

TABLE 64.5 Clinical Uses of Principal Biomedical Lasers

Ar-ion laser Pigmented (vascular) soft-tissue ablation in gynecology; general and oral sugery; otolaryngology; vascular lesion coagulation in dermatology; retinal coagulation in ophthalmology	*CO_2 laser* Soft-tissue ablation — dissection and bulk tissue removal in dermatology; gynecology; general, oral, plastic, and neurosurgery; otolaryngology; podiatry; urology
Nd:YAG laser Soft-tissue, particularly pigmented vascular tissue, ablation — dissection and bulk tissue removal — in dermatology; gastroenterology; gynecology; general, arthroscopic, neuro-plastic, and thoracic surgery; urology; posterior capsulotomy (ophthalmology) with pulsed 1.064 millimicron and ocular lens	*GaAlAs diode laser* Pigmented (vascular) soft-tissue ablation — dissection and bulk removal in gynecology; gastroenterology, general, surgery, and urology; FDA approval for otolaryngology and thoracic surgery pending

Defining Terms

Biomedical Laser Radiation Ranges

Infrared (IR) radiation: The portion of the electromagnetic spectrum within the wavelength range 760 nm–1 mm, with the regions 760 nm–1.400 μm and 1.400–10.00 μm, respectively, called the near- and mid-IR regions.

Ultraviolet (UV) radiation: The portion of the electromagnetic spectrum within the wavelength range 100–400 nm.

Visible (VIS) radiation: The portion of the electromagnetic spectrum within the wavelength range 400–760 nm.

Laser Medium Nomenclature

Argon fluoride (ArF): Argon fluoride eximer laser (an eximer is a diatomic molecule which can exist only in an excited state).

Ar ion: Argon ion.

CO_2: Carbon dioxide.

Cr:Al_20_3: Ruby laser.

Er:YAG: Erbium yttrium aluminum garnet.

GaAlAs: Gallium aluminum laser.

HeNe: Helium neon laser.

Ho:YAG: Holmium yttrium aluminum garnet.

Nd:YAG: Neodymium yttrium aluminum garnet.

Optical Fiber Nomenclature

Ag halide: Silver halide, halide ion, typically bromine (Br) and chlorine (Cl).

Fused silica: Fused SiO_2.

References

Bass L.S., Moazami N., Pocsidio J. et al. (1992). Change in type I collagen following laser welding. *Lasers Surg. Med.* 12: 500.

Costello A.J., Johnson D.E., and Bolton D.M. (1992). Nd:YAG laser ablation of the prostate as a treatment for benign prostate hypertrophy. *Lasers Surg. Med.* 12: 121.

Fitzpatrick R.E. (1993). Comparison of the Q-switched ruby, Nd:YAG, and alexandrite lasers in tattoo removal. *Lasers Surg. Med.* (Suppl.) 6: 52.

Gannot I., Dror J., Calderon S. et al. (1994). Flexible waveguides for IR laser radiation and surgery applications. *Lasers Surg. Med.* 14: 184.

Garrison B.J. and Srinivasan R. (1985). Laser ablation of organic polymers: microscopic models for photochemical and thermal processes. *J. Appl. Physiol.* 58: 2909.

Grossweiner L.I. (1989). Photophysics. In K.C. Smith (Ed.), *The Science of Photobiology*, pp. 1–47. New York, Plenum.

Hale G.M. and Querry M.R. (1973). Optical constants of water in the 200 nm to 200 μm wavelength region. *Appl. Opt.* 12: 555.

Hibst R., Bende T., and Schröder D. (1992). Wet corneal ablation by Er:YAG laser radiation. *Lasers Surg. Med.* (Suppl.) 4: 56.

Hill R.A., Le M.T., Yashiro H. et al. (1993). Ab-interno erbium (Er:YAG) laser schlerostomy with iridotomy in dutch cross rabbits. *Lasers Surg. Med.* 13: 559.

Judy M.M., Matthews J.L., Aronoff B.L. et al. (1993a). Soft tissue studies with 805 nm diode laser radiation: thermal effects with contact tips and comparison with effects of 1064 nm. Nd:YAG laser radiation. *Lasers Surg. Med.* 13: 528.

Judy M.M., Matthews J.L., Gardetto W.W. et al. (1993b). Side firing laser-fiber technology for minimally invasive transurethral treatment of benign prostate hyperplasia. *Proc. Soc. Photo-Opt. Instr. Eng. (SPIE)* 1982: 86.

Judy M.M., Matthews J.L., Goodson J.R. et al. (1992). Thermal effects in tissues from simultaneous coaxial CO_2 and Nd:YAG laser beams. *Lasers Surg. Med.* 12: 222.

Katoh N. and Peyman G.A. (1988). Effects of laser wavelengths on experimental retinal detachments and retinal vessels. *Jpn. J. Ophthalmol.* 32: 196.

Keye W.R., Matson G.A., and Dixon J. (1983). The use of the argon laser in treatment of experimental endometriosis. *Fertil. Steril.* 39: 26.

Lakowicz J.R. (1983). *Principles of Fluorescence Spectroscopy.* New York, Plenum.

Lengyel B.A. (1971). *Lasers.* New York, John Wiley.

Levi L. (1980). *Applied Optics*, Vol. 2. New York, John Wiley.

Li Z.Z., Code J.E., and Van de Merve W.P. (1992). Er:YAG laser ablation of enamel and dentin of human teeth: determination of ablation rates at various fluences and pulse repetition rates. *Lasers Surg. Med.* 12: 625.

Maes K.E. and Sherk H.H. (1994). Bone and meniscal ablation using the erbium YAG laser. *Lasers Surg. Med.* (Suppl.) 6: 31.

Mahadevan A., Mitchel M.F., Silva E. et al. (1993). Study of the fluorescence properties of normal and neoplastic human cervical tissue. *Lasers Surg. Med.* 13: 647.

McGillis S.T., Bailin P.L., Fitzpatrick R.E. et al. (1994). Successful treatments of blue, green, brown and reddish-brown tattoos with the Q-switched alexandrite laser. *Laser Surg. Med.* (Suppl.) 6: 52.

Meijering L.J.T., VanGermert M.J.C., Gijsbers G.H.M. et al. (1993). Limits of radial time constants to approximate thermal response of tissue. *Lasers Surg. Med.* 13: 685.

Merberg G.N. (1993). Current status of infrared fiberoptics for medical laser power delivery. *Lasers Surg. Med.* 13: 572.

Miller I.D. and Veitch A.R. (1993). Optical modeling of light distributions in skin tissue following laser irradiation. *Lasers Surg. Med.* 13: 565.

Mordon S., Beacco C., Rotteleur G. et al. (1993). Relation between skin surface temperature and minimal blanching during argon, Nd:YAG 532, and cw dye 585 laser therapy of port-wine stains. *Lasers Surg. Med.* 13: 124.

Motamedi M., Torres J.H., Cammack T. et al. (1993). Thermodynamics of cw laser interaction with prostatic tissue: effects of simultaneous cooling on lesion size. *Lasers Surg. Med.* (Suppl.) 5: 64.

Oz M.C., Chuck R.S., Johnson J.P. et al. (1989). Indocyanine green dye-enhanced welding with a diode laser. *Surg. Forum* 40: 316.

Rastegar S., Jacques S.C., Motamedi M. et al. (1992). Theoretical analysis of high-power diode laser (810 nm) and Nd:YAG laser (1064 nm) for coagulation of tissue: predictions for prostate coagulation. *Proc. Soc. Photo-Opt. Instr. Eng. (SPIE)* 1646: 150.

Schroder T., Brackett K., and Joffe S. (1987). An experimental study of effects of electrocautery and various lasers on gastrointestinal tissue. *Surgery* 101: 691.

Spikes J.D. (1989). Photosensitization. In K.C. Smith (Ed.), *The Science of Photobiology*, 2nd ed., pp. 79–110. New York, Plenum.

Van de Hulst H.C. (1957). *Light Scattering by Small Particles.* New York, John Wiley.

Van Saarloos P.P. and Constable I.J. (1993). Improved eximer laser photorefractive keratectomy system. *Lasers Surg. Med.* 13: 189.

White A., Handler P., and Smith E.L. (1968). *Principles of Biochemistry*, 4th ed. New York, McGraw-Hill.

Further Information

Current research on the optical, thermal, and photochemical interactions of radiation and their effect on biologic tissues, are published routinely in the journals: *Laser in Medicine and Surgery, Lasers in the Life Sciences,* and *Photochemistry Photobiology* and to a lesser extent in *Applied Optics and Optical Engineering.*

Clinical evaluations of biomedical laser applications appear in *Lasers and Medicine and Surgery* and in journals devoted to clinical specialties such as *Journal of General Surgery, Journal of Urology, Journal of Gastroenterological Surgery.*

The annual symposium proceedings of the biomedical section of the Society of Photo-Optical Instrumentation Engineers (SPIE) contain descriptions of new and current research on application of lasers and optics in biomedicine.

The book *Lasers* (a second edition by Bela A. Lengyel), although published in 1971, remains a valuable resource on the fundamental physics of lasers—gas, dye solid-state, and semiconducting diode. A more recent book, *The Laser Guidebook* by Jeffrey Hecht, published in 1992, emphasizes the technical characteristics of the gas, diode, solid-state, and semiconducting diode lasers.

The *Journal of Applied Physics, Physical Review Letters,* and *Applied Physics Letters* carry descriptions of the newest advances and experimental phenomena in lasers and optics.

The book *Safety with Lasers and Other Optical Sources* by David Sliney and Myron Wolbarsht, published in 1980, remains a very valuable resource on matters of safety in laser use.

Laser safety standards for the United States are given for all laser uses and types in the American National Standard (ANSI) Z136.1-1993, Safe Use of Lasers.

65

Instrumentation for Cell Mechanics

Nathan J. Sniadecki
Christopher S. Chen
University of Pennsylvania

65.1 Background

Mechanical forces are essential to life at the microscale — from tethering at the junctions between cells that compose a tissue to externally applied loads arising in the cellular environment. Consider the perturbations from acoustic sounds that affect the mechanosensors on auditory hair cells in the inner ear, the contractile forces that a dividing cell imparts on itself in order to split into two daughter cells during cytokinesis, or the bone and muscle loss that occurs from the reduced loads in microgravity [1,2]. Mechanical forces are particularly important in the cardiovascular and musculoskeletal systems [3]. Increased shear stress in the blood flow leads to the dilation and restructuring of blood vessels [4]. The immune response of leukocytes requires that they adhere to and transmigrate through the endothelial barrier of blood vessels [5]. The forces required between leukocytes and endothelium, and between neighboring endothelial cells, in order to execute such complex events have become an important avenue of research [6,7]. Arterial hypertension causes the underlying vessel walls to constrict, preventing local aneurysms and vessel failure [3]. Long-term exposure to such hypertension leads to increased thickening and stiffing of the vessel walls causing vessel stenosis. In the skeletal system, exercise-induced compressive forces increase bone and cartilage mass and strength while subnormal stresses, from bedrest, immobilization, or space travel, results in decreased bone mass [2]. Despite the clear demonstration that mechanical forces are an essential factor in the daily life of many cells and tissues, the underlying question remains to understand how these forces exert their effects.

A key insight to these areas of study has been that nearly all of the adaptive processes are regulated at the cellular level. That is, many of the tissue responses to forces are actually cellular responses. The contraction and hyperproliferation of smooth muscle cells embedded within the arteries in response to hypertension causes the vessel wall constriction and thickening. Changes in bone mass result from both changes in the production of new bone cells and the metabolic capacity of the existing bone cells. For example, mesenchymal stems cells have been found to differentiate into bone-producing osteoblasts if they are allowed to experience mechanical stresses generated between the individual cells and their local surroundings, but become fat-storing adipocytes when such stresses are eliminated [8]. Thus, an understanding of the importance of forces to medicine and biology must first derive from a better characterization of the forces acting at the single cell level. To begin to explore and characterize the forces experienced and generated by cells (cellular forces), engineers are taking a two-pronged approach. First, they are developing a better understanding of the cell as a mechanical object; and second, they are employing new tools for analyzing cellular forces in the micro and nanoscale.

The measurement of cellular forces is a difficult task because cells are active. That is, they continually change and adapt their mechanical structure in response to their surroundings. The primary mechanical elements in cells are polymers of proteins — in particular, actin, tubulin, and intermediate filament proteins — that are collectively called the cytoskeleton. These cytoskeletal scaffolding structures are continually disassembling and reassembling, realigning, renetworking, contracting, and lengthening. Perhaps one of the most fascinating and simultaneously challenging aspects of characterizing these mechanical rearrangements is that they often occur in direct response to mechanical perturbations. If one pulls on a corner of a cell to measure its material response, it will adjust to the perturbation with reinforcement at the point of applied force. If one places a cell on a soft substrate, it will adjust its shape to achieve a balance between its contractile forces generated by its cytoskeleton and the adhesion forces at its extracellular foundation. The dynamic response of cells to such forces makes the characterization of the natural state of cells difficult. Nonetheless, one of the major goals of mechanobiology is not only to characterize cellular mechanics, but also to identify the mechanism by which cells sense, transduce, and respond to mechanical forces. Whether the mechanosensor itself is an individual protein, a network of structures, or some novel control process remains to be determined. Due to the intimate interaction between the properties of the cell and the techniques used to measure them, we will first provide a brief introduction to the mechanics of the cells themselves, followed by the techniques used to measure the forces that they generate. In addition, since cells are quite small, we will provide a brief discussion of scaling laws and emphasize the technical challenges associated with measuring forces at this length scale.

65.2 Cellular Mechanics

The behavior and function of a cell is dependent to a large degree on the cytoskeleton which consists of three polymer filament systems — microfilaments, intermediate filaments, and microtubules. Acting together as a system, these cytoskeleton filament proteins serve as the scaffolding in the cytoplasm of the cell that (1) supports the delicate cell membrane, (2) tethers and secures organelles in position or guide their transport through the cytoplasm, and (3) in conjunction with various motor proteins, the machinery that provides force necessary for locomotion or protrusion formation [1]. Microfilaments are helical polymers formed from actin that organize into parallel bundles for filopodia extensions and contractile stress fibers or into extensively cross-linked networks at the leading edge of a migrating cell and throughout the cell cortex that supports the cell membrane (Figure 65.1b). Microtubules have tubulin subunits and form long, hollow cylinders that emanate from a single centrosome, which is located near the nucleus (Figure 65.1c). Of the three types of cytoskeletal filaments, microtubules have the higher bending resistance and act to resist compression [9]. Intermediate filaments form extensive networks within the cytoplasm that extend circumferentially from the meshwork structure that surrounds the nucleus (Figure 65.1d). These rope-like filaments are easy to bend but difficult to break. They are particularly predominant in the cytoplasm of cells that are subject to mechanical stress, which highlights their role in tissue-strengthening [10]. Since these three subcellular structures have distinct mechanical properties and varying concentrations between cells, the measured cellular forces and mechanical properties of a particular cell may not be the same as the next cell.

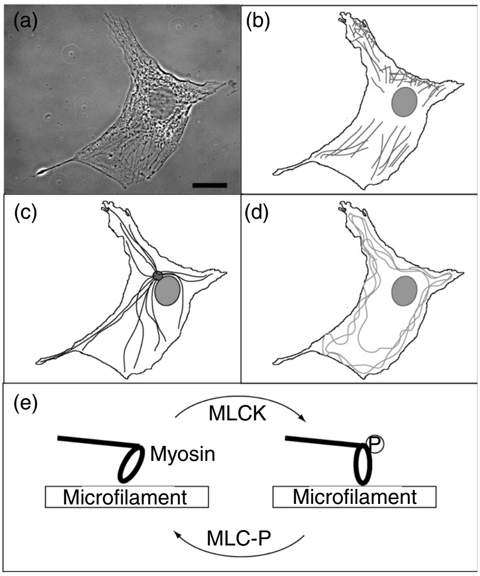


FIGURE 65.1 Cells have cytoskeletal structures that determine their force generation machinery. (a) Phase contract image of 3T3 fibroblast (bar: 10 μm). The diagram shows the general shape and location of (b) microfilaments, (c) microtubules, and (d) intermediate filaments. (e) Schematic of conventional actomyosin motor for cellular forces.

Microfilaments, when coupled with the motor protein myosin, generate the cellular forces that influence the function of the cell and surrounding tissue. Often, skeletal muscle cells come to mind when one thinks of cellular force generation. Cardiac muscle, striated muscle, smooth muscle, and myoepithelial cells, all of which originate from a common precursor myoblast cell, employ a contractile system involving actin as the filament and myosin as the motor protein. Myosin binds to the microfilaments and moves along it in a step-wise linear ratchet mechanism as a means to generate contractile force. Although well studied in muscle, the same actomyosin contractile apparatus found in skeletal muscle cells is present in nearly all cell types. Myosin changes its structure during cycles of phosphorylation, regulated by myosin light chain kinase (MLCK) and myosin light chain phosphatase (MLC-P), to create power strokes that advance the head of the protein along the filament (Figure 65.1e). Each cycle advances the myosin head about 5 nm along the actin filament and produces an average force of 3 to 4 pN [11]. The interaction of myosin and actin is not a constant machine but instead is cycled as dictated by the upstream signaling pathways that regulate MLCK and MLC-P.

A highly researched aspect of cellular forces that highlight their dynamic nature is cell locomotion. These forces, aptly named traction forces, are medically relevant for the metastasis of cancer and occurs when a sessile cell develops the ability to migrate from the tumor and into the bloodstream. Additionally, embryonic development, tissue formation, and wound healing exemplify the physiological relevance of cell migration. The mechanism that generates cell locomotion depends on the coordination of changes in cell shape, via restructuring of the cytoskeletal filaments, and shifting of adhesion sites to the extracellular matrix. When a cell moves forward, it extends a protrusion at the leading edge via microtubule formation and actin polymerization to create new adhesion sites, which are called focal adhesions or focal contacts (Figure 65.2). These nonuniformly distributed, punctate structures link the cytoskeletal filaments and the motor proteins to the substrate and are present at both ends of a migrating cell. After forming new attachments, the cell contracts to move the cell body forward by disassembling contacts at the back end. The locomotion is continued in a treadmill fashion of front protrusion and rear contraction. On account of its dynamic response, a cell does not have uniform mechanical properties and certainly cannot be regarded as a Hookean material with time-independent and linear material properties. As a result, a classical continuum model for the cell has not provided significant insight into the basis for the mechanical properties of a cell.

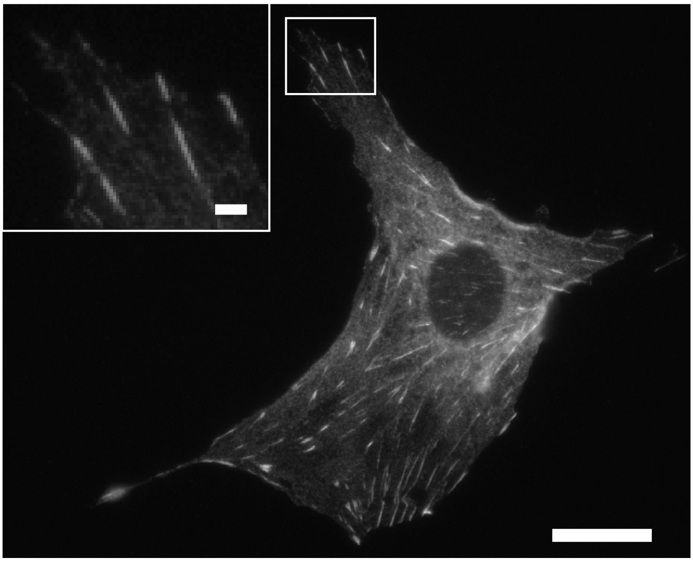


FIGURE 65.2 Immunofluorescence staining of focal adhesions (bar: 10 μm, inset bar: 2 μm).

To understand both the signaling pathways that control these cytoskeletal processes and the resultant mechanics of cell force-generation, researchers have begun to develop methods to measure these forces. To coordinate a global response with other cells of the tissue to a particular stimulus, there exists a complex mechanism of communication between the cells within a tissue. Often cellular signaling, or communication, is mediated by the release of soluble chemicals or growth factors. When these chemicals diffuse across populations of cells, each cell detects the signal via specific receptors and responds accordingly. One of the most challenging aspects of studying cellular mechanics, however, is that cells are also able to detect physical changes even under a constant chemical environment, and respond with changes in cellular function or mechanical state. The sensing and translation of mechanical signals into a functional biochemical signal is known as mechanotransduction. One of the sites in which mechanotransduction occurs is at focal adhesions. A critical engineering challenge in understanding how cells sense and control cellular forces is to characterize the nanonewton forces at these adhesion sites, and correlate these forces with the intercellular signaling response of mechanotransduction.

Based on these insights about the cells and their intracellular structure, it is clear that how one chooses to measure cellular forces can affect the measurements themselves. Since the mechanical properties of a cell are non-Hookean, one can consider the cell to have a history or a memory. Similar to hysteresis of materials, the loadings of forces on a cell have a time-dependent response and so a cell is often modeled as a viscoelastic material — a soft glassy material. As a result, the procedural conditions of a particular experiment must be examined for a time-dependent response. Thus, the observer needs to consider whether to measure the range of forces exerted by one cell to obtain the direct time-response but with a limited data population, or to interrogate a large number of cells with many cellular forces and only report the average value, thereby disregarding the effect that transitions between loading conditions have on a cell. These considerations should act as a simple warning that one must treat each new system, device, and resulting data on cellular mechanics with a healthy degree of guarded skepticism. Despite such concerns, investigators have begun to develop numerous tools that ultimately will address these issues.

65.3 Scaling Laws

In addition to the difficulty in measuring the nanonewton-scale forces that cells exert on the extracellular matrix or at cellular interconnects, the spots where these forces are applied are subcellular (micrometer and nanometer-scale) in dimension. Microscale detection is required in order to measure the different traction forces that are applied at the front vs. the back of a migrating cells. Moreover, since mechanical stresses appear to be sensed and applied at individual focal adhesions (ranging in size from 0.1 to 2 μm^2), it is pertinent to have spatial resolution in the submicron range (Figure 65.2). To obtain the microscale measurement power for studying cellular mechanics, researchers have turned to new techniques and tools. However, in their development, one must consider the impact of scaling laws on the design of the tool. At the micrometer or nanometer scale, the ratio of surface area to volume dramatically differs from the length scale to which we are accustomed. Thus, surface forces dominate over body forces since the former scales with the inverse of the square of the length (L^{-2}) while the later scales with the inverse of the length to the third power (L^{-3}). For example, adhesion forces and fluid shear forces are often more critical to the function of a cell than those of gravity [2,4]. The microscale forces that compose the environment of a cell are difficult to measure with the types of tools that are typically used at the macroscale.

Not only must the designer of these new tools consider the types of forces they want to measure, but also how they will transfer the data to some readable form. Using a microscope to read the measurements from the tool is a noninvasive technique that does not require direct connections to the substrate. However, the technique does require that the substrate be optically transparent, which limits materials available for fabrication, and has optical limitations in resolving below hundreds of nanometers. Despite the limitations, coupling microscopy with a force sensor does provide improved measurement read-out capabilities over other techniques such as electronics, which have sensitivity limitations due to the low signal to noise ratio from thermal and charge fluctuations in the aqueous environment and integration complexity in constructing the electrical components on the same substrate as the force sensors.

In scaling down to the cellular level, the development of the measuring instruments becomes dependent on experimental materials and microfabrication. Typical hard materials used in strain gauges and springs do not bend under nanonewton loadings with the same displacement that is required for measurement sensitivity. On account of this, soft materials are employed in the construction of the microsensors, even though these thin-film materials are not as fully characterized as their bulk material counterparts. Additionally, as the surface to volume ratio increases at the microscale, the effect of the different chemical composition at the surface of the material, such as the native oxide layers of iron, copper, or silicon, may have more dramatic effects on the overall material properties. In building devices with these materials, the microfabrication techniques used must have good reliability for repeatable and uniform measurements on the device. Even though the equipment used in microfabrication is engineered to deposit material with uniform properties and thickness across the device, tolerance issues are still pertinent because of the topological effect that a micrometer defect can have on the environment that the cell senses. Most of the microsensors are made one at a time or in limited batches and consistency in the fabrication methods is critical for repeatable measurements. In conclusion, the devices detailed in the following sections are powerful measurement tools for detecting the nanoscale cellular forces but are still prototypes, in which consistency in their construction is important to corroborating the scientific discoveries that they provide.

65.4 Measurement of Cellular Force

Studying the forces that cells exert on their microenvironment and their corresponding biological mechanisms generally involve culturing cells on flexible substrates that the cells physically deform when applying their contraction or traction forces. When the stiffness of the substrate has been characterized, then optical measurement of the substrate distortion reports the cellular force. A relationship between force and displacement holds whether the substrate is a film of polyacrylamide gel that distorts under the force of a contracting cell adhered to it or silicone microcantilevers that are deflected under the forces of a migrating cell. In the early 1980s, Albert Harris first pioneered the method of measuring cellular forces on thin films of silicone that wrinkled upon the force of the adherent cells and has since evolved into devices that use microfabrication techniques to obtain improved precision of their force sensors.

65.4.1 Membrane Wrinkling

The thin membranes of liquid silicon rubber were cross-linked when exposed briefly to flame so that a thin skin of rubber was cured to ~1 μm thickness on top of the remaining liquid rubber that served as the lubricant layer between the glass coverslip [12,13]. Cells could be cultured on the silicone rubber, which is optically transparent and nontoxic, and as they spread on the skin surface, the adhesion forces they applied to the skin were strong enough to produce wrinkles and fold in the skin (Figure 65.3a). Directly underneath the cell, the wrinkles were circumferential with the cell boundary indicating that compressive forces created the folds in the membrane. At areas surrounding the adherent cell, the wrinkles projected out along radial lines from the cell boundary along the axes of tension forces. No observation of the cell pushing against the silicone membrane has been observed. This technique was a breakthrough in that cellular forces had not been experimentally observed and that qualitative measurement of the different regions of compression and tension could be observed simultaneously.

The membrane wrinkling technique has recently been improved upon with an additional fabrication step to reduce the stiffness of the substrate for increased wrinkles and folds and semi-quantitative

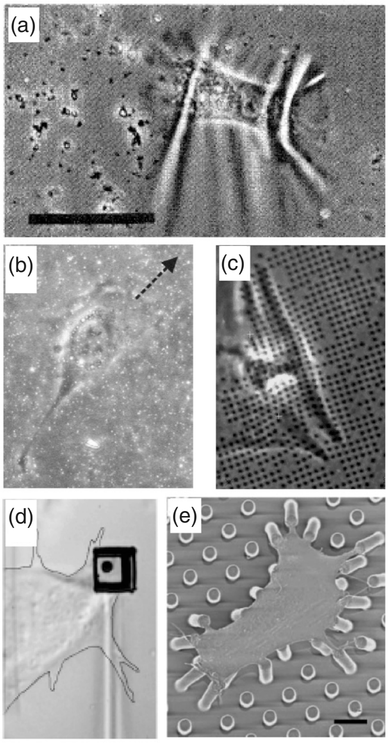


FIGURE 65.3 Techniques for the measurement of cellular forces. (a) Adherent fibroblast cell exerts forces strong enough to wrinkle the underlying silicone rubber membrane (black bar: 50 μm). (Reproduced from Harris, A.K., P. Wild, and D. Stopak, *Science*, 1980, **208**: 177–179.) (b) Traction forces from migrating fibroblast (arrow indicates direction) are measured from the displacement of fluorescent microparticles embedded in a polyacrylamide substrate. (Reproduced from Munevar, S., Y.-L. Wang, and M. Dembo, *Biophys. J.*, 2001, **80**: 1744–1757.) (c) Contracting fibroblast distorts the regular array of micropatterned fluorescent dots. (Reproduced from Balaban, N.Q. et al., *Nat. Cell Biol.*, 2001, **3**: 466–472.) (d) Bending of horizontal microcantilever locally reports the traction force of a subcellular region during fibroblast migration. (Reproduced from Galbraith, C.G. and M.P. Sheetz, *Proc. Natl Acad. Sci., USA*, 1997, **94**: 9114–9118.) (e) Local contraction forces of smooth muscle cell are measured with an array of vertical elastomeric microcantilevers that deflect under cellular forces (black bar: 10 μm). (From Tan, J.L. et al., *Proc. Natl Acad. Sci., USA*, 2003, **100**: 1484–1489. With permission.)

measurement of cellular forces in the hundreds of nanonewtons range. After the flame curing, the membranes were exposed to UV irradiation to weaken the cross-linking of the silicon sheets [14,15]. Applying a known tip-force from a glass pipette on the surface of the membrane and measuring the resultant wrinkles correlated the distortion and force relationship. Specifically, the researchers determined that the length of the wrinkles formed from the pipette was linear with the applied force and called it the "wrinkle stiffness." Using this new technique, they observed that cytokinesis occurs through increased contractility at the equator of the cell, near the cleavage furrow. The traction force drops as the two daughter cells pinch apart. The newly formed cells migrate away from each other resulting in an increase in traction wrinkles until a strong enough force is generated to rupture the intercellular junction. The observed elastic recoil when the daughter cells break their junction causes them to rapidly separate and there is a relaxation in the surrounding wrinkles. Furthermore, the increased number of wrinkles in this technique allows for the measure of the different subcellular forces that occur during live cell migration. At the lamellipodium of a migrating cell, the wrinkles were radial and remain anchored to spots, possibly focal adhesion, as the cell advanced forward. Once these spots were located at the boundary between the lamellipodium and cell body, the wrinkles transitioned to compressive wrinkles. At the rear of the cell, the wrinkle forces diminished slower than the decreasing cell contact area so that the shear stress of the cell increased to pull it forward. The forces that occur at different regions of a cells attachment to the substrate reveal the coordination between pulling forces at the front of the cell and detachment forces that act against the migrating cell.

The membrane wrinkling technique is sensitive to cellular forces and can monitor the force changes at regions of interest within the adhesion area of a cell over time, but it is only a qualitative technique. It does not have adequate subcellular force resolution to measure the applied forces at the focal adhesions. Quantification of the force by means of the "wrinkle stiffness" is not an accurate measurement of forces due to the nonlinear lengthening of the wrinkles when forces are applied at multiple locations on the membrane. Moreover, the chaotic buckling of the membrane has a low repeatability, which makes matching the wrinkle patterns or lengths between experiments inaccurate.

65.4.2 Traction Force Microscopy

To address these issues, traction force microscopy, a technique employing a nonwrinkling elastic substrate, was developed for cell mechanics [16]. The device layout is similar to Harris et al. in that a thin, highly compliant polymer membrane is cured on a glass coverslip on which cells are cultured, except that the membrane is not allowed to wrinkle. In addition to silicone rubber, polyacrylamide membranes have been used as the flexible membrane for cell attachment [17,18]. Instead of wrinkles, fluorescent beads with nanometer diameter were embedded into the material during the fabrication to act as displacement markers (Figure 65.3b). Fixing the sides of the membrane to the edges of the coverslip enables a prestress to be added to the membrane, which suppresses the wrinkling but enables adequate flexibility to allow in-plane traction forces to create visible displacements of the beads. Subtracting the position of the beads under the forces that the cell exerts and the position once the cell was removed from the surface determined the small movements of the beads between the two states, that is, relative displacement field. The corresponding force mapping of the cell is translated from the displacement field, which involves complex mathematical methods requiring the use of a supercomputer. The results provide spatial resolution of $\sim 5\ \mu$m to measure the forces at smaller areas underneath the cell.

In obtaining the corresponding force map, the beads do not move as an ideal spring, in which the displacement is directly proportional to the applied force. Instead, many beads move in response to a single traction force because of the continuous membrane and their movement diminishes as a function of distance from the point of traction. As a result, many force mappings may be possible solutions for the measured displacement field. Appropriate constraints must be applied to the calculation for the solution to converge to a proper solution. Additionally, the displacement beads are discrete markers that are randomly seeded with a nonuniform density, resulting in the lack of displacement information in regions of the cell. To postulate on the magnitude and direction of the forces in the area between beads, a grid meshing approximation is superimposed on the cell area during the force calculation in order to solve for the force and displacement relationship at all regions. In fact, the placement of mesh nodes in these sparse areas leads to an ill-posed problem in solving for the force map because often more force points are introduced than there are displacement data due to the random seeding of beads underneath the cell area. Despite these limitations, the solution for the membrane displacement is well addressed in linear elastic theory [19]. The membrane can be regarded as a semi-infinite space of an incompressible, elastic material with tangential forces applied only at the boundary plane. Under these assumptions, the displacement field, $d(\mathbf{m})$, and the stress field, $T(r)$ are related by an integral relation:

$$d_i(\mathbf{m}) = \iint G_{ij}(\mathbf{m} - \mathbf{r}) T_j(\mathbf{r})\, \mathrm{d}\mathbf{r} \tag{65.1}$$

where $i, j \leq 2$ for the two-dimensional half-space and $G_{ij}(\mathbf{m} - \mathbf{r})$ is Green's function that relates the displacement at position **m** resulting from the point force at position **r**. Obtaining the stress field requires inverting Equation 65.1, which is not always a unique solution because often there are not enough beads to determine all the force points. To address this problem, regularization schemes are used to apply additional criteria in selecting the solution of the inversion operation. These criteria include incorporating the constraint that the sum of all of the traction forces must balance, that the forces are only applied at the limited points of focal adhesions, and that the least complex solution be used.

In contrast to the random seeding of beads, microfabricated regular arrays of fluorescent beads have been imprinted onto the elastomeric substrate for improved force tracking [23]. The deformation of the marker pattern on the substrate is readily observed under the microscopy during the recording of a cell's forces (Figure 65.3c). The patterns are formed with Si and GaAs molds to create sub-micron spot diameters with 2 to 30 μm spacing. The calculation for the force mapping is similar to the random seeding but with significant reduction in the number of possible solutions due to the uniform density of displacement markers throughout the cell area. The simplification of the problem makes the calculation readily attainable on a standard PC. Moreover, the regular pattern improved measurement power of the technique to a force resolution of 2 nN.

65.4.3 Micro-Cantilever Force Sensors

In the previous methods, the use of a continuous membrane for measuring cell forces has the inherent disadvantage that the discrete forces applied at the focal adhesions are convoluted with distribution of displacements. Since the force calculation is not direct, constraints and selection criteria are required in order to solve for the appropriate force mapping. The lack of a direct, linear technique to transduce the physical substrate deformation into unique traction force readings has necessitated the use of microfabricated devices to measure cellular forces. An innovative approach is the use of microcantilevers that act as force transducers. The first demonstration of these sensors is a horizontal cantilever fabricated on a silicon wafer where the cell bends the cantilever in the plane of the traction force as it migrates across it (Figure 65.3d) [20,21]. Since the sensor is mechanically decoupled from the substrate, the deflection of the cantilever directly reports only the local force. The simple spring equation relates the visually measured deflection of the cantilever beam, δ to the cellular traction force:

$$F = K\delta \tag{65.2}$$

where K is the measured spring constant for the cantilever. The devices are constructed out of a polysilicon thin-film that is deposited on top of a phosphosilicate glass sacrificial layer. Once the sacrificial layer is etched away, the beam is freestanding and fully deflected under the force of a cell. These fabrication steps are labor-intensive and expensive, hence these devices are often reused between experiments. Even though this technique has quick force calculation, the horizontal design of the cantilever restricts the measurements to forces along one axis and only a single location on the cell.

Modifying the design to a high-density array of vertical cantilevers improved both the spatial resolution of the force sensor and the scope of possible experiments [22]. With each cantilever placed perpendicular to the plane of traction forces, the spacing between each sensor is significantly reduced (Figure 65.3e). These devices are made from silicone rubber that has cylindrical cantilevers formed from a microfabricated mold. The cost per device is inexpensive once the reusable mold has been built and so the devices are disposable. The tips of the cantilevers are coated with extracellular matrix proteins so that cells attach and spread across several cantilevers. The bending of the posts is easily measured under a microscope as the cells probe the tips and apply traction forces. As with the horizontal design, the deflection of the posts is related by a simple relationship between force and displacement:

$$F = \left(\frac{3EI}{L^3}\right)\delta \tag{65.3}$$

where E is the modulus of elasticity of the silicone rubber, I is the moment of inertia, and L is the length of the cantilevers. However, the deflection of the post is not limited to one axis, the force reported is a true vector quantity in which force mappings are possible with an equivalent resolution to those from traction force microscopy. With the close proximity between sensors and measuring independence between them, the array of vertical cantilevers can examine cells at a higher population density than previous methods. Moreover, the technique allows for more relevant studies than previously possible because the forces of large monolayers of cells can be measured. This technique does expose the cell to a topology that is not akin to *in vitro* conditions, which may have an affect on its biological response.

65.5 Conclusion

The mechanical force that cells experience in the environment directly regulate their function in healthy tissue. Through the sensing of these forces, cells interact with these mechanical signals through biological responses and mechanical force generation of their cytoskeletal structures and motor proteins. The engineering of deformable substrates to measure the cellular forces has provided powerful insight into the protein interactions associated with mechanotransduction. However, despite these advances, these devices have several issues that need to be addressed in order to overcome their limitations. First, one needs to consider how the cell reacts to the new environment that the tooling presents it. The nonplanar topology and high compliance of the vertical microcantilever substrate may cause the cell to react to an environment that is physiologically irrelevant. Additionally, chemical composition of substrate or deposited extracellular matrix may have a direct effect on what signaling pathways are activated during its mechanical response. Second, since the devices used in cellular mechanics studies are prototypes, they may be lacking in adequate calibration between samples or quality control in device fabrication. These variations are significant if the cell population studied is not sufficiently large, as in the case of expensive or labor-intensive techniques. Lastly, the construction of these devices needs to be simple enough so that widespread use is possible. A large collective effort can be used to screen the numerous protein interactions that occur during the mechanotransduction signaling pathways. In this manner, the understanding of the interaction between mechanical forces and biological response can provide valuable insight into the treatment of diseased states of tissue or cancer.

To achieve this goal, there are many future directions that the techniques described can be advanced to. Foremost is the integration of cellular mechanics instrumentation with other fluorescent microscopy techniques, such as fluorescent recovery after photobleaching (FRAP), GFP protein-labeling, and fluorescent resonant emission transfer (FRET). These optical techniques allow one to detect proteins at the single molecular level, and in combination with force mapping, provide a correlation between molecular activity and observable mechanics. The incorporation of nanotechnology materials or devices may provide powerful new sensors that improve both spatial resolution and force measurement. Since the size of a focal adhesion is ten to hundreds of nanometers and the force of a single actomyosin motor is few piconewtons, the ability to resolve these structures would provide greater insight into the mechanical behavior of cells. Additionally, the constructions of three-dimensional measurement techniques, be it in gels or more complex sensing devices, would extend the current two-dimensional understanding of force mechanics into an environment more pertinent to cellular interactions in living tissue. One early attempt has been made where two traction force substrates have been used to sandwich a cell while providing some understanding of three-dimensional forces, the substrates are still planar and constrain how the cell organizes its cytoskeleton and adhesions along those planes. Lastly, strong exploration into the development of devices or techniques that are usable for *in vivo* studies of mechanotransduction would open new areas of treatment for diseases in the cardiovascular and skeletal systems.

Acknowledgments

This work was in part funded by NIH grants, EB00262 and HL73305. NJS was supported by the NRSA — Ruth Kirschtein Postdoctoral Fellowship.

References

[1] Alberts, B. et al., *Molecular Biology of the Cell.* 4th ed., 2002, New York, NY: Garland Science.

[2] Cowin, S.C., On mechanosensation in bone under microgravity. *Bone*, 1998, **22**: 119S–125S.

[3] Chen, C.S., J. Tan, and J. Tien, Mechanotrandsuction at cell–matrix and cell–cell contacts *Ann. Rev. Biomed. Eng.*, 2004, **6**: 275–302.

[4] Davies, P.F., Flow-mediate endothelial mechanotransduction. *Phys. Rev.*, 1995, **75**: 519–560.

[5] Johnson-Leger, C., M. Aurrand-Lions, and B.A. Imhof, The parting of the endothelium: miracle, or simply a junctional affair? *J. Cell Sci.*, 2000, **113**: 921–933.

[6] Worthylake, R.A. and K. Burridge, Leukocyte transendothelial migration: orchestrating the underlying molecular machinery. *Curr. Opin. Cell Biol.*, 2001, **13**: 569–577.

[7] Dudek, S.M. and J.G.N. Garcia, Cytoskeletal regulation of pulmonary vascular permeability. *J. Appl. Physiol.*, 2001, **91**: 1487–1500.

[8] McBeath, R. et al., Cell shape, cytoskeletal tension, and RhoA regulate stem cell lineage commitment. *Develop. Cell*, 2004, **6**: 483–495.

[9] Ingber, D.E., Tensegrity I. Cell structure and hierarchical systems biology. *J. Cell Sci.*, 2003, **116**: 1157–1173.

[10] Couloumbe, P.A. and P. Wong, Cytoplasmic intermediate filaments revealed as dynamic and multipurpose scaffolds. *Nat. Cell Biol.*, 2004, **6**: 699–706.

[11] Finer, J.T., R.M. Simmons, and J.A. Spudich, Single myosin molecule mechanics: piconewton forces and nanometre steps. *Nature*, 1994, **368**: 113–119.

[12] Harris, A.K., P. Wild, and D. Stopak, Silicone rubber substrata: a new wrinkle in the study of cell locomotion. *Science*, 1980, **208**: 177–179.

[13] Harris, A.K., Tissue culture cells on deformable substrata: biomechanical implications. *J. Biomech. Eng.*, 1984, **106**: 19–24.

[14] Burton, K. and D.L. Taylor, Traction forces of cytokinesis measured with optically modified elastic substrate. *Nat. Cell Biol.*, 1997, **385**: 450–454.

[15] Burton, K., J.H. Park, and D.L. Taylor, Keratocytes generate traction froces in two phase. *Mol. Biol. Cell*, 1999, **10**: 3745–3769.

[16] Lee, J. et al., Traction forces generated by locomoting keratocytes. *J. Cell Biol.*, 1994, **127**: 1957–1964.

[17] Dembo, M. and Y.-L. Wang, Stresses at the cell-to-substrate interface during locomotion of fibroblasts. *Biophys. J.*, 1999, **76**: 2307–2316.

[18] Munevar, S., Y.-L. Wang, and M. Dembo, Traction force microscopy of migrating normal and H-ras transformed 3T3 fibroblasts. *Biophys. J.*, 2001, **80**: 1744–1757.

[19] Dembo, M. et al., Imaging the traction stresses exerted by locomoting cells with the elastic substratum method. *Biophys. J.*, 1996, **70**: 2008–2022.

[20] Balaban, N.Q. et al., Force and focal adhesion assembly: a close relationship studied using elastic micropatterned substrates. *Nat. Cell Biol.*, 2001, **3**: 466–472.

[21] Galbraith, C.G. and M.P. Sheetz, A micromachined device provides a new bend on fibroblast traction forces. *Proc. Natl Acad. Sci. USA*, 1997, **94**: 9114–9118.

[22] Galbraith, C.G. and M.P. Sheetz, Keratocytes pull with similar forces on their dorsal and ventral surfaces. *J. Cell Biol.*, 1999, **147**: 1313–1323.

[23] Tan, J.L. et al., Cells lying on a bed of microneedles: an approach to isolate mechanical force. *Proc. Natl Acad. Sci. USA*, 2003. **100**: 1484–1489.

66

Blood Glucose Monitoring

David D. Cunningham
Abbott Diagnostics

The availability of blood glucose monitoring devices for home use has significantly impacted the treatment of diabetes with the American Diabetes Association currently recommending that Type 1 insulin-dependent diabetic individuals perform blood glucose testing four times per day. Less than optimal outcomes are associated with high and low blood glucose levels. Injection of too much insulin without enough food lowers blood sugar into the hypoglycemic range, glucose below 60 mg/dL, resulting in mild confusion or in more severe cases loss of consciousness, seizure, and coma. On the other hand, long-term high blood sugar levels lead to diabetic complications such as eye, kidney, heart, nerve, or blood vessel disease [The Diabetes Control and Complications Trial Research Group, 1993]. Complications were tracked in a large clinical study showing that an additional 5 years of life, 8 years of sight, 6 years free from kidney disease, and 6 years free of amputations can be expected for a diabetic following tight glucose control vs. the standard regimen [The Diabetes Control and Complications Trial Research Group, 1996]. This compelling need for simple, accurate glucose measurements has lead to continuous improvements in sample test strips, electronic meters, and sample acquisition techniques. Some of the landmarks in glucose testing are shown in Table 66.1. Glucose monitoring systems are now available from a number of companies through pharmacy and mail-order outlets without a prescription. The remainder of the chapter comprises a history of technical developments with an explanation of the principles behind optical and electrochemical meters including examples of the chemical reactions used in commercial products.

TABLE 66.1 Landmarks in Glucose Monitoring

1941 — Effervescent tablet test for glucose in urine
1956 — Dip and read test strip for glucose in urine
1964 — Dry reagent blood glucose test strip requiring timing, wash step, and visual comparison to a color chart
1970 — Meter to read reflected light from a test strip, designed for use in the doctor's office
1978 — Major medical literature publications on home blood glucose monitoring with portable meters
1981 — Finger lancing device automatically lances and retracts tip
1986 — Electrochemical test strip and small meter in the form of a pen
1997 — Multiple test strip package for easy loading into meter
2001 — Integrated alternate-site glucose monitoring for painless, one-step testing

66.1 Medicine and Historical Methods

Diabetes is an ancient disease that was once identified by the attraction of ants to the urine of an affected individual. Later, physicians would often rely on the sweet taste of the urine in diagnosing the disease. Once the chemical reducing properties of glucose were discovered, solutions of a copper salt and dye, typically *o*-toluidine, were used for laboratory tests, and by the 1940s the reagents had been formulated into tablets for use in test tubes of urine. More specific tests were developed using glucose oxidase which could be impregnated on a dry paper strip. The reaction of glucose with glucose oxidase produces hydrogen peroxide which can subsequently react with a colorless dye precursor in the presence of hydrogen peroxide to form a visible color (see Equation 66.3). The first enzyme-based test strips required the addition of the sample to the strip for 1 min and subsequent washing of the strip. Visual comparison of the color on the test strip to the color on a chart was required to estimate the glucose concentration. However, measurement of glucose in urine is not adequate since only after the blood glucose level is very high for several hours does glucose "spill-over" into the urine. Other physiological fluids such as sweat and tears are not suitable because the glucose level is much lower than in blood.

Whole blood contains hemoglobin inside the red blood cells that can interfere with the measurement of color on a test strip. In order to prevent staining of the test strip with red blood cells, an ethyl cellulose layer was applied over the enzyme and dye impregnated paper on a plastic support [Mast, 1967]. Previously, in a commercially available test strip, the enzymes and dye were incorporated into a homogeneous water-resistant film that prevented penetration of red blood cells into the test strips and enabled their easy removal upon washing [Rey et al., 1971]. Through various generations of products, the formulations of the strips were improved to eliminate the washing/wiping steps and electronic meters were developed to measure the color.

66.2 Development of Colorimetric Test Strips and Optical Reflectance Meters

Optically based strips are generally constructed with various layers which provide a support function, a reflective function, an analytical function, and a sample-spreading function as illustrated in Figure 66.1. The support function serves as a foundation for the dry reagent and may also contain the reflective function. Otherwise, insoluble, reflective, or scattering materials such as TiO_2, $BaSO_4$, MgO, or ZnO are added to the dry reagent formulation. The analytical function contains the active enzyme. The reaction schemes used in several commercial products are described in greater detail in the following paragraphs. The spreading function must rapidly disperse the sample laterally after application and quickly form

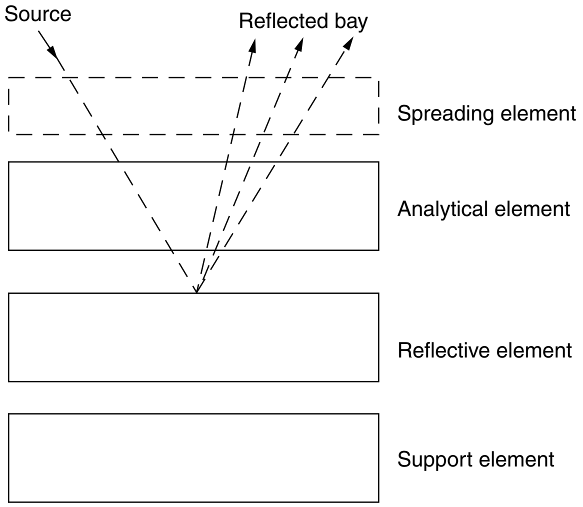


FIGURE 66.1 Basic functions of a reflectance-based test strip. (From Henning T.P. and Cunningham, D.D. 1998, *Commercial Biosensors*, John Wiley & Sons, pp. 3–46. With permission.)

a uniform sample concentration on the analytically active portion of the strip. Swellable films and semipermeable membranes, particularly glass fiber fleece has been used to spread and separate plasma from whole blood. Upon formation of the colored reaction product, the amount of diffuse light reflected from the analytical portion of the strip decreases according to the following equation:

$$\%R = (I_u/I_s)R_s \tag{66.1}$$

where I_u is the reflected light from the sample, I_s is the reflected light from a standard, and R_s is the percent reflectivity of the standard. The Kubelka–Munk equation gives the relationship in a more useful form.

$$C\,\alpha K/S = (1-R)^2/2R \tag{66.2}$$

where C is concentration, K is the absorption coefficient, S is the scattering coefficient, and R is the percent reflectance divided by 100.

The analytical function of the strip is based on an enzyme reaction with glucose and subsequent color forming reactions. Although the most stable enzyme is chosen for product development, some loss in activity occurs during manufacturing due to factors such as pH, temperature, physical sheer stress, organic solvents, and various other denaturing actions or agents. Additional inactivation occurs during the storage of the product. In general, sufficient enzyme and other reagents are incorporated into the strip so that the assay reactions near completion in a conveniently short time. Reagent formulations often include thickening agents, builders, emulsifiers, dispersion agents, pigments, plasticisers, pore formers, wetting agents and the like. These materials provide a uniform reaction layer required for good precision and accuracy. The cost of the materials in the strip must be low since it is only used once.

The glucose oxidase/peroxidase reaction scheme used in the Lifescan ONE TOUCH® [Phillips et al., 1990] and SureStep™ test strips follows. Glucose oxidase catalyzes the oxidation of glucose forming gluconic acid and hydrogen peroxide. The oxygen concentration in blood (ca. 0.3 m*M*) is much lower than the glucose concentration (3 to 35 m*M*), hence oxygen from the atmosphere must diffuse into the test strip to bring the reaction to completion. Peroxidase catalyzes the reaction of the hydrogen peroxide with 3-methyl-2-benzothiazolinone hydrazone (MBTH) and 3-dimethylaminobenzoic acid (DMAB).

A naphthalene sulfonic acid salt replaces DMAB in the SureStep strip.

$$\begin{gathered}\text{glucose} + \text{oxygen} \xrightarrow{\text{GOx}} \text{gluconic acid} + H_2O_2 \\ H_2O_2 + \text{MBTH} + \text{DMAB} \xrightarrow{\text{peroxidase}} \text{MBTH–DMAB (blue)}\end{gathered} \tag{66.3}$$

The hexokinase reaction scheme used in the Bayer GLUCOMETER ENCORE™ test strip is shown below. Hexokinase, ATP, and magnesium react with glucose to produce glucose-6-phosphate. The glucose-6-phosphate reacts with glucose-6-phosphate dehydrogenase and NAD$^+$ to produce NADH. The NADH then reacts with diaphorase and reduces the tetrazolium indicator to produce a brown compound (formazan). The reaction sequence requires three enzymes but is insensitive to oxygen.

$$\text{glucose} + \text{ATP} \xrightarrow[\text{Mg}^{+2}]{\text{HK}} \text{G-6-P} + \text{ADP} \tag{66.4}$$

$$\begin{gathered}\text{G-6-P} + \text{NAD}^+ \xrightarrow{\text{G-6-PDH}} \text{6-PG} + \text{NADH} + \text{H}^+ \\ \text{NADH} + \text{tetrazolium} \xrightarrow{\text{diaphorase}} \text{formazan (brown)} + \text{NAD}^+\end{gathered}$$

The reaction scheme used in the Roche Accu-Chek® Instant™ strip is shown below [Hoenes et al., 1995]. *Bis*-(2-hydroxy-ethyl)-(4-hydroximinocyclohex-2,5-dienylidene) ammonium chloride (BHEHD) is reduced by glucose to the corresponding hydroxylamine derivative and further to the corresponding diamine under the catalytic action of glucose oxidase. Note that while oxygen is not required in the reaction, oxygen in the sample may compete with the intended reaction creating an oxygen dependency. The diamine reacts with a 2,18-phosphomolybdic acid salt to form molybdenum blue.

$$\begin{gathered}\text{glucose} + \text{BHEHD} \xrightarrow{\text{GOx}} \text{diamine} \\ P_2Mo_{18}O_{62}^{-6} + \text{diamine} \longrightarrow MoO_{2.0}(OH) \text{ to } MoO_{2.5}(OH)_{0.5}\text{(molybdenum blue)}\end{gathered} \tag{66.5}$$

The reaction scheme used in the Roche Accu-Chek® Easy™ Test strip is shown below [Freitag, 1990]. Glucose oxidase reacts with ferricyanide and forms potassium ferric ferrocyanide (Prussian Blue). Again, oxygen is not required but may compete with the intended reaction.

$$\begin{aligned}&\text{glucose} + \text{GOD(ox)} \rightarrow \text{GOD (red)} \\ &\text{GOD (red)} + [Fe(CN)_6]^{-3} \rightarrow \text{GOD (ox)} + [Fe(CN)_6]^{-4} \\ &3[Fe(CN)_6]^{-4} + 4FeCl_3 \rightarrow Fe_4[Fe(CN)_6]_3 \text{ Prussian Blue}\end{aligned} \tag{66.6}$$

Optical test strips and reflectance meters typically require 3–15 μL of blood and read out an answer in 10–30 sec. A significant technical consideration in the development of a product is the measurement

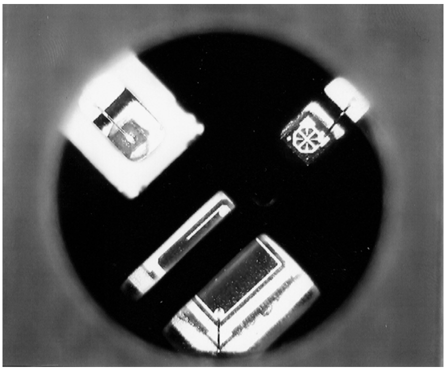


FIGURE 66.2 Photograph of light emitting diodes and photodetector on the OneTouch meter. Photodetector at bottom (blue), 635 nm light emitting diode at top left, 700 nm light emitting diode at top right. Optics viewed through the 4.5 mm hole in the strip after removal of the reagent membrane. Courtesy of John Grace.

of samples spanning the range of red blood cell concentrations (percent hematocrit) typically found in whole blood. Common hematocrit and glucose ranges are 30–55% and 40–500 mg/dL (2.2–28 m*M*), respectively. The Lifescan ONE TOUCH® meter contains two light emitting diodes (635 and 700 nm) which allows measurement of the color due to red blood cells and the color due to the dye. Reflectance measurements from both LEDs are measured with a single photodetector as shown in Figure 66.2. All glucose meters measure the detector signal at various timepoints and if the curve shape is not within reasonable limits an error message is generated. Some meters measure and correct for ambient temperature. Of course, optical systems are subject to interference from ambient light conditions and may not work in direct sunlight. Optical systems have gradually lost market share to electrochemical systems which were introduced commercially in 1987. Optical test strips generally require a larger blood sample and take longer to produce the result. Presently, optical reflectance meters are more costly to manufacture, require larger batteries, and are more difficult to calibrate than electrochemical meters.

66.3 Emergence of Electrochemical Strips

Electrochemical systems are based on the reaction of an electrochemically active mediator with an enzyme. The mediator is oxidized at a solid electrode with an applied positive potential. Electrons will flow between the mediator and electrode surface when a minimum energy is attained. The energy of the electrons in the mediator is fixed based on the chemical structure but the energy of the electrons in the solid electrode can changed by applying a voltage between the working electrode and a second electrode. The rate of the electron transfer reaction between the mediator and a working electrode surface is given by the Butler–Volmer equation [Bard and Faulkner, 1980]. When the potential is large enough the mediator reaching the electrode reacts rapidly and the reaction becomes diffusion controlled. The current from a diffusion limited reaction follows the Cottrell equation,

$$i = (nFAD^{1/2}C)/(\pi^{1/2}t^{1/2}) \tag{66.7}$$

where i is current, n is number of electrons, F is Faradays constant, A is electrode area, C is the concentration, D is the diffusion coefficient, and t is time. The current from a diffusion-controlled electrochemical reaction will decay away as the reciprocal square root of time. This means that the maximum electrochemical signal occurs at short times as opposed to color forming reactions where the color becomes more intense with time. The electrochemical method relies on measuring the current from the electron transfer between the electrode and the mediator. However, when a potential is first applied to the electrode the dipole moments of solvent molecules will align with the electric field on the surface of the electrode causing a current to flow. Thus, at very short times this charging current interferes with the analytical measurement. Electrochemical sensors generally apply a potential to the electrode surface and measure the current after the charging current has decayed sufficiently. With small volumes of sample, coulometric analysis can be used to measure the current required for complete consumption of glucose.

The reaction scheme used in the first commercial electrochemical test strip from MediSense is shown below. Electron transfer rates between the reduced form of glucose oxidase and ferricinium ion derivatives are very rapid compared with the unwanted side-reaction with oxygen [Cass et al., 1984]. Electrochemical oxidation of ferrocene is performed at 0.6 V. Oxidation of interferences, such as ascorbic acid and acetaminophen present in blood, are corrected for by measuring the current at a second electrode on the strip that does not contain glucose oxidase.

$$\begin{aligned}&\text{glucose} + \text{GOx (oxidized)} \rightarrow \text{gluconolactone} + \text{Gox (reduced)}\\&\text{GOx (reduced)} + \text{ferricinium}^{+} \rightarrow \text{GOx (oxidized)} + \text{ferrocene}\\&\text{ferrocene} \rightarrow \text{ferricinium}^{+} + \text{electron (reaction at solid electrode surface)}\end{aligned} \tag{66.8}$$

The reaction scheme used in the Abbott Laboratories MediSense Products Precision-Xtra and Sof-Tact test strips follows. The glucose dehydrogenase (GDH) enzyme does not react with oxygen and the phenanthroline quinine mediator can be oxidized at 0.2 V, which is below the oxidation potential of most interfering substances.

$$\begin{aligned}&\text{glucose} + \text{GDH/NAD}^{+} \rightarrow \text{GDH/NADH} + \text{gluconolactone}\\&\text{GDH/NADH} + \text{PQ} \rightarrow \text{GDH/NAD}^{+} + \text{PQH}_2\\&\text{PQH}_2 \rightarrow \text{PQ} + \text{electrons (reaction at solid electrode surface)}\end{aligned} \tag{66.9}$$

The working electrode on most commercially available electrochemical strips is made by screen printing a conductive carbon ink on a plastic substrate, however, a more expensive noble metal foil is also used. Many of the chemistries described above are used on more than one brand of strip from the company. The package insert provided with the test strips describes the test principle and composition. Generally, test strips are manufactured, tested, and assigned a calibration code. The calibration code provided with each package of strips must be manually entered into the meter by the user. However, some high-end meters now automatically read the calibration code from the strips. Meters designed for use in the hospital have bar code readers to download calibration, quality control, and patient information. Test strips are supplied in bottles or individual foil wrappers to protect them from moisture over their shelf-life, typically about 1 year. The task of opening and inserting individual test strips into the meter has been minimized by packaging multiple test strips in the form of a disk shaped cartridge or a drum that is placed into the meter.

66.4 Improvements in User Interactions with the System and Alternate Site Testing

Both Type 1 and Type 2 diabetic individuals do not currently test as often as recommended by physicians so systems developed in the last few years have aimed to improve compliance with physician recommendations while maintaining accuracy. Historically, the biggest source of errors in glucose testing involved interaction of the user with the system. Blood is typically collected by lancing the edge of the end of the finger to a depth of about 1.5 mm. Squeezing or milking is required to produce a hanging drop of blood. The target area on test strips is clearly identified by design. A common problem is smearing a drop of blood on top of a strip resulting in a thinner than normal layer of blood over part of the strip and a low

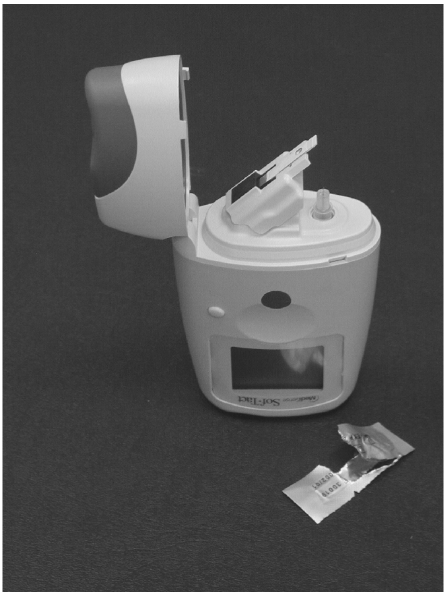


FIGURE 66.3 Sof-Tact meter with cover opened to load test strip and lancet. Test strip is inserted into an electrical connector. The hole in the opposite end of the test strip allows the lancet to pass through. The white cylindrical lancet is loaded into the lancet tip holder in the housing. To perform a test, the cover is closed, the gray area of the cover placed against the skin and blue button depressed.

reading. Many strips now require that the blood drop be applied to the end or side of the strip where capillary action is used to fill the strip. Partial filling can be detected electrochemically or by observation of the fill window on the strip. The small capillary space in the TheraSense electrochemical strip requires only 300 nl of blood.

Progressively thinner diameter lancets have come to the market with current sizes typically in the 28 to 31 gauge range. Most lancets are manufactured with three grinding steps to give a tri-level point. After loading the lancet into the lancing device, a spring system automatically lances and retracts the point. The depth of lancing is commonly adjustable through the use of several settings on the device or use of a different end-piece cap. Unfortunately, the high density of nerve endings on the finger make the process painful and some diabetic individuals do not test as often as they should due to the pain and residual soreness caused by fingersticks. Recently, lancing devices have been designed to lance and apply pressure on the skin of body sites other than the finger, a process termed "**alternate site testing**." Use of alternate site sampling lead to the realization that capillary blood from alternate sites can have slightly different glucose and hematocrit values than blood from a fingerstick due to the more arterial nature of blood in the fingertips. The pain associated with lancing alternate body sites is typically rated as painless most of the time and less painful than a fingerstick over 90% of the time. A low volume test strip, typically one microliter or less, is required to measure the small blood samples obtained from alternate sites. Some care and technique is required to obtain an adequate amount of blood and transfer it into the strips when using small blood samples.

One alternate site device, the Abbott/MediSense Sof-Tact meter, automatically extracts and transfers blood to the test strip (see Figure 66.3). The device contains a vacuum pump, a lancing device, and a test strip that is automatically indexed over the lancet wound after lancing. The vacuum turns off after sufficient blood enters the strip to make an electrical connection. The key factors and practical limits of blood extraction using a vacuum combined with skin stretching were investigated to assure that sufficient blood could be obtained for testing [Cunningham et al., 2002]. The amount of blood extracted increases with the application of heat or vacuum prior to lancing, the level of vacuum, the depth of lancing, the time

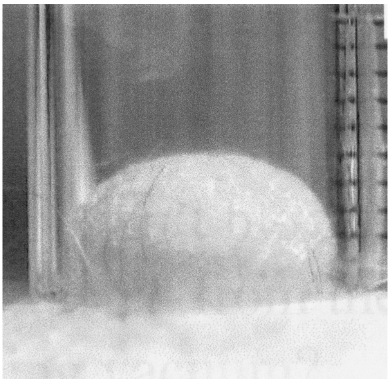


FIGURE 66.4 Photograph of skin on the forearm stretching up into a glass tube upon application of vacuum. Markings on tube at right in 1 mm increments. Courtesy of Douglas Young.

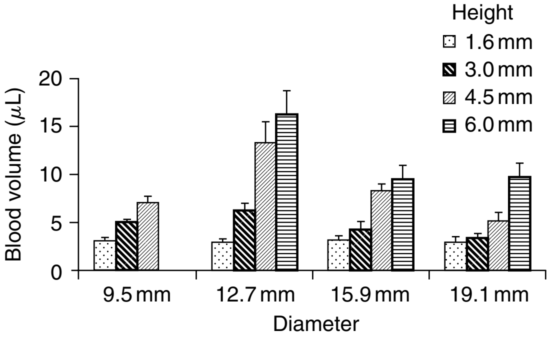


FIGURE 66.5 Effect of skin stretching by vacuum on blood volume extracted from lancet wounds on the forearm. Mean blood volume ± SE in 30 sec with −7.5 psig vacuum for nosepieces of different inner diameter and inside step height. (From Cunningham D.D. et al., 2002. *J. Appl. Physiol.* 92: 1089–1096. With permission.)

of collection, and the amount of skin stretching (see Figure 66.4). Particularly important is the diameter and height that skin is allowed to stretch into a nosepiece after the application of a vacuum as shown in Figure 66.5. A vacuum combined with skin stretching increases blood extraction by increasing the lancet wound opening, increasing the blood available for extraction by vasodilatation, and reducing the venous return of blood through the capillaries. The electrochemical test strip used with the meter can be inserted into a secondary support and used with a fingerstick sample when the battery is low.

The size of a meter is often determined by the size of the display although electrochemical meters can be made smaller than reflectance meters. The size and shape of one electrochemical meter, with a relatively small display, is indistinguishable from a standard ink pen. All meters store recent test results in memory and many allow downloading of the results to a computer. Advanced software functions are supplied with some meters to allow entry of exercise, food, and insulin doses, and a PDA-meter combination is on the market. Two combined insulin dosing-glucose meters are available. One combines an insulin injection pen with an electrochemical glucose meter, and the other combines a continuous insulin infusion pump with an electrochemical glucose meter using telemetry for downloading the glucose measurements into the insulin pump memory. The variety of meters available in the market is mainly driven by the need to satisfy the desires of various customer segments which are driven by different factors, such as cost, ease of use, or incorporation of a specific design or functional feature.

66.5 Future Directions

Several approaches to continuous glucose sensing are being actively pursued based on the desire to obtain better glucose control through a combination of sensing and insulin administration. The most advanced is an electrochemical needle sensor that is inserted through the skin into the subcutaneous fat layer [Feldman et al., 2003]. A second approach is to porate the skin and extract interstitial fluid for measurements with a miniature sensor [Gebhart et al., 2003]. With either of these approaches, infection becomes a concern after a few days. Longer term sensing may involve surgical implantation of a battery-operated unit although many issues remain with the long-term sensor stability and the biocompatibility of various materials of construction. One 24-h device, the GlucoWatch based on transdermal reverese iontophoresis [Kurnick et al., 1998] gained FDA approval but acceptance of the device in the market has been poor due to the need to calibrate the device with multiple fingersticks and poor precision and accuracy. A number of noninvasive spectroscopic approaches have been described, however, the amount of clinical data reported to date is very limited [Khalil, 1999]. Continuous sensing devices coupled with insulin delivery will almost certainly have a significant impact on the treatment of diabetes in the future [Siegel and Ziaie, 2004]. Less certain is the timing for the market launch of specific devices, the form and function of winning technologies, and the realization of commercial success.

Defining Terms

Alternate site testing: Lancing sites other than the finger to obtain blood in a less painful manner. The small volume of blood obtained from alternate sites requires use of a test strip requiring one microliter or less of blood.

Type 1 Diabetes: The immune system destroys insulin-producing islet cells in the pancreas, usually in children and young adults, hence regular injections of insulin are required (also referred to as juvenile diabetes).

Type 2 Diabetes: A complex disease based on gradual resistance to insulin and diminished production of insulin. Treatment often progresses from oral medications to insulin injections as disease progresses. Also referred to as adult onset diabetes and noninsulin dependent diabetes mellitus (NIDDM).

References

Bard A.J. and Faulkner L.R. 1980. *Electrochemical Methods*, John Wiley & Sons, New York, pp. 103, 143.

Cass A., Davis G., Francis G., Hill H., Aston W., Higgins I., Plotkin E., Scott L., and Turner A. 1984. Ferrocene-mediated enzyme electrode for amperometric determination of glucose. *Anal. Chem.*, 56: 667.

The Diabetes Control and Complications Trial Research Group, 1993. 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.* 329: 977.

The Diabetes Control and Complications Trial Research Group, 1996. Lifetime benefits and costs of intensive therapy as practiced in the diabetes control and complications Trial. *J. Amer. Med. Assoc.* 276: 1409.

Feldman B., Brazg R., Schwartz S., and Weinstein R. 2003. A continuous glucose sensor based on wired enzyme technology — results from a 3-day trial in patients with type 1 diabetes. *Diabetes Technol. Ther.* 5: 769.

Freitag H. 1990. Method and reagent for determination of an analyte via enzymatic means using a ferricyanide/ferric compound system. U.S. Patent 4,929,545.

Gebhart S., Faupel M., Fowler R., Kapsner C., Lincoln D., McGee V., Pasqua J., Steed L., Wangsness M., Xu F., and Vanstory M. 2003. Glucose sensing in transdermal body fluid collected under continuous vacuum pressure via micropores in the stratum corneum. *Diabetes Technol. Ther.* 5: 159.

Hoenes J., Wielinger H., and Unkrig V. 1995. Use of a soluble salt of a heteropoly acid for the determination of an analyte, a corresponding method of determination as well as a suitable agent thereof. U.S. Patent 5,382,523

Khalil O.S. 1999. Spectroscopic and clinical aspects of noninvasive glucose measurements. *Clin. Chem.* 45: 165.

Kurnik R.T., Berner B., Tamada J., and Potts R.O. 1998. Design and simulation of a reverse iontophoretic glucose monitoring device. *J. Electrochem. Soc.* 145: 4199.

Mast R.L. 1967. Test article for the detection of glucose. U.S. Patent 3,298,789.

Phillips R., McGarraugh G., Jurik F., and Underwood R. 1990. Minimum procedure system for the determination of analytes. U.S. Patent 4,935,346.

Rey H., Rieckman P., Wiellager H., and Rittersdorf W. 1971 Diagnostic agent. U.S. Patent 3,630,957.

Siegel R.A. and Ziaie B. 2004. Biosensing and drug delivery at the microscale. *Adv. Drug Deliv. Rev.* 56: 121.

Further Reading

Ervin K.R. and Kiser E.J. 1999. Issues and implications in the selection of blood glucose monitoring technologies. *Diabetes Technol. Ther.* 1: 3.

Henning T.P. and Cunningham D.D. 1998. Biosensors for personal diabetes management, in *Commercial Biosensors*, Ramsey G., Ed., John Wiley & Sons, pp. 3–46.

Test results of glucose meters are often compared with results from a reference method and presented in the form of a Clark Error Grid that defines zones with different clinical implications. Clarke W.L., Cox D.C., Gonder-Frederick L.A., Carter W. and Pohl S.L. 1987. Evaluating clinical accuracy of systems for self-Monitoring of blood glucose. *Diabetes Care* 10: 622–628.

Error Grid Analysis has recently been extended for evaluation of continuous glucose monitoring sensors. Kovatchev B.P., Gonder-Frederick L.A., Cox D.J., and Clarke W.L. 2004. Evaluating the accuracy of continuous glucose-monitoring sensors. *Diabetes Care* 27: 1922.

Reviews and descriptions of many marketed products are available on-line at: www.childrenwithdiabetes.com.

Interviews of several people involved with the initial development of the blood glucose meters are available on-line at: www.mendosa.com/history.htm.

67

Atomic Force Microscopy: Probing Biomolecular Interactions

Christopher M. Yip
University of Toronto

67.1 Introduction

Discerning and understanding structure–function relationships is often predicated on our ability to measure these properties on a variety of length scales. Fundamentally, nanotechnology and nanoscience might be arguably based on the precept that we need to understand how interactions occur at the atomic and molecular length scales if we are to truly understand how to manipulate processes and structures and ultimately control physical/chemical/electronic properties on more bulk macroscopic length scales. There is a clear need to understanding the pathways and functional hierarchy involved in the development of complex architectures from their simple building blocks. In order to study such phenomena at such a basic level, we need tools capable of performing measurements on these same length scales. If we can couple these capabilities with the ability to map these attributes against a real-space image of such structures, in real-time and hopefully, under real-world conditions, this would provide the researcher with a particularly powerful and compelling set of approaches to characterizing interactions and structures.

Powerful functional imaging tools such as single molecule fluorescence and nonlinear optical microscopies such as CARS and SHG through to the various electron microscopies (SEM/TEM/STEM) provide a powerful suite of tools for characterizing phenomena under a vast range of conditions and situations. What many of these techniques lack, however, is the ability to acquire true real-space, real-time information about surface structures on near-molecular length scales, and, in the case of many of these techniques, in the absence of specific labeling strategies. Atomic force microscopy (AFM), or more correctly, scanning probe microscopy (SPM) has come into the forefront as one of the most powerful tools for characterizing molecular scale phenomena and interactions and in particular, their contribution to the development of macroscopic mechanical properties, structures, and ultimately function.

This review will explore some of the recent advances in scanning probe microscopy, including the fundamentals of SPM, where it has been applied in the context of biomolecular structures and functions — from single molecules to large aggregates and complexes — and introduce some new innovations in the field of correlated imaging tools designed to address many of the key limitations of this family of techniques.

67.2 Background

Scanning probe microscopy is founded on a fundamentally simple principle — by raster-scanning a sharp tip over a surface, and monitoring tip–sample interactions, which can range in scope from repulsive to attractive forces to local variations in temperature and viscoelasticity, it is possible to generate real-space images of surfaces with near-molecular scale (and in some cases, atomic scale) resolution. One can reasonably describe these images as isosurfaces of a parameter as a function of (x, y, z) space.

Since its inception in the mid-1980s, SPM has become a very well-accepted technique for characterizing surfaces and interfacial processes with nanometer-scale resolution and precision [Hansma et al., 1988; Lillehei and Bottomley, 2000; Poggi et al., 2002, 2004]. Emerging from efforts in the semi-conductor and physics fields, SPM has perhaps made its greatest impact in the biological sciences and the fields of soft materials [Engel and Muller, 2000]. What has really driven its use in these fields has been its perhaps unique ability to acquire such high resolution data, both spatial and most recently force, in real-time and often *in situ*. This growth has been fostered by a wealth of SPM-based imaging modes, including intermittent contact or tapping mode [Moller et al., 1999], and recently force spectroscopy and force volume imaging techniques [Florin et al., 1994b] [Rief et al., 1997b; Heinz and Hoh, 1999b; Oesterfelt et al., 1999], [Brown and Hoh, 1997; A-Hassan et al., 1998; Walch et al., 2000]. These attributes are particularly compelling for the study of protein assembly at surfaces, ranging from polymers through metals to model-supported planar lipid bilayers and live cell membranes [Pelling et al., 2004].

67.3 SPM Basics

Similar to a technique known as stylus profilometry, including very early work by Young on the "Topographiner" [Young et al., 1971], scanning probe microscopy is a rather simple concept. As you raster-scan a sharp tip and a surface past each other, you monitor any number of tip–surface interactions. One can then generate a surface contour map that reflects relative differences in interaction intensity as a function of surface position. Precise control over the tip–sample separation distance through the use of piezoelectric scanners and sophisticated feedback control schemes is what provides the SPM technique with its high spatial and force resolution.

Based on the scanning tunneling microscope (STM), which operates on the principle of measuring the tunneling current between two conducting surfaces separated by a very small distance. [Binnig et al., 1982],

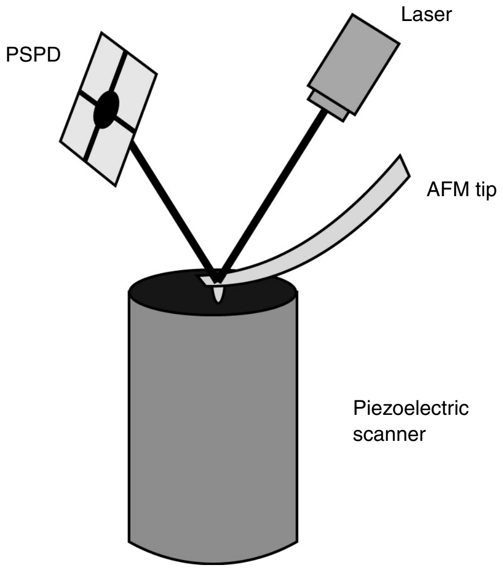


FIGURE 67.1 In AFM, a typically pyramidal silicon nitride tip is positioned over a sample. The relative motion between the tip and sample is controlled by a piezoelectric scanner. In this image, the sample is mounted to the scanner so the sample moves relative to a fixed tip. The reflection of a laser focused onto the back of the AFM tip is monitored on a four-quadrant position sensitive photodiode (PSPD). As the AFM tip is raster-scanned over the sample surface, variations in tip-sample interactions result in (vertical and lateral) deflection of the tip. This deflection is reflected in movement of the laser spot on the PSPD and is used to produce a three-dimensional topographical image of the surface.

atomic force microscopy is predicated on mapping local variations in the intermolecular and interatomic forces between the tip and the sample being scanned [Binnig et al., 1986]. In a conventional AFM, the surface is scanned with a nominally atomically sharp tip, typically pyramidal in shape, which is mounted on the underside of an extremely sensitive cantilever. The theoretical force sensitivity of these tips is on the order of 10^{-14} newtons (N), although practical limitations reduce this value to $\sim 10^{-10}$ N. The resolution of an SPM is highly dependent on the nature of the sample, with near-atomic scale resolution often achievable on atomically flat surfaces (crystals) while soft, and often mobile interfaces, such as cells or membranes, are often challenging to image with a typical resolution in these cases of ∼5 to 10 nm, depending on what you are imaging.

The relative motion of the tip and sample is controlled through the use of piezoelectric crystal scanners. The user sets the desired applied force (or amplitude dampening in the case of the intermittent contact imaging techniques). Deviations from these set point values are picked up as *error* signals on a four-quadrant position sensitive photodetector (PSPD), and then fed into the main computer (Figure 67.1). The error signal provided to the instrument is then used to generate a feedback signal that is used as the input to the feedback control software. The tip–sample separation distance is then dynamically changed in real-time and adjusted according to the error signal. While detection of the deflection signal is the simplest feedback signal, there are a host of other feedback signals that could be used to control the tip–sample mapping, including tip oscillation (amplitude/phase).

67.4 Imaging Mechanisms

67.4.1 Contact

During imaging, the AFM tracks gradients in interaction forces, either attractive or repulsive, between the tip and the surface (Figure 67.2). Similar to how the scanning tunneling microscope mapped out

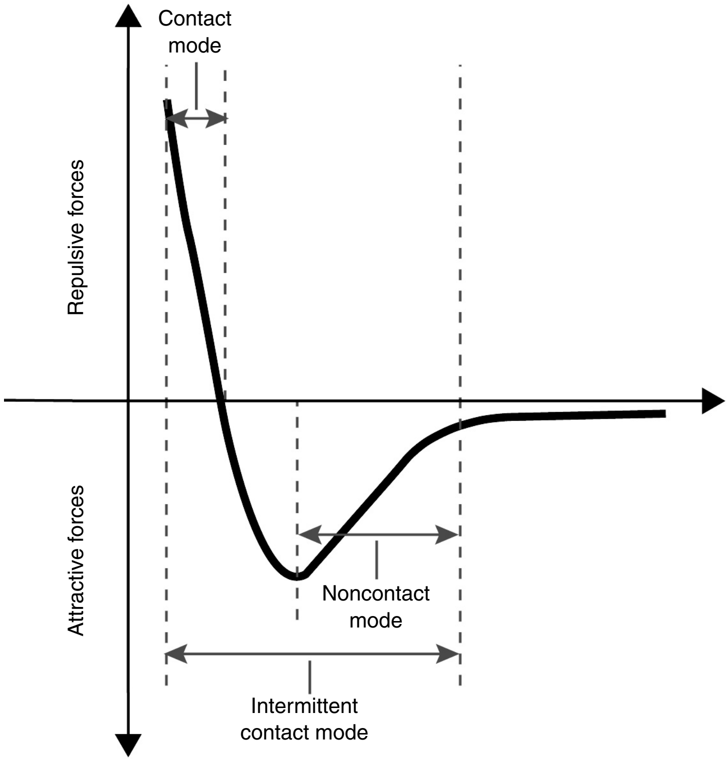


FIGURE 67.2 Tip-sample interaction forces vs tip-sample separation distance. In AFM, the instrument can operate in a number of different modes, based on the nature of the forces felt by the tip as it approaches the surface. When the instrument is operated in *non-contact* mode, the tip-sample separation is maintained so that the tip only feel an attractive interaction with the surface. In *contact* mode, the tip-sample interaction is repulsive. In *intermittent contact* mode, the tip alternates between sensing attractive and repulsive interactions with the surface. This is often achieved by vertically oscillating the tip during scanning.

local variations in tip–sample tunneling current, the AFM uses this force gradient to generate an iso-force surface image. In contact mode imaging, the tip–sample interaction is maintained at a specific, user defined load. It is this operating mode that arguably provides the best resolution for imaging of surfaces and structures. It also provides direct access to so-called friction force imaging where transient twisting of the cantilever during scanning can be used to develop maps of relative surface friction [Magonov and Reneker, 1997; Paige, 2003]. The ability to quantitate such data is limited due to difficulties in determining the torsional stiffness of the cantilevers, a determination that is further exacerbated by the shape of the cantilever. In contact mode imaging, this image represents either a constant attractive, or repulsive, tip–sample force, that is chosen by the user. Incorrect selection of this load can result in damage to the surface when the applied force is higher than what the surface can withstand, or poor tracking when the chosen set point load is too low. Subtle manipulation of these imaging forces affords the user the unique ability to both probe local structure and determine the response of the structure to the applied force.

67.4.2 Noncontact

In noncontact mode imaging, the AFM tip is actively oscillated near its resonance frequency at a distance of tens to hundreds of Angstroms away from sample surface. The resulting image represents an isosurface corresponding to regions of constant amplitude dampening. As the forces between the tip and the surface are very small, noncontact mode AFM is ideally suited for imaging softer samples such as proteins, surfactants, or membranes. In this mode, one often uses cantilevers with a higher spring constant that those employed during normal contact mode imaging. The net result is a very small feedback signal, which can make instrument control difficult and imaging challenging [Dinte et al., 1996; Lvov et al., 1998].

67.4.3 Intermittent Contact

This method, in which the tip alternates from the repulsive to the attractive regions of the tip–sample interaction curve, has become the method of choice currently for most AFM-basing imaging. In early contact mode work, it was quickly realized that poorly adhering molecules could be rapidly displaced by the sweeping motion of the AFM cantilever. This "snow-plow" effect has been largely ameliorated by vertically oscillating the tip during imaging, which removes (to a large extent) the lateral forces present during contact mode imaging. As the vertical oscillations occur at a drive frequency that is several orders of magnitude higher than the actual raster-scanning frequency, it is possible to obtain comparable lateral and vertical resolution as the continuous contact techniques. Since one detects the relative damping of the tip's free vertical oscillation during imaging, an intermittent contact mode AFM image can be viewed as an iso-energy dissipation landscape. Intermittent contact imaging provides access to other imaging modes, including phase imaging, which measures the phase shift between the applied and detected tip oscillations. This derivative signal is particularly useful for tracking spatial distributions of the relative modulus, viscoelasticity, and adhesive characteristics of surfaces, and has proven to be very powerful for studying polymeric materials [Fritzsche and Henderson, 1997; Hansma et al., 1997; Magonov et al., 1997; Magonov and Reneker, 1997; Magonov and Heaton, 1998; Noy et al., 1998b; Magonov and Godovsky, 1999; Nagao and Dvorak, 1999; Holland and Marchant, 2000; Paige, 2003] [Winkler et al., 1996; Brandsch et al., 1997; Czajkowsky et al., 1998; Noy et al., 1998a; Walch et al., 2000; Opdahl et al., 2001; Scott and Bhushan, 2003]. Recent work has shown that phase imaging is particularly useful for studying biological systems, including adsorbed proteins and supported lipid bilayers, even in the absence of topographic contrast [Argaman et al., 1997; Holland and Marchant, 2000; Krol et al., 2000; Deleu et al., 2001].

In intermittent contact imaging, the cantilever can be oscillated either acoustically or magnetically. In the first case, the cantilever is vertically oscillated by a piezoelectric crystal typically mounted under the cantilever. In air, this is typically a single resonance frequency. In fluid imaging, coupling of the cantilever motion with the fluid, and the fluid cell, can result in a complex power spectrum with multiple apparent resonant peaks. In this case, choosing the appropriate peak to operate with can be difficult and experience is often the best guide. Selection of the appropriate cantilever for intermittent contact imaging will depend on the physical imaging environment (air/fluid). In air, one typically uses the so-called diving board tips, which have a relatively high resonance peak of $\sim$250 kHz (depending on the manufacturer). In fluid, viscous coupling between the tip and the surrounding fluid results in an increase in the apparent resonant frequency of the tip. This allow for the use of the conventional V-shaped contact-mode cantilevers. In magnetic mode, the AFM tip/cantilever assembly can be placed in an oscillating magnetic field [Lindsay et al., 1993; Florin et al., 1994a; Han et al., 1996]. In this case, the silicon nitride AFM tip is coated with a thin magnetic film and it is the interaction of this film with the field that induces the tip oscillations. This is a fundamentally cleaner approach; however, there can be issues, including the tip coating affecting the spatial resolution of the AFM, the quality of the coating, and the nature of the sample.

67.4.4 Applications

The breadth of possible applications for scanning probe microscopy seems almost endless. As has been described earlier, the concepts underlying the instrument itself are, arguably, quite simple and to date, the software that drives the instrumentation and the technology itself is effectively turnkey. This does not mean that SPM itself is a very simple tool — it is critical that the user has a good grasp of the physical principles that underpin how an SPM image itself is generated. Similarly, the user must have a good understanding of the nature of their samples, and how they might behave during imaging — an aspect that is of particular interest to those investigating cellular phenomena. In the following sections, we will explore how SPM/AFM-based investigations have provided novel insights into the structure and function of biomolecular assemblies. We focus on a few specific areas, rather than attempting to cover the whole scope of the field. A careful look at the recent reviews by Bottomley et al., will give the reader a sense of the range of topics that are being studied by this and related techniques [Lillehei and Bottomley, 2000, 2001; Poggi et al., 2002, 2004].

SPM has made inroads in a number of different arenas, which can be separated into several key areas (1) imaging; (2) force spectroscopy; and (3) nanomechanical property measurement. We note that it would be difficult to cover all possible applications of this technique and we will restrict our focus to *in situ* studies of biomolecular systems.

67.5 Imaging

The real-space imaging capabilities, coupled with the ability to simultaneously display derivative images, such as phase (viscoelasticity), friction, and temperature is perhaps the most attractive attribute of the scanning probe microscope. In the case of biomolecules, it is the ability to perform such imaging but in buffer media, under a variety of solution conditions and temperatures, in real-time that has really powered the acceptance of this technique by the biomedical community [Conway et al., 2000; Moradian-Oldak et al., 2000; Oesterhelt et al., 2000a; Rochet et al., 2000; Trottier et al., 2000]. Such capabilities are allowing researchers to gain a glimpse of the mechanics and dynamics of protein assembly and function, and the role of extrinsic factors, such as pH, temperature, or other ligands on these processes [Thompson et al., 2000]. For example, *in situ* SPM has been used successfully to visualize and characterize voltage and pH-dependent conformational changes in two-dimensional arrays of OmpF [Moller et al., 1999] while a number of groups have used *in situ* SPM to characterize transcription and DNA-complex formation [Hun Seong et al., 2002; Mukherjee et al., 2002; Seong et al., 2002; Tahirov et al., 2002; Rivetti et al., 2003].

The raster-scanning action of the tip does, however, complicate *in situ* imaging. At the very least, if a process occurs faster than the time required to capture a single image, the SPM may in fact miss the event, or worse, in fact create an artifact associated with the motion of the object. The magnitude of this effect obviously depends on the kinetics of the processes under study. Accordingly, there can be a significant time lag between the first and last scan lines in an SPM image. This is particularly important when one is viewing live cell data where scan times are necessarily slow (∼1 Hz) [Cho et al., 2002; Jena, 2002]. New advances in tip and controller technology are now helping to improve the stability of the instrument under high scan-rate conditions. For instance, special cantilevers are often required when one begins to approach TV scan rates in the SPM. These may include tips with active piezoelectric elements. A particularly useful, and low-cost option for improving time resolution is to simply disable one of the scanning directions so that the AFM image is a compilation of line scans taken at the same location as a function of time [Petsev et al., 2000]. This would therefore generate an isosurface wherein one of the image axes reflects time and not position.

67.6 Crystallography

In situ SPM has been used with great success to study the mechanisms associated with crystal growth [Ward, 2001], from amino acids [Manne et al., 1993], to zeolite crystallization [Agger et al., 2003], and biomineralization [Costa and Maquis, 1998; Teng et al., 1998; Wen et al., 2000]. For protein crystals, studies have ranged from early investigations of lysozyme [Durbin and Feher, 1996], to insulin [Yip et al., 2000], antibodies [Kuznetsov et al., 2000; Plomp et al., 2003], and recently the mechanisms of protein crystal repair [Plomp et al., 2003]. The advantages of SPM for characterizing protein crystallization mechanisms have been very well elucidated in a review by McPherson et al. [2000]. It is worth mentioning that the high spatial and temporal resolution capabilities of the SPM are ideal for examining and measuring the thermodynamic parameters for these processes. These range from local variations in free energy and their correlation with conventional models of crystal growth to a recent study of apoferritin in which step advancement rates were correlated with the product of the density of surface kink sites and the frequency of attachment [Yau and Vekilov, 2000].

A particular challenge for SPM studies of crystal growth is that interpreting the data paradoxically often requires that one already have a known crystal structure or a related isoform for comparison of packing motifs and molecular orientation. Recently the focus has shifted toward understanding the growth process. The ability to perform extended duration *in situ* imaging presents the crystallographer with the unique opportunity of directly determining the mechanisms and kinetics of crystal nucleation and growth. [McPherson et al., 2000; Yau et al., 2000; Yau and Vekilov, 2000; Day et al., 2001; Ko et al., 2001; Kuznetsov et al., 2001a–c; Lucas et al., 2001; McPherson et al., 2001; Yau et al., 2001; Yau and Vekilov, 2001; Chen and Vekilov, 2002; Malkin et al., 2002; Plomp et al., 2002, 2003]. This is arguably a consequence of the inability of the SPM to acquire true real-space three-dimensional images of the interior regions of the proteins. Often, the proteins will appear as amorphous blob to the AFM, even when packed into a lattice and it is therefore difficult to assign a specific secondary structure to the protein. In related work, dissolution studies of small molecule crystals have been very enlightening. Recent work by Danesh et al., resolved the difference between various crystal polymorphs including face-specific dissolution rates for drug candidates [Danesh et al., 2000a, b, 2001] while Guo et al. [2002] examined the effect of specific proteins on the crystallization of calcium oxalate monohydrate. In a particularly interesting study, Frincu et al. [2004] investigated cholesterol crystallization from bile solutions using calcite as a model substrate. In this chapter, the authors were able to use *in situ* SPM to characterize the role of specific substrate interactions in driving the initial nucleation events associated with cholesterol crystallization. Extended-duration imaging allowed the researchers to characterize the growth rates and the onset of Ostwald ripening under physiological conditions. They were able to confirm their observations and models by calculating the interfacial energies associated with the attachment of the cholesterol crystal to the calcite substrate. It is this rather powerful combination of *in situ* real-time characterization with theoretical modeling that has made *in situ* SPM a particularly compelling technique for studying self-assembly at interfaces.

67.6.1 Protein Aggregation and Fibril Formation

In a related context, the self-assembly of proteins into fibrillar motifs has been an area of active research for many years, owing in large part to the putative links to diseases such as Alzheimer's, Huntingtin's, and even diabetes in the context of *in vitro* insulin fibril formation [Waugh et al., 1950; Foster et al., 1951]. *In situ* studies of aggregation and fibrillogenesis by SPM have included collagen [Baselt et al., 1993; Cotterill et al., 1993; Gale et al., 1995; Watanabe et al., 1997; Taatjes et al., 1999], and spider silk [Li et al., 1994; Gould et al., 1999; Miller et al., 1999; Oroudjev et al., 2002]. The clinical implications of fibril and plaque formation and the fact that *in situ* SPM is perhaps the only means of acquiring real-space information on these processes and structures that clinically cannot be easily assayed has driven recent investigations of insulin amyloid polypeptide (IAPP), amylin, beta-amyloid, and synuclein [Harper et al., 1997a, b; Yang et al., 1999; Huang et al., 2000; Roher et al., 2000; McLaurin et al., 2002; Parbhu et al., 2002; Yip et al., 2002; Gorman et al., 2003].

Perhaps driven more by an applied technology perspective, *in situ* SPM has provided unique insights into the role of the nucleating substrate on directing the kinetics, orientation, and structure of the emerging fibril. As noted by Kowalewski in their investigation of beta-amyloid formation on different surfaces, chemically and structurally dissimilar substrate may in fact facilitate growth biasing the apparent kinetics and orientation of the aggregate [Kowalewski and Holtzman, 1999]. Since SPM imaging requires a supporting substrate, there is often a tacit assumption that this surface is passive and would not adversely influence the aggregation or growth process. However, what has become immediately obvious from a number of studies is that these surfaces can, and do, alter the nucleation and growth patterns [Yang et al., 2002; Wang et al., 2003]. Characterization, either theoretical or experimental using the *in situ* capabilities of the SPM, will, in principle, identify how the local physical/electronic/chemical nature of the surface will drive fibril formation [Sherrat et al., 2004]. However, it is clearly important that one be aware of this substrate-directing effect. One must ensure that appropriate controls were in place, or performed, so that the aggregate as seen by the SPM is clearly the responsible agent for nucleation. All of this certainly brings up the questions of (1) is the aggregate observed by these *in situ* tools truly the causative agent; (2) what role is the substrate playing in the aggregation or assembly pathway. The first point is a particularly compelling one when it concerns studies of protein adsorption and assembly. While the SPM can certainly resolve nanometer-sized objects, there always remains a question as to whether the object resolved by SPM is the smallest stable structure or whether there may be a solution species that is in fact smaller. Correlating solution with surface self-assembly mechanisms and structures, especially in the context of biomolecular complexes and phenomena, can be challenging and often one must resort to complementary, corroborative tools such as light scattering.

67.6.2 Membrane Protein Structure and Assemblies

One area in which scanning probe microscopy has made a significant impact has been in the structural characterization of membrane dynamics and protein–membrane interactions and assembly. Supported planar lipid bilayers are particularly attractive as model cell membranes [Sackmann, 1996] and recent work has provided very detailed insights of their local dynamics and structure [Dufrene and Lee, 2000; Jass et al., 2000; Leonenko et al., 2000; Richter et al., 2003], as well as the dynamics of domain formation [Rinia et al., 1999; McKiernan et al., 2000; Yuan et al., 2000; Giocondi et al., 2001b]. Exploiting the *in situ* high resolution imaging capabilities of the SPM, workers have been able to follow thermal phase transitions (gel–fluid) in supported bilayers [Giocondi et al., 2001b; Muresan et al., 2001; Tokumasu et al., 2002], and Langmuir–Blodgett films [Nielsen et al., 2000]. Recently, thermal transitions in mixed composition supported bilayers have been studied by *in situ* SPM [Giocondi et al., 2001a; Giocondi and Le Grimellec, 2004] where the so-called ripple phase domains were seen to form as the system entered the gel–fluid coexistence regime [Leidy et al., 2002].

In situ SPM has also been particularly useful for investigating the dynamics of the so-called lipid raft structures. For example, Rinia et al., investigated the role of cholesterol in the formation of rafts using a complex mixture of dioleoylphosphatidylcholine (DOPC), sphingomyelin (SpM), and cholesterol as the model membrane [Rinia and de Kruijff, 2001]. They were able to demonstrate that the room temperature phase separation seen in the SpM/DOPC bilayers, in the absence of cholesterol, at room temperature was simply a consequence of the gel-state SpM and fluid-state DOPC domains. As the cholesterol content increased, the authors reported the formation of SpM/cholesterol-rich domains or "lipid rafts" within the (DOPC) fluid domains. In related work, Van Duyl et al. [2003] observed similar domain formation for (1 : 1) SpM/DOPC SPBs containing 30 mol% cholesterol [van Duyl et al., 2003].

The effect of dynamically changing the cholesterol levels on raft formation and structure was reported by Lawrence et al. [2003]. By adding either water-soluble cholesterol or methyl-β-cyclodextrin (Mβ-CD), a cholesterol-sequestering agent, the authors were able to directly resolve the effect of adding or removing cholesterol on domain structure and dynamics, including a biphasic response to cholesterol level that was seen as a transient formation of raft domains as the cholesterol level was reduced.

The relative ease with which SPBs can be formed has prompted studies of reconstituted membrane proteins, including ion channels and transmembrane receptors [Lal et al., 1993; Puu et al., 1995, 2000; Takeyasu et al., 1996; Neff et al., 1997; Bayburt et al., 1998; Rinia et al., 2000; Fotiadis et al., 2001; Yuan and Johnston, 2001; Slade et al., 2002]. The premise here is that the SPB provides a membrane-mimicking environment for the protein allowing it to adopt a nominally native orientation at the bilayer surface. It is difficult to know a priori which way the receptor molecules will be oriented in the final supported bilayer since reconstitution occurs via freeze–thaw or sonication into the vesicle/liposome suspension [Radler et al., 1995; Puu and Gustafson, 1997; Jass et al., 2000; Reviakine and Brisson, 2000]. Often one relies on a statistical analysis of local surface topographies, which presumes that there is a distinct difference in the size and shape of the extra- and intracellular domains.

The SPBs have also been used as effective substrates in a vast number of AFM studies of the interactions between protein molecules and membrane surfaces. These studies have included membrane-active and membrane-associated proteins. Peptide-induced changes in membrane morphology and membrane disruption has been directly observed in SPBs in the presence of the amphipathic peptides; filipin, amphotericin B, and mellitin [Santos et al., 1998; Steinem et al., 2000; Milhaud et al., 2002]. In related work, the N-terminal domain of the capsid protein cleavage product of the flock house virus (FHV), has found to cause the formation of interdigitated domains upon exposure of the supported lipid bilayers to the soluble peptide [Janshoff et al., 1999].

Peptide–membrane interactions are also thought to be critical to the mechanism of neurodegenerative diseases such as Alzheimer's (AD) and Parkinson's (PD). For example, studies of α-synuclein with supported lipid bilayers revealed the gradual formation and growth of defects within the SPB [Jo et al., 2000]. Interestingly, the use of a mutant form of the -synuclein protein revealed a qualitatively slower rate of bilayer disruption. We conducted an analogous experiment to investigate the interaction between the amyloid-β (Aβ) peptide with SPBs prepared from a total brain lipid mixture [Yip and McLaurin, 2001]. Interestingly, *in situ* SPM revealed that the association of monomeric Aβ1-40 peptide with the SPB's resulted in rapid formation of fibrils followed by membrane disruption. Control experiments performed with pure component DMPC bilayers revealed similar membrane disruption however the mechanism was qualitatively different with the formation of amorphous aggregates rather than well-formed fibrils.

67.7 Force Spectroscopy

67.7.1 Fundamentals

Although most often used for imaging, by disabling the x- and y-scan directions and monitoring the tip deflection in the z-direction, the AFM is capable of measuring protein–protein and ligand–receptor binding forces, often with sub-piconewton resolution. The ability to detect such low forces is due to the low spring constant of the AFM cantilever (0.60 to 0.06 N/m). In these AFM force curve measurements, the tip is modeled as a Hookian spring whereby the amount of tip deflection (Δz) is directly related to the attractive/repulsive forces (F) acting on the tip through the tip spring constant (k). At the start of the force curve, the AFM tip is held at a null position of zero deflection out of contact with the sample surface. The tip–sample separation distance is gradually reduced and then enlarged using a triangular voltage cycle applied to the piezoelectric scanner. This will bring the tip into and out of contact with the sample surface. As the piezo extends, the sample surface contacts the AFM tip causing the tip to deflect upward until a maximum applied force is reached and the scanner then begins to retract. We should note that when the gradient of the attractive force between the tip and sample exceeds the spring constant of the tip, the tip will "jump" into contact with the sample surface. As the scanner retracts, the upward tip deflection is reduced until it reaches the null position. As the sample continues to move away from the tip, attractive forces between the tip and the surface hold the tip in contact with the surface and the tip begins to deflect in the opposite direction. The tip continues to deflect downward until the restoring force of the tip cantilever overcomes the attractive forces and the tip jumps out of contact with the sample surface (E), thereby providing us with an estimate of the tip–sample unbinding force, given as:

$$F = -k\Delta z$$

This force spectroscopy approach has found application ranging from mapping effect of varying ionic strength on the interactions between charged surfaces [Butt, 1991; Ducker et al., 1991; Senden and Drummond, 1995; Bowen et al., 1998; Liu et al., 2001; Tulpar et al., 2001; Lokar and Ducker, 2002; Mosley et al., 2003; Lokar and Ducker, 2004] [Ducker and Cook, 1990; Ducker et al., 1991, 1994; Butt et al., 1995; Manne and Gaub, 1997; Toikka and Hayes, 1997; Zhang et al., 1997; Hodges, 2002], to studying electrostatic forces at crystal surfaces [Danesh et al., 2000c; Muster and Prestidge, 2002].

67.7.2 Single Molecule Force Spectroscopy

The idea of mapping forces at surfaces rapidly lead to the concept of chemically modifying the SPM tips with ligands so that specific intermolecular interactions can be measured — *single molecule force spectroscopy* [Noy et al., 1995]. In principle, if we can measure the forces associated with the binding of a ligand to its complementary receptor, we may be able to correlate these forces with association energies [Leckband, 2000]. By tethering a ligand of interest, in the correct orientation, to the force microscope tip and bringing the now-modified tip into contact with an appropriately functionalized surface, one can now conceivably directly measure the attractive and repulsive intermolecular forces between single molecules as a function of the tip–sample separation distance. The vertical tip jump during pull-off can be used to estimate the interaction force, which can be related to the number of binding sites, adhesive contact area, and the molecular packing density of the bound molecules. In the case of biomolecular systems, multiple intermolecular interactions exist and both dissociation and (re)association events may occur on the time scale of the experiment resulting in broad retraction curve with discrete, possibly quantized, pull-off events. This approach has been used to investigate a host of interaction forces between biomolecules [Florin et al., 1994b; Hinterdorfer et al., 1996b; Rief et al., 1997a; Smith and Radford, 2000], and DNA–nucleotide interactions [Lee et al., 1994].

Although estimates of the adhesive interaction forces may be obtained from the vertical tip excursions during the retraction phase of the force curve, during pull-off, the width and shape of the retraction curve reflects entropically unfavorable molecular unfolding and elongation processes.

Although simple in principle, it was soon recognized that the force spectroscopy experiment was highly sensitive to sampling conditions. For example, it is now well recognized that the dynamics of the measurement will significantly influence the shape of the unbinding curve. It is well known that the rate of ligand–receptor dissociation increases with force resulting in a logarithmic dependence of the unbinding force with rate [Bell, 1978] and studies have shown that single molecule techniques, such as AFM, clearly sample an interaction energy landscape [Strunz et al., 2000]. It is therefore clear that forces measured by the AFM cannot be trivially related to binding affinities [Merkel et al., 1999]. Beyond these simple sampling rate dependence relationships, we must also be aware of the dynamics of the tip motion during the acquisition phase of the measurement. In particular, when these interactions are mapped in fluid media, one must consider the hydrodynamic drag associated with the (rapid) motion of the tip through the fluid [Janovjak et al., 2004]. This drag effect can be considerable when factored into the interaction force determination.

Another key consideration is that in single molecule force spectroscopy, the ligands of interest are necessarily immobilized at force microscope tips and sample surfaces. In principle, this approach will allow one to directly measure or evaluate the spatial relationship between the ligand and its corresponding receptor site. For correct binding to occur, the ligands of interest must be correctly oriented, have the appropriate secondary and tertiary structure, and be sufficiently flexible (or have sufficiently high unrestricted mobility) that they can bind correctly. An appropriate immobilization strategy would therefore require a priori information about the ligand's sequence, conformation, and the location of the binding sites [Wagner, 1998; Wadu-Mesthrige et al., 2000]. Strategies that have worked in the past include N-nitrilo-triacetic acid linkages [Schmitt et al., 2000] and His-tags to preferentially orient ligands at surface [Ill et al., 1993; Thomson et al., 1999]. More recent efforts have focused on the use of polyethylene glycol tethers to help extend the ligands away from the tip [Kada et al., 2001; Nevo et al., 2003; Stroh et al., 2004], a strategy that has proven to be quite reliable and robust [Hinterdorfer et al., 1996a; Raab et al., 1999; Schmidt et al., 1999; Baumgartner et al., 2000a,b].

67.7.3 Force Volume

Acquiring force curves at each point on an image plane provides a means of acquiring so-called force volume maps, a data-intensive imaging approach capable of providing a map of relative adhesion forces and charge densities across surfaces [Gad et al., 1997; Radmacher, 1997; Heinz and Hoh, 1999b]

[Heinz and Hoh, 1999a; Shellenberger and Logan, 2002]. This approach has been used successfully to examine polymer surfaces and surfaces under fluid [Mizes et al., 1991; van der Werf et al., 1994], as well as live cells [Gad et al., 1997; Nagao and Dvorak, 1998; Walch et al., 2000].

67.7.4 Pulsed Force Mode

As indicated earlier, force volume measurements are very time-consuming and this has led to the development of pulsed force mode imaging [Rosa-Zeiser et al., 1997]. Capable of rapidly acquiring topographic, elasticity, and adhesion data, pulsed force mode operates by sampling selected regions of the force–distance curve during contact-mode imaging. During image scanning, an additional sinusoidal oscillation imparted to the tip brings the tip in- and out-of contact with the surface at each point of the image. Careful analysis of the pulsed force spectrum can yield details about surface elasticity and adhesion [Okabe et al., 2000; Zhang et al., 2000a,b; Fujihira et al., 2001; Schneider et al., 2002; Kresz et al., 2004; Stenert et al., 2004]. Compared with the ~Hz sample rates present in conventional force volume imaging, in pulsed force mode, spectra are acquired on kHz sampling rates. Although this helps to resolve the issue related to the speed of data acquisition, one must clearly consider the possibilities associated (possible) rate-dependence of the adhesion forces, and as indicated in the previous section, the hydrodynamic forces would play a larger role.

67.8 Binding Forces

As discussed earlier, force spectroscopy samples regions of an energy landscape wherein the strength of a bond (and its lifetime) is highly dependent on the rate with which the spectra are collected [Evans and Ritchie, 1999] [Strunz et al., 2000]. At low loading rates, intermolecular bonds have long lifetimes but exhibit small unbinding forces, while at high loading rates, the same bonds will have shorter lifetimes and larger unbinding forces. In the case of biomolecular complexes, since multiple interactions are involved in stabilizing the binding interface, the dissociation pathway of a ligand–receptor complex will exhibit a number of unbinding energy barriers. This would suggest that one could in fact, sample any number of dissociation pathways, each with its own set of transitional bonding interactions.

For the majority of single molecule force microscopy studies, individual ligands have been either randomly adsorbed onto or directly attached to the AFM tip through covalent bond formation. Covalent binding of a molecule to the tip offers a more stable "anchor" during force measurements as a covalent bond is ~10 times stronger than a typical ligand–receptor bond [Grandbois et al., 1999]. Covalent binding also facilitates oriented attachment of the ligand as compared to random adsorption where the orientation of the ligand on the tip surface must be statistically inferred. These advantages are tempered with the challenges present in immobilizing molecules to surfaces such as the AFM tip. As mentioned earlier, oriented ligands have been tethered covalently to AFM tips through use of flexible poly(ethylene-glycol) (PEG)-linkers [Hinterdorfer et al., 1996b]. In this way, the peptide or ligand is extended away from the tip surface, which provides it with sufficient flexibility and conformational freedom for it to reorient and sample conformational space. Heterobifunctional PEG derivatives have provided the necessary synthetic flexibility for coupling a host of different ligands to the AFM tips. [Haselgrubler et al., 1995; Hinterdorfer et al., 1996b] [Willemsen et al., 1998] [Raab et al., 1999; Baumgartner et al., 2000a; Kada et al., 2001; Wielert-Badt et al., 2002; Nevo et al., 2003].

The field of single molecule force spectroscopy comprises two theme areas — the first pertains to mapping or measuring forces between discrete molecules. For example, a number of groups have investigated antibody–antigen interactions [Ros et al., 1998; Allen et al., 1999] and have shown that these unbinding forces may correlate with thermal dissociation rates [Schwesinger et al., 2000]. Force spectroscopy has been used to study the energetics of protein adsorption [Gergely et al., 2000]. Although the high force sensitivity of this approach is exceptionally attractive, it is equally important to recognize key experimental considerations, including the use of appropriate controls. Recently, a number of computational approaches, including steered molecular dynamics [Lu and Schulten, 1999; Marszalek et al., 1999; Baerga-Ortiz et al., 2000; Lu and Schulten, 2000; Isralewitz et al., 2001; Altmann et al., 2002; Gao et al., 2002a,b, 2003; Carrion-Vazquez et al., 2003], Monte Carlo simulations [Clementi et al., 1999], and graphical energy function analyses [Qian and Shapiro, 1999] have been used to simulate these dissociation experiments.

Force spectroscopy is also being applied to study protein-unfolding pathways. It was recognized in the early work that was done during the retraction phase of the AFM force curve, the molecule is subjected to a high tensile stress, and can undergo reversible elongation and unfolding. Careful control over the applied load (and the degree of extension) will allow one to probe molecular elasticity and the energetics involved in the unfolding/folding process [Vesenka et al., 1993; Engel et al., 1999; Fisher et al., 1999; Fotiadis et al., 2002] [Müller et al., 1998; Oesterhelt et al., 2000b; Rief et al., 2000; Best and Clarke, 2002; Muller et al., 2002; Oberhauser et al., 2002; Oroudjev et al., 2002; Rief and Grubmuller, 2002; Zhang et al., 2002; Carrion-Vazquez et al., 2003; Hertadi et al., 2003; Kellermayer et al., 2003; Williams et al., 2003; Janovjak et al., 2004; Schwaiger et al., 2004]. Past studies have included investigations of titin [Oberhauser et al., 2001], IgG phenotypes [Carrion-Vazquez et al., 1999], various polysaccharides [Marszalek et al., 1999], and spider silk proteins [Becker et al., 2003].

By bringing the AFM tip into contact with the surface-adsorbed molecules, and carefully controlling the rate and extent of withdrawal from the surface, it is now possible to resolve transitions that may be ascribed to unfolding of individual protein domains. Others have employed this "forced unfolding" approach to look at spectrin [Rief et al., 1999; Lenne et al., 2000], lysozyme [Yang et al., 2000], and DNA [Clausen-Schaumann et al., 2000]. Caution needs to be exercised during such experiments. Often the protein of interest is allowed to simply absorb to the substrate to form a film. Force curves performed on these films are then conducted in random locations and the retraction phase of the curve analyzed for elongation and unbinding events. This is a highly statistical approach and somewhat problematic. In such a configuration, the general premise is that the tip will bind to the protein somewhere and that if enough samples are acquired, there will be a statistically relevant number of curves that will exhibit the anticipated number of unbinding and unfolding events. What is fundamentally challenging here is that there is no a priori means of knowing where the tip will bind to the protein, which would obviously affect its ability to under extension, and it is difficult to assess the interactions between the protein and the supporting substrate or possibly other entangled proteins.

Where single molecule imaging comes to the forefront is in the combination of imaging and single molecule force spectroscopy. In the past, force spectroscopy has relied heavily on random sampling of the immobilized proteins, often without direct imaging of the selected protein. Recently, Raab et al. combined dynamic force microscopy, wherein a magnetically coated AFM tip is oscillated in close proximity to a surface by analternating magnetic field. This enabled the researchers to apply what they termed "recognition imaging" to facilitate mapping of individual molecular recognition sites on a surface [Raab et al., 1999]. In recognition imaging, specific binding events are detected through dampening of the amplitude of oscillation of the ligand-modified tip due to specific binding of the antibody on the tip to an antigen on the surface. The resulting AFM antibody–antigen recognition image will display regions of enhanced contrast that can be identified as possible binding sites or domains. In an excellent demonstration of the coupled imaging and force spectroscopy, Oesterhelt et al., studied the unfolding of bacteriorhodopsin by directly adsorbing native purple membrane to a surface, imaging the trimeric structure of the BR, and then carefully pulling on a selected molecule [Oesterhelt et al., 2000a]. This allowed them to resolve the force required to destabilize the BR helices from the membrane and by reimaging the same area, show that extraction occurred two helices at a time.

Computationally, these phenomena are most often modeled as worm-like chains [Zhang and Evans, 2001]. To assess what exactly "forced unfolding" involves, Paci and Karplus examined the role of topology and energetics on protein unfolding via externally applied forces and compared it against the more traditional thermal unfolding pathways [Paci and Karplus, 2000].

67.8.1 Mechanical Properties

The use of the AFM/SPM as a nanomechanical tester has certainly blossomed. For example, over the past 10 years, AFM-based nanoindentation has been used to determine the elastic modulus of polymers [Weisenhorn et al., 1993], biomolecules [Vinckier et al., 1996; Laney et al., 1997; Lekka et al., 1999; Parbhu et al., 1999; Suda et al., 1999] [Cuenot et al., 2000], cellular and tissue surfaces [Shroff et al., 1995; Mathur et al., 2000; Velegol and Logan, 2002; Touhami et al., 2003; Alhadlaq et al., 2004; Ebenstein and Pruitt, 2004], pharmaceutical solids [Liao and Wiedmann, 2004] and even teeth [Balooch et al., 2004]. What is particularly challenging in these applications is the need for careful consideration when extrapolating bulk moduli against the nanoindentation data. Often the classical models need to be adjusted in order to compensate for the small (nanometer) contact areas involved in the indentation [Landman et al., 1990]. A particularly important consideration with AFM-based nanoindentation is the sampling geometry. While traditional indentation instrumentation applies a purely vertical load on the sample, by virtue of the cantilever arrangement of the AFM system, there is also a lateral component to the indentation load. This leads to an asymmetry in the indentation profile. This asymmetry can make it difficult to compare AFM-based nanoindentation with traditional approaches using a center-loaded system. Often this effect is nullified by the use of a spherical tip with a well-defined geometry; however, this entails a further compromise in the ability to perform imaging prior to the indentation process. This effect has been extensively covered in the literature, especially in the context of polymer blends and composite materials [Van Landringham et al., 1997a–c, 1999; Bogetti et al., 1999; Bischel et al., 2000]. Other considerations include the relative stiffness of the AFM cantilever, the magnitude of the applied load, tip shape that plays a significant role in the indentation process, and possibly the dwell-time. In many cases, the relatively soft cantilever will allow one to perform more precise modulus measurements including the ability to image prior to, and immediately after, an indentation measurement. At an even more pragmatic level, determining the stiffness both in-plane and torsional, of the cantilever can be challenging, with approaches ranging from the traditional end-mass to new techniques based on thermal noise and resonant frequency shifts [Cleveland et al., 1993; Hutter and Bechhoefer, 1993; Bogdanovic et al., 2000]. Accurate determination of these values is essential in order for the correct assessment of the local stiffness to be made.

67.8.2 Coupled Imaging

While AFM/SPM is certainly a powerful tool for following structure and dynamics at surfaces under a wide variety of conditions, it can only provide relative information within a given imaging frame. It similarly cannot confirm (easily) that the structure being imaged is in fact the protein of interest. It could in fact be said that SPM images are artefactual until proven otherwise. This confirmation step can involve some *in situ* control, which might be a change in pH or T, or introduction of another ligand or reagent that would cause a change in the same that could be resolved by the SPM. Absent an *in situ* control, or in fact as an adjunct, careful shape/volume analysis is often conducted to characterize specific features in a sample. Image analysis and correlation tools and techniques are often exploited for postacquisition analysis. There has always been an obvious need to techniques or tools that can provide this complementary information, ideally in a form that could be readily integrated into the SPM.

Optical imaging represents perhaps the best tool for integration with scanning probe microscopy. This is motivated by the realization that there are a host of very powerful single molecule optical imaging techniques capable of addressing many of the key limitations of SPM, such as the ability to resolve dynamic events on millisecond time scales. Recent advances in confocal laser scanning (CLSM) and total internal reflectance fluorescence (TIRFM) techniques have enabled single molecule detection with subdiffraction limited images [Ambrose et al., 1999; Sako et al., 2000a,b; Osborne et al., 2001; Ludes and Wirth, 2002; Sako and Uyemura, 2002; Borisenko et al., 2003; Cannone et al., 2003; Michalet et al., 2003; Wakelin and Bagshaw, 2003].

67.8.3 Near-Field — SNOM/NSOM

In the family of scanning probe microscopes, perhaps the best example of an integrated optical-SPM system are the scanning near-field (or near-field scanning) optical microscopes (SNOM or NSOM), which use near-field excitation of the sample to obtained subdiffraction limited images with spatial resolution comparable to conventional scanning probe microscopes [Muramatsu et al., 1995; Sekatskii et al., 2000; de Lange et al., 2001; Edidin 2001; Harris 2003]. NSOM has been used successfully in single molecule studies of dyes [Betzig and Chichester, 1993], proteins [Moers et al., 1995; Garcia-Parajo et al., 1999; van Hulst et al., 2000], and the structure of lignin and ion channels [Ianoul et al., 2004; Micic et al., 2004]. NSOM imaging has also provided insights into ligand-induced clustering of the ErbB2 receptor, a member of the epidermal growth factor (EGF) receptor tyrosine kinase family, in the membrane of live cells [Nagy et al., 1999]. Fluorescence lifetime imaging by NSOM has been used to examine the energy and electron-transfer processes of the light harvesting complex (LHC II) [Hosaka and Saiki, 2001; Sommer and Franke, 2002] in intact photosynthetic membranes [Dunn et al., 1994]. NSOM has also been used to monitor the fluorescence resonance energy transfer (FRET) between single pairs of donor and acceptor fluorophores on dsDNA molecules [Ha et al., 1996]. Challenges that face the NSOM community arguably lie in the robust design of the imaging tips [Burgos et al., 2003; Prikulis et al., 2003].

67.8.4 Evanescent-Wave — TIRF (Total Internal Reflection Fluorescence Microscopy)

Time resolved single molecule imaging can be difficult and in the case of the AFM, one may question whether the local phenomena imaged by AFM is specific to that particular imaging location. This is especially true for studies of dynamic phenomena since the scanning action of the AFM tip effectively acts to increase mass transfer into the imaging volume. Recently, combined AFM/TIRF techniques have been used to study force transmission [Mathur et al., 2000] and single-particle manipulation [Nishida et al., 2002] in cells. These studies helped to address a particularly challenging aspect of scanning probe microscopy, which was that SPM/AFM can only (realistically) infer data about the upper surface of structures and that data on the underside of a structure, for instance the focal adhesions of a cell, are largely invisible to the SPM tip. In the case of cell adhesion, one might be interested in how a cell responds to a local stress applied to its apical surface by monitoring changes in focal adhesion density and size. Using a combined AFM-TIRF system, it then becomes possible to directly interrogate the basal surface of the cell (by TIRF) while applying a load or examining the surface topography of the cell by *in situ* AFM. We recently reported on the design and use of a AFM — objective-based TIRF-based instrument for the study of supported bilayer systems [Shaw et al., 2003]. By coupling these two instruments together we were able to identify unequivocally the gel and fluid domains in a mixed dPOPC/dPPC system. What was particularly compelling was the observation of ~10 to 20% difference in the lateral dimension of the features as resolved by TIRF and AFM. While this likely reflects the inherent diffraction limited nature of TIRFM, we can in fact use the AFM data to confirm the real-space size of the structures that are responsible for the fluorescence image contrast. This combined system also provided another interesting insight. The nonuniform fluorescence intensity across the domains resolved by TIRF may reflect a nonuniform distribution of NBD-PC within dPOPC. It may also be linked to the time required to capture a TIRF image relative to the AFM imaging. At a typical scan rate of 2 Hz, it would require ~4 min to capture a conventional 512 × 512 pixel AFM image, compared with the ~30 frame/sec video imaging rate of the TIRF camera system. As such the TIRFM system represents an excellent means of visualizing and capturing data that occur on time scales faster than what can be readily resolved by the AFM. This further suggests that the differences in fluorescence intensity may reflect real-time fluctuations in the structure of the lipid bilayer that are not detected (or detectable) by AFM imaging. In a particularly intriguing experiment that used TIRF as an excitation source rather than in an imaging mode, Hugel and others were able to measure the effect of a conformational change on the relative stiffness of a photosensitive polymer

[Hugel et al., 2002]. By irradiating the sample *in situ*, they were able to initiate a *cis–trans* conformational change that resulted in a change in the backbone conformation of the polymer.

67.9 Summary

As can be readily seen in the brief survey of the SPM field, it is clearly expanding both in terms of technique and range of applications. The systems are becoming more ubiquitous and certainly more approachable by the general user; however, what is clearly important is that care must be taken in data interpretation, instrument control, and sample preparation. For example, early studies of intermolecular forces often did not exercise the same level of control over their sampling conditions as is commonplace today and this clearly impacts critical analysis of the resulting force spectra. Recognizing the limitations of the tools and hopefully developing strategies that help to overcome these limitations represent a key goal for many SPM users.

New innovations in integrated single molecule correlated functional imaging tools will certainly continue to drive advances in this technology. As we have seen, fluorescence imaging, either as NSOM/TIRF/CSLM, when coupled with SPM provides an excellent *in situ* tool for characterizing biomolecular interactions and phenomena. Unfortunately, such a scheme requires specific labeling strategies and it would be preferable to effect such measurements in the absence of a label. Recent work has focused on near-field vibrational microscopy to acquire both IR and Raman spectra on nanometer length scales [Knoll et al., 2001; Hillenbrand et al., 2002; Anderson and Gaimari, 2003], while a combined Raman-SPM system was used to characterize the surface of an insect compound eye [Anderson and Gaimari, 2003]. These exciting new developments are clear evidence that the field of SPM is not a mature one but one that in fact continues to accelerate.

References

A-Hassan, E., Heinz, W.F., Antonik, M., D'Costa, N.P., Nageswaran, S., Schoenenberger, C.-A., and Hoh, J.H. (1998). Relative microelastic mapping of living cells by atomic force microscopy. *Biophys. J.* 74: 1564–1578.

Agger, J.R., Hanif, N., Cundy, C.S., Wade, A.P., Dennison, S., Rawlinson, P.A., and Anderson, M.W. (2003). Silicalite crystal growth investigated by atomic force microscopy. *J. Am. Chem. Soc.* 125: 830–839.

Alhadlaq, A., Elisseeff, J.H., Hong, L., Williams, C.G., Caplan, A.I., Sharma, B., Kopher, R.A., Tomkoria, S., Lennon, D.P., Lopez, A. et al. (2004). Adult stem cell driven genesis of human-shaped articular condyle. *Ann. Biomed. Eng.* 32: 911–923.

Allen, S., Davies, J., Davies, M.C., Dawkes, A.C., Roberts, C.J., Tendler, S.J., and Williams, P.M. (1999). The influence of epitope availability on atomic-force microscope studies of antigen-antibody interactions. *Biochem. J.* 341: 173–178.

Altmann, S.M., Grunberg, R.G., Lenne, P.F., Ylanne, J., Raae, A., Herbert, K., Saraste, M., Nilges, M., and Horber, J.K. (2002). Pathways and intermediates in forced unfolding of spectrin repeats. *Structure (Camb.)* 10: 1085–1096.

Ambrose, W.P., Goodwin, P.M., and Nolan, J.P. (1999). Single-molecule detection with total internal reflectance excitation: comparing signal-to-background and total signals in different geometries. *Cytometry* 36: 224–231.

Anderson, M.S. and Gaimari, S.D. (2003). Raman-atomic force microscopy of the ommatidial surfaces of dipteran compound eyes. *J. Struct. Biol.* 142: 364–368.

Argaman, M., Golan, R., Thomson, N.H., and Hansma, H.G. (1997). Phase imaging of moving DNA molecules and DNA molecules replicated in the atomic force microscope. *Nucleic Acids Res.* 25: 4379–4384.

Baerga-Ortiz, A., Rezaie, A.R., and Komives, E.A. (2000). Electrostatic dependence of the thrombin-thrombomodulin interaction. *J. Mol. Biol.* 296: 651–658.

Balooch, G., Marshall, G.W., Marshall, S.J., Warren, O.L., Asif, S.A., and Balooch, M. (2004). Evaluation of a new modulus mapping technique to investigate microstructural features of human teeth. *J. Biomech.* 37: 1223–1232.

Baselt, D.R., Revel, J.P., and Baldeschwieler, J.D. (1993). Subfibrillar structure of type I collagen observed by atomic force microscopy. *Biophys. J.* 65: 2644–2655.

Baumgartner, W., Hinterdorfer, P., Ness, W., Raab, A., Vestweber, D., Schindler, H., and Drenckhahn, D. (2000a). Cadherin interaction probed by atomic force microscopy. *Proc. Natl Acad. Sci. USA* 97: 4005–4010.

Baumgartner, W., Hinterdorfer, P., and Schindler, H. (2000b). Data analysis of interaction forces measured with the atomic force microscope. *Ultramicroscopy* 82: 85–95.

Bayburt, T.H., Carlson, J.W., and Sligar, S.G. (1998). Reconstitution and imaging of a membrane protein in a nanometer-size phospholipid bilayer. *J. Struct. Biol.* 123: 37–44.

Becker, N., Oroudjev, E., Mutz, S., Cleveland, J.P., Hansma, P.K., Hayashi, C.Y., Makarov, D.E., and Hansma, H.G. (2003). Molecular nanosprings in spider capture-silk threads. *Nat. Mater.* 2: 278–283.

Bell, G.I. (1978). Models for the specific adhesion of cells to cells. *Science* 200: 618–627.

Best, R.B. and Clarke, J. (2002). What can atomic force microscopy tell us about protein folding? *Chem. Commun. (Camb.)*: 183–192.

Betzig, E. and Chichester, R.J. (1993). Single molecules observed by near-field scanning optical microscopy. *Science* 262: 1422–1425.

Binnig, G., Quate, C.F., and Gerber, C. (1986). Atomic force microscope. *Phys. Rev. Lett.* 56: 930–933.

Binnig, G., Rohrer, H., Gerber, C., and Weibel, E. (1982). Tunneling through a controllable vacuum gap. *Rev. Modern Phys.* 59: 178–180.

Bischel, M.S., Van Landringham, M.R., Eduljee, R.F., Gillespie, J.W.J., and Schultz, J.M. (2000). On the use of nanoscale indentation with the AFM in the identification of phases in blends on linear low density polyethylene and high density polyethylene. *J. Mat. Sci.* 35: 221–228.

Bogdanovic, G., Meurk, A., and Rutland, M.W. (2000). Tip friction–torsional spring constant determination. *Colloids Surf. B Biointerfaces* 19: 397–405.

Bogetti, T.A., Wang, T., Van Landringham, M.R., Eduljee, R.F., and Gillespie, J.W.J. (1999). Characterization of nanoscale property variations in polymer composite systems: Part 2 — Finite element modeling. *Composites Part A* 30: 85–94.

Borisenko, V., Lougheed, T., Hesse, J., Fureder-Kitzmuller, E., Fertig, N., Behrends, J.C., Woolley, G.A., and Schutz, G.J. (2003). Simultaneous optical and electrical recording of single gramicidin channels. *Biophys. J.* 84: 612–622.

Bowen, W.R., Hilal, N., Lovitt, R.W., and Wright, C.J. (1998). Direct measurement of interactions between adsorbed protein layers using an atomic force microscope. *J. Colloid. Interface Sci.* 197: 348–352.

Brandsch, R., Bar, G., and Whangbo, M.-H. (1997). On the factors affecting the contrast of height and phase images in tapping mode atomic force microscopy. *Langmuir* 13: 6349–6353.

Brown, H.G. and Hoh, J.H. (1997). Entropic exclusion by neurofilament sidearms: a mechanism for maintaining interfilament spacing. *Biochemistry* 36: 15035–15040.

Burgos, P., Lu, Z., Ianoul, A., Hnatovsky, C., Viriot, M.L., Johnston, L.J., and Taylor, R.S. (2003). Near-field scanning optical microscopy probes: a comparison of pulled and double-etched bent NSOM probes for fluorescence imaging of biological samples. *J. Microsc.* 211: 37–47.

Butt, H. (1991). Measuring electrostatic, van der Waals, and hydration forces in electrolyte solutions with an atomic force microscope. *Biophys. J.* 60: 1438–1444.

Butt, H.-J., Jaschke, M., and Ducker, W. (1995. Measuring surface forces in aqueous electrolyte solution with the atomic force microscopy. *Bioelectrochem. Bioenerg.* 38: 191–201.

Cannone, F., Chirico, G., and Diaspro, A. (2003). Two-photon interactions at single fluorescent molecule level. *J. Biomed. Opt.* 8: 391–395.

Carrion-Vazquez, M., Li, H., Lu, H., Marszalek, P.E., Oberhauser, A.F., and Fernandez, J.M. (2003). The mechanical stability of ubiquitin is linkage dependent. *Nat. Struct. Biol.*

Carrion-Vazquez, M., Marszalek, P.E., Oberhauser, A.F., and Fernandez, J.M. (1999). Atomic force microscopy captures length phenotypes in single proteins. *Proc. Natl Acad. Sci. USA* 96: 11288–11292.

Chen, K. and Vekilov, P.G. (2002). Evidence for the surface-diffusion mechanism of solution crystallization from molecular-level observations with ferritin. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 66: 021606.

Cho, S.J., Quinn, A.S., Stromer, M.H., Dash, S., Cho, J., Taatjes, D.J., and Jena, B.P. (2002). Structure and dynamics of the fusion pore in live cells. *Cell Biol. Int.* 26: 35–42.

Clausen-Schaumann, H., Rief, M., Tolksdorf, C., and Gaub, H.E. (2000). Mechanical stability of single DNA molecules. *Biophys. J.* 78: 1997–2007.

Clementi, C., Carloni, P., and Maritan, A. (1999). Protein design is a key factor for subunit–subunit association. *Proc. Natl Acad. Sci. USA* 96: 9616–9621.

Cleveland, J.P., Manne, S., Bocek, D., and Hansma, P.K. (1993). A nondestructive method for determining the spring constant of cantilevers for scanning force microscopy. *Rev. Sci. Instrum.* 64: 403–405.

Conway, K.A., Harper, J.D., and Lansbury, P.T., Jr. (2000). Fibrils formed *in vitro* from alpha-synuclein and two mutant forms linked to Parkinson's disease are typical amyloid. *Biochemistry* 39: 2552–2563.

Costa, N. and Maquis, P.M. (1998). Biomimetic processing of calcium phosphate coating. *Med. Eng. Phys.* 20: 602–606.

Cotterill, G.F., Fergusson, J.A., Gani, J.S., and Burns, G.F. (1993). Scanning tunnelling microscopy of collagen I reveals filament bundles to be arranged in a left-handed helix. *Biochem. Biophys. Res. Commun.* 194: 973–977.

Cuenot, S., Demoustier-Champagne, S., and Nysten, B. (2000). Elastic modulus of polypyrrole nanotubes. *Phys. Rev. Lett.* 85: 1690–1693.

Czajkowsky, D.M., Allen, M.J., Elings, V., and Shao, Z. (1998). Direct visualization of surface charge in aqueous solution. *Ultramicroscopy* 74: 1–5.

Danesh, A., Chen, X., Davies, M.C., Roberts, C.J., Sanders, G.H., Tendler, S.J., Williams, P.M., and Wilkins, M.J. (2000a). The discrimination of drug polymorphic forms from single crystals using atomic force microscopy. *Pharm. Res.* 17: 887–890.

Danesh, A., Chen, X., Davies, M.C., Roberts, C.J., Sanders, G.H.W., Tendler, S.J.B., and Williams, P.M. (2000b). Polymorphic discrimination using atomic force microscopy: distinguishing between two polymorphs of the drug cimetidine. *Langmuir* 16: 866–870.

Danesh, A., Connell, S.D., Davies, M.C., Roberts, C.J., Tendler, S.J., Williams, P.M., and Wilkins, M.J. (2001). An *in situ* dissolution study of aspirin crystal planes (100) and (001) by atomic force microscopy. *Pharm. Res.* 18: 299–303.

Danesh, A., Davies, M.C., Hinder, S.J., Roberts, C.J., Tendler, S.J., Williams, P.M., and Wilkins, M.J. (2000c). Surface characterization of aspirin crystal planes by dynamic chemical force microscopy. *Anal. Chem.* 72: 3419–3422.

Day, J., Kuznetsov, Y.G., Larson, S.B., Greenwood, A., and McPherson, A. (2001). Biophysical studies on the RNA cores of satellite tobacco mosaic virus. *Biophys. J.* 80: 2364–2371.

de Lange, F., Cambi, A., Huijbens, R., de Bakker, B., Rensen, W., Garcia-Parajo, M., van Hulst, N., and Figdor, C.G. (2001). Cell biology beyond the diffraction limit: near-field scanning optical microscopy. *J. Cell Sci.* 114: 4153–4160.

Deleu, M., Nott, K., Brasseur, R., Jacques, P., Thonart, P., and Dufrene, Y.F. (2001). Imaging mixed lipid monolayers by dynamic atomic force microscopy. *Biochim. Biophys. Acta* 1513: 55–62.

Dinte, B.P., Watson, G.S., Dobson, J.F., and Myhra, S. (1996). Artefacts in non-contact mode force microscopy: the role of adsorbed moisture. *Ultramicroscopy* 63: 115–124.

Ducker, W.A. and Cook, R.F. (1990). Rapid measurement of static and dynamic surface forces. *Appl. Phys. Lett.* 56: 2408–2410.

Ducker, W.A., Senden, T.J., and Pashley, R.M. (1991). Direct measurement of colloidal forces using an atomic force microscope. *Nature* 353: 239–241.

Ducker, W.A., Xu, Z., and Israelachvili, J.N. (1994). Measurements of hydrophobic and DLVO forces in bubble-surface interactions in aqueous solutions. *Langmuir* 10: 3279–3289.

Dufrene, Y.F. and Lee, G.U. (2000). Advances in the characterization of supported lipid films with the atomic force microscope. *Biochim. Biophys. Acta* 1509: 14–41.

Dunn, R.C., Holtom, G.R., Mets, L., and Xie, X.S. (1994). Near-field imaging and fluorescence lifetime measurement of light harvesting complexes in intact photosynthetic membranes. *J. Phys. Chem.* 98: 3094–3098.

Durbin, S.D. and Feher, G. (1996). Protein crystallization. *Annu. Rev. Phys. Chem.* 47: 171–204.

Ebenstein, D.M. and Pruitt, L.A. (2004). Nanoindentation of soft hydrated materials for application to vascular tissues. *J. Biomed. Mater. Res.* 69A: 222–232.

Edidin, M. (2001). Near-field scanning optical microscopy, a siren call to biology. *Traffic* 2: 797–803.

Engel, A., Gaub, H.E., and Muller, D.J. (1999). Atomic force microscopy: a forceful way with single molecules. *Curr. Biol.* 9: R133–136.

Engel, A. and Muller, D.J. (2000). Observing single biomolecules at work with the atomic force microscope. *Nat. Struct. Biol.* 7: 715–718.

Evans, E. and Ritchie, K. (1999). Strength of a weak bond connecting flexible polymer chains. *Biophys. J.* 76: 2439–2447.

Fisher, T.E., Marszalek, P.E., Oberhauser, A.F., Carrion-Vazquez, M., and Fernandez, J.M. (1999). The micro-mechanics of single molecules studied with atomic force microscopy. *J. Physiol.* 520: 5–14.

Florin, E.-L., Radmacher, M., Fleck, B., and Gaub, H.E. (1994a). Atomic force microscope with magnetic force modulation. *Rev. Sci. Instrum.* 65: 639–643.

Florin, E.L., Moy, V.T., and Gaub, H.E. (1994b). Adhesion forces between individual ligand–receptor pairs. *Science* 264: 415–417.

Foster, G.E., Macdonald, J., and Smart, J.V. (1951). The assay of insulin *in vitro* by fibril formation and precipitation. *J. Pharm. Pharmacol.* 3: 897–904.

Fotiadis, D., Jeno, P., Mini, T., Wirtz, S., Muller, S.A., Fraysse, L., Kjellbom, P., and Engel, A. (2001). Structural characterization of two aquaporins isolated from native spinach leaf plasma membranes. *J. Biol. Chem.* 276: 1707–1714.

Fotiadis, D., Scheuring, S., Muller, S.A., Engel, A., and Muller, D.J. (2002). Imaging and manipulation of biological structures with the AFM. *Micron* 33: 385–397.

Frincu, M.C., Fleming, S.D., Rohl, A.L., and Swift, J.A. (2004). The epitaxial growth of cholesterol crystals from bile solutions on calcite substrates. *J. Am. Chem. Soc.* 126: 7915–7924.

Fritzsche, W. and Henderson, E. (1997). Mapping elasticity of rehydrated metaphase chromosomes by scanning force microscopy. *Ultramicroscopy* 69: 191–200.

Fujihira, M., Furugori, M., Akiba, U., and Tani, Y. (2001). Study of microcontact printed patterns by chemical force microscopy. *Ultramicroscopy* 86: 75–83.

Gad, M., Itoh, A., and Ikai, A. (1997). Mapping cell wall polysaccharides of living microbial cells using atomic force microscopy. *Cell. Biol. Int.* 21: 697–706.

Gale, M., Pollanen, M.S., Markiewicz, P., and Goh, M.C. (1995). Sequential assembly of collagen revealed by atomic force microscopy. *Biophys. J.* 68: 2124–2128.

Gao, M., Craig, D., Lequin, O., Campbell, I.D., Vogel, V., and Schulten, K. (2003). Structure and functional significance of mechanically unfolded fibronectin type III1 intermediates. *Proc. Natl Acad. Sci. USA* 100: 14784–14789.

Gao, M., Craig, D., Vogel, V., and Schulten, K. (2002a). Identifying unfolding intermediates of FN-III(10) by steered molecular dynamics. *J. Mol. Biol.* 323: 939–950.

Gao, M., Wilmanns, M., and Schulten, K. (2002b). Steered molecular dynamics studies of titin i1 domain unfolding. *Biophys. J.* 83: 3435–3445.

Garcia-Parajo, M.F., Veerman, J.A., Segers-Nolten, G.M., de Grooth, B.G., Greve, J., and van Hulst, N.F. (1999). Visualising individual green fluorescent proteins with a near field optical microscope. *Cytometry* 36: 239–246.

Gergely, C., Voegel, J., Schaaf, P., Senger, B., Maaloum, M., Horber, J.K., and Hemmerle, J. (2000). Unbinding process of adsorbed proteins under external stress studied by atomic force microscopy spectroscopy. *Proc. Natl Acad. Sci. USA* 97: 10802–10807.

Giocondi, M.-C., Vie, V., Lesniewska, E., Milhiet, P.-E., Zinke-Allmang, M., and Le Grimellec, C. (2001a). Phase topology and growth of single domains in lipid bilayers. *Langmuir* 17: 1653–1659.

Giocondi, M.C. and Le Grimellec, C. (2004). Temperature dependence of the surface topography in dimyristoylphosphatidylcholine/distearoylphosphatidylcholine multibilayers. *Biophys. J.* 86: 2218–2230.

Giocondi, M.C., Pacheco, L., Milhiet, P.E., and Le Grimellec, C. (2001b). Temperature dependence of the topology of supported dimirystoyl-distearoyl phosphatidylcholine bilayers. *Ultramicroscopy* 86: 151–157.

Gorman, P.M., Yip, C.M., Fraser, P.E., and Chakrabartty, A. (2003). Alternate aggregation pathways of the Alzheimer beta-amyloid peptide: abeta association kinetics at endosomal pH. *J. Mol. Biol.* 325: 743–757.

Gould, S.A., Tran, K.T., Spagna, J.C., Moore, A.M., and Shulman, J.B. (1999). Short and long range order of the morphology of silk from Latrodectus hesperus (Black Widow) as characterized by atomic force microscopy. *Int. J. Biol. Macromol.* 24: 151–157.

Grandbois, M., Beyer, M., Rief, M., Clausen-Schaumann, H., and Gaub, H.E. (1999). How strong is a covalent bond? *Science* 283: 1727–1730.

Guo, S., Ward, M.D., and Wesson, J.A. (2002). Direct visualization of calcium oxalate monohydrate crystallization and dissolution with atomic force microscopy and the role of polymeric additives. *Langmuir* 18: 4282–4291.

Ha, T., Enderle, T., Ogletree, D.F., Chemla, D.S., Selvin, P.R., and Weiss, S. (1996). Probing the interaction between two single molecules: fluorescence resonance energy transfer between a single donor and a single acceptor. *Proc. Natl Acad. Sci. USA* 93: 6264–6268.

Han, W., Lindsay, S.M., and Jing, T. (1996). A magnetically driven oscillating probe microscope for operation in liquids. *Appl. Phys. Lett.* 69: 4111–4113.

Hansma, H.G., Kim, K.J., Laney, D.E., Garcia, R.A., Argaman, M., Allen, M.J., and Parsons, S.M. (1997). Properties of biomolecules measured from atomic force microscope images: a review. *J. Struct. Biol.* 119: 99–108.

Hansma, P.K., Elings, V., Marti, O., and Bracker, C.E. (1988). Scanning tunneling microscopy and atomic force microscopy: application to biology and technology. *Science* 242: 209–216.

Harper, J.D., Lieber, C.M., and Lansbury, P.T., Jr. (1997a). Atomic force microscopic imaging of seeded fibril formation and fibril branching by the Alzheimer's disease amyloid-beta protein. *Chem. Biol.* 4: 951–959.

Harper, J.D., Wong, S.S., Lieber, C.M., and Lansbury, P.T. (1997b). Observation of metastable Abeta amyloid protofibrils by atomic force microscopy. *Chem. Biol.* 4: 119–125.

Harris, C.M. (2003). Shedding light on NSOM. *Anal. Chem.* 75: 223A–228A.

Haselgrubler, T., Amerstorfer, A., Schindler, H., and Gruber, H.J. (1995). Synthesis and applications of a new poly(ethylene glycol) derivative for the crosslinking of amines with thiols. *Bioconjug. Chem.* 6: 242–248.

Heinz, W.F. and Hoh, J.H. (1999a). Relative surface charge density mapping with the atomic force microscope. *Biophys. J.* 76: 528–538.

Heinz, W.F. and Hoh, J.H. (1999b). Spatially resolved force spectroscopy of biological surfaces using the atomic force microscope. *Trends Biotechnol.* 17: 143–150.

Hertadi, R., Gruswitz, F., Silver, L., Koide, A., Koide, S., Arakawa, H., and Ikai, A. (2003). Unfolding mechanics of multiple OspA substructures investigated with single molecule force spectroscopy. *J. Mol. Biol.* 333: 993–1002.

Hillenbrand, R., Taubner, T., and Keilmann, F. (2002). Phonon-enhanced light matter interaction at the nanometre scale. *Nature* 418: 159–162.

Hinterdorfer, P., Baumgartner, W., Gruber, H.J., and Schilcher, K. (1996a). Detection and localization of individual antibody-antigen recognition events by atomic force microscopy. *Proc. Natl Acad. Sci. USA* 93: 3477–3481.

Hinterdorfer, P., Baumgartner, W., Gruber, H.J., Schilcher, K., and Schindler, H. (1996b). Detection and localization of individual antibody-antigen recognition events by atomic force microscopy. *Proc. Natl Acad. Sci. USA* 93: 3477–3481.

Hodges, C.S. (2002). Measuring forces with the AFM: polymeric surfaces in liquids. *Adv. Colloid Interface Sci.* 99: 13–75.

Holland, N.B. and Marchant, R.E. (2000). Individual plasma proteins detected on rough biomaterials by phase imaging AFM. *J. Biomed. Mater. Res.* 51: 307–315.

Hosaka, N. and Saiki, T. (2001). Near-field fluorescence imaging of single molecules with a resolution in the range of 10 nm. *J. Microsc.* 202: 362–364.

Huang, T.H., Yang, D.S., Plaskos, N.P., Go, S., Yip, C.M., Fraser, P.E., and Chakrabartty, A. (2000). Structural studies of soluble oligomers of the Alzheimer beta-amyloid peptide. *J. Mol. Biol.* 297: 73–87.

Hugel, T., Holland, N.B., Cattani, A., Moroder, L., Seitz, M., and Gaub, H.E. (2002). Single-molecule optomechanical cycle. *Science* 296: 1103–1106.

Hun Seong, G., Kobatake, E., Miura, K., Nakazawa, A., and Aizawa, M. (2002). Direct atomic force microscopy visualization of integration host factor-induced DNA bending structure of the promoter regulatory region on the Pseudomonas TOL plasmid. *Biochem. Biophys. Res. Commun.* 291: 361–366.

Hutter, J.L. and Bechhoefer, J. (1993). Calibration of atomic-force microscope tips. *Rev. Sci. Instrum.* 64: 1868–1873.

Ianoul, A., Street, M., Grant, D., Pezacki, J., Taylor, R.S., and Johnston, L.J. (2004). Near-field scanning fluorescence microscopy study of ion channel clusters in cardiac myocyte membranes. *Biophys. J.*

Ill, C.R., Keivens, V.M., Hale, J.E., Nakamura, K.K., Jue, R.A., Cheng, S., Melcher, E.D., Drake, B., and Smith, M.C. (1993). A COOH-terminal peptide confers regiospecific orientation and facilitates atomic force microscopy of an IgG1. *Biophys. J.* 64: 919–924.

Isralewitz, B., Gao, M., and Schulten, K. (2001). Steered molecular dynamics and mechanical functions of proteins. *Curr. Opin. Struct. Biol.* 11: 224–230.

Janovjak, H., Struckmeier, J., and Muller, D.J. (2004). Hydrodynamic effects in fast AFM single-molecule force measurements. *Eur. Biophys. J.*

Janshoff, A., Bong, D.T., Steinem, C., Johnson, J.E., and Ghadiri, M.R. (1999). An animal virus-derived peptide switches membrane morphology: possible relevance to nodaviral tranfection processes. *Biochemistry* 38: 5328–5336.

Jass, J., Tjarnhage, T., and Puu, G. (2000). From liposomes to supported, planar bilayer structures on hydrophilic and hydrophobic surfaces: an atomic force microscopy study. *Biophys. J.* 79: 3153–3163.

Jena, B.P. (2002). Fusion pore in live cells. *News Physiol. Sci.* 17: 219–222.

Jo, E., McLaurin, J., Yip, C.M., St George-Hyslop, P., and Fraser, P.E. (2000). α-synuclein membrane interactions and lipid specificity. *J. Biol. Chem.* 275: 34328–34334.

Kada, G., Blayney, L., Jeyakumar, L.H., Kienberger, F., Pastushenko, V.P., Fleischer, S., Schindler, H., Lai, F.A., and Hinterdorfer, P. (2001). Recognition force microscopy/spectroscopy of ion channels: applications to the skeletal muscle Ca2+ release channel (RYR1). *Ultramicroscopy* 86: 129–137.

Kellermayer, M.S., Bustamante, C., and Granzier, H.L. (2003). Mechanics and structure of titin oligomers explored with atomic force microscopy. *Biochim. Biophys. Acta* 1604: 105–114.

Knoll, A., Magerle, R., and Krausch, G. (2001). Tapping mode atomic force microscopy on polymers: where is the true sample surface? *Macromolecules* 34: 4159–4165.

Ko, T.P., Kuznetsov, Y.G., Malkin, A.J., Day, J., and McPherson, A. (2001). X-ray diffraction and atomic force microscopy analysis of twinned crystals: rhombohedral canavalin. *Acta Crystallogr. D Biol. Crystallogr.* 57: 829–839.

Kowalewski, T. and Holtzman, D.M. (1999). *In situ* atomic force microscopy study of Alzheimer's beta-amyloid peptide on different substrates: new insights into mechanism of beta-sheet formation. *Proc. Natl Acad. Sci. USA* 96: 3688–3693.

Kresz, N., Kokavecz, J., Smausz, T., Hopp, B., Csete, A., Hild, S., and Marti, O. (2004). Investigation of pulsed laser deposited crystalline PTFE thin layer with pulsed force mode AFM. *Thin Solid Films* 453–454: 239–244.

Krol, S., Ross, M., Sieber, M., Kunneke, S., Galla, H.J., and Janshoff, A. (2000). Formation of three-dimensional protein-lipid aggregates in monolayer films induced by surfactant protein B. *Biophys. J.* 79: 904–918.

Kuznetsov, Y.G., Larson, S.B., Day, J., Greenwood, A., and McPherson, A. (2001a). Structural transitions of satellite tobacco mosaic virus particles. *Virology* 284: 223–234.

Kuznetsov, Y.G., Malkin, A.J., Lucas, R.W., and McPherson, A. (2000). Atomic force microscopy studies of icosahedral virus crystal growth. *Colloids Surf. B Biointerfaces* 19: 333–346.

Kuznetsov, Y.G., Malkin, A.J., Lucas, R.W., Plomp, M., and McPherson, A. (2001b). Imaging of viruses by atomic force microscopy. *J. Gen. Virol.* 82: 2025–2034.

Kuznetsov, Y.G., Malkin, A.J., and McPherson, A. (2001c). Self-repair of biological fibers catalyzed by the surface of a virus crystal. *Proteins* 44: 392–396.

Lal, R., Kim, H., Garavito, R.M., and Arnsdorf, M.F. (1993). Imaging of reconstituted biological channels at molecular resolution by atomic force microscopy. *Am. J. Physiol.* 265: C851–C856.

Landman, U., Luedtke, W.D., Burnham, N.A., and Colton, R.J. (1990). Atomistic mechanisms and dynamics of adhesion, nanoindentation, and fracture. *Science* 248: 454–461.

Laney, D.E., Garcia, R.A., Parsons, S.M., and Hansma, H.G. (1997). Changes in the elastic properties of cholinergic synaptic vesicles as measured by atomic force microscopy. *Biophys. J.* 72: 806–813.

Lawrence, J.C., Saslowsky, D.E., Edwardson, J.M., and Henderson, R.M. (2003). Real-time analysis of the effects of cholesterol on lipid raft behavior using atomic force microscopy. *Biophys. J.* 84: 1827–1832.

Leckband, D. (2000). Measuring the forces that control protein interactions. *Annu. Rev. Biophys. Biomol. Struct.* 29: 1–26.

Lee, G.U., Chrisey, L.A., and Colton, R.J. (1994). Direct measurement of the forces between complementary strands of DNA. *Science* 266: 771–773.

Leidy, C., Kaasgaard, T., Crowe, J.H., Mouritsen, O.G., and Jorgensen, K. (2002). Ripples and the formation of anisotropic lipid domains: imaging two-component supported double bilayers by atomic force microscopy. *Biophys. J.* 83: 2625–2633.

Lekka, M., Laidler, P., Gil, D., Lekki, J., Stachura, Z., and Hrynkiewicz, A.Z. (1999). Elasticity of normal and cancerous human bladder cells studied by scanning force microscopy. *Eur. Biophys. J.* 28: 312–316.

Lenne, P.F., Raae, A.J., Altmann, S.M., Saraste, M., and Horber, J.K. (2000). States and transitions during forced unfolding of a single spectrin repeat. *FEBS Lett.* 476: 124–128.

Leonenko, Z.V., Carnini, A., and Cramb, D.T. (2000). Supported planar bilayer formation by vesicle fusion: the interaction of phospholipid vesicles with surfaces and the effect of gramicidin on bilayer properties using atomic force microscopy. *Biochim. Biophys. Acta* 1509: 131–147.

Li, S.F., McGhie, A.J., and Tang, S.L. (1994). New internal structure of spider dragline silk revealed by atomic force microscopy. *Biophys. J.* 66: 1209–1212.

Liao, X. and Wiedmann, T.S. (2004). Characterization of pharmaceutical solids by scanning probe microscopy. *J. Pharm. Sci.* 93: 2250–2258.

Lillehei, P.T. and Bottomley, L.A. (2000). Scanning probe microscopy. *Anal. Chem.* 72: 189R–196R.

Lillehei, P.T. and Bottomley, L.A. (2001). Scanning force microscopy of nucleic acid complexes. *Meth. Enzymol.* 340: 234–251.

Lindsay, S.M., Lyubchenko Yu, L., Tao, N.J., Li, Y.Q., Oden, P.I., Derose, J.A., and Pan, J. (1993). Scanning tunneling microscopy and atomic force microscopy studies of biomaterials at a liquid–solid interface. *J. Vac. Sci. Technol. A* 11: 808–815.

Liu, J.-F., Min, G., and Ducker, W.A. (2001). AFM study of cationic surfactants and cationic polyelectrolyutes at the silica-water interface. *Langmuir* 17: 4895–4903.

Lokar, W.J. and Ducker, W.A. (2002). Proximal adsorption of dodecyltrimethylammonium bromide to the silica-electrolyte solution interface. *Langmuir* 18: 3167–3175.

Lokar, W.J. and Ducker, W.A. (2004). Proximal adsorption at glass surfaces: ionic strength, pH, chain length effects. *Langmuir* 20: 378–388.

Lu, H. and Schulten, K. (1999). Steered molecular dynamics simulations of force-induced protein domain unfolding. *Proteins* 35: 453–463.

Lu, H. and Schulten, K. (2000). The key event in force-induced unfolding of Titin's immunoglobulin domains. *Biophys. J.* 79: 51–65.

Lucas, R.W., Kuznetsov, Y.G., Larson, S.B., and McPherson, A. (2001). Crystallization of Brome mosaic virus and T = 1 Brome mosaic virus particles following a structural transition. *Virology* 286: 290–303.

Ludes, M.D. and Wirth, M.J. (2002). Single-molecule resolution and fluorescence imaging of mixed-mode sorption of a dye at the interface of C18 and acetonitrile/water. *Anal. Chem.* 74: 386–393.

Lvov, Y., Onda, M., Ariga, K., and Kunitake, T. (1998). Ultrathin films of charged polysaccharides assembled alternately with linear polyions. *J. Biomater. Sci. Polym. Ed.* 9: 345–355.

Magonov, S. and Godovsky, Y. (1999). Atomic force microscopy, part 8: visualization of granular nanostructure in crystalline polymers. *Am. Lab.* 1999: 52–58.

Magonov, S. and Heaton, M.G. (1998). Atomic force microscopy, part 6: recent developments in AFM of polymers. *Am. Lab.* 30.

Magonov, S.N., Elings, V., and Whangbo, M.-H. (1997). Phase imaging and stiffness in tapping mode AFM. *Surface Sci.* 375: L385–L391.

Magonov, S.N. and Reneker, D.H. (1997). Characterization of polymer surfaces with atomic force microscopy. *Annu. Rev. Mater. Sci.* 27: 175–222.

Malkin, A.J., Plomp, M., and McPherson, A. (2002). Application of atomic force microscopy to studies of surface processes in virus crystallization and structural biology. *Acta Crystallogr. D Biol. Crystallogr.* 58: 1617–1621.

Manne, S., Cleveland, J.P., Stucky, G.D., and Hansma, P.K. (1993). Lattice resolution and solution kinetics on surfaces of amino acid crystals: an atomic force microscope study. *J. Cryst. Growth (Netherlands)* 130: 333–340.

Manne, S. and Gaub, H.E. (1997). Force microscopy: measurement of local interfacial forces and surface stresses. *Curr. Opin. Colloid Interface Sci.* 2: 145–152.

Marszalek, P.E., Lu, H., Li, H., Carrion-Vazquez, M., Oberhauser, A.F., Schulten, K., and Fernandez, J.M. (1999). Mechanical unfolding intermediates in titin modules. *Nature* 402: 100–103.

Mathur, A.B., Truskey, G.A., and Reichert, W.M. (2000). Atomic force and total internal reflection fluorescence microscopy for the study of force transmission in endothelial cells. *Biophys. J.* 78: 1725–1735.

McKiernan, A.E., Ratto, T.V., and Longo, M.L. (2000). Domain growth, shapes, and topology in cationic lipid bilayers on mica by fluorescence and atomic force microscopy. *Biophys. J.* 79: 2605–2615.

McLaurin, J., Darabie, A.A., and Morrison, M.R. (2002). Cholesterol, a modulator of membrane-associated Abeta-fibrillogenesis. *Ann. NY Acad. Sci.* 977: 376–383.

McPherson, A., Malkin, A.J., Kuznetsov, Y.G., and Plomp, M. (2001). Atomic force microscopy applications in macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* 57: 1053–1060.

McPherson, A., Malkin, A.J., and Kuznetsov Yu, G. (2000). Atomic force microscopy in the study of macromolecular crystal growth. *Annu. Rev. Biophys. Biomol. Struct.* 29: 361–410.

Merkel, R., Nassoy, P., Leung, A., Ritchie, K., and Evans, E. (1999). Energy landscapes of receptor–ligand bonds explored with dynamic force spectroscopy. *Nature* 397: 50–53.

Michalet, X., Kapanidis, A.N., Laurence, T., Pinaud, F., Doose, S., Pflughoefft, M., and Weiss, S. (2003). The power and prospects of fluorescence microscopies and spectroscopies. *Annu. Rev. Biophys. Biomol. Struct.* 32: 161–182.

Micic, M., Radotic, K., Jeremic, M., Djikanovic, D., and Kammer, S.B. (2004). Study of the lignin model compound supramolecular structure by combination of near-field scanning optical microscopy and atomic force microscopy. *Colloids Surf. B Biointerfaces* 34: 33–40.

Milhaud, J., Ponsinet, V., Takashi, M., and Michels, B. (2002). Interactions of the drug amphotericin B with phospholipid membranes containing or not ergosterol: new insight into the role of ergosterol. *Biochim. Biophys. Acta* 1558: 95–108.

Miller, L.D., Putthanarat, S., Eby, R.K., and Adams, W.W. (1999). Investigation of the nanofibrillar morphology in silk fibers by small angle X-ray scattering and atomic force microscopy. *Int. J. Biol. Macromol.* 24: 159–165.

Mizes, H.A., Loh, K.-G., Miller, R.J.D., Ahujy, S.K., and Grabowski, G.A. (1991). Submicron probe of polymer adhesion with atomic force microscopy. Dependence on topography and material inhomogenities. *Appl. Phys. Lett.* 59: 2901–2903.

Moers, M.H., Ruiter, A.G., Jalocha, A., and van Hulst, N.F. (1995). Detection of fluorescence in situ hybridization on human metaphase chromosomes by near-field scanning optical microscopy. *Ultramicroscopy* 61: 279–283.

Moller, C., Allen, M., Elings, V., Engel, A., and Muller, D.J. (1999). Tapping-mode atomic force microscopy produces faithful high-resolution images of protein surfaces. *Biophys. J.* 77: 1150–1158.

Moradian-Oldak, J., Paine, M.L., Lei, Y.P., Fincham, A.G., and Snead, M.L. (2000). Self-assembly properties of recombinant engineered amelogenin proteins analyzed by dynamic light scattering and atomic force microscopy. *J. Struct. Biol.* 131: 27–37.

Mosley, L.M., Hunter, K.A., and Ducker, W.A. (2003). Forces between colloid particles in natural waters. *Environ. Sci. Technol.* 37: 3303–3308.

Mukherjee, S., Brieba, L.G., and Sousa, R. (2002). Structural transitions mediating transcription initiation by T7 RNA polymerase. *Cell* 110: 81–91.

Müller, D.J., Fotiadis, D., and Engel, A. (1998). Mapping flexible protein domains at subnanometer resolution with the atomic force microscope. *FEBS Lett.* 430: 105–111.

Muller, D.J., Kessler, M., Oesterhelt, F., Moller, C., Oesterhelt, D., and Gaub, H. (2002). Stability of bacteriorhodopsin alpha-helices and loops analyzed by single-molecule force spectroscopy. *Biophys. J.* 83: 3578–3588.

Muramatsu, H., Chiba, N., Umemoto, T., Homma, K., Nakajima, K., Ataka, T., Ohta, S., Kusumi, A., and Fujihira, M. (1995). Development of near-field optic/atomic force microscope for biological materials in aqueous solutions. *Ultramicroscopy* 61: 265–269.

Muresan, A.S., Diamant, H., and Lee, K.Y. (2001). Effect of temperature and composition on the formation of nanoscale compartments in phospholipid membranes. *J. Am. Chem. Soc.* 123: 6951–6952.

Muster, T.H. and Prestidge, C.A. (2002). Face specific surface properties of pharmaceutical crystals. *J. Pharm. Sci.* 91: 1432–1444.

Nagao, E. and Dvorak, J.A. (1998). An integrated approach to the study of living cells by atomic force microscopy. *J. Microsc.* 191: 8–19.

Nagao, E. and Dvorak, J.A. (1999). Phase imaging by atomic force microscopy: analysis of living homoiothermic vertebrate cells. *Biophys. J.* 76: 3289–3297.

Nagy, P., Jenei, A., Kirsch, A.K., Szollosi, J., Damjanovich, S., and Jovin, T.M. (1999). Activation-dependent clustering of the erbB2 receptor tyrosine kinase detected by scanning near-field optical microscopy. *J. Cell. Sci.* 112: 1733–1741.

Neff, D., Tripathi, S., Middendorf, K., Stahlberg, H., Butt, H.J., Bamberg, E., and Dencher, N.A. (1997). Chloroplast F0F1 ATP synthase imaged by atomic force microscopy. *J. Struct. Biol.* 119: 139–148.

Nevo, R., Stroh, C., Kienberger, F., Kaftan, D., Brumfeld, V., Elbaum, M., Reich, Z., and Hinterdorfer, P. (2003). A molecular switch between alternative conformational states in the complex of Ran and importin beta1. *Nat. Struct. Biol.* 10: 553–557.

Nielsen, L.K., Bjornholm, T., and Mouritsen, O.G. (2000). Fluctuations caught in the act. *Nature* 404: 352.

Nishida, S., Funabashi, Y., and Ikai, A. (2002). Combination of AFM with an objective-type total internal reflection fluorescence microscope (TIRFM) for nanomanipulation of single cells. *Ultramicroscopy* 91: 269–274.

Noy, A., Frisbie, C.D., Rozsnyai, L.F., Wrighton, M.S., and Leiber, C.M. (1995). Chemical force microscopy: exploiting chemically-modified tips to quantify adhesion, friction, and functional group distributions in molecular assemblies. *J. Am. Chem. Soc.* 117: 7943–7951.

Noy, A., Sanders, C.H., Vezenov, D.V., Wong, S.S., and Lieber, C.M. (1998a). Chemically-sensitive imaging in tapping mode by chemical force microscopy: relationship between phase lag and adhesion. *Langmuir* 14: 1508–1511.

Noy, A., Sanders, C.H., Vezenov, D.V., Wong, S.S., and Lieber, C.M. (1998b). Chemically-sensitive imaging in tapping mode by chemical force microscopy: relationship between phase lag and adhesion. *Langmuir* 14: 1508–1511.

Oberhauser, A.F., Badilla-Fernandez, C., Carrion-Vazquez, M., and Fernandez, J.M. (2002). The mechanical hierarchies of fibronectin observed with single-molecule AFM. *J. Mol. Biol.* 319: 433–447.

Oberhauser, A.F., Hansma, P.K., Carrion-Vazquez, M., and Fernandez, J.M. (2001). Stepwise unfolding of titin under force-clamp atomic force microscopy. *Proc. Natl Acad. Sci. USA* 98: 468–472.

Oesterfelt, F., Rief, M., and Gaub, H.E. (1999). Single molecule force spectroscopy by AFM indicates helical structure of poly(ethylene-glycol) in water. *New J. Phys.* 1: 6.1–6.11.

Oesterhelt, F., Oesterhelt, D., Pfeiffer, M., Engel, A., Gaub, H.E., and Muller, D.J. (2000a). Unfolding pathways of individual bacteriorhodopsins. *Science* 288: 143–146.

Oesterhelt, F., Oesterhelt, D., Pfeiffer, M., Engel, A., Gaub, H.E., and Müller, D.J. (2000b). Unfolding pathways of individual bacteriorhodopsins. *Science* 288: 143–146.

Okabe, Y., Furugori, M., Tani, Y., Akiba, U., and Fujihira, M. (2000). Chemical force microscopy of microcontact-printed self-assembled monolayers by pulsed-force-mode atomic force microscopy. *Ultramicroscopy* 82: 203–212.

Opdahl, A., Hoffer, S., Mailhot, B., and Somorjai, G.A. (2001). Polymer surface science. *Chem. Rec.* 1: 101–122.

Oroudjev, E., Soares, J., Arcdiacono, S., Thompson, J.B., Fossey, S.A., and Hansma, H.G. (2002). Segmented nanofibers of spider dragline silk: atomic force microscopy and single-molecule force spectroscopy. *Proc. Natl Acad. Sci. USA* 99: 6460–6465.

Osborne, M.A., Barnes, C.L., Balasubramanian, S., and Klenerman, D. (2001). Probing DNA surface attachment and local environment using single molecule spectroscopy. *J. Phys. Chem. B.* 105: 3120–3126.

Paci, E. and Karplus, M. (2000). Unfolding proteins by external forces and temperature: the importance of topology and energetics. *Proc. Natl Acad. Sci. USA* 97: 6521–6526.

Paige, M.F. (2003). A comparison of atomic force microscope friction and phase imaging for the characterization of an immiscible polystyrene/poly(methyl methacrylate) blend film. *Polymer* 44: 6345–6352.

Parbhu, A., Lin, H., Thimm, J., and Lal, R. (2002). Imaging real-time aggregation of amyloid beta protein (1-42) by atomic force microscopy. *Peptides* 23: 1265–1270.

Parbhu, A.N., Bryson, W.G., and Lal, R. (1999). Disulfide bonds in the outer layer of keratin fibers confer higher mechanical rigidity: correlative nano-indentation and elasticity measurement with an AFM. *Biochemistry* 38: 11755–11761.

Pelling, A.E., Sehati, S., Gralla, E.B., Valentine, J.S., and Gimzewski, J.K. (2004). Local nanomechanical motion of the cell wall of Saccharomyces cerevisiae. *Science* 305: 1147–1150.

Petsev, D.N., Thomas, B.R., Yau, S., and Vekilov, P.G. (2000). Interactions and aggregation of apoferritin molecules in solution: effects of added electrolytes. *Biophys. J.* 78: 2060–2069.

Plomp, M., McPherson, A., and Malkin, A.J. (2003). Repair of impurity-poisoned protein crystal surfaces. *Proteins* 50: 486–495.

Plomp, M., Rice, M.K., Wagner, E.K., McPherson, A., and Malkin, A.J. (2002). Rapid visualization at high resolution of pathogens by atomic force microscopy: structural studies of herpes simplex virus-1. *Am. J. Pathol.* 160: 1959–1966.

Poggi, M.A., Bottomley, L.A., and Lillehei, P.T. (2002). Scanning probe microscopy. *Anal. Chem.* 74: 2851–2862.

Poggi, M.A., Gadsby, E.D., Bottomley, L.A., King, W.P., Oroudjev, E., and Hansma, H. (2004). Scanning probe microscopy. *Anal. Chem.* 76: 3429–3444.

Prikulis, J., Murty, K.V., Olin, H., and Kall, M. (2003). Large-area topography analysis and near-field Raman spectroscopy using bent fibre probes. *J. Microsc.* 210: 269–273.

Puu, G., Artursson, E., Gustafson, I., Lundstrom, M., and Jass, J. (2000). Distribution and stability of membrane proteins in lipid membranes on solid supports. *Biosens. Bioelectron.* 15: 31–41.

Puu, G. and Gustafson, I. (1997). Planar lipid bilayers on solid supports from liposomes — factors of importance for kinetics and stability. *Biochim. Biophys. Acta* 1327: 149–161.

Puu, G., Gustafson, I., Artursson, E., and Ohlsson, P.A. (1995). Retained activities of some membrane proteins in stable lipid bilayers on a solid support. *Biosens. Bioelectron.* 10: 463–476.

Qian, H. and Shapiro, B.E. (1999). Graphical method for force analysis: macromolecular mechanics with atomic force microscopy. *Proteins* 37: 576–581.

Raab, A., Han, W., Badt, D., Smith-Gill, S.J., Lindsay, S.M., Schindler, H., and Hinterdorfer, P. (1999). Antibody recognition imaging by force microscopy. *Nat. Biotechnol.* 17: 901–905.

Radler, J., Strey, H., and Sackmann, E. (1995). Phenomenology and kinetics of lipid bilayer spreading on hydrophilic surfaces. *Langmuir* 11: 4539–4548.

Radmacher, M. (1997). Measuring the elastic properties of biological samples with the AFM. *IEEE Eng. Med. Biol. Mag.* 16: 47–57.

Reviakine, I. and Brisson, A. (2000). Formation of supported phospholipid bilayers from unilamellar vesicles investigated by atomic force microscopy. *Langmuir* 16: 1806–1815.

Richter, R., Mukhopadhyay, A., and Brisson, A. (2003). Pathways of lipid vesicle deposition on solid surfaces: a combined QCM-D and AFM study. *Biophys. J.* 85: 3035–3047.

Rief, M., Gautel, M., and Gaub, H.E. (2000). Unfolding forces of titin and fibronectin domains directly measured by AFM. *Adv. Exp. Med. Biol.* 481: 129–136; discussion 137–141.

Rief, M., Gautel, M., Oesterhelt, F., Fernandez, J.M., and Gaub, H.E. (1997a). Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science* 276: 1109–1112.

Rief, M. and Grubmuller, H. (2002). Force spectroscopy of single biomolecules. *Chemphyschem* 3: 255–261.

Rief, M., Oesterhelt, F., Heymann, B., and Gaub, H.E. (1997b). Single molecule force spectroscopy on polysaccharides by atomic force microscopy. *Science* 275: 1295–1297.

Rief, M., Pascual, J., Saraste, M., and Gaub, H.E. (1999). Single molecule force spectroscopy of spectrin repeats: low unfolding forces in helix bundles. *J. Mol. Biol.* 286: 553–561.

Rinia, H.A. and de Kruijff, B. (2001). Imaging domains in model membranes with atomic force microscopy. *FEBS Lett.* 504: 194–199.

Rinia, H.A., Demel, R.A., van der Eerden, J.P., and de Kruijff, B. (1999). Blistering of langmuir-blodgett bilayers containing anionic phospholipids as observed by atomic force microscopy. *Biophys. J.* 77: 1683–1693.

Rinia, H.A., Kik, R.A., Demel, R.A., Snel, M.M.E., Killian, J.A., van Der Eerden, J.P.J.M., and de Kruijff, B. (2000). Visualization of highly ordered striated domains induced by transmembrane peptides in supported phosphatidylcholine bilayers. *Biochemistry* 39: 5852–5858.

Rivetti, C., Vannini, N., and Cellai, S. (2003). Imaging transcription complexes with the Atomic Force Microscope. *Ital. J. Biochem.* 52: 98–103.

Rochet, J.C., Conway, K.A., and Lansbury, P.T., Jr. (2000). Inhibition of fibrillization and accumulation of prefibrillar oligomers in mixtures of human and mouse alpha-synuclein. *Biochemistry* 39: 10619–10626.

Roher, A.E., Baudry, J., Chaney, M.O., Kuo, Y.M., Stine, W.B., and Emmerling, M.R. (2000). Oligomerization and fibril assembly of the mayloid-beta protein. *Biochim. Biophys. Acta* 1502: 31–43.

Ros, R., Schwesinger, F., Anselmetti, D., Kubon, M., Schafer, R., Pluckthun, A., and Tiefenauer, L. (1998). Antigen binding forces of individually addressed single-chain Fv antibody molecules. *Proc. Natl Acad. Sci. USA* 95: 7402–7405.

Rosa-Zeiser, A., Weilandt, E., Hild, S., and Marti, O. (1997). The simultaneous measurement of elastic, electrostatic and adhesive properties by scanning force microscopy: pulsed force mode operation. *Meas. Sci. Technol.* 8: 1333–1338.

Sackmann, E. (1996). Supported membranes: scientific and practical applications. *Science* 271: 43–48.

Sako, Y., Hibino, K., Miyauchi, T., Miyamoto, Y., Ueda, M., and Yanagida, T. (2000a). Single-molecule imaging of signaling molecules in living cells. *Single Mol.* 2: 159–163.

Sako, Y., Minoghchi, S., and Yanagida, T. (2000b). Single-molecule imaging of EGFR signalling on the surface of living cells. *Nat. Cell. Biol.* 2: 168–172.

Sako, Y. and Uyemura, T. (2002). Total internal reflection fluorescence microscopy for single-molecule imaging in living cells. *Cell Struct. Funct.* 27: 357–365.

Santos, N.C., Ter-Ovanesyan, E., Zasadzinski, J.A., Prieto, M., and Castanho, M.A. (1998). Filipin-induced lesions in planar phospholipid bilayers imaged by atomic force microscopy. *Biophys. J.* 75: 1869–1873.

Schmidt, T., Hinterdorfer, P., and Schindler, H. (1999). Microscopy for recognition of individual biomolecules. *Microsc. Res. Tech.* 44: 339–346.

Schmitt, L., Ludwig, M., Gaub, H.E., and Tampe, R. (2000). A metal-chelating microscopy tip as a new toolbox for single-molecule experiments by atomic force microscopy. *Biophys. J.* 78: 3275–3285.

Schneider, M., Zhu, M., Papastavrou, G., Akari, S., and Mohwald, H. (2002). Chemical pulsed-force microscopy of single polyethyleneimine molecules in aqueous solution. *Langmuir* 18: 602–606.

Schwaiger, I., Kardinal, A., Schleicher, M., Noegel, A.A., and Rief, M. (2004). A mechanical unfolding intermediate in an actin-crosslinking protein. *Nat. Struct. Mol. Biol.* 11: 81–85.

Schwesinger, F., Ros, R., Strunz, T., Anselmetti, D., Guntherodt, H.J., Honegger, A., Jermutus, L., Tiefenauer, L., and Pluckthun, A. (2000). Unbinding forces of single antibody-antigen complexes correlate with their thermal dissociation rates. *Proc. Natl Acad. Sci. USA* 97: 9972–9977.

Scott, W.W. and Bhushan, B. (2003). Use of phase imaging in atomic force microscopy for measurement of viscoelastic contrast in polymer nanocomposites and molecularly thick lubricant films. *Ultramicroscopy* 97: 151–169.

Sekatskii, S.K., Shubeita, G.T., and Dietler, G. (2000). Time-gated scanning near-field optical microscopy. *Appl. Phys. Lett.* 77: 2089–2091.

Senden, T.J. and Drummond, C.J. (1995). Surface chemistry and tip-sample interactions in atomicforce microscopy. *Colloids and Surfaces* 94.

Seong, G.H., Yanagida, Y., Aizawa, M., and Kobatake, E. (2002). Atomic force microscopy identification of transcription factor NFkappaB bound to streptavidin-pin-holding DNA probe. *Anal. Biochem.* 309: 241–247.

Shaw, J.E., Slade, A., and Yip, C.M. (2003). Simultaneous in situ total internal reflectance fluorescence/atomic force microscopy studies of DPPC/dPOPC microdomains in supported planar lipid bilayers. *J. Am. Chem. Soc.* 125: 111838–111839.

Shellenberger, K. and Logan, B.E. (2002). Effect of molecular scale roughness of glass beads on colloidal and bacterial deposition. *Environ. Sci. Technol.* 36: 184–189.

Sherrat, M.J., Holmes, D.F., Shuttleworth, C.A., and Kielty, C.M. (2004). Substrate-dependent morphology of supramolecular assemblies: fibrillin and type-IV collagen microfibrils. *Biophys. J.* 86: 3211–3222.

Shroff, S.G., Saner, D.R., and Lal, R. (1995). Dynamic micromechanical properties of cultured rat atrial myocytes measured by atomic force microscopy. *Am. J. Physiol.* 269: C286–C292.

Slade, A., Luh, J., Ho, S., and Yip, C.M. (2002). Single molecule imaging of supported planar lipid bilayer — reconstituted human insulin receptors by in situ scanning probe microscopy. *J. Struct. Biol.* 137: 283–291.

Smith, D.A. and Radford, S.E. (2000). Protein folding: pulling back the frontiers. *Curr. Biol.* 10: R662–R664.

Sommer, A.P. and Franke, R.P. (2002). Near-field optical analysis of living cells in vitro. *J. Proteome Res.* 1: 111–114.

Steinem, C., Galla, H.-J., and Janshoff, A. (2000). *Phys. Chem. Chem. Phys.* 2: 4580–4585.

Stenert, M., Döring, A., and Bandermann, F. (2004). Poly(methyl methacrylate)-block-polystyrene and polystyrene-block-poly (*n*-butyl acrylate) as compatibilizers in PMMA/PnBA blends. *e-Polymers* 15: 1–16.

Stroh, C.M., Ebner, A., Geretschlager, M., Freudenthaler, G., Kienberger, F., Kamruzzahan, A.S., Smith-Gill, S.J., Gruber, H.J., and Hinterdorfer, P. (2004). Simultaneous topography and recognition imaging using force microscopy. *Biophys. J.* 87: 1981–1990.

Strunz, T., Oroszlan, K., Schumakovitch, I., Guntherodt, H.J., and Hegner, M. (2000). Model energy landscapes and the force-induced dissociation of ligand-receptor bonds. *Biophys. J.* 79.

Suda, H., Sasaki, Y.C., Oishi, N., Hiraoka, N., and Sutoh, K. (1999). Elasticity of mutant myosin subfragment-1 arranged on a functional silver surface. *Biochem. Biophys. Res. Commun.* 261: 276–282.

Taatjes, D.J., Quinn, A.S., and Bovill, E.G. (1999). Imaging of collagen type III in fluid by atomic force microscopy. *Microsc. Res. Tech.* 44: 347–352.

Tahirov, T.H., Sato, K., Ichikawa-Iwata, E., Sasaki, M., Inoue-Bungo, T., Shiina, M., Kimura, K., Takata, S., Fujikawa, A., Morii, H., et al. (2002). Mechanism of c-Myb-C/EBP beta cooperation from separated sites on a promoter. *Cell* 108: 57–70.

Takeyasu, K., Omote, H., Nettikadan, S., Tokumasu, F., Iwamoto-Kihara, A., and Futai, M. (1996). Molecular imaging of *Escherichia coli* F0F1-ATPase in reconstituted membranes using atomic force microscopy. *FEBS Lett.* 392: 110–113.

Teng, H.H., Dove, P.M., Orme, C.A., and De Yoreo, J.J. (1998). Thermodynamics of calcite growth: baseline for understanding biomineral formation. *Science* 282: 724–727.

Thompson, J.B., Paloczi, G.T., Kindt, J.H., Michenfelder, M., Smith, B.L., Stucky, G., Morse, D.E., and Hansma, P.K. (2000). Direct observation of the transition from calcite to aragonite growth as induced by abalone shell proteins. *Biophys. J.* 79: 3307–3312.

Thomson, N.H., Smith, B.L., Almqvist, N., Schmitt, L., Kashlev, M., Kool, E.T., and Hansma, P.K. (1999). Oriented, active Escherichia coli RNA polymerase: an atomic force microscope study. *Biophys. J.* 76: 1024–1033.

Toikka, G. and Hayes, R.A. (1997). Direct measurement of colloidal forces between mica and silica in aqueous electrolyte. *J. Colloid Interface Sci.* 191: 102–109.

Tokumasu, F., Jin, A.J., and Dvorak, J.A. (2002). Lipid membrane phase behaviour elucidated in real time by controlled environment atomic force microscopy. *J. Electron Microsc. (Tokyo)* 51: 1–9.

Touhami, A., Nysten, B., and Dufrene, Y.F. (2003). Nanoscale mapping of the elasticity of microbial cells by atomic force microscopy. *Langmuir* 19: 1745–1751.

Trottier, M., Mat-Arip, Y., Zhang, C., Chen, C., Sheng, S., Shao, Z., and Guo, P. (2000). Probing the structure of monomers and dimers of the bacterial virus phi29 hexamer RNA complex by chemical modification. *Rna* 6: 1257–1266.

Tulpar, A., Subramaniam, V., and Ducker, W.A. (2001). Decay lengths in double-layer forces in solutions of partly associated ions. *Langmuir* 17: 8451–8454.

van der Werf, K.O., Putman, C.A., de Grooth, B.G., and Greve, J. (1994). Adhesion force imaging in air and liquid by adhesion mode atomic force microscopy. *Appl. Phys. Lett.* 65: 1195–1197.

van Duyl, B.Y., Ganchev, D., Chupin, V., de Kruijff, B., and Killian, J.A. (2003). Sphingomyelin is much more effective than saturated phosphatidylcholine in excluding unsaturated phosphatidylcholine from domains formed with cholesterol. *FEBS Lett.* 547: 101–106.

van Hulst, N.F., Veerman, J.A., Garcia-Parajo, M.F., and Kuipers, J. (2000). Analysis of individual (macro)molecules and proteins using near-field optics. *J. Chem. Phys.* 112: 7799–7810.

Van Landringham, M.R., Dagastine, R.R., Eduljee, R.F., McCullough, R.L., and Gillespie, J.W.J. (1999). Characterization of nanoscale property variations in polymer composite systems: Part 1 — Experimental results. *Composites Part A* 30.

Van Landringham, M.R., McKnight, S.H., Palmese, G.R., Bogetti, T.A., Eduljee, R.F., and Gillespie, J.W.J. (1997a). Characterization of interphase regions using atomic force microscopy. *Mat. Res. Soc. Symp. Proc.* 458: 313–318.

Van Landringham, M.R., McKnight, S.H., Palmese, G.R., Eduljee, R.F., Gillespie, J.W.J., and McCullough, R.L. (1997b). Relating polymer indentation behavior to elastic modulus using atomic force microscopy. *Mat. Res. Soc. Symp. Proc.* 440: 195–200.

Van Landringham, M.R., McKnight, S.H., Palmese, G.R., Huang, X., Bogetti, T.A., Eduljee, R.F., and Gillespie, J.W.J. (1997c). Nanoscale indentation of polymer systems using the atomic force microscope. *J. Adhesion* 64: 31–59.

Velegol, S.B. and Logan, B.E. (2002). Contributions of bacterial surface polymers, electrostatics, and cell elasticity to the shape of AFM force curves. *Langmuir* 18: 5256–5262.

Vesenka, J., Manne, S., Giberson, R., Marsh, T., and Henderson, E. (1993). Collidal gold particles as an incompressible atomic force microscope imaging standard for assessing the compressibility of biomolecules. *Biochem. J.* 65: 992–997.

Vinckier, A., Dumortier, C., Engelborghs, Y., and Hellemans, L. (1996). Dynamical and mechanical study of immobilized microtubules with atomic force microscopy. *J. Vac. Sci. Technol.* 14: 1427–1431.

Wadu-Mesthrige, K., Amro, N.A., and Liu, G.Y. (2000). Immobilization of proteins on self-assembled monolayers. *Scanning* 22: 380–388.

Wagner, P. (1998). Immobilization strategies for biological scanning probe microscopy. *FEBS Lett.* 430: 112–115.

Wakelin, S. and Bagshaw, C.R. (2003). A prism combination for near isotropic fluorescence excitation by total internal reflection. *J. Microsc.* 209: 143–148.

Walch, M., Ziegler, U., and Groscurth, P. (2000). Effect of streptolysin O on the microelasticity of human platelets analyzed by atomic force microscopy. *Ultramicroscopy* 82: 259–267.

Wang, Z., Zhou, C., Wang, C., Wan, L., Fang, X., and Bai, C. (2003). AFM and STM study of beta-amyloid aggregation on graphite. *Ultramicroscopy* 97: 73–79.

Ward, M.D. (2001). Bulk crystals to surfaces: combining X-ray diffraction and atomic force microscopy to probe the structure and formation of crystal interfaces. *Chem. Rev.* 2001: 1697–1725.

Watanabe, M., Kobayashi, M., Fujita, Y., Senga, K., Mizutani, H., Ueda, M., and Hoshino, T. (1997). Association of type VI collagen with D-periodic collagen fibrils in developing tail tendons of mice. *Arch. Histol. Cytol.* 60: 427–434.

Waugh, D.F., Thompson, R.E., and Weimer, R.J. (1950). Assay of insulin *in vitro* by fibril elongation and precipitation. *J. Biol. Chem.* 185: 85–95.

Weisenhorn, A.L., Khorsandi, M., Kasas, S., Gotzos, V., and Butt, H.-J. (1993). Deformation and height anomaly of soft surfaces studied with an AFM. *Nanotechnology* 4: 106–113.

Wen, H.B., Moradian-Oldak, J., Zhong, J.P., Greenspan, D.C., and Fincham, A.G. (2000). Effects of amelogenin on the transforming surface microstructures of Bioglass in a calcifying solution. *J. Biomed. Mater. Res.* 52: 762–773.

Wielert-Badt, S., Hinterdorfer, P., Gruber, H.J., Lin, J.T., Badt, D., Wimmer, B., Schindler, H., and Kinne, R.K. (2002). Single molecule recognition of protein binding epitopes in brush border membranes by force microscopy. *Biophys. J.* 82: 2767–2774.

Willemsen, O.H., Snel, M.M., van der Werf, K.O., de Grooth, B.G., Greve, J., Hinterdorfer, P., Gruber, H.J., Schindler, H., van Kooyk, Y., and Figdor, C.G. (1998). Simultaneous height and adhesion imaging of antibody-antigen interactions by atomic force microscopy. *Biophys. J.* 75: 2220–2228.

Williams, P.M., Fowler, S.B., Best, R.B., Toca-Herrera, J.L., Scott, K.A., Steward, A., and Clarke, J. (2003). Hidden complexity in the mechanical properties of titin. *Nature* 422: 446–449.

Winkler, R.G., Spatz, J.P., Sheiko, S., Moller, M., Reineker, P., and Marti, O. (1996). Imaging material properties by resonant tapping-force microscopy: a model investigation. *Phys. Rev. B* 54: 8908–8912.

Yang, D.S., Yip, C.M., Huang, T.H., Chakrabartty, A., and Fraser, P.E. (1999). Manipulating the amyloid-beta aggregation pathway with chemical chaperones. *J. Biol. Chem.* 274: 32970–32974.

Yang, G., Cecconi, C., Baase, W.A., Vetter, I.R., Breyer, W.A., Haack, J.A., Matthews, B.W., Dahlquist, F.W., and Bustamante, C. (2000). Solid-state synthesis and mechanical unfolding of polymers of T4 lysozyme. *Proc. Natl Acad. Sci. USA* 97: 139–144.

Yang, G., Woodhouse, K.A., and Yip, C.M. (2002). Substrate-facilitated assembly of elastin-like peptides: studies by variable-temperature *in situ* atomic force microscopy. *J. Am. Chem. Soc.* 124: 10648–10649.

Yau, S.T., Petsev, D.N., Thomas, B.R., and Vekilov, P.G. (2000). Molecular-level thermodynamic and kinetic parameters for the self-assembly of apoferritin molecules into crystals. *J. Mol. Biol.* 303: 667–678.

Yau, S.T., Thomas, B.R., Galkin, O., Gliko, O., and Vekilov, P.G. (2001). Molecular mechanisms of microheterogeneity-induced defect formation in ferritin crystallization. *Proteins* 43: 343–352.

Yau, S.T. and Vekilov, P.G. (2000). Quasi-planar nucleus structure in apoferritin crystallization. *Nature* 406: 494–497.

Yau, S.T. and Vekilov, P.G. (2001). Direct observation of nucleus structure and nucleation pathways in apoferritin crystallization. *J. Am. Chem. Soc.* 123: 1080–1089.

Yip, C.M., Brader, M.L., Frank, B.H., DeFelippis, M.R., and Ward, M.D. (2000). Structural studies of a crystalline insulin analog complex with protamine by atomic force microscopy. *Biophys. J.* 78: 466–473.

Yip, C.M., Darabie, A.A., and McLaurin, J. (2002). Abeta42-peptide assembly on lipid bilayers. *J. Mol. Biol.* 318: 97–107.

Yip, C.M. and McLaurin, J. (2001). Amyloid-beta peptide assembly: a critical step in fibrillogenesis and membrane disruption. *Biophys. J.* 80: 1359–1371.

Young, R., Ward, J., and Scire, F. (1971). The topografiner: an instrument for measuring surface microtopography. *Rev. Sci. Instr.* 43: 999.

Yuan, C., Chen, A., Kolb, P., and Moy, V.T. (2000). Energy landscape of streptavidin-biotin complexes measured by atomic force microscopy. *Biochemistry* 39: 10219–10223.

Yuan, C. and Johnston, L.J. (2001). Atomic force microscopy studies of ganglioside GM1 domains in phosphatidylcholine and phosphatidylcholine/cholesterol bilayers. *Biophys. J.* 81: 1059–1069.

Zhang, B. and Evans, J.S. (2001). Modeling AFM-induced PEVK extension and the reversible unfolding of Ig/FNIII domains in single and multiple titin molecules. *Biophys. J.* 80: 597–605.

Zhang, B., Wustman, B.A., Morse, D., and Evans, J.S. (2002). Model peptide studies of sequence regions in the elastomeric biomineralization protein, Lustrin A. I. The C-domain consensus-PG-, -NVNCT-motif. *Biopolymers* 63: 358–369.

Zhang, H., Grim, P.C.M., Vosch, T., Wiesler, U.-M., Berresheim, A.J., Mullen, K., and De Schryver, F.C. (2000a). Discrimination of dendrimer aggregates on mica based on adhesion force: a pulsed force mode atomic force microscopy study. *Langmuir* 16: 9294–9298.

Zhang, H., Grim, P.C.M., Vosch, T., Wiesler, U.-M., Berresheim, A.J., Mullen, K., and De Schryver, F.C. (2000b). Discrimination of dendrimer aggregates on mica based on adhesion force: a pulsed mode atomic force microscopy study. *Langmuir* 16: 9294–9298.

Zhang, J., Uchida, E., Yuama, Y., and Ikada, Y. (1997). Electrostatic interaction between ionic polymer grafted surfaces studied by atomic force microscopy. *J. Colloid Interface Sci.* 188: 431–438.

68

Parenteral Infusion Devices

Gregory I. Voss
Robert D. Butterfield
IVAC Corporation

The circulatory system is the body's primary pathway for both the distribution of oxygen and other nutrients and the removal of carbon dioxide and other waste products. Since the entire blood supply in a healthy adult completely circulates within 60 sec, substances introduced into the circulatory system are distributed rapidly. Thus intravenous (IV) and intraarterial access routes provide an effective pathway for the delivery of fluid, blood, and medicants to a patient's vital organs. Consequently, about 80% of hospitalized patients receive infusion therapy. Peripheral and central veins are used for the majority of infusions. Umbilical artery delivery (in neonates), enteral delivery of nutrients, and epidural delivery of anesthetics and analgesics comprise smaller patient populations. A variety of devices can be used to provide flow through an intravenous catheter. An intravenous delivery system typically consists of three major components (1) fluid or drug reservoir, (2) catheter system for transferring the fluid or drug from the reservoir into the vasculature through a venipuncture, and (3) device for regulation and/or generating flow (see Figure 68.1).

This chapter is separated into five sections. Section 68.1 describes the clinical needs associated with intravenous drug delivery that determine device performance criteria. Section 68.2 reviews the principles of flow through a tube; Section 68.3 introduces the underlying electromechanical principles for flow regulation and/or generation and their ability to meet the clinical performance criteria. Section 68.4 reviews complications associated with intravenous therapy, and Section 68.5 concludes with a short list of articles providing more detailed information.

68.1 Performance Criteria for IV Infusion Devices

The IV pathway provides an excellent route for continuous drug therapy. The ideal delivery system regulates drug concentration in the body to achieve and maintain a desired result. When the drug's effect

FIGURE 68.1 A typical IV infusion system.

cannot be monitored directly, it is frequently assumed that a specific blood concentration or infusion rate will achieve the therapeutic objective. Although underinfusion may not provide sufficient therapy, overinfusion can produce even more serious toxic side effects.

The therapeutic range and risks associated with under- and overinfusion are highly drug and patient dependent. Intravenous delivery of fluids and electrolytes often does not require very accurate regulation. Low-risk patients can generally tolerate well infusion rate variability of ±30% for fluids. In some situations, however, specifically for fluid-restricted patients, prolonged under- or overinfusion of fluids can compromise the patient's cardiovascular and renal systems.

The infusion of many drugs, especially potent cardioactive agents, requires high accuracy. For example, post-coronary-artery-bypass-graft patients commonly receive sodium nitroprusside to lower arterial blood pressure. Hypertension, associated with underinfusion, subjects the graft sutures to higher stress with an increased risk for internal bleeding. Hypotension associated with overinfusion can compromise the cardiovascular state of the patient. Nitroprusside's potency, short onset delay, and short half-life (30 to 180 sec) provide for very tight control, enabling the clinician to quickly respond to the many events that alter the patient's arterial pressure. The fast response of drugs such as nitroprusside creates a need for short-term flow uniformity as well as long-term accuracy.

The British Department of Health employs *Trumpet curves* in their Health Equipment Information reports to compare flow uniformity of infusion pumps. For a prescribed flow rate, the trumpet curve is the plot of the maximum and minimum measured percentage flow rate error as a function of the accumulation interval (Figure 68.2). Flow is measured gravimetrically in 30-sec blocks for 1 h. These blocks are summed to produce 120-sec, 300-sec, and other longer total accumulation intervals. Though the 120-sec window may not detect flow variations important in delivery of the fastest acting agents, the trumpet curve provides a helpful means for performance comparison among infusion devices. Additional statistical information such as standard deviations may be derived from the basic trumpet flow measurements.

The short half-life of certain pharmacologic agents and the clotting reaction time of blood during periods of stagnant flow require that fluid flow be maintained without significant interruption. Specifically, concern has been expressed in the literature that the infusion of sodium nitroprusside and other short

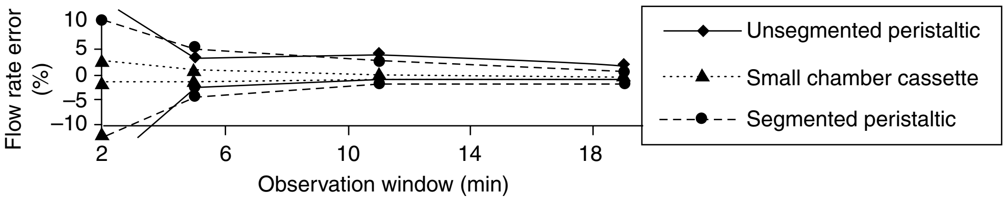


FIGURE 68.2 Trumpet curve for several representative large volume infusion pumps operated a 5 ml/h. Note that peristaltic pumps were designed for low risk patients.

half-life drugs occur without interruption exceeding 20 sec. Thus, minimization of false alarms and rapid detection of occlusions are important aspects of maintaining a constant vascular concentration. Accidental occlusions of the IV line due to improper positioning of stopcocks or clamps, kinked tubing, and clotted catheters are common.

Occlusions between pump and patient present a secondary complication in maintaining serum drug concentration. Until detected, the pump will infuse, storing fluid in the delivery set. When the occlusion is eliminated, the stored volume is delivered to the patient in a bolus. With concentrated pharmaceutic agents, this bolus can produce a large perturbation in the patient's status.

Occlusions of the pump intake also interrupt delivery. If detection is delayed, inadequate flow can result. During an intake occlusion, in some pump designs removal of the delivery set can produce abrupt aspiration of blood. This event may precipitate clotting and cause injury to the infusion site.

The common practice of delivering multiple drugs through a single venous access port produces an additional challenge to maintaining uniform drug concentration. Although some mixing will occur in the venous access catheter, fluid in the catheter more closely resembles a first-in/first-out digital queue: during delivery, drugs from the various infusion devices mix at the catheter input, an equivalent fluid volume discharges from the outlet. Rate changes and flow nonuniformity cause the mass flow of drugs at the outlet to differ from those at the input. Consider a venous access catheter with a volume of 2 ml and a total flow of 10 ml/h. Due to the digital queue phenomenon, an incremental change in the intake flow rate of an individual drug will not appear at the output for 12 min. In addition, changing flow rates for one drug will cause short-term marked swings in the delivery rate of drugs using the same access catheter. When the delay becomes significantly larger than the time constant for a drug that is titrated to a measurable patient response, titration becomes extremely difficult leading to large oscillations.

As discussed, the performance requirements for drug delivery vary with multiple factors: drug, fluid restriction, and patient risk. Thus the delivery of potent agents to fluid-restricted patients at risk require the highest performance standards defined by flow rate accuracy, flow rate uniformity, and ability to minimize risk of IV-site complications. These performance requirements need to be appropriately balanced with the device cost and the impact on clinician productivity.

68.2 Flow through an IV Delivery System

The physical properties associated with the flow of fluids through cylindrical tubes provide the foundation for understanding flow through a catheter into the vasculature. Hagen–Poiseuille's equation for laminar flow of a Newtonian fluid through a rigid tube states

$$Q = \pi \cdot r^4 \cdot \frac{(P_1 - P_2)}{8 \cdot \eta \cdot L} \tag{68.1}$$

where Q is the flow; P_1 and P_2 are the pressures at the inlet and outlet of the tube, respectively; L and r are the length and internal radius of the tube, respectively; and η is fluid viscosity. Although many drug delivery systems do not strictly meet the flow conditions for precise application of the laminar flow

TABLE 68.1 Resistance Measurements for Catheter Components Used for Infusion

Component	Length, cm	Flow Resistance, Fluid Ohm, mmHg/(l/h)
Standard administration set	91–213	4.3–5.3
Extension tube for CVP monitoring	15	15.5
19-gauge epidural catheter	91	290.4–497.1
18-gauge needle	6–9	14.1–17.9
23-gauge needle	2.5–9	165.2–344.0
25-gauge needle	1.5–4.0	525.1–1412.0
Vicra Quick-Cath Catheter 18-gauge	5	12.9
Extension set with 0.22 μm air-eliminating filter		623.0
0.2 μm filter		555.0

Note: Mean values are presented over a range of infusions (100, 200, and 300 ml/h) and sample size ($n = 10$).

equation, it does provide insight into the relationship between flow and pressure in a catheter. The fluid analog of Ohms Law describes the resistance to flow under constant flow conditions:

$$R = \frac{P_1 - P_2}{Q} \tag{68.2}$$

Thus, resistance to flow through a tube correlates directly with catheter length and fluid viscosity and inversely with the fourth power of catheter diameter. For steady flow, the delivery system can be modeled as a series of resistors representing each component, including administration set, access catheter, and circulatory system. When dynamic aspects of the delivery system are considered, a more detailed model including catheter and venous compliance, fluid inertia, and turbulent flow is required. Flow resistance may be defined with units of mmHg/(l/h), so that 1 fluid ohm $= 4.8 \times 10^{-11}$ Pa sec/m^3. Studies determining flow resistance for several catheter components with distilled water for flow rates of 100, 200, and 300 ml/h appear in Table 68.1.

68.3 Intravenous Infusion Devices

From Hagen–Poiselluie's equation, two general approaches to intravenous infusion become apparent. First, a hydrostatic pressure gradient can be used with adjustment of delivery system resistance controlling flow rate. Complications such as partial obstructions result in reduced flow which may be detected by an automatic flow monitor. Second, a constant displacement flow source can be used. Now complications may be detected by monitoring elevated fluid pressure and/or flow resistance. At the risk of overgeneralization, the relative strengths of each approach will be presented.

68.3.1 Gravity Flow/Resistance Regulation

The simplest means for providing regulated flow employs gravity as the driving force with a roller clamp as controlled resistance. Placement of the fluid reservoir 60 to 100 cm above the patient's right atrium provides a hydrostatic pressure gradient P_h equal to 1.34 mmHg/cm of elevation. The modest physiologic mean pressure in the veins, P_v, minimally reduces the net hydrostatic pressure gradient. The equation for flow becomes

$$Q = \frac{P_\mathrm{h} - P_\mathrm{v}}{R_\mathrm{mfr} + R_\mathrm{n}} \tag{68.3}$$

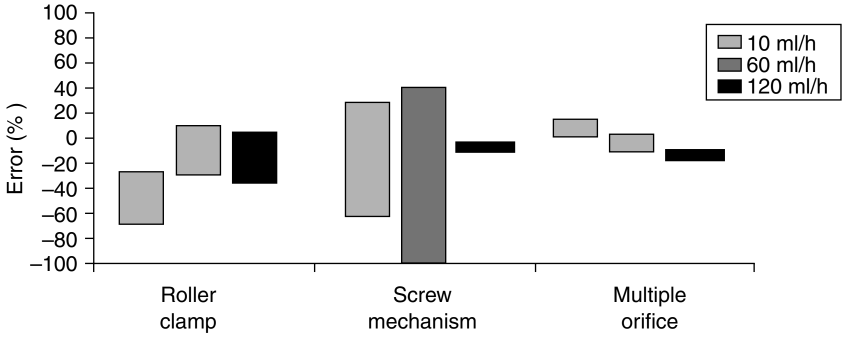


FIGURE 68.3 Drift in flow rate (mean ± standard deviation) over a 4-h period for three mechanical flow regulators at initial flow rates of 10, 60, and 120 ml/h with distilled water at constant hydrostatic pressure gradient.

where R_{mfr} and R_{n} are the resistance to flow through the mechanical flow regulator and the remainder of the delivery system, respectively. Replacing the variables with representative values for an infusion of 5% saline solution into a healthy adult at 100 ml/h yields

$$100\ \mathrm{ml/h} = \frac{(68 - 8)\ \mathrm{mmHg}}{(550 + 50)\ \mathrm{mmHg/(l/h)}} \tag{68.4}$$

Gravity flow cannot be used for arterial infusions since the higher vascular pressure exceeds available hydrostatic pressure.

Flow stability in a gravity infusion system is subject to variations in hydrostatic and venous pressure as well as catheter resistance. However, the most important factor is the change in flow regulator resistance caused by viscoelastic creep of the tubing wall (see Figure 68.3). Caution must be used in assuming that a preset flow regulator setting will accurately provide a predetermined rate. The clinician typically estimates flow rate by counting the frequency of drops falling through an in-line drip-forming chamber, adjusting the clamp to obtain the desired drop rate. The cross-sectional area of the drip chamber orifice is the major determinant of drop volume. Various manufacturers provide minidrip sets designed for pediatric (e.g., 60 drops/ml) and regular sets designed for adult (10 to 20 drops/ml) patients. Tolerances on the drip chamber can cause a 3% error in minidrip sets and a 17% error in regular sets at 125 ml/h flow rate with 5% dextrose in water. Mean drop size for rapid rates increased by as much as 25% over the size of drops which form slowly. In addition, variation in the specific gravity and surface tension of fluids can provide an additional large source of drop size variability.

Some mechanical flow regulating devices incorporate the principle of a Starling resistor. In a Starling device, resistance is proportional to hydrostatic pressure gradient. Thus, the device provides a negative feedback mechanism to reduce flow variation as the available pressure gradient changes with time.

Mechanical flow regulators comprise the largest segment of intravenous infusion systems, providing the simplest means of operation. Patient transport is simple, since these devices require no electric power. Mechanical flow regulators are most useful where the patient is not fluid restricted and the acceptable therapeutic rate range of the drug is relatively wide with minimal risk of serious adverse sequelae. The most common use for these systems is the administration of fluids and electrolytes.

68.3.2 Volumetric Infusion Pumps

Active pumping infusion devices combine electronics with a mechanism to generate flow. These devices have higher performance standards than simple gravity flow regulators. The Association for the Advancement of Medical Instrumentation (AAMI) recommends that long-term rate accuracy for infusion pumps remain within ±10% of the set rate for general infusion and, for the more demanding applications, that long-term flow remain within ±5%. Such requirements typically extend to those agents with narrow therapeutic indices and/or low flow rates, such as the neonatal population or other fluid-restricted patients. The British Department of Health has established three main categories for hospital-based infusion devices: neonatal infusions, high-risk infusions, and low-risk infusions. Infusion control for neonates requires the highest performance standards, because their size severely restricts fluid volume. A fourth category, ambulatory infusion, pertains to pumps worn by patients.

68.3.3 Controllers

These devices automate the process of adjusting the mechanical flow regulator. The most common controllers utilize sensors to count the number of drops passing through the drip chamber to provide flow feedback for automatic rate adjustment. Flow rate accuracy remains limited by the rate and viscosity dependence of drop size. Delivery set motion associated with ambulation and improper angulation of the drip chamber can also hinder accurate rate detection.

An alternative to the drop counter is a volumetric metering chamber. A McGaw Corporation controller delivery set uses a rigid chamber divided by a flexible membrane. Instrument-controlled valves allow fluid to fill one chamber from the fluid reservoir, displacing the membrane driving the fluid from the second chamber toward the patient. When inlet and outlet valves reverse state, the second chamber is filled while the first chamber delivers to the patient. The frequency of state change determines the average flow rate. Volumetric accuracy demands primarily on the dimensional tolerances of the chamber. Although volumetric controllers may provide greater accuracy than drop-counting controllers, their disposables are inherently more complex, and maximum flow is still limited by head height and system resistance.

Beyond improvements in flow rate accuracy, controllers should provide an added level of patient safety by quickly detecting IV-site complications. The IVAC Corporation has developed a series of controllers employing pulsed modulated flow providing for monitoring of flow resistance as well as improved accuracy.

The maximum flow rate achieved by gravimetric based infusion systems can become limited by R_n and by concurrent infusion from other sources through the same catheter. In drop-counting devices, flow rate uniformity suffers at low flow rates from the discrete nature of the drop detector.

In contrast with infusion controllers, pumps generate flow by mechanized displacement of the contents of a volumetric chamber. Typical designs provide high flow rate accuracy and uniformity for a wide rate range (0.1 to 1000.0 ml/h) of infusion rates. Rate error correlates directly with effective chamber volume, which, in turn, depends on both instrument and disposable repeatability, precision, and stability under varying load. Stepper or servo-controlled dc motors are typically used to provide the driving force for the fluid. At low flow rates, dc motors usually operate in a discrete stepping mode. On average, each step propels a small quanta of fluid toward the patient. Flow rate uniformity therefore is a function of both the average volume per quanta and the variation in volume. Mechanism factors influencing rate uniformity include: stepping resolution, gearing and activator geometries, volumetric chamber coupling geometry, and chamber elasticity. When the quanta volume is not inherently uniform over the mechanism's cycle, software control has been used to compensate for the variation.

68.3.4 Syringe Pumps

These pumps employ a syringe as both reservoir and volumetric pumping chamber. A precision leadscrew is used to produce constant linear advancement of the syringe plunger. Except for those ambulatory systems that utilize specific microsyringes, pumps generally accept syringes ranging in size from 5 to 100 ml. Flow rate accuracy and uniformity are determined by both mechanism displacement characteristics and tolerance on the internal syringe diameter. Since syringe mechanisms can generate a specified linear travel with less than 1% error, the manufacturing tolerance on the internal cross-sectional area of the syringe largely determines flow rate accuracy. Although syringes can be manufactured to tighter tolerances, standard plastic syringes provide long-term accuracy of ±5%. Flow rate uniformity, however, can benefit

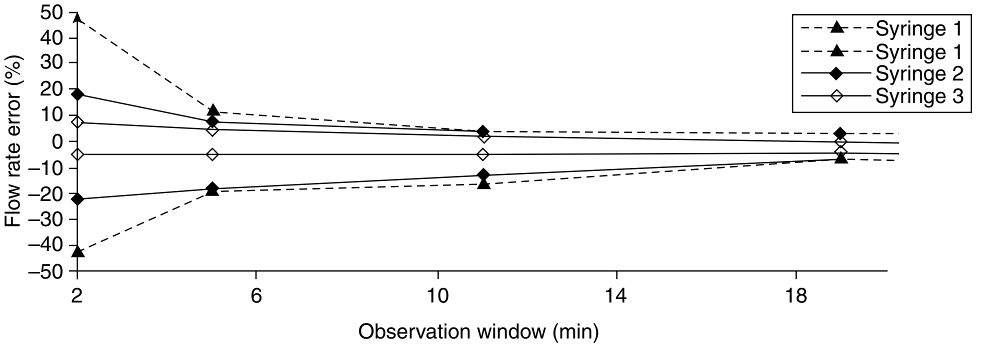


FIGURE 68.4 Effect of syringe type on Trumpet curve of a syringe pump at 1 ml/h.

from the ability to select syringe size (see Figure 68.4). Since many syringes have similar stroke length, diameter variation provides control of volume. Also the linear advancement per step is typically fixed. Therefore selection of a lower-volume syringe provides smaller-volume quanta. This allows tradeoffs among drug concentration, flow rate, and duration of flow per syringe. Slack in the gear train and drive shaft coupling as well as plunger slip cause rate inaccuracies during the initial stages of delivery (see Figure 68.5a).

Since the syringe volumes are typically much smaller than reservoirs used with other infusion devices, syringe pumps generally deliver drugs in either fluid-restricted environments or for short duration. With high-quality syringes, flow rate uniformity in syringe pumps is generally superior to that accomplished by other infusion pumps. With the drug reservoir enclosed within the device, syringe pumps manage patient transport well, including the operating room environment.

Cassette pumps conceptually mimic the piston type action of the syringe pump but provide an automated means of repeatedly emptying and refilling the cassette. The process of refilling the cassette in single piston devices requires an interruption in flow (see Figure 68.5b). The length of interruption relative to the drug's half-life determines the impact of the refill period on hemodynamic stability. To eliminate the interruption caused by refill, dual piston devices alternate refill and delivery states, providing nearly continuous output. Others implement cassettes with very small volumes which can refill in less than a second (see Figure 68.2). Tight control of the internal cross-sectional area of the pumping chamber provides exceptional flow rate accuracy. Manufacturers have recently developed remarkably small cassette pumps that can still generate the full spectrum of infusion rate (0.1 to 999.0 ml/h). These systems combine pumping chamber, inlet and outlet valving, pressure sensing, and air detection into a single complex component.

Peristaltic pumps operate on a short segment of the IV tubing. Peristaltic pumps can be separated into two subtypes. Rotary peristaltic mechanisms operate by compressing the pumping segment against the rotor housing with rollers mounted on the housing. With rotation, the rollers push fluid from the container through the tubing toward the patient. At least one of the rollers completely occludes the tubing against the housing at all times precluding free flow from the reservoir to the patient. During a portion of the revolution, two rollers trap fluid in the intervening pumping segment. The captured volume between the rollers determines volumetric accuracy. Linear peristaltic pumps hold the pumping segment in a channel pressed against a rigid backing plate. An array of cam-driven actuators sequentially occlude the segment starting with the section nearest the reservoir forcing fluid toward the patient with a sinusoidal wave action. In a typical design using uniform motor step intervals, a characteristic flow wave resembling a positively biased sine wave is produced (see Figure 68.5c).

Infusion pumps provide significant advantages over both manual flow regulators and controllers in several categories. Infusion pumps can provide accurate delivery over a wide range of infusion rates (0.1 to 999.0 ml/h). Neither elevated system resistance nor distal line pressure limit the maximum infusion rate. Infusion pumps can support a wider range of applications including arterial infusions, spinal and

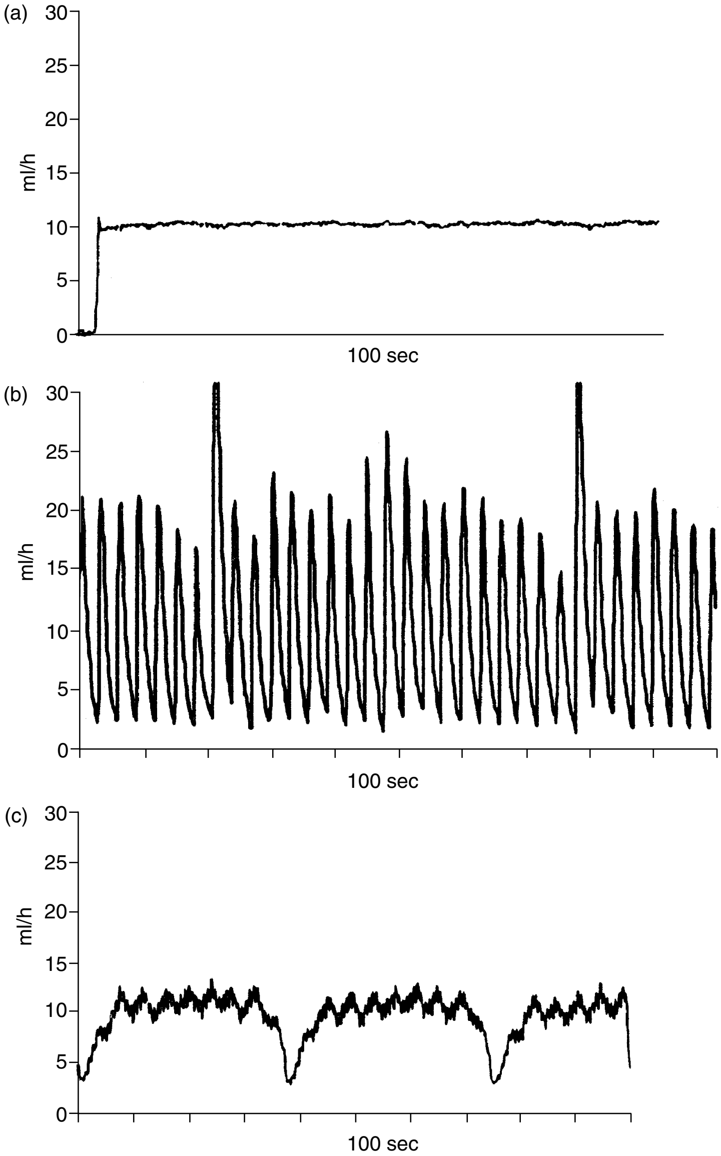


FIGURE 68.5 Continuous flow pattern for a representative, (a) syringe, (b) cassette, and (c) linear peristaltic pump at 10 ml/h.

epidural infusions, and infusions into pulmonary artery or central venous catheters. Flow rate accuracy of infusion pumps is highly dependent on the segment employed as the pumping chamber (see Figure 68.2). Incorporating special syringes or pumping segments can significantly improve flow rate accuracy (see Figure 68.6). Both manufacturing tolerances and segment material composition significantly dictate flow rate accuracy. Time- and temperature-related properties of the pumping segment further impact long-term drift in flow rate.

68.4 Managing Occlusions of the Delivery System

One of the most common problems in managing an IV delivery system is the rapid detection of occlusion in the delivery system. With a complete occlusion, the resistance to flow approaches infinity. In this condition,

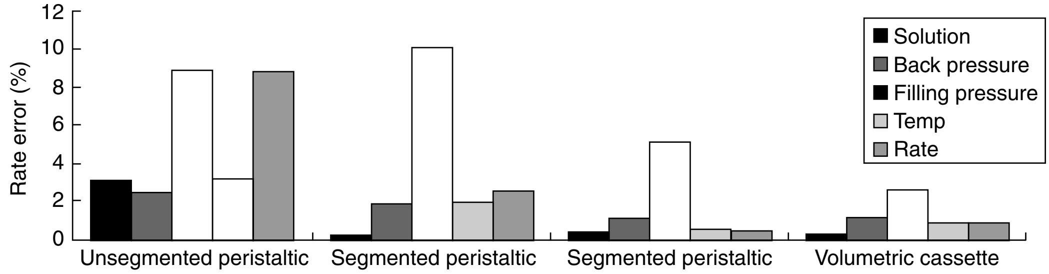


FIGURE 68.6 Impact of 5 variables on flow rate accuracy in 4 different infusion pumps. Variables tested included Solution: Distilled water and 25% dextrose in water; Back pressure: −100 and 300 mmHg; Pumping Segment Filling Pressure: −30 inches of water and +30 inches of water; Temperature: 10°C and 40°C; and Infusion rate: 5 ml/h and 500 ml/h. Note: first and second peristaltic mechanism qualified for low risk patients, while the third peristaltic device qualified for high-risk patients.

gravimetric-based devices cease to generate flow. Mechanical flow regulators have no mechanism for adverse event detection and thus must rely on the clinician to identify an occlusion as part of routine patient care. Electronic controllers sense the absence of flow and alarm in response to their inability to sustain the desired flow rate.

The problem of rapidly detecting an occlusion in an infusion pump is more complex. Upstream occlusions that occur between the fluid reservoir and the pumping mechanism impact the system quite differently than downstream occlusions which occur between the pump and the patient. When an occlusion occurs downstream from an infusion pump, the pump continues to propel fluid into the section of tubing between the pump and the occlusion. The time rate of pressure rise in that section increases in direct proportion to flow rate and inversely with tubing compliance (compliance, C, is the volume increase in a closed tube per mmHg pressure applied). The most common approach to detecting downstream occlusion requires a pressure transducer immediately below the pumping mechanism. These devices generate an alarm when either the mean pressure or rate of change in pressure exceeds a threshold. For pressure-limited designs, the time to downstream alarm (TTA) may be estimated as

$$\text{TTA} = \frac{P_{\text{alarm}} \cdot C_{\text{delivery-set}}}{\text{flow rate}} \tag{68.5}$$

Using a representative tubing compliance of 1 μl/mmHg, flow rate of 1 ml/h, and a fixed alarm threshold set of 500 mmHg, the time to alarm becomes

$$\text{TTA} = \frac{500_{\text{mmHg}} \cdot 1000\ \text{ml/mmHg}}{1\ \text{ml/h}} = 30\ \text{min} \tag{68.6}$$

where TTA is the time from occlusion to alarm detection. Pressure-based detection algorithms depend on accuracy and stability of the sensing system. Lowering the threshold on absolute or relative pressure for occlusion alarm reduces the TTA, but at the cost of increasing the likelihood of false alarms. Patient movement, patient-to-pump height variations, and other clinical circumstances can cause wide perturbations in line pressure. To optimize the balance between fast TTA and minimal false alarms, some infusion pumps allow the alarm threshold to be set by the clinician or be automatically shifted upward in response to alarms; other pumps attempt to optimize performance by varying pressure alarm thresholds with flow rate.

A second approach to detection of downstream occlusions uses motor torque as an indirect measure of the load seen by the pumping mechanism. Although this approach eliminates the need for a pressure sensor, it introduces additional sources for error including friction in the gear mechanism or pumping mechanism that requires additional safety margins to protect against false alarms. In syringe pumps, where the coefficient of static friction of the syringe bunge (rubber end of the syringe plunger) against the syringe wall can be substantial, occlusion detection can exceed 1 h at low flow rates.

Direct, continuous measurement of downstream flow resistance may provide a monitoring modality which overcomes the disadvantages of pressure-based alarm systems, especially at low infusion rates. Such a monitoring system would have the added advantage of performance unaffected by flow rate, hydrostatic pressure variations, and motion artifacts.

Upstream occlusions can cause large negative pressures as the pumping mechanism generates a vacuum on the upstream tubing segment. The tube may collapse and the vacuum may pull air through the tubing walls or form cavitation bubbles. A pressure sensor situated above the mechanism or a pressure sensor below the mechanism synchronized with filling of the pumping chamber can detect the vacuum associated with an upstream occlusion. Optical or ultrasound transducers, situated below the mechanism, can detect air bubbles in the catheter, and air-eliminating filters can remove air, preventing large air emboli from being introduced into the patient.

Some of the most serious complications of IV therapy occur at the venipuncture site; these include extravasation, postinfusion phlebitis (and thrombophlebitis), IV-related infections, ecchymosis, and hematomas. Other problems that do not occur as frequently include speed shock and allergic reactions.

Extravasation (or infiltration) is the inadvertent perfusion of infusate into the interstitial tissue. Reported percentage of patients to whom extravasation has occurred ranges from 10% to over 25%. Tissue damage does not occur frequently, but the consequences can be severe, including skin necrosis requiring significant plastic and reconstructive surgery and amputation of limbs. The frequency of extravasation injury correlates with age, state of consciousness, and venous circulation of the patient as well as the type, location, and placement of the intravenous cannula. Drugs that have high osmolality, vessicant properties, or the ability to induce ischemia correlate with frequency of extravasation injury. Neonatal and pediatric patients who possess limited communication skills, constantly move, and have small veins that are difficult to cannulate require superior vigilance to protect against extravasation.

Since interstitial tissue provides a greater resistance to fluid flow than the venous pathway, infusion devices with accurate and precise pressure monitoring systems have been used to detect small pressure increases due to extravasation. To successfully implement this technique requires diligence by the clinician, since patient movement, flow rate, catheter resistance, and venous pressure variations can obscure the small pressure variations resulting from the extravasation. Others have investigated the ability of a pumping mechanism to withdraw blood as indicative of problems in a patent line. The catheter tip, however, may be partially in and out of the vein such that infiltration occurs yet blood can be withdrawn from the patient. A vein might also collapse under negative pressure in a patent line without successful blood withdrawal. Techniques currently being investigated which monitor infusion impedance (resistance and compliance) show promise for assisting in the detection of extravasation.

When a catheter tip wedges into the internal lining of the vein wall, it is considered positional. With the fluid path restricted by the vein wall, increases in line resistance may indicate a positional catheter. With patient movement, for example wrist flexation, the catheter may move in and out of the positional state. Since a positional catheter is thought to be more prone toward extravasation than other catheters, early detection of a positional catheter and appropriate adjustment of catheter position may be helpful in reducing the frequency of extravasation.

Postinfusion phlebitis is acute inflammation of a vein used for IV infusion. The chief characteristic is a reddened area or red streak that follows the course of the vein with tenderness, warmth, and edema at the venipuncture site. The vein, which normally is very compliant, also hardens. Phlebitis positively correlates with infusion rate and with the infusion of vesicants.

Fluid overload and speed shock result from the accidental administration of a large fluid volume over a short interval. Speed shock associates more frequently with the delivery of potent medications, rather than fluids. These problems most commonly occur with manually regulated IV systems, which do not provide the safety features of instrumented lines. Many IV sets designed for instrumented operation will free-flow when the set is removed from the instrument without manual clamping. To protect against this possibility, some sets are automatically placed in the occluded state on disengagement. Although an apparent advantage, reliance on such automatic devices may create a false sense of security and lead to manual errors with sets not incorporating these features.

68.5 Summary

Intravenous infusion has become the mode of choice for delivery of a large class of fluids and drugs both in hospital and alternative care settings. Modern infusion devices provide the clinician with a wide array of choices for performing intravenous therapy. Selection of the appropriate device for a specified application requires understanding of drug pharmacology and pharmacokinetics, fluid mechanics, and device design and performance characteristics. Continuing improvements in performance, safety, and cost of these systems will allow even broader utilization of intravenous delivery in a variety of settings.

References

Association for the Advancement of Medical Instrumentation (1992). *Standard for Infusion Devices.* Arlington.

Bohony J. (1993). Nine common intravenous complications and what to do about them. *Am. J. Nursing* 10: 45.

British Department of Health (1990). *Evaluation of Infusion Pumps and Controllers.* HEI Report #198.

Glass P.S.A., Jacobs J.R., Reves J.G. (1991). Technology for continuous infusions in anesthesia. Continuous Infusions in Anesthesia. *Int. Anesthesiol. Clin.* 29: 39.

MacCara M. (1983). Extravasation: A hazard of intravenous therapy. *Drug Intell. Clin. Pharm.* 17: 713.

Further Information

Peter Glass provides a strong rationale for intravenous therapy including pharmacokinetic and pharmacodynamic bases for continuous delivery. Clinical complications around intravenous therapy are well summarized by MacCara [1983] and Bohony [1993]. The AAMI Standard for Infusion Devices provides a comprehensive means of evaluating infusion device technology, and the British Department of Health OHEI Report #198 provides a competitive analysis of pumps and controllers.

69

Clinical Laboratory: Separation and Spectral Methods

Richard L. Roa
Baylor University Medical Center

The purpose of the clinical laboratory is to analyze body fluids and tissues for specific substances of interest and to report the results in a form which is of value to clinicians in the diagnosis and treatment of disease. A large range of tests has been developed to achieve this purpose. Four terms commonly used to describe tests are **accuracy, precision, sensitivity,** and **specificity.** An accurate test, on average, yields true values. Precision is the ability of a test to produce identical results upon repeated trials. Sensitivity is a measure of how small an amount of substance can be measured. Specificity is the degree to which a test measures the substance of interest without being affected by other substances which may be present in greater amounts.

The first step in many laboratory tests is to separate the material of interest from other substances. This may be accomplished through extraction, filtration, and centrifugation. Another step is derivatization, in which the substance of interest is chemically altered through addition of reagents to change it into a substance which is easily measured. For example, one method for measuring glucose is to add otoluidine which, under proper conditions, forms a green-colored solution with an absorption maximum at 630 nm. Separation and derivatization both improve the specificity required of good tests.

69.1 Separation Methods

Centrifuges are used to separate materials on the basis of their relative densities. The most common use in the laboratory is the separation of cells and platelets from the liquid part of the blood. This requires a relative centrifugal force (RCF) of roughly 1000 g (1000 times the force of gravity) for a period of 10 min. Relative centrifugal force is a function of the speed of rotation and the distance of the sample from the center of rotation as stated in Equation 69.1

$$\text{RCF} = (1.12 \times 10^{-5})\, r(\text{rpm})^2 \tag{69.1}$$

where RCF is the relative centrifugal force in g, and r is the radius in cm.

Some mixtures require higher g-loads in order to achieve separation in a reasonable period of time. Special rotors contain the sample tubes inside a smooth container, which minimizes air resistance to allow faster rotational speeds. Refrigerated units maintain the samples at a cool temperature throughout long high-speed runs which could lead to sample heating due to air friction on the rotor. Ultracentrifuges operate at speeds on the order of 100,000 rpm and provide relative centrifugal forces of up to 600,000 g. These usually require vacuum pumps to remove the air which would otherwise retard the rotation and heat the rotor.

69.2 Chromatographic Separations

Chromatographic separations depend upon the different rates at which various substances moving in a stream (mobile phase) are retarded by a stationary material (stationary phase) as they pass over it. The mobile phase can be a volatilized sample transported by an inert carrier gas such as helium or a liquid transported by an organic solvent such as acetone. Stationary phases are quite diverse depending upon the separation being made, but most are contained within a long, thin tube (column). Liquid stationary phases may be used by coating them onto inert packing materials. When a sample is introduced into a chromatographic column, it is carried through it by the mobile phase. As it passes through the column, the substances which have greater affinity for the stationary phase fall behind those with less affinity. The separated substances may be detected as individual peaks by a suitable detector placed at the end of the chromatographic column.

69.3 Gas Chromatography

The most common instrumental chromatographic method used in the clinical laboratory is the gas–liquid chromatograph. In this system the mobile phase is a gas, and the stationary phase is a liquid coated onto either an inert support material, in the case of a packed column, or the inner walls of a very thin tube, in the case of a capillary column. Capillary columns have the greatest resolving power but cannot handle large sample quantities. The sample is injected into a small heated chamber at the beginning of the column, where it is volatilized if it is not already a gaseous sample. The sample is then carried through the column by an inert carrier gas, typically helium or nitrogen. The column is completely housed within an oven. Many gas chromatographs allow for the oven temperature to be programmed to slowly increase for a set time after the sample injection is made. This produces peaks which are spread more uniformly over time.

Four detection methods commonly used with gas chromatography are thermal conductivity, flame ionization, nitrogen/phosphorous, and mass spectrometry. The thermal conductivity detector takes advantage of variations in thermal conductivity between the carrier gas and the gas being measured. A heated filament immersed in the gas leaving the chromatographic column is part of a Wheatstone bridge circuit. Small variations in the conductivity of the gas cause changes in the resistance of the filament, which are recorded. The flame ionization detector measures the current between two plates with a voltage applied between them. When an organic material appears in the flame, ions which contribute to the current are formed. The NP detector, or nitrogen/phosphorous detector, is a modified flame ionization detector (see Figure 69.1) which is particularly sensitive to nitrogen- and phosphorous-containing compounds.

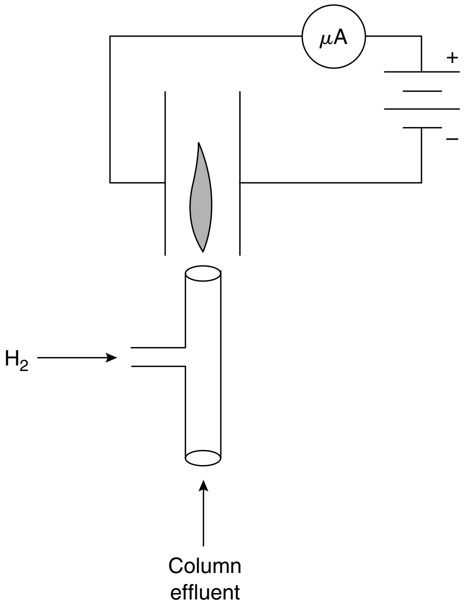


FIGURE 69.1 Flame ionization detector. Organic compounds in the column effluent are ionized in the flame, producing a current proportional to the amount of the compound present.

Mass spectrometry (MS) provides excellent sensitivity and selectivity. The concept behind these devices is that the volatilized sample molecules are broken into ionized fragments which are then passed through a mass analyzer that separates the fragments according to their mass/charge (m/z) ratios. A mass spectrum, which is a plot of the relative abundance of the various fragments versus m/z, is produced. The mass spectrum is characteristic of the molecule sampled. The mass analyzer most commonly used with gas chromatographs is the quadrupole detector, which consists of four rods that have dc and RF voltages applied to them. The m/z spectrum can be scanned by appropriate changes in the applied voltages. The detector operates in a manner similar to that of a photomultiplier tube except that the collision of the charged particles with the cathode begins the electron cascade, resulting in a measurable electric pulse for each charged particle captured. The MS must operate in a high vacuum, which requires good pumps and a porous barrier between the GC and MS that limits the amount of carrier gas entering the MS.

69.4 High-Performance Liquid Chromatography

In liquid chromatography, the mobile phase is liquid. High-performance liquid chromatography (HPLC) refers to systems which obtain excellent resolution in a reasonable time by forcing the mobile phase at high pressure through a long thin column. The most common pumps used are pistons driven by asymmetrical cams. By using two such pumps in parallel and operating out of phase, pressure fluctuations can be minimized. Typical pressures are 350–1500 psi, though the pressure may be as high as 10,000 psi. Flow rates are in the 1–10 ml/min range.

A common method for placing a sample onto the column is with a loop injector, consisting of a loop of tubing which is filled with the sample. By a rotation of the loop, it is brought in series with the column, and the sample is carried onto the column. A UV/visible spectrophotometer is often used as a detector for this method. A mercury arc lamp with the 254-nm emission isolated is useful for detection of aromatic compounds, while diode array detectors allow a complete spectrum from 190 to 600 nm in 10 msec. This provides for detection and identification of compounds as they come off the column. Fluorescent, electrochemical, and mass analyzer detectors are also used.

69.5 Basis for Spectral Methods

Spectral methods rely on the absorption or emission of electromagnetic radiation by the sample of interest. Electromagnetic radiation is often described in terms of frequency or wavelength. Wavelengths are those obtained in a vacuum and may be calculated with the formula

$$\lambda = c/\upsilon \tag{69.2}$$

where λ is the wavelength in meters, c the speed of light in vacuum (3×10^8 m/sec), and ν the frequency in Hz.

The frequency range of interest for most clinical laboratory work consists of the visible (390–780 nm) and the ultraviolet or UV (180–390 nm) ranges. Many substances absorb different wavelengths preferentially. When this occurs in the visible region, they are colored. In general, the color of a substance is the complement of the color it absorbs, for example, absorption in the blue produces a yellow color. For a given wavelength or bandwidth, transmittance is defined as

$$T = \frac{I_t}{I_i} \tag{69.3}$$

where T is the transmittance ratio (often expressed as %), I_i the incident light intensity, and I_t the transmitted light intensity. Absorbance is defined as

$$A = -\log_{10} 1/T \tag{69.4}$$

Under suitable conditions, the absorbance of a solution with an absorbing compound dissolved in it is proportional to the concentration of that compound as well as the path length of light through it. This relationship is expressed by Beer's law:

$$A = abc \tag{69.5}$$

where A is the absorbance, a the a constant, b the path length, and c the concentration.

A number of situations may cause deviations from Beer's law, such as high concentration or mixtures of compounds which absorb at the wavelength of interest. From an instrumental standpoint, the primary causes are stray light and excessive spectral bandwidth. Stray light refers to any light reaching the detector other than light from the desired pass-band which has passed through sample. Sources of stray light may include room light leaking into the detection chamber, scatter from the cuvette, and undesired **fluorescence**.

A typical spectrophotometer consists of a light source, some form of wavelength selection, and a detector for measuring the light transmitted through the samples. There is no single light source that covers the entire visible and UV spectrum. The source most commonly used for the visible part of the spectrum is the tungsten–halogen lamp, which provides continuous radiation over the range of 360 to 950 nm. The deuterium lamp has become the standard for much UV work. It covers the range from 220 to 360 nm. Instruments which cover the entire UV/visible range use both lamps with a means for switching from one lamp to the other at a wavelength of approximately 360 nm (Figure 69.2).

Wavelength selection is accomplished with filters, prisms, and diffraction gratings. Specially designed interference filters can provide bandwidths as small as 5 nm. These are useful for instruments which do not need to scan a range of wavelengths. Prisms produce a nonlinear dispersion of wavelengths with the longer wavelengths closer together than the shorter ones. Since the light must pass through the prism material, they must be made of quartz for UV work. Diffraction gratings are surfaces with 1000 to 3000 grooves/mm cut into them. They may be transmissive or reflective; the reflective ones are more popular since there is no attenuation of light by the material. They produce a linear dispersion. By proper selection of slit widths, pass bands of 0.1 nm are commonly achieved.

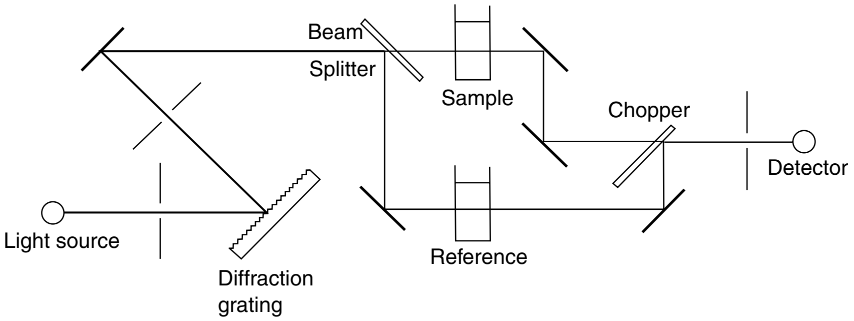


FIGURE 69.2 Dual-beam spectrophotometer. The diffraction grating is rotated to select the desired wavelength. The beam splitter consists of a half-silvered mirror which passes half the light while reflecting the other half. A rotating mirror with cut-out sections (chopper) alternately directs one beam and then the other to the detector.

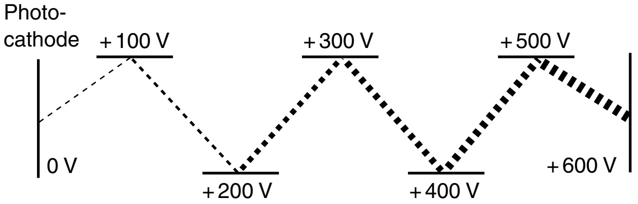


FIGURE 69.3 Photomultiplier tube. Incident photons cause the photocathode to emit electrons which collide with the first dynode which emits additional electrons. Multiple dynodes provide sufficient gain to produce an easily measurable electric pulse from a single photon.

The most common detector is the photomultiplier tube, which consists of a photosensitive cathode that emits electrons in proportion to the intensity of light striking it (Figure 69.3). A series of 10–15 dynodes, each at 50–100 V greater potential than the preceding one, produce an electron amplification of 4–6 per stage. Overall gains are typically a million or more. Photomultiplier tubes respond quickly and cover the entire spectral range. They require a high voltage supply and can be damaged if exposed to room light while the high voltage is applied.

69.6 Fluorometry

Certain molecules absorb a photon's energy and then emit a photon with less energy (longer wavelength). When the reemission occurs in less than 10^{-8} sec, the process is known as fluorescence. This physical process provides the means for assays which are 10 to 100 times as sensitive as those based on absorption measurements. This increase in sensitivity is largely because the light measured is all from the sample of interest. A dim light is easily measured against a black background, while it may be lost if added to an already bright background.

Fluorometers and spectrofluorometers are very similar to photometers and spectrophotometers but with two major differences. Fluorometers and spectrofluorometers use two monochrometers, one for excitation light and one for emitted light. By proper selection of the bandpass regions, all the light used to excite the sample can be blocked from the detector, assuring that the detector sees only fluorescence. The other difference is that the detector is aligned off-axis, commonly at 90°, from the excitation source. At this angle, scatter is minimal, which helps ensure a dark background for the measured fluorescence. Some spectrofluorometers use polarization filters both on the input and output light beams, which allows for fluorescence polarization studies (Figure 69.4). An intense light source in the visible-to-UV range is

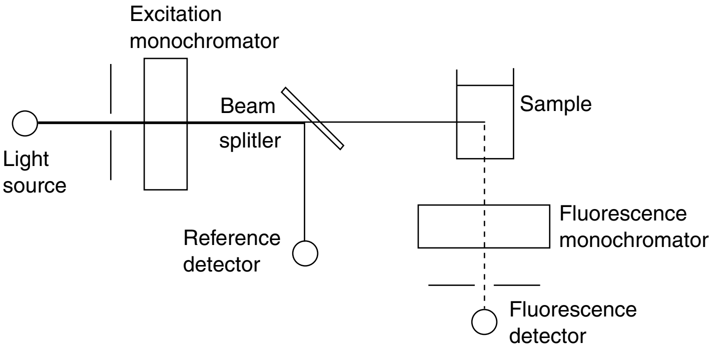


FIGURE 69.4 Spectrofluorometer. Fluorescence methods can be extremely sensitive to the low background interference. Since the detector is off-axis from the incident light and a second monochromator blocks light of wavelengths illuminating the sample, virtually no signal reaches the detector other than the desired fluorescence.

desirable. A common source is the xenon or mercury arc lamps, which provide a continuum of radiation over this range.

69.7 Flame Photometry

Flame photometry is used to measure sodium, potassium, and lithium in body fluids. When these elements are heated in a flame they emit characteristic wavelengths of light. The major emission lines are 589 nm (yellow) for sodium, 767 nm (violet) for potassium, and 671 nm (red) for lithium. An atomizer introduces a fine mist of the sample into a flame. For routine laboratory use, a propane and compressed air flame is adequate. High-quality interference filters with narrow pass bands are often used to isolate the major emission lines. The narrow band pass is necessary to maximize the signal-to-noise ratio. Since it is impossible to maintain stable aspiration, atomization, and flame characteristics, it is necessary to use an internal standard of known concentration while making measurements of unknowns. In this way the ratio of the unknown sample's emission to the internal standard's emission remains stable even as the total signal fluctuates. An internal standard is usually an element which is found in very low concentration in the sample fluid. By adding a high concentration of this element to the sample, its concentration can be known to a high degree of accuracy. Lithium, potassium, and cesium all may be used as internal standards depending upon the particular assay being conducted.

69.8 Atomic Absorption Spectroscopy

Atomic absorption spectroscopy is based on the fact that just as metal elements have unique emission lines, they have identical absorption lines when in a gaseous or dissociated state. The atomic absorption spectrometer takes advantage of these physical characteristics in a clever manner, producing an instrument with approximately 100 times the sensitivity of a flame photometer for similar elements. The sample is aspirated into a flame, where the majority of the atoms of the element being measured remain in the ground state, where they are capable of absorbing light at their characteristic wavelengths. An intense source of exactly these wavelengths is produced by a hollow cathode lamp. These lamps are constructed so that the cathode is made from the element to be measured, and the lamps are filled with a low pressure of argon or neon gas. When a current is passed through the lamp, metal atoms are sputtered off the cathode and collide with the argon or neon in the tube, producing emission of the characteristic wavelengths. A monochromator and photodetector complete the system.

Light reaching the detector is a combination of that which is emitted by the sample (undesirable) and light from the hollow cathode lamp which was not absorbed by the sample in the flame (desirable).

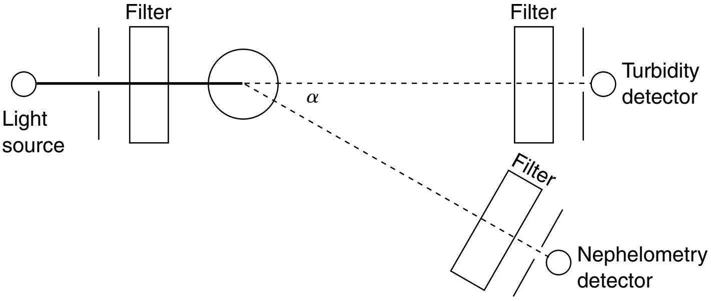


FIGURE 69.5 Nephelometer. Light scattered by large molecules is measured at an angle α away from the axis of incident light. The filters select the wavelength range desired and block undesired fluorescence. When $\alpha = 0$, the technique is known as turbidimetry.

By pulsing the light from the lamp either by directly pulsing the lamp or with a chopper, and using a detector which is sensitive to ac signals and insensitive to dc signals, the undesirable emission signal is eliminated. Each element to be measured requires a lamp with that element present in the cathode. Multielement lamps have been developed to minimize the number of lamps required. Atomic absorption spectrophotometers may be either single beam or double beam; the double-beam instruments have greater stability.

There are various flameless methods for atomic absorption spectroscopy in which the burner is replaced with a method for vaporizing the element of interest without a flame. The graphite furnace which heats the sample to 2700° consists of a hollow graphite tube which is heated by passing a large current through it. The sample is placed within the tube, and the light beam is passed through it while the sample is heated.

69.9 Turbidimetry and Nephelometry

Light scattering by particles in solution is directly proportional to both concentration and molecular weight of the particles. For small molecules the scattering is insignificant, but for proteins, immunoglobulins, immune complexes, and other large particles, light scattering can be an effective method for the detection and measurement of particle concentration. For a given wavelength λ of light and particle size d, scattering is described as Raleigh ($d < \lambda/10$), Raleigh-Debye (dÝλ), or Mie ($d > 10\lambda$). For particles that are small compared to the wavelength, the scattering is equal in all directions. However, as the particle size becomes larger than the wavelength of light, it becomes preferentially scattered in the forward direction. Light-scattering techniques are widely used to detect the formation of antigen–antibody complexes in immunoassays.

When light scattering is measured by the attenuation of a beam of light through a solution, it is called **turbidimetry.** This is essentially the same as absorption measurements with a photometer except that a large pass-band is acceptable. When maximum sensitivity is required a different method is used — direct measurement of the scattered light with a detector placed at an angle to the central beam. This method is called **nephelometry**. A typical nephelometer will have a light source, filter, sample cuvette, and detector set at an angle to the incident beam (Figure 69.5).

Defining Terms

Accuracy: The degree to which the average value of repeated measurements approximate the true value being measured.

Fluorescence: Emission of light by an atom or molecule following absorption of a photon by greater energy. Emission normally occurs within 10^{-8} of absorption.

Nephelometry: Measurement of the amount of light scattered by particles suspended in a fluid.

Precision: A measure of test reproducibility.

Sensitivity: A measure of how small an amount or concentration of an analyte can be detected.
Specificity: A measure of how well a test detects the intended analyte without being "fooled" by other substances in the sample.
Turbidimetry: Measurement of the attenuation of a light beam due to light lost to scattering by particles suspended in a fluid.

References

[1] Burtis C.A. and Ashwood E.R. (Eds.) 1994. *Tietz Textbook of Clinical Chemistry*, 2nd ed., Philadelphia, W.B. Saunders.
[2] Hicks M.R., Haven M.C., and Schenken J.R. et al. (Eds.) 1987. *Laboratory Instrumentation*, 3rd ed., Philadelphia, Lippincott.
[3] Kaplan L.A. and Pesce A.J. (Eds.) 1989. *Clinical Chemistry: Theory, Analysis, and Correlation*, 2rd ed., St. Louis, Mosby.
[4] Tietz N.W. (Ed.) 1987. *Fundamentals of Clinical Chemistry*, 3rd ed., Philadelphia, W.B. Saunders.
[5] Ward J.M., Lehmann C.A., and Leiken A.M. 1994. *Clinical Laboratory Instrumentation and Automation: Principles, Applications, and Selection*, Philadelphia, W.B. Saunders.

70

Clinical Laboratory: Nonspectral Methods and Automation

Richard L. Roa
Baylor University Medical Center

70.1 Particle Counting and Identification

The Coulter principle was the first major advance in automating blood cell counts. The cells to be counted are drawn through a small aperture between two fluid compartments, and the electric impedance between the two compartments is monitored (see Figure 70.1). As cells pass through the aperture, the impedance increases in proportion to the volume of the cell, allowing large numbers of cells to be counted and sized rapidly. Red cells are counted by pulling diluted blood through the aperture. Since red cells greatly outnumber white cells, the contribution of white cells to the red cell count is usually neglected. White cells are counted by first destroying the red cells and using a more concentrated sample.

Modern cell counters using the Coulter principle often use **hydrodynamic focusing** to improve the performance of the instrument. A sheath fluid is introduced which flows along the outside of a channel with the sample stream inside it. By maintaining laminar flow conditions and narrowing the channel, the sample stream is focused into a very thin column with the cells in single file. This eliminates problems with cells flowing along the side of the aperture or sticking to it and minimizes problems with having more than one cell in the aperture at a time.

Flow cytometry is a method for characterizing, counting, and separating cells which are suspended in a fluid. The basic flow cytometer uses hydrodynamic focusing to produce a very thin stream of fluid containing cells moving in single file through a quartz flow chamber (Figure 70.2). The cells are characterized on the basis of their scattering and fluorescent properties. This simultaneous measurement of

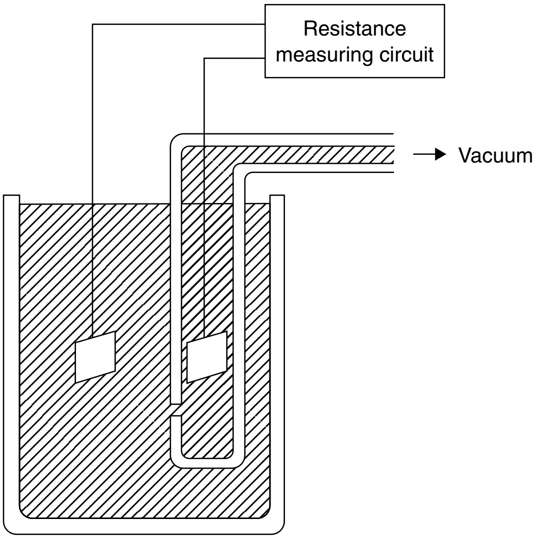


FIGURE 70.1 Coulter method. Blood cells are surrounded by an insulating membrane, which makes them nonconductive. The resistance of electrolyte-filled channel will increase slightly as cells flow through it. This resistance variation yields both the total number of cells which flow through the channel and the volume of each cell.

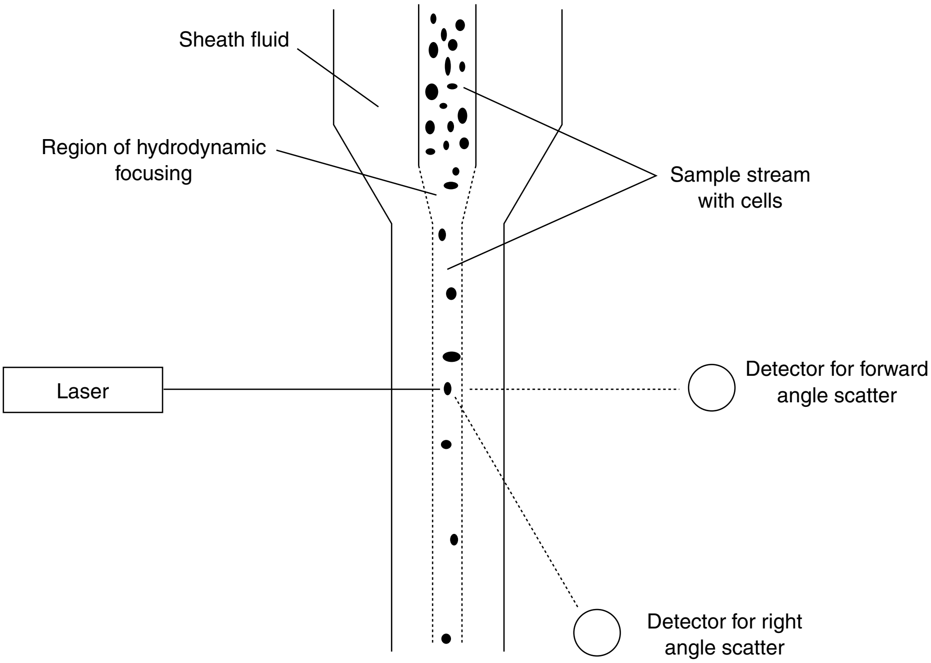

Sheath fluid
Region of hydrodynamic focusing
Sample stream with cells
Laser
Detector for forward angle scatter
Detector for right angle scatter

FIGURE 70.2 Flow cytometer. By combining hydrodynamic focusing, state-of-the-art optics, fluorescent labels, and high-speed computing, large numbers of cells can be characterized and sorted automatically.

scattering and fluorescence is accomplished with a sophisticated optical system that detects light from the sample both at the wavelength of the excitation source (scattering) as well as at longer wavelengths (fluorescence) at more than one angle. Analysis of these measurements produces parameters related to the cells' size, granularity, and natural or tagged fluorescence. High-pressure mercury or xenon arc lamps can be used as light sources, but the argon laser (488 nm) is the preferred source for high-performance instruments.

One of the more interesting features of this technology is that particular cells may be selected at rates that allow collection of quantities of particular cell types adequate for further chemical testing. This is accomplished by breaking the outgoing stream into a series of tiny droplets using piezoelectric vibration. By charging the stream of droplets and then using deflection plates controlled by the cell analyzer, the cells of interest can be diverted into collection vessels.

The development of monoclonal antibodies coupled with flow cytometry allows for quantitation of T and B cells to assess the status of the immune system as well as characterization of leukemias, lymphomas, and other disorders.

70.2 Electrochemical Methods

Electrochemical methods are increasingly popular in the clinical laboratory, for measurement not only of electrolytes, blood gases, and pH but also of simple compounds such as glucose. **Potentiometry** is a method in which a voltage is developed across electrochemical cells as shown in Figure 70.3. This voltage is measured with little or no current flow.

Ideally, one would like to measure all potentials between the reference solution in the indicator electrode and the test solution. Unfortunately there is no way to do that. Interface potentials develop across any metal-liquid boundary, across liquid junctions, and across the ion-selective membrane. The key to making potentiometric measurements is to ensure that all the potentials are constant and do not vary with the composition of the test solution except for the potential of interest across the ion-selective membrane. By maintaining the solutions within the electrodes constant, the potential between these solutions and the metal electrodes immersed in them is constant. The liquid junction is a structure which severely limits bulk flow of the solution but allows free passage of all ions between the solutions. The reference electrode commonly is filled with saturated KCl, which produces a small, constant liquid-junction potential. Thus, any change in the measured voltage (V) is due to a change in the ion concentration in the test solution for which the membrane is selective.

The potential which develops across an ion-selective membrane is given by the Nernst equation:

$$V = \left(\frac{RT}{zF}\right) \ln \frac{a_2}{a_1} \tag{70.1}$$

where R is the gas constant $= 8.314$ J/K mol, $T =$ the temperature in K, $z =$ the ionization number, $F =$ the Faraday constant $= 9.649 \times 10^4$ C/mol, $a_n =$ the activity of ion in solution n. When one of the

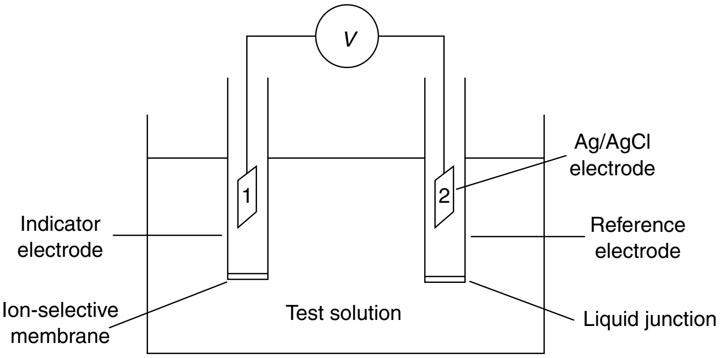


FIGURE 70.3 Electrochemical cell.

solutions is a reference solution, this equation can be rewritten in a convenient form as

$$V = V_0 + \frac{N}{z} \log_{10} a \tag{70.2}$$

where V_0 is the constant voltage due to reference solution and N the Nernst slope Ý 59 mV/decade at room temperature. The actual Nernst slope is usually slightly less than the theoretical value. Thus, the typical pH meter has two calibration controls. One adjusts the offset to account for the value of V_0, and the other adjusts the range to account for both temperature effects and deviations from the theoretical Nernst slope.

70.3 Ion-Specific Electrodes

Ion-selective electrodes use membranes which are permeable only to the ion being measured. To the extent that this can be done, the specificity of the electrode can be very high. One way of overcoming a lack of specificity for certain electrodes is to make multiple simultaneous measurement of several ions which include the most important interfering ones. A simple algorithm can then make corrections for the interfering effects. This technique is used in some commercial electrolyte analyzers. A partial list of the ions that can be measured with ion-selective electrodes includes H^+ (pH), Na^+, K^+, Li^+, Ca^{++}, Cl^-, F^-, NH_4^+, and CO_2.

NH_4^+, and CO_2 are both measured with a modified ion-selective electrode. They use a pH electrode modified with a thin layer of a solution (sodium bicarbonate for CO_2 and ammonium chloride for NH_4^+) whose pH varies depending on the concentration of ammonium ions or CO_2 it is equilibrated with. A thin membrane holds the solution against the pH glass electrode and provides for equilibration with the sample solution. Note that the CO_2 electrode in Figure 70.4 is a combination electrode. This means that both the reference and indicating electrodes have been combined into one unit. Most pH electrodes are made as combination electrodes.

The Clark electrode measures pO_2 by measuring the current developed by an electrode with an applied voltage rather than a voltage measurement. This is an example of **amperometry.** In this electrode a voltage

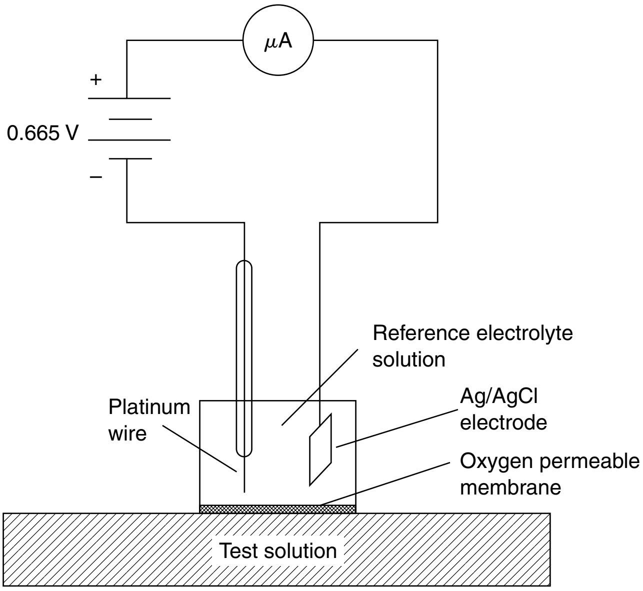


FIGURE 70.4 Clark electrode.

of approximately −0.65 V is applied to a platinum electrode relative to a Ag/AgCl electrode in an electrolyte solution. The reaction

$$O_2 + 2H^+ + 2e^- \rightarrow H_2O_2$$

proceeds at a rate proportional to the partial pressure of oxygen in the solution. The electrons involved in this reaction form a current which is proportional to the rate of the reaction and thus to the pO_2 in the solution.

70.4 Radioactive Methods

Isotopes are atoms which have identical atomic number (number of protons) but different atomic mass numbers (protons + neutrons). Since they have the same number of electrons in the neutral atom, they have identical chemical properties. This provides an ideal method for labeling molecules in a way that allows for detection at extremely low concentrations. Labeling with radioactive isotopes is extensively used in radioimmunoassays where the amount of antigen bound to specific antibodies is measured. The details of radioactive decay are complex, but for our purposes there are three types of emission from decaying nuclei: *alpha, beta,* and **gamma radiation**. Alpha particles are made up of two neutrons and two protons (helium nucleus). Alpha emitters are rarely used in the clinical laboratory. Beta emission consists of electrons or positrons emitted from the nucleus. They have a continuous range of energies up to a maximum value characteristic of the isotope. **Beta radiation** is highly interactive with matter and cannot penetrate very far in most materials. Gamma radiation is a high-energy form of electromagnetic radiation. This type of radiation may be continuous, discrete, or mixed depending on the details of the decay process. It has greater penetrating ability than beta radiation (see Figure 70.5).

The kinetic energy spectrum of emitted radiation is characteristic of the isotope. The energy is commonly measured in electron volts (eV). One electron volt is the energy acquired by an electron falling through a potential of 1 V. The isotopes commonly used in the clinical laboratory have energy spectra which range from 18 keV to 3.6 MeV.

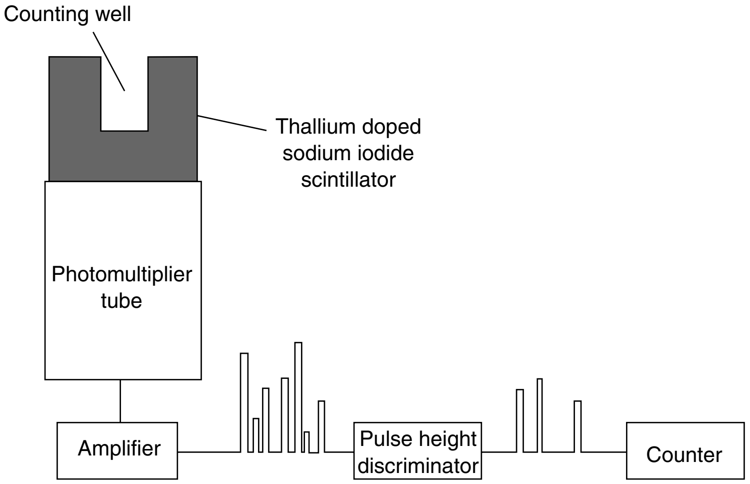

Counting well
Thallium doped sodium iodide scintillator
Photomultiplier tube
Amplifier
Pulse height discriminator
Counter

FIGURE 70.5 Gamma counted. The intensity of the light flash produced when a gamma photon interacts with a scintillator is proportional to the energy of the photon. The photomultiplier tube converts these light flashes into electric pulses which can be selected according to size (gamma energy) and counted.

The activity of a quantity of radioactive isotope is defined as the number of disintegrations per second which occur. The usual units are the curie (Ci), which is defined as 3.7×10^{10} dps, and the becquerel (Bq), defined as 1 dps. Specific activity for a given isotope is defined as activity per unit mass of the isotope.

The rate of decay for a given isotope is characterized by the decay constant λ, which is the proportion of the isotope which decays in unit time. Thus, the rate of loss of radioactive isotope is governed by the equation

$$\frac{\mathrm{d}N}{\mathrm{d}t} = -\lambda N \tag{70.3}$$

where N is the amount of radioactive isotope present at time t. The solution to this differential equation is:

$$N = N_0 \mathrm{e}^{-\lambda t} \tag{70.4}$$

It can easily be shown that the amount of radioactive isotope present will be reduced by half after time

$$t_{1/2} = \frac{0.693}{\lambda} \tag{70.5}$$

This is known as the half-life for the isotope and can vary widely; for example, carbon-14 has a half-life of 5760 years, and iodine-131 has a half-life of 8.1 days.

The most common method for detection of radiation in the clinical laboratory is by **scintillation**. This is the conversion of radiation energy into photons in the visible or near-UV range. These are detected with photomultiplier tubes.

For gamma radiation, the scintillating crystal is made of sodium iodide doped with about 1% thallium, producing 20 to 30 photons for each electron-volt of energy absorbed. The photomultiplier tube and amplifier circuit produce voltage pulses proportional to the energy of the absorbed radiation. These voltage pulses are usually passed through a pulse-height analyzer which eliminates pulses outside a preset energy range (window). Multichannel analyzers can discriminate between two or more isotopes if they have well-separated energy maxima. There generally will be some spill down of counts from the higher-energy isotope into the lower-energy isotope's window, but this effect can be corrected with a simple algorithm. Multiple well detectors with up to 64 detectors in an array are available which increase the throughput for counting systems greatly. Counters using the sodium iodide crystal scintillator are referred to as gamma counters or well counters.

The lower energy and short penetration ability of beta particles requires a scintillator in direct contact with the decaying isotope. This is accomplished by dissolving or suspending the sample in a liquid fluor. Counters which use this technique are called beta counters or liquid scintillation counters.

Liquid scintillation counters use two photomultiplier tubes with a coincidence circuit that prevents counting of events seen by only one of the tubes. In this way, false counts due to chemiluminescence and noise in the phototube are greatly reduced. Quenching is a problem in all liquid scintillation counters. Quenching is any process which reduces the efficiency of the scintillation counting process, where efficiency is defined as

$$\text{Efficiency} = \text{counts per minute/decays per minute} \tag{70.6}$$

A number of techniques have been developed that automatically correct for quenching effects to produce estimates of true decays per minute from the raw counts. Currently there is a trend away from beta-emitting isotopic labels, but these assays are still used in many laboratories.

70.5 Coagulation Timers

Screening for and diagnosis of coagulation disorders is accomplished by assays that determine how long it takes for blood to clot following initiation of the clotting cascade by various reagents. A variety of instruments have been designed to automate this procedure. In addition to increasing the speed and throughput of such testing, these instruments improve the reproducibility of such tests. All the instruments provide precise introduction of reagents, accurate timing circuits, and temperature control. They differ in the method for detecting clot formation. One of the older methods still in use is to dip a small metal hook into the blood sample repeatedly and lift it a few millimeters above the surface. The electric resistance between the hook and the sample is measured, and when fibrin filaments form, they produce a conductive pathway which is detected as clot formation. Other systems detect the increase in viscosity due to fibrin formation or the scattering due to the large polymerized molecules formed. Absorption and fluorescence spectroscopy can also be used for clot detection.

70.6 Osmometers

The **colligative properties** of a solution are a function of the number of solute particles present regardless of size or identity. Increased solute concentration causes an increase in osmotic pressure and boiling point and a decrease in vapor pressure and freezing point. Measuring these changes provides information on the total solute concentration regardless of type. The most accurate and popular method used in clinical laboratories is the measurement of freezing point depression. With this method, the sample is supercooled to a few degrees below 0°C while being stirred gently. Freezing is then initiated by vigorous stirring. The heat of fusion quickly brings the solution to a slushy state where an equilibrium exists between ice and liquid, ensuring that the temperature is at the freezing point. This temperature is measured. A solute concentration of 1 osmol/kg water produces a freezing point depression of 1.858°C. The measured temperature depression is easily calibrated in units of milliosmols/kg water.

The vapor pressure depression method has the advantage of smaller sample size. However, it is not as precise as the freezing point method and cannot measure the contribution of volatile solutes such as ethanol. This method is not used as widely as the freezing point depression method in clinical laboratories.

Osmolality of blood is primarily due to electrolytes such as Na^+ and Cl^-. Proteins with molecular weights of 30,000 or more atomic mass units (amu) contribute very little to total osmolality due to their smaller numbers (a single Na^+ ion contributes just as much to osmotic pressure as a large protein molecule). However, the contribution to osmolality made by proteins is of great interest when monitoring conditions leading to pulmonary edema. This value is known as colloid osmotic pressure, or oncotic pressure, and is measured with a membrane permeable to water and all molecules smaller than about 30,000 amu. By placing a reference saline solution on one side and the unknown sample on the other, an osmotic pressure is developed across the membrane. This pressure is measured with a pressure transducer and can be related to the true colloid osmotic pressure through a calibration procedure using known standards.

70.7 Automation

Improvements in technology coupled with increased demand for laboratory tests as well as pressures to reduce costs have led to the rapid development of highly automated laboratory instruments. Typical automated instruments contain mechanisms for measuring, mixing, and transport of samples and reagents, measurement systems, and one or more microprocessors to control the entire system. In addition to system control, the computer systems store calibration curves, match test results to specimen IDs, and generate reports. Automated instruments are dedicated to complete blood counts, coagulation studies, microbiology assays, and immunochemistry, as well as high-volume instruments used in clinical chemistry laboratories. The chemistry analyzers tend to fall into one of four classes: continuous flow, centrifugal, pack-based, and dry-slide-based systems. The continuous flow systems pass successive samples and reagents through a single set of tubing, where they are directed to appropriate mixing, dialyzing, and measuring stations. Carry-over from one sample to the next is minimized by the introduction of air bubbles and wash solution between samples.

Centrifugal analyzers use plastic rotors which serve as reservoirs for samples and reagents and also as cuvettes for optical measurements. Spinning the plastic rotor mixes, incubates, and transports the test solution into the cuvette portion of the rotor, where the optical measurements are made while the rotor is spinning.

Pack-based systems are those in which each test uses a special pack with the proper reagents and sample preservation devices built-in. The sample is automatically introduced into as many packs as tests required. The packs are then processed sequentially.

Dry chemistry analyzers use no liquid reagents. The reagents and other sample preparation methods are layered onto a slide. The liquid sample is placed on the slide, and after a period of time the color developed is read by reflectance photometry. Ion-selective electrodes have been incorporated into the same slide format.

There are a number of technological innovations found in many of the automated instruments. One innovation is the use of fiberoptic bundles to channel excitation energy toward the sample as well as transmitted, reflected, or emitted light away from the sample to the detectors. This provides a great deal of flexibility in instrument layout. Multiwavelength analysis using a spinning filter wheel or diode array detectors is commonly found. The computers associated with these instruments allow for innovative improvements in the assays. For instance, when many analytes are being analyzed from one sample, the interference effects of one analyte on the measurement of another can be predicted and corrected before the final report is printed.

70.8 Trends in Laboratory Instrumentation

Predicting the future direction of laboratory instrumentation is difficult, but there seem to be some clear trends. Decentralization of the laboratory functions will continue with more instruments being located in or around ICUs, operating rooms, emergency rooms, and physician offices. More electrochemistry-based tests will be developed. The flame photometer is already being replaced with ion-selective electrode methods. Instruments which analyze whole blood rather than **plasma** or **serum** will reduce the amount of time required for sample preparation and will further encourage testing away from the central laboratory. Dry reagent methods increasingly will replace wet chemistry methods. Radioimmunoassays will continue to decline with the increasing use of methods for performing immunoassays that do not rely upon radioisotopes such as enzyme-linked fluorescent assays.

Defining Terms

Alpha radiation: Particulate radiation consisting of a helium nucleus emitted from a decaying anucleus.

Amperometry: Measurements based on current flow produced in an electrochemical cell by an applied voltage.

Beta radiation: Particulate radiation consisting of an electron or positron emitted from a decaying nucleus.

Colligative properties: Physical properties that depend on the number of molecules present rather than on their individual properties.

Gamma radiation: Electromagnetic radiation emitted from an atom undergoing nuclear decay.

Hydrodynamic focusing: A process in which a fluid stream is first surrounded by a second fluid and then narrowed to a thin stream by a narrowing of the channel.

Isotopes: Atoms with the same number of protons but differing numbers of neutrons.

Plasma: The liquid portion of blood.

Potentiometry: Measurement of the potential produced by electrochemical cells under equilibrium conditions with no current flow.

Scintillation: The conversion of the kinetic energy of a charged particle or photon to a flash of light.

Serum: The liquid portion of blood remaining after clotting has occurred.

References

Burtis C.A. and Ashwood E.R. (Eds.) 1994. *Tietz Textbook of Clinical Chemistry,* 2nd ed., Philadelphia, Saunders Company.

Hicks M.R., Haven M.C., Schenken J.R. et al. (Eds.) 1987. *Laboratory Instrumentation*, 3rd ed., Philadelphia, Lippincott Company, 1987.

Kaplan L.A. and Pesce A.J. (Eds.) 1989. *Clinical Chemistry: Theory, Analysis, and Correlation*, 2nd ed., St. Louis, Mosby.

Tietz N.W. (Ed.). 1987. *Fundamentals of Clinical Chemistry,* 3rd ed., Philadelphia, W.B. Saunders.

Ward J.M., Lehmann C.A., and Leiken A.M. 1994. *Clinical Laboratory Instrumentation and Automation: Principles, Applications, and Selection*, Philadelphia, W.B. Saunders.

71

Noninvasive Optical Monitoring

Ross Flewelling
Nellcor Incorporation

Optical measures of physiologic status are attractive because they can provide a simple, noninvasive, yet real-time assessment of medical condition. Noninvasive optical monitoring is taken here to mean the use of visible or near-infrared light to directly assess the internal physiologic status of a person without the need of extracting a blood of tissue sample or using a catheter. Liquid water strongly absorbs ultraviolet and infrared radiation, and thus these spectral regions are useful only for analyzing thin surface layers or respiratory gases, neither of which will be the subject of this review. Instead, it is the visible and near-infrared portions of the electromagnetic spectrum that provide a unique "optical window" into the human body, opening new vistas for noninvasive monitoring technologies.

Various molecules in the human body possess distinctive spectral absorption characteristics in the visible or near-infrared spectral regions and therefore make optical monitoring possible. The most strongly absorbing molecules at physiologic concentrations are the hemoglobins, myoglobins, **cytochromes**, melanins, carotenes, and bilirubin (see Figure 71.1 for some examples). Perhaps less appreciated are the less distinctive and weakly absorbing yet ubiquitous materials possessing spectral characteristics in the near-infrared: water, fat, proteins, and sugars. Simple optical methods are now available to quantitatively and noninvasively measure some of these compounds directly in intact tissue. The most successful methods to date have used hemoglobins to assess the oxygen content of blood, cytochromes to assess the respiratory status of cells, and possibly near-infrared to assess endogenous concentrations of metabolytes, including glucose.

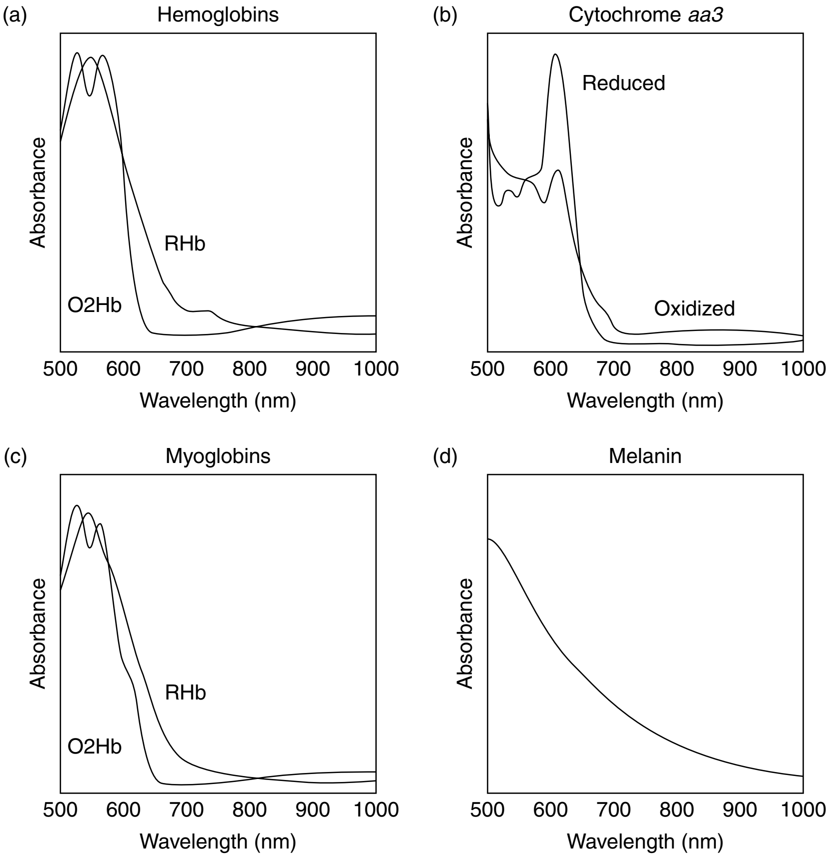


FIGURE 71.1 Absorption spectra of some endogenous biologic materials. (a) hemoglobins, (b) cytochrome *aa3*, (c) myoglobins, and (d) melanin.

71.1 Oximetry and Pulse Oximetry

Failure to provide adequate oxygen to tissues — **hypoxia** — can in a matter of minutes result in reduced work capacity of muscles, depressed mental activity, and ultimately cell death. It is therefore of considerable interest to reliably and accurately determine the amount of oxygen in blood or tissues. **Oximetry** is the determination of the oxygen content of blood of tissues, normally by optical means. In the clinical laboratory the oxygen content of whole blood can be determined by a bench-top cooximeter or blood gas analyzer. But the need for timely clinical information and the desire to minimize the inconvenience and cost of extracting a blood sample and later analyze it in the lab has led to the search for alternative noninvasive optical methods. Since the 1930s, attempts have been made to use multiple wavelengths of light to arrive at a complete spectral characterization of a tissue. These approaches, although somewhat successful, have remained of limited utility owing to the awkward instrumentation and unreliable results.

It was not until the invention of **pulse oximetry** in the 1970s and its commercial development and application in the 1980s that noninvasive oximetry became practical. Pulse oximetry is an extremely easy-to-use, noninvasive, and accurate measurement of real-time arterial oxygen saturation. Pulse oximetry is now used routinely in clinical practice, has become a standard of care in all U.S. operating rooms, and is increasingly used wherever critical patients are found. The explosive growth of this new technology and its considerable utility led John Severinghaus and Poul Astrup [1986] in an excellent historical review to conclude that pulse oximetry was "arguably the most significant technological advance ever made in monitoring the well-being and safety of patients during anesthesia, recovery and critical care."

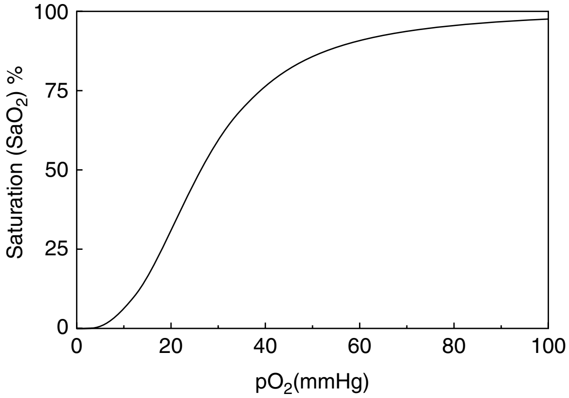


FIGURE 71.2 Hemoglobin oxygen dissociation curve showing the sigmoidal relationship between the partial pressure of oxygen and the oxygen saturation of blood. The curve is given approximately by $\%SaO_2 = 100\%/[1+P_{50}/pO_2^n]$, with $n = 2.8$ and $P_{50} = 26$ mmHg.

71.1.1 Background

The partial pressure of oxygen (pO_2) in tissues need only be about 3 mmHg to support basic metabolic demands. This tissue level, however, requires capillary pO_2 to be near 40 mmHg, with a corresponding arterial pO_2 of about 95 mmHg. Most of the oxygen carried by blood is stored in red blood cells reversibly bound to hemoglobin molecules. Oxygen saturation (SaO_2) is defined as the percentage of hemoglobin-bound oxygen compared to the total amount of hemoglobin available for reversible oxygen binding. The relationship between the oxygen partial pressure in blood and the oxygen saturation of blood is given by the hemoglobin oxygen dissociation curve as shown in Figure 71.2. The higher the pO_2 in blood, the higher the SaO_2. But due to the highly cooperative binding of four oxygen molecules to each hemoglobin molecule, the oxygen binding curve is sigmoidal, and consequently the SaO_2 value is particularly sensitive to dangerously low pO_2 levels. With a normal arterial blood pO_2 above 90 mmHg, the oxygen saturation should be at least 95%, and a pulse oximeter can readily verify a safe oxygen level. If oxygen content falls, say to a pO_2 below 40 mmHg, metabolic needs may not be met, and the corresponding oxygen saturation will drop below 80%. Pulse oximetry therefore provides a direct measure of oxygen sufficiency and will alert the clinician to any danger of imminent hypoxia in a patient.

Although endogenous molecular oxygen is not optically observable, hemoglobin serves as an oxygen-sensitive "dye" such that when oxygen reversibly binds to the iron atom in the large heme prosthetic group, the electron distribution of the heme is shifted, producing a significant color change. The optical absorption of hemoglobin in its oxygenated and deoxygenated states is shown in Figure 71.1. Fully oxygenated blood absorbs strongly in the blue and appears bright red; deoxygenated blood absorbs through the visible region and is very dark (appearing blue when observed through tissue due to light scattering effects). Thus the optical absorption spectra of oxyhemaglobin (O_2Hb) and "reduced" deoxyhemoglobin (RHb) differ substantially, and this difference provides the basis for spectroscopic determinations of the proportion of the two hemoglobin states. In addition to these two normal functional hemoglobins, there are also **dysfunctional hemoglobins** — carboxyhemoglobin, methemoglobin, and sulhemoglobin — which are spectroscopically distinct but do not bind oxygen reversibly. Oxygen saturation is therefore defined in Equation 71.1 only in terms of the **functional saturation** with respect to O_2Hb and RHb:

$$S_aO_2 = \frac{\mathrm{O_2Hb}}{\mathrm{RHb} + O_2\mathrm{Hb}} \times 100\% \qquad (71.1)$$

Cooximeters are bench-top analyzers that accept whole blood samples and utilize four or more wavelengths of monochromatic light, typically between 500 and 650 nm, to spectroscopically determine the various individual hemoglobins in the sample. If a blood sample can be provided, this spectroscopic method is accurate and reliable. Attempts to make an equivalent quantitative analysis noninvasively through intact tissue have been fraught with difficulty. The problem has been to contend with the wide variation in scattering and nonspecific absorption properties of very complex heterogeneous tissue. One of the more successful approaches, marketed by Hewlett–Packard, used eight optical wavelengths transmitted through the pinna of the ear. In this approach a "bloodless" measurement is first obtained by squeezing as much blood as possible from an area of tissue; the arterial blood is then allowed to flow back, and the oxygen saturation is determined by analyzing the change in the spectral absorbance characteristics of the tissue. While this method works fairly well, it is cumbersome, operator dependent, and does not always work well on poorly perfused or highly pigmented subjects.

In the early 1970s, Takuo Aoyagi recognized that most of the interfering nonspecific tissue effects could be eliminated by utilizing only the change in the signal during an arterial pulse. Although an early prototype was built in Japan, it was not until the refinements in implementation and application by Biox (now Ohmeda) and Nellcor Incorporated in the 1980s that the technology became widely adopted as a safety monitor for critical care use.

71.1.2 Theory

Pulse oximetry is based on the fractional change in light transmission during an arterial pulse at two different wavelengths. In this method the fractional change in the signal is due only to the arterial blood itself, and therefore the complicated nonpulsatile and highly variable optical characteristics of tissue are eliminated. In a typical configuration, light at two different wavelengths illuminating one side of a finger will be detected on the other side, after having traversed the intervening vascular tissues (Figure 71.3). The transmission of light at each wavelength is a function of the thickness, color, and structure of the skin, tissue, bone, blood, and other material through which the light passes. The absorbance of light by a sample is defined as the negative logarithm of the ratio of the light intensity in the presence of the sample (I) to that without (I_0): $A = -\log(I/I_0)$. According to the **Beer–Lambert law,** the absorbance of a sample at a given wavelength with a molar absorptivity (ϵ) is directly proportional to both the concentration (c) and pathlength (l) of the absorbing material: $A = \epsilon cl$. (In actuality, biologic tissue is highly scattering, and the Beer–Lambert law is only approximately correct; see the references for further elaboration). Visible or near-infrared light passing through about one centimeter of tissue (e.g., a finger) will be attenuated by about one or two orders of magnitude for a typical emitter–detector geometry, corresponding to an effective optical density (OD) of 1 to 2 OD (the detected light intensity is decreased

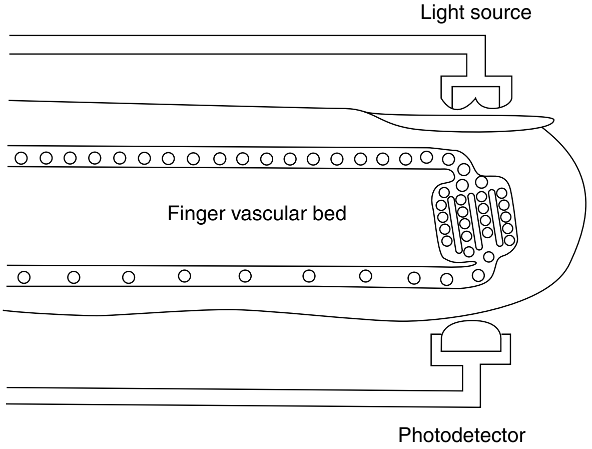


FIGURE 71.3 Typical pulse oximeter sensing configuration on a finger. Light at two different wavelengths is emitted by the source, diffusely scattered through the finger, and detected on the opposite side by a photodetector.

by one order of magnitude for each OD unit). Although hemoglobin in the blood is the single strongest absorbing molecule, most of the total attenuation is due to the scattering of light away from the detector by the highly heterogeneous tissue. Since human tissue contains about 7% blood, and since blood contains typically about 14 g/dL hemoglobin, the effective hemoglobin concentration in tissue is about 1 g/dL ($\sim$150 μM). At the wavelengths used for pulse oximetry (650–950 nm), the oxy- and deoxyhemoglobin molar absorptivities fall in the range of 100–1000 $M^{-1}\text{cm}^{-1}$, and consequently hemoglobin accounts for less than 0.2 OD of the total observed optical density. Of this amount, perhaps only 10% is pulsatile, and consequently pulse signals of only a few percent are ultimately measured, at times even one-tenth of this.

A mathematical model for pulse oximetry begins by considering light at two wavelengths, λ_1 and λ_2, passing through tissue and being detected at a distant location as in Figure 71.3. At each wavelength the total light attenuation is described by four different component absorbances: oxyhemoglobin in the blood (concentration c_o, molar absorptivity ϵ_o, and effective pathlength l_o), "reduced" deoxyhemoglobin in the blood (concentration c_r,molar absorptivity ϵ_r, and effective pathlength l_r), specific variable absorbances that are not from the arterial blood (concentration c_x, molar absorptivity ϵ_x, and effective pathlength l_x), and all other non-specific sources of optical attenuation, combined as A_y, which can include light scattering, geometric factors, and characteristics of the emitter and detector elements. The total absorbance at the two wavelengths can then be written:

$$\begin{cases} A_{\lambda_1} = \epsilon_{o_1} c_o l_o + \epsilon_{r_1} c_r l_r + \epsilon_{x_1} c_x l_x + A_{y_1} \\ A_{\lambda_2} = \epsilon_{o_2} c_o l_o + \epsilon_{r_2} c_r l_r + \epsilon_{x_2} c_x l_x + A_{y_2} \end{cases} \tag{71.2}$$

The blood volume change due to the arterial pulse results in a modulation of the measured absorbances. By taking the time rate of change of the absorbances, the two last terms in each equation are effectively zero, since the concentration and effective pathlength of absorbing material outside the arterial blood do not change during a pulse $[\mathrm{d}(c_x l_x)/\mathrm{d}t = 0]$, and all the nonspecific effects on light attenuation are also effectively invariant on the time scale of a cardiac cycle ($\mathrm{d}A_y/\mathrm{d}t = 0$). Since the extinction coefficients are constant, and the blood concentrations are constant on the time scale of a pulse, the time-dependent changes in the absorbances at the two wavelengths can be assigned entirely to the change in the blood pathlength ($\mathrm{d}l_o/\mathrm{d}t$ and $\mathrm{d}l_r/\mathrm{d}t$). With the additional assumption that these two blood pathlength changes are equivalent (or more generally, their ratio is a constant), the ratio R of the time rate of change of the absorbance at wavelength 1 to that at wavelength 2 reduces to the following:

$$R = \frac{\mathrm{d}A_{\lambda_1}/\mathrm{d}t}{\mathrm{d}A_{\lambda_2}/\mathrm{d}t} = \frac{-\mathrm{d}\log(I_1/I_o)/\mathrm{d}t}{-\mathrm{d}\log(I_2/I_o)/\mathrm{d}t} = \frac{(\Delta I_1/I_1)}{(\Delta I_2/I_2)} = \frac{\epsilon_{o_1} c_o + \epsilon_{r_1} c_r}{\epsilon_{o_2} c_o + \epsilon_{r_2} c_r} \tag{71.3}$$

Observing that functional oxygen saturation is given by $S = c_o/(c_o + c_r)$, and that $(1 - S) = c_r/(c_o + c_r)$, the oxygen saturation can then be written in terms of the ratio R as follows

$$S = \frac{\epsilon_{r1} - \epsilon_{r2} R}{(\epsilon_{r1} - \epsilon_{o1}) - (\epsilon_{r2} - \epsilon_{o2})R} \tag{71.4}$$

Equation 71.4 provides the desired relationship between the experimentally determined ratio R and the clinically desired oxygen saturation S. In actual use, commonly available LEDs are used as the light sources, typically a red LED near 660 nm and a near-infrared LED selected in the range 890 to 950 nm. Such LEDs are not monochromatic light sources, typically with bandwidths between 20 and 50 nm, and therefore standard molar absorptivities for hemoglobin cannot be used directly in Equation 71.4. Further, the simple model presented above is only approximately true; for example, the two wavelengths do not necessarily have the exact same pathlength changes, and second-order scattering effects have been ignored. Consequently the relationship between S and R is instead determined empirically by fitting the clinical data to a generalized function of the form $S = (a - bR)/(c - dR)$. The final empirical calibration will ultimately depend on the details of an individual sensor design, but these variations can be determined

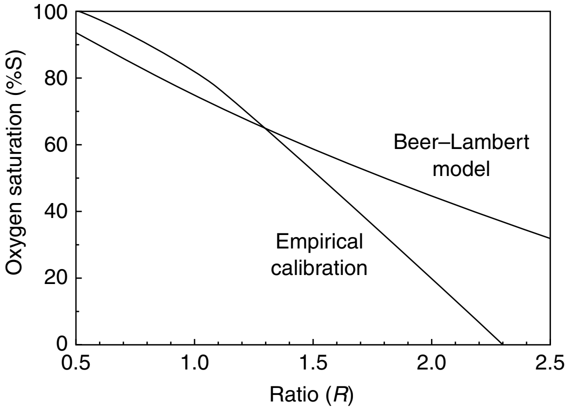


FIGURE 71.4 Relationship between the measured ratio of fractional changes in light intensity at two wavelengths, R, and the oxygen saturation S. Beer–Lambert model is from Equation 71.4 with $\epsilon_{o1} = 100$, $\epsilon_{o2} = 300$, $\epsilon_{r1} = 800$, and $\epsilon_{r2} = 200$. Empirical calibration is based on %S $= 100\% \times (a - bR)/(c - dR)$ with $a = 1000$, $b = 550$, $c = 900$, and $d = 350$, with a linear extrapolation below 70%.

for each sensor and included in unique calibration parameters. A typical empirical calibration for R vs. S is shown in Figure 71.4, together with the curve that standard molar absorptivities would predict.

In this way the measurement of the ratio of the fractional change in signal intensity of the two LEDs is used along with the empirically determined calibration equation to obtain a beat-by-beat measurement of the arterial oxygen saturation in a perfused tissue — continuously, noninvasively, and to an accuracy of a few percent.

71.1.3 Application and Future Directions

Pulse oximetry is now routinely used in nearly all operating rooms and critical care areas in the United States and increasingly throughout the world. It has become so pervasive and useful that it is now being called the “fifth” vital sign (for an excellent review of practical aspects and clinical applications of the technology see Kelleher [1989]).

The principal advantages of pulse oximetry are that it provides continuous, accurate, and reliable monitoring of arterial oxygen saturation on nearly all patients, utilizing a variety of convenient sensors, reusable as well as disposable. Single-patient-use adhesive sensors can easily be applied to fingers for adults and children and to arms for legs or neonates. Surface reflectance sensors have also been developed based on the same principles and offer a wider choice for sensor location, though they tend to be less accurate and prone to more types of interference.

Limitations of pulse oximetry include sensitivity to high levels of optical or electric interference, errors due to high concentrations of dysfunctional hemoglobins (methemoglobin or carboxyhemoglobin) or interference from physiologic dyes (such as methylene blue). Other important factors, such as total hemoglobin content, fetal hemoglobin, or sickle cell trait, have little or no effect on the measurement except under extreme conditions. Performance can also be compromised by poor signal quality, as may occur for poorly perfused tissues with weak pulse amplitudes or by motion artifact.

Hardware and software advances continue to provide more sensitive signal detection and filtering capabilities, allowing pulse oximeters to work better on more ambulatory patients. Already some pulse oximeters incorporate ECG synchronization for improved signal processing. A pulse oximeter for use in labor and delivery is currently under active development by several research groups and companies. A likely implementation may include use of a reflectance surface sensor for the fetal head to monitor the adequacy of fetal oxygenation. This application is still in active development, and clinical utility remains to be demonstrated.

71.2 Nonpulsatile Spectroscopy

71.2.1 Background

Nonpulsatile optical spectroscopy has been used for more than half a century for noninvasive medical assessment, such as in the use of multiwavelength tissue analysis for oximetry and skin reflectance measurement for bilirubin assessment in jaundiced neonates. These early applications have found some limited use, but with modest impact. Recent investigations into new nonpulsatile spectroscopy methods for assessment of deep-tissue oxygenation (e.g., cerebral oxygen monitoring), for evaluation of respiratory status at the cellular level, and for the detection of other critical analytes, such as glucose, may yet prove more fruitful. The former applications have led to spectroscopic studies of cytochromes in tissues, and the latter has led to considerable work into new approaches in near-infrared analysis of intact tissues.

71.2.2 Cytochrome Spectroscopy

Cytochromes are electron-transporting, heme-containing proteins found in the inner membranes of mitochondria and are required in the process of oxidative phosphorylation to convert metabolytes and oxygen into CO_2 and high-energy phosphates. In this metabolic process the cytochromes are reversibly oxidized and reduced, and consequently the oxidation–reduction states of cytochromes *c* and aa_3 in particular are direct measures of the respiratory condition of the cell. Changes in the absorption spectra of these molecules, particularly near 600 and 830 nm for cytochrome aa_3, accompany this shift. By monitoring these spectral changes, the cytochrome oxidation state in the tissues can be determined (see, e.g., Jöbsis [1977] and Jöbsis et al. [1977]). As with all nonpulsatile approaches, the difficulty is to remove the dependence of the measurements on the various nonspecific absorbing materials and highly variable scattering effects of the tissue. To date, instruments designed to measure cytochrome spectral changes can successfully track relative changes in brain oxygenation, but absolute quantitation has not yet been demonstrated.

71.2.3 Near-Infrared Spectroscopy and Glucose Monitoring

Near-infrared (NIR), the spectral region between 780 and 3000 nm, is characterized by broad and overlapping spectral peaks produced by the overtones and combinations of infrared vibrational modes. Figure 71.5 shows typical NIR absorption spectra of fat, water, and starch. Exploitation of this spectral region for *in vivo*

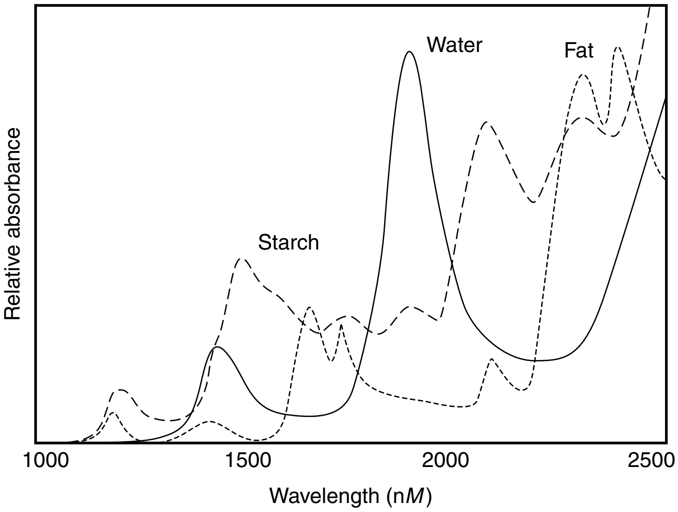


FIGURE 71.5 Typical near-infrared absorption spectra of several biologic materials.

analysis has been hindered by the same complexities of nonpulsatile tissue spectroscopy described above and is further confounded by the very broad and indistinct spectral features characteristic of the NIR. Despite these difficulties, NIR spectroscopy has garnered considerable attention, since it may enable the analysis of common analytes.

Karl Norris and coworkers pioneered the practical application of NIR spectroscopy, using it to evaluate water, fat, and sugar content of agricultural products (see Osborne et al. [1993] and Burns and Cuirczak [1992]). The further development of sophisticated **multivariate analysis** techniques, together with new scattering models (e.g., Kubelka–Munk theory) and high-performance instrumentation, further extended the application of NIR methods. Over the past decade, many research groups and companies have touted the use of NIR techniques for medical monitoring, such as for determining the relative fat, protein, and water content of tissue, and more recently for noninvasive glucose measurement. The body composition analyses are useful but crude and are mainly limited to applications in nutrition and sports medicine. Noninvasive glucose monitoring, however, is of considerable interest.

More than 2 million diabetics in the United States lance their fingers three to six times a day to obtain a drop of blood for chemical glucose determination. The ability of these individuals to control their glucose levels, and the quality of their life generally, would dramatically improve if a simple, noninvasive method for determining blood glucose levels could be developed. Among the noninvasive optical methods proposed for this purpose are optical rotation, NIR analysis, and Raman spectroscopy. The first two have received the most attention. Optical rotation methods aim to exploit the small optical rotation of polarized light by glucose. To measure physiologic glucose levels in a 1-cm thick sample to an accuracy of 25 mg/dL would require instrumentation that can reliably detect an optical rotation of at least 1 millidegree. Finding an appropriate *in vivo* optical path for such measurements has proved most difficult, with most approaches looking to use either the aqueous humor or the anterior chamber of the eye [Coté et al., 1992; Rabinovitch et al., 1982]. Although several groups have developed laboratory analyzers that can measure such a small effect, so far *in vivo* measurement has not been demonstrated, due both to unwanted scattering and optical activity of biomaterials in the optical path and to the inherent difficulty in developing a practical instrument with the required sensitivity.

NIR methods for noninvasive glucose determination are particularly attractive, although the task is formidable. Glucose has spectral characteristics near 1500 nm and in the 2000 to 2500 nm band where many other compounds also absorb, and the magnitude of the glucose absorbance in biologic samples is typically two orders of magnitude lower than those of water, fat, or protein. The normal detection limit for NIR spectroscopy is on the order of one part in 10^3, whereas a change of 25 mg/dL in glucose concentration corresponds to an absorbance change of 10^{-4} to 10^{-5}. In fact, the temperature dependence of the NIR absorption of water alone is at least an order of magnitude greater than the signal from glucose in solution. Indeed, some have suggested that the apparent glucose signature in complex NIR spectra may actually be the secondary effect of glucose on the water.

Sophisticated chemometric (particularly multivariate analysis) methods have been employed to try to extract the glucose signal out of the noise (for methods reviews see Martens and Næs [1989] and Haaland [1992]). Several groups have reported using multivariate techniques to quantitate glucose in whole blood samples, with encouraging results [Haaland et al., 1992]. And despite all theoretical disputations to the contrary, some groups claim the successful application of these multivariate analysis methods to noninvasive *in vivo* glucose determination in patients [Robinson et al., 1992]. Yet even with the many groups working in this area, much of the work remains unpublished, and few if any of the reports have been independently validated.

71.2.4 Time-Resolved Spectroscopy

The fundamental problem in making quantitative optical measurements through intact tissue is dealing with the complex scattering phenomena. This scattering makes it difficult to determine the effective pathlength for the light, and therefore attempts to use the Beer–Lambert law, or even to determine a consistent empirical calibration, continue to be thwarted. Application of new techniques in time-resolved spectroscopy may be able to tackle this problem. Thinking of light as a packet of photons, if a single packet from a light source is sent through tissue, then a distant receiver will detected a photon distribution over time — the photons least scattered arriving first and the photons most scattered arriving later. In principle, the first photons arriving at the detector passed directly through the tissue. For these first photons the distance between the emitter and the detector is fixed and known, and the Beer–Lambert law should apply, permitting determination of an *absolute* concentration for an absorbing component. The difficulty in this is, first, that the measurement time scale must be on the order of the photon transit time (subnanosec), and second, that the number of photons getting through without scattering will be extremely small, and therefore the detector must be exquisitely sensitive. Although these considerable technical problems have been overcome in the laboratory, their implementation in a practical instrument applied to a real subject remains to be demonstrated. This same approach is also being investigated for noninvasive optical imaging, since the unscattered photons should produce sharp images (see Chance et al. [1988], Chance [1991], and Yoo and Alfano [1989]).

71.3 Conclusions

The remarkable success of pulse oximetry has established noninvasive optical monitoring of vital physiologic functions as a modality of considerable value. Hardware and algorithm advances in pulse oximetry are beginning to broaden its use outside the traditional operating room and critical care areas. Other promising applications of noninvasive optical monitoring are emerging, such as for measuring deep tissue oxygen levels, determining cellular metabolic status, or for quantitative determination of other important physiologic parameters such as blood glucose. Although these latter applications are not yet practical, they may ultimately impact noninvasive clinical monitoring just as dramatically as pulse oximetry.

Defining Terms

Beer–Lambert law: Principle stating that the optical absorbance of a substance is proportional to both the concentration of the substance and the pathlength of the sample.

Cytochromes: Heme-containing proteins found in the membranes of mitochondria and required for oxidative phosphorylation, with characteristic optical absorbance spectra.

Dysfunctional hemoglobins: Those hemoglobin species that cannot reversibly bind oxygen (carboxyhemoglobin, methemoglobin, and sulfhemoglobin).

Functional saturation: The ratio of oxygenated hemoglobin to total nondysfunctional hemoglobins (oxyhemoglobin plus deoxyhemoglobin).

Hypoxia: Inadequate oxygen supply to tissues necessary to maintain metabolic activity.

Multivariate analysis: Empirical models developed to relate multiple spectral intensities from many calibration samples to known analyte concentrations, resulting in an optimal set of calibration parameters.

Oximetry: The determination of blood or tissue oxygen content, generally by optical means.

Pulse oximetry: The determination of functional oxygen saturation of pulsatile arterial blood by ratiometric measurement of tissue optical absorbance changes.

References

Burns, D.A. and Ciurczak, E.W. (Eds.). (1992). *Handbook of Near-Infrared Analysis.* New York, Marcel Dekker.

Chance, B. (1991). Optical method. *Annu. Rev. Biophys. Biophys. Chem.* 20: 1.

Chance, B., Leigh, J.S., Miyake, H. et al. (1988). Comparison of time-resolved and -unresolved measurements of deoxyhemoglobin in brain. *Proc. Natl Acad. Sci. USA* 85: 4971.

Coté G.L., Fox M.D., and Northrop, R.B. (1992). Noninvasive optical polarimetric glucose sensing using a true phase measurement technique. *IEEE Trans. Biomed. Eng.* 39: 752.

Haaland, D.M. (1992). Multivariate calibration methods applied to the quantitative analysis of infrared spectra. In P.C. Jurs (Ed.), *Computer-Enhanced Analytical Spectroscopy*, Vol. 3, pp. 1–30. New York, Plenum Press.

Haaland, D.M., Robinson, M.R., Koepp, G.W., et al. (1992). Reagentless near-infrared determination of glucose in whole blood using multivariate calibration. *Appl. Spectros.* 46: 1575.

Jöbsis, F.F. (1977). Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. *Science* 198: 1264.

Jöbsis, F.F., Keizer, L.H., LaManna, J.C. et al. (1977). Reflectance spectrophotometry of cytochrome *aa*$_3$ *in vivo. J. Appl. Physiol.* 43: 858.

Kelleher, J.F. (1989). Pulse oximetry. *J. Clin. Monit.* 5: 37.

Martens, H. and Næs, T. (1989). *Multivariate Calibration.* New York, John Wiley & Sons.

Osborne, B.G., Fearn, T., and Hindle, P.H. (1993). *Practical NIR Spectroscopy with Applications in Food and Beverage Analysis.* Essex, England, Longman Scientific & Technical.

Payne, J.P. and Severinghaus, J.W. (Eds.). (1986). *Pulse Oximetry.* New York, Springer-Verlag.

Rabinovitch, B., March, W.F., and Adams, R.L. (1982). Noninvasive glucose monitoring of the aqueous humor of the eye: Part I. Measurement of very small optical rotations. *Diabetes Care* 5: 254.

Robinson, M.R., Eaton, R.P., Haaland, D.M. et al. (1992). Noninvasive glucose monitoring in diabetic patients: A preliminary evaluation. *Clin. Chem.* 38: 1618.

Severinghaus, J.W. and Astrup, P.B. (1986). History of blood gas analysis. VI. Oximetry. *J. Clin. Monit.* 2: 135.

Severinghaus, J.W. and Honda, Y. (1987a). History of blood gas analysis. VII. Pulse oximetry. *J. Clin. Monit.* 3: 135.

Severinghaus, J.W. and Honda, Y. (1987b). Pulse oximetry. *Int. Anesthesiol. Clin.* 25: 205.

Severinghaus, J.W. and Kelleher, J.F. (1992). Recent developments in pulse oximetry. *Anesthesiology* 76: 1018.

Tremper, K.K. and Barker, S.J. (1989). Pulse oximetry. *Anesthesiology* 70: 98.

Wukitsch, M.W., Petterson, M.T., Tobler, D.R. et al. (1988). Pulse oximetry: Analysis of theory, technology, and practice. *J. Clin. Monit.* 4: 290.

Yoo, K.M. and Alfano, R.R. (1989). Photon localization in a disordered multilayered system. *Phys. Rev. B* 39: 5806.

Further Information

Two collections of papers on pulse oximetry include a book edited by J.P. Payne and J.W. Severinghaus, *Pulse Oximetry* (New York, Springer-Verlag, 1986), and a journal collection — *International Anesthesiology Clinics* (25, 1987). For technical reviews of pulse oximetry, see J.A. Pologe's, 1987 "Pulse Oximetry" (*Int. Anesthesiol. Clin.* 25: 137), Kevin K. Tremper and Steven J. Barker's, 1989 "Pulse Oximetry" (*Anesthesiology* 70: 98), and Michael W. Wukitsch, Michael T. Patterson, David R. Tobler, and coworkers' 1988 "Pulse Oximetry: Analysis of Theory, Technology, and Practice" (*J. Clin. Monit.* 4: 290).

For a review of practical and clinical applications of pulse oximetry, see the excellent review by Joseph K. Kelleher (1989) and John Severinghaus and Joseph F. Kelleher (1992). John Severinghaus and Yoshiyuki Honda have written several excellent histories of pulse oximetry (1987a, 1987b).

For an overview of applied near-infrared spectroscopy, see Donald A. Burns and Emil W. Ciurczak (1992) and B.G. Osborne, T. Fearn, and P.H. Hindle (1993). For a good overview of multivariate methods, see Harald Martens and Tormod Næs (1989).

72

Medical Instruments and Devices Used in the Home

Bruce R. Bowman
Edward Schuck
EdenTec Corporation

72.1 Scope of the Market for Home Medical Devices

The market for medical devices used in the home and alternative sites has increased dramatically in the last 10 years and has reached an overall estimated size of more than $1.6 billion [FIND/SVP, 1992]. In the past, hospitals have been thought of as the only places to treat sick patients. But with the major emphasis on reducing healthcare costs, increasing numbers of sicker patients move from hospitals to their homes. Treating sicker patients outside the hospital places additional challenges on medical device design and patient use. Equipment designed for hospital use can usually rely on trained clinical personnel to support the devices. Outside the hospital, the patient and/or family members must be able to use the equipment, requiring these devices to have a different set of design and safety features. This chapter will identify some of the major segments using medical devices in the home and discuss important design considerations associated with home use.

Table 72.1 outlines market segments where devices and products are used to treat patients outside the hospital [FIND/SVP, 1992]. The durable medical equipment market is the most established market providing aids for patients to improve access and mobility. These devices are usually not life supporting or sustaining, but in many cases they can make the difference in allowing a patient to be able to function outside a hospital or nursing or skilled facility. Other market segments listed employ generally more sophisticated solutions to clinical problems. These will be discussed by category of use.

The incontinence and ostomy area of products is one of the largest market segments and is growing in direct relationship to our aging society. Whereas sanitary pads and colostomy bags are not very "high-tech," well-designed aids can have a tremendous impact on the comfort and independence of these patients. Other solutions to incontinence are technically more sophisticated, such as use of electric stimulation of the **sphincter** muscles through an implanted device or a miniature stimulator inserted as an anal or vaginal plug to maintain continence [Wall et al., 1993].

Many forms of equipment are included in the Respiratory segment. These devices include those that maintain life support as well as those that monitor patients' respiratory function. These patients, with proper medical support, can function outside the hospital at a significant reduction in cost and increased patient comfort [Pierson, 1994]. One area of this segment, infant apnea monitors, provides parents or caregivers the cardio/respiratory status of an at-risk infant so that intervention (**CPR**, etc.) can be initiated if the baby has a life-threatening event. The infant monitor shown in Figure 72.1 is an example of a patient monitor designed for home use and will be discussed in more detail later in this chapter. Pulse oximetry

TABLE 72.1 Major Market Segments Outside Hospitals

Market segment	Estimated equipment size 1991	Device examples
Durable medical equipment	\$373 M*	Specialty beds, wheelchairs, toilet aids, ambulatory aids
Incontinence and **ostomy** products	\$600 M*	Sanitary pads, electrical stimulators, **colostomy** bags
Respiratory equipment	\$180 M*	Oxygen therapy, portable ventilators, nasal CPAP, monitors, **apnea** monitors
Drug infusion, drug measurement	\$300 M	Infusion pumps, access ports, patient-controlled analgesia (PCA), glucose measurement, implantable pumps
Pain control and functional stimulation	\$140 M	**Transcutaneous electrical nerve stimulation (TENS)**, **functional electrical nerve stimulation (FES)**

* *Source:* FIND/SVP (1992). The Market for Home Care Products, a Market Intelligence Report. New York.

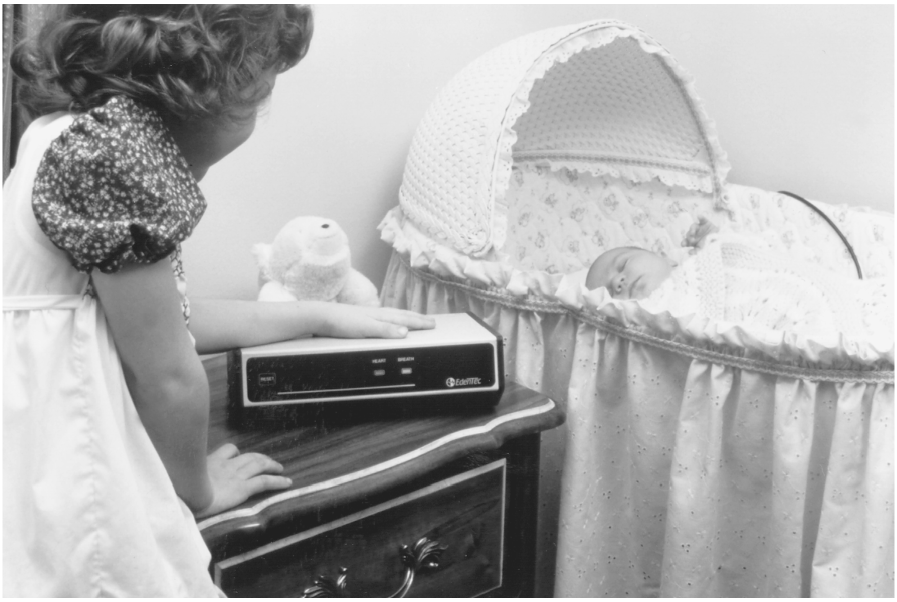


FIGURE 72.1 Infant apnea monitor used in a typical home setting. (Photo courtesy of EdenTec Corporation.)

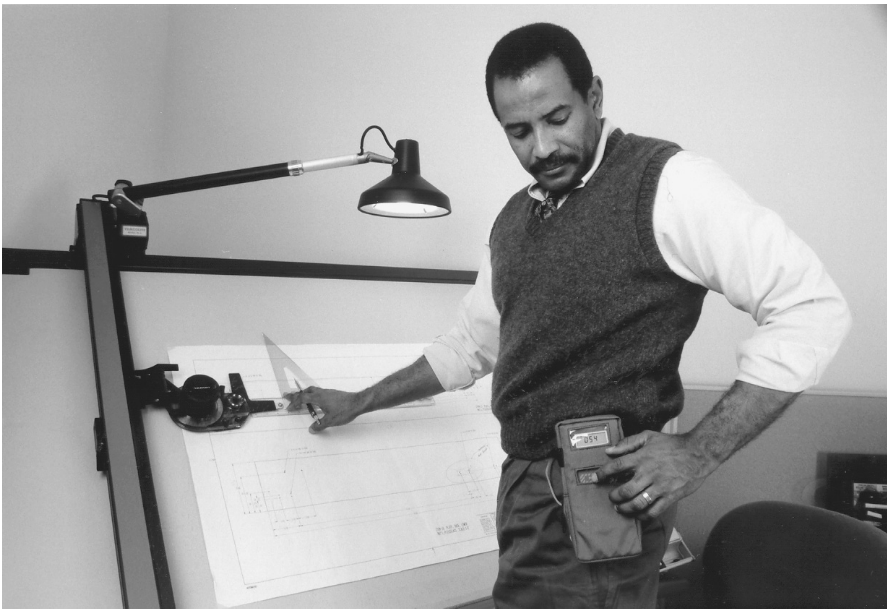


FIGURE 72.2 Portable drug pump used throughout the day. (Photo courtesy of Pharmacia Deltec Inc.)

monitors are also going home with patients. They are used to measure noninvasively the oxygen level of patients receiving supplemental oxygen or ventilator-dependent patients to determine if they are being properly ventilated.

Portable infusion pumps are an integral part of providing antibiotics, pain management, **chemotherapy**, and **parenteral** and **enteral nutrition**. The pump shown in Figure 72.2 is an example of technology that allows the patient to move about freely while receiving sometimes lengthy drug therapy. Implantable drug pumps are also available for special long-term therapy needs.

Pain control using electric stimulation in place of drug therapy continues to be an increasing market. The delivery of small electric impulses to block pain is continuing to gain medical acceptance for treatment outside the hospital setting. A different form of electric stimulation called functional electric stimulation (FES) applies short pulses of electric current to the nerves that control weak or paralyzed muscles. This topic is covered as a separate chapter in this book.

Growth of the homecare business has created problems in overall healthcare costs since a corresponding decrease in hospital utilization has not yet occurred. In the future, however, increased homecare will necessarily result in reassessment and downsizing in the corresponding hospital segment. There will be clear areas of growth and areas of consolidation in the new era of healthcare reform. It would appear, however, that homecare has a bright future of continued growth.

72.2 Unique Challenges to the Design and Implementation of High-Tech Homecare Devices

What are some of the unique requirements of devices that could allow more sophisticated equipment to go home with ordinary people of varied educational levels without compromising their care? Even though each type of clinical problem has different requirements for the equipment that must go home with the patient, certain common qualities must be inherent in most devices used in the home. Three areas to consider when equipment is used outside of the hospital are that the device (1) must provide a positive clinical outcome, (2) must be safe and easy to use, and (3) must be user-friendly enough so that it will be used.

72.2.1 The Device Must Provide a Positive Clinical Outcome

Devices cannot be developed any longer just because new technology becomes available. They must solve the problem for which they were intended and make a significant clinical difference in the outcome or management of the patient while saving money. These realities are being driven by those who reimburse for devices, as well as by the FDA as part of the submission for approval to market a new device.

72.2.2 The Device Must Be Safe to Use

Homecare devices may need to be even *more* reliable and even *safer* than hospital devices. We often think of hospitals as having the best quality and most expensive devices that money can buy. In addition to having the best equipment to monitor patients, hospitals have nurses and aids that keep an eye on patients so that equipment problems may be quickly discovered by the staff. A failure in the home may go unnoticed until it is too late. Thus systems for home use really need extra reliability with automatic backup systems and early warning signals.

Safety issues can take on a different significance depending on the intended use of the device. Certain safety issues are important regardless of whether the device is a critical device such as an implanted **cardiac pacemaker** or a noncritical device such as a bed-wetting alarm. No device should be able to cause harm to the patient regardless of how well or poorly it may be performing its intended clinical duties. Devices must be safe when exposed to all the typical environmental conditions to which the device could be exposed while being operated by the entire range of possible users of varied education and while exposed to siblings and other untrained friends or relatives. For instance, a bed-wetting alarm should not cause skin burns under the sensor if a glass of water spills on the control box. This type of safety issue must be addressed even when it significantly affects the final cost to the consumer.

Other safety issues are not obviously differentiated as to being actual safety issues or simply nuisances or inconveniences to the user. It is very important for the designer to properly define these issues; although some safety features can be included with little or no extra cost, other safety features may be very costly to implement. It may be a nuisance for the patient using a TENS pain control stimulator to have the device inadvertently turned off when its on/off switch is bumped while watching TV. In this case, the patient only experiences a momentary cessation of pain control until the unit is turned back on. But it could mean injuries or death to the same patient driving an automobile who becomes startled when his TENS unit inadvertently turns on and he causes an accident.

Reliability issues can also be mere inconveniences or major safety issues. Medical devices should be free of design and materials defects so that they can perform their intended functions reliably. Once again, reliability does not necessarily need to be expensive and often can be obtained with good design. Critical devices, that is, devices that could cause death or serious injury if they stopped operating properly, may need to have redundant systems for backup, which likely will increase cost.

72.2.3 The Device Must Be Designed So That It *Will* Be Used

A great deal of money is being spent in healthcare on devices for patients that end up not being used. There are numerous reasons for this happening including that the wrong device was prescribed for the patient's problem in the first place; the device works, but it has too many false alarms; the device often fails to operate properly; it is cumbersome to use or difficult to operate or too uncomfortable to wear.

72.2.3.1 Ease of Use

User-friendliness is one of the most important features in encouraging a device to be used. Technological sophistication may be just as necessary in areas that allow ease of use as in attaining accuracy and reliability in the device. The key is that the technologic sophistication be transparent to the user so that the device does not intimidate the user. Transparent features such as automatic calibration or automatic sensitivity adjustment may help allow successful use of a device that would otherwise be too complicated.

Notions of what makes a device easy to use, however, need to be thoroughly tested with the patient population intended for the device. Caution needs to be taken in defining what "simple" means to different people. A VCR may be simple to the designer because all features can be programmed with one button, but it may not be simple to users if they have to remember that it takes two long pushes and one short to get into the clock-setting program.

Convenience for the user is also extremely important in encouraging use of a device. Applications that require devices to be portable must certainly be light enough to be carried. Size is almost always important for anything that must fit within the average household. Either a device must be able to be left in place in the home or it must be easy to set up, clean, and put away. Equipment design can make the difference between the patient appropriately using the equipment or deciding that it is just too much hassle to bother.

72.2.3.2 Reliability

Users must also have confidence in the reliability of the device being used and must have confidence that if it is not working properly, the device will tell them that something is wrong. Frequent breakdowns or false alarms will result in frustration and ultimately in reduced compliance. Eventually patients will stop using the device altogether. Most often, reliability can be designed into a product with little or no extra cost in manufacturing, and everything that can be done at no cost to enhance reliability should be done. It is very important, however, to understand what level of additional reliability involving extra cost is necessary for product acceptance. Reliability can always be added by duplicated backup systems, but the market or application may not warrant such an approach. Critical devices which are implanted, such as cardiac pacemakers, have much greater reliability requirements, since they involve not only patient frustration but also safety.

72.2.3.3 Cost Reimbursement

Devices must be paid for before the patient can realize the opportunity to use new, effective equipment. Devices are usually paid for by one of two means. First, they are covered on an American Medical Association Current Procedural Terminology Code (**CPT-code**) which covers the medical, surgical, and diagnostic services provided by physicians. The CPT-codes are usually priced out by Medicare to establish a baseline reimbursement level. Private carriers usually establish a similar or different level of reimbursement based on regional or other considerations. Gaining new CPT-codes for new devices can take a great deal of time and effort. The second method is to cover the procedure and device under a **capitated fee** where the hospital is reimbursed a lump sum for a procedure including the device, hospital, homecare, and physician fees.

Every effort should be made to design devices to be low cost. Device cost is being scrutinized more and more by those who reimburse. It is easy to state, however, that a device needs to be inexpensive. Unfortunately the reality is that healthcare reforms and new regulations by the FDA are making medical devices more costly to develop, to obtain regulatory approvals for [FDA, 1993], and to manufacture.

72.2.3.4 Professional Medical Service Support

The more technically sophisticated a device is, the more crucial that homecare support and education be a part of a program. In fact, in many cases, such support and education are as important as the device itself.

Medical service can be offered by numerous homecare service companies. Typically these companies purchase the equipment instead of the patient, and a monthly fee is charged for use of the equipment along with all the necessary service. The homecare company then must obtain reimbursement from third-party payers. Some of the services offered by the homecare company include training on how to use the equipment, CPR training, transporting the equipment to the home, servicing/repairing equipment, monthly visits, and providing on-call service 24 h a day. The homecare provider must also be able to provide feedback to the treating physician on progress of the treatment. This feedback may include how well the equipment is working, the patient's medical status, and compliance of the patient.

72.3 Infant Monitor Example

Many infants are being monitored in the home using apnea monitors because they have been identified with breathing problems [Kelly, 1992]. These include newborn premature babies who have **apnea of prematurity** [Henderson-Smart, 1992; NIH, 1987], siblings of babies who have died of **sudden infant death syndrome (SIDS)** [Hunt, 1992; NIH, 1987], or infants who have had an **apparent life-threatening episode (ALTE)** related to lack of adequate respiration [Kahn et al., 1992; NIH, 1987]. Rather than keeping infants in the hospital for a problem that they may soon outgrow (1–6 months), doctors often discharge them from the hospital with an infant apnea monitor that measures the duration of breathing pauses and heart rate and sounds an alarm if either parameter crosses limits prescribed by the doctor.

Infant apnea monitors are among the most sophisticated devices used routinely in the home. These devices utilize microprocessor control, sophisticated breath-direction and artifact rejection firmware algorithms, and internal memory that keeps track of use of the device as well as recording occurrence of events and the physiologic waveforms associated with the events. The memory contents can be downloaded directly to computer or sent via modem remotely where a complete 45-day report can be provided to the referring physician (see Figure 72.3).

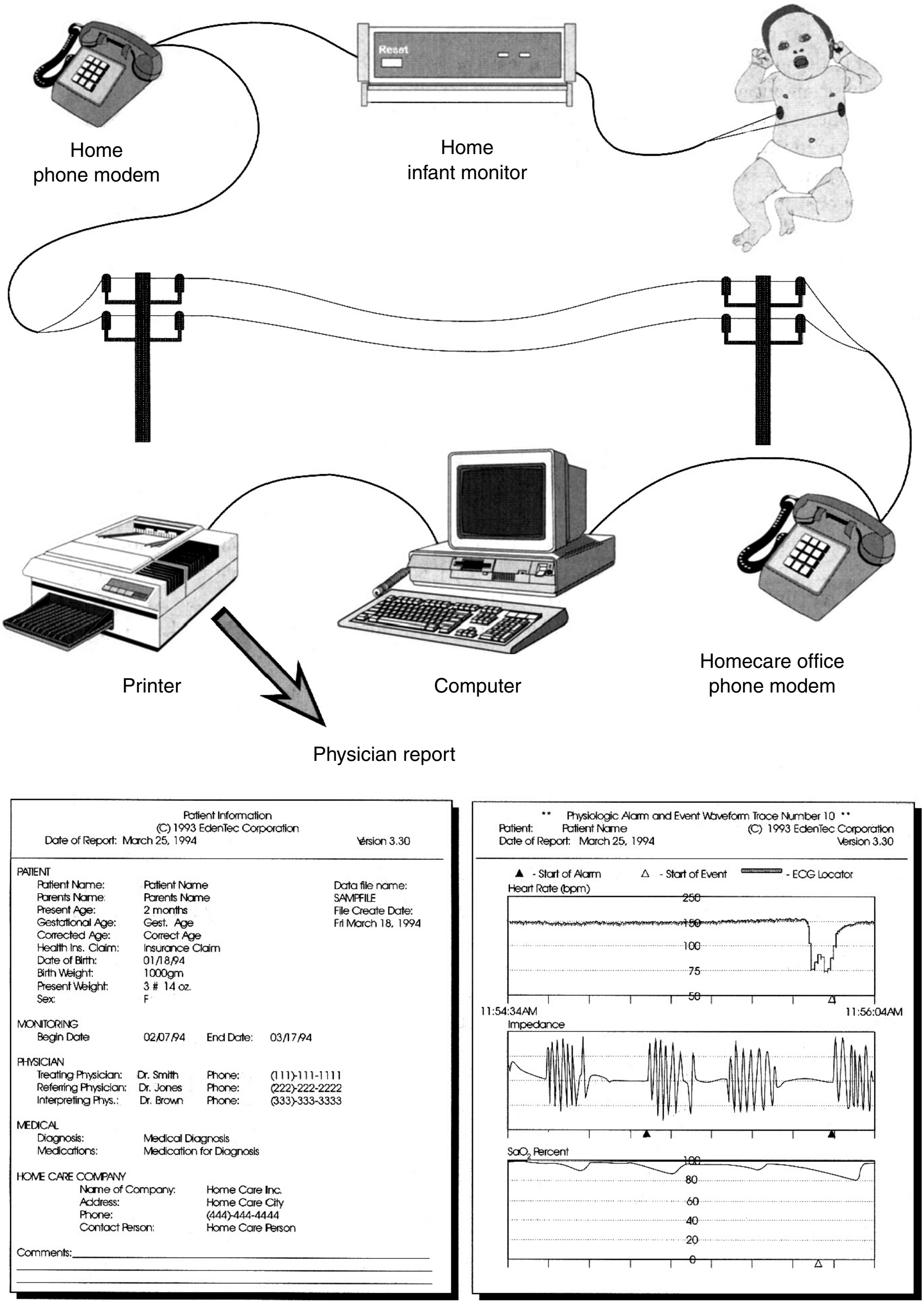
Most apnea monitors measure breathing effort through impedance pneumography. A small (100–200 μA) high-frequency (25–100 kHz) constant-current train of pulses is applied across the chest between a pair of electrodes. The voltage needed to drive the current is measured, and thereby the effective impedance between the electrodes can be calculated. Impedance across the chest increases as the chest expands and decreases as the chest contracts with each breath. The impedance change with each breath can be as low as 0.2 Ω on top of an electrode base impedance of 2000 Ω, creating some interesting signal-to-noise challenges. Furthermore, motion artifact and blood volume changes in the heart and chest can cause impedance changes of 0.6 Ω or more that can look just like breathing. Through the same pair of electrodes, heart rate is monitored by picking up the **electrocardiogram (ECG)** [AAMI, 1988].

Because the impedance technique basically measures the motion of the chest, this technique can only be used to monitor **central apnea** or lack of breathing effort. Another less common apnea in infants called **obstructive apnea** results when an obstruction of the airway blocks air from flowing in spite of breathing effort. Obstructive apnea cannot be monitored using impedance pneumography [Kelly, 1992].

There is a very broad socioeconomic and educational spectrum of parents or caregivers who may be monitoring their infants with an apnea monitor. This creates an incredible challenge for the design of the device so that it is easy enough to be used by a variety of caregivers. It also puts special requirements on the homecare service company that must be able to respond to these patients within a matter of minutes, 24 h a day.

The user-friendly monitor shown in Figure 72.1 uses a two-button operation, the on/off switch, and a reset switch. The visual alarm indicators are invisible behind a back-lit panel except when an actual alarm occurs. A word describing the alarm then appears. By not showing all nine possible alarm conditions unless an alarm occurs, parent confusion and anxiety is minimized. Numerous safety features are built into the unit, some of which are noticeable but many of which are internal to the operation of the monitor. One useful safety feature is the self-check. When the device is turned on, each alarm LED lights in sequence, and the unit beeps once indicating that the self-check was completed successfully. This gives users the opportunity to confirm that all the alarm visual indicators and the audible indicator are working and provides added confidence for users leaving their baby on the monitor. A dual-level battery alarm gives an early warning that the battery will soon need charging. The weak battery alarm allows users to reset the monitor and continue monitoring their babies for several more hours before depleting the battery to the charge battery level where the monitor must be attached to the ac battery charger/adapter. This allows parents the freedom to leave their homes for a few hours knowing that their child can continue to be monitored.

A multistage alarm reduces the risk of parents sleeping through an alarm. Most parents are sleep-deprived with a new baby. Consequently, it can be easy for parents in a nearby room to sleep through a monitor alarm even when the monitor sounds at 85 dB. A three-stage alarm helps to reduce this risk. After



Patient Information
(C) 1993 EdenTec Corporation
Date of Report: March 25, 1994 Version 3.30

PATIENT

Patient Name:	Patient Name	Data file name:
Parents Name:	Parents Name	SAMPFILE
Present Age:	2 months	File Create Date:
Gestational Age:	Gest. Age	Fri March 18, 1994
Corrected Age:	Correct Age	
Health Ins. Claim:	Insurance Claim	
Date of Birth:	01/18/94	
Birth Weight:	1000gm	
Present Weight:	3 # 14 oz.	
Sex:	F	

MONITORING

Begin Date	02/07/94	End Date:	03/17/94

PHYSICIAN

Treating Physician:	Dr. Smith	Phone:	(111)-111-1111
Referring Physician:	Dr. Jones	Phone:	(222)-222-2222
Interpreting Phys.:	Dr. Brown	Phone:	(333)-333-3333

MEDICAL

Diagnosis:	Medical Diagnosis
Medications:	Medication for Diagnosis

HOME CARE COMPANY

Name of Company:	Home Care Inc.
Address:	Home Care City
Phone:	(444)-444-4444
Contact Person:	Home Care Person

Comments:______________________________

** Physiologic Alarm and Event Waveform Trace Number 10 **
Patient: Patient Name (C) 1993 EdenTec Corporation
Date of Report: March 25, 1994 Version 3.30

FIGURE 72.3 Infant apnea monitor with memory allows data to be sent by modem to generate physician report. (Drawing courtesy of EdenTec Corporation.)

10 sec of sounding at 1 beep/sec, the alarm switches to 3 beeps/sec for the next 10 sec. Finally, if an alarm has not resolved itself after 20 sec, the alarm switches to 6 beeps/sec. Each stage of alarm sounds more intense than the previous one and offers the chance of jolting parents out of even the deepest sleep.

The physician always prescribes what alarm settings should be used by the homecare service company when setting up the monitor. As a newborn matures, these settings may need to be adjusted. Sometimes the parents can be relied upon for making these setting changes. To allow both accessibility to these switches as well as to keep them safe from unauthorized tampering from a helping brother or sister, a special tamper-resistant-adjustment procedure is utilized. Two simultaneous actions are required in order to adjust the alarm limit settings. The reset button must be continually pressed on the front of the unit while changing settings on the back of the unit. Heart rate levels are set in beats per minute, and apnea duration is set in single-second increments. Rather than using easy-to-set push-button switches, "pen-set" switches are used which require a pen or other sharp implement to make the change. If the proper switch adjustment procedure is not followed, the monitor alarms continuously and displays a switch alarm until the settings are returned to their original settings. A similar technique is used for turning the monitor off. The reset button must first be pressed and then the on/off switch turned to the off position. Violation of this procedure will result in a switch alarm.

Other safety features are internal to the monitor and are transparent to the user. The monitor's alarm is designed to be normally on from the moment the device is turned on. Active circuitry controlled by the microprocessor turns the alarm off when there are no active alarm conditions. If anything hangs up the processor or if any of a number of components fail, the alarm will not turn off and will remain on in a fail-safe mode. This "alarm on unless turned off" technique is also used in a remote alarm unit for parents with their baby in a distant room. If a wire breakage occurs between the monitor and the remote alarm unit, or a connector pulls loose, or a component fails, the remote alarm no longer is turned off by the monitor and it alarms in a fail-safe condition.

Switches, connectors, and wires are prone to fail. One way to circumvent this potential safety issue is use of switches with a separate line for each possible setting. The monitor continuously polls every switch line of each switch element to check that "exactly" one switch position is making contact. This guards against misreading bad switch elements, a switch inadvertently being set between two positions, or a bad connector or cable. Violation of the "exactly one contact condition" results in a switch alarm.

It is difficult to manage an apnea monitoring program in rural areas where the monitoring family may be a hundred miles or more away from the homecare service company. There are numerous ways to become frustrated with the equipment and stop using the monitor. Therefore, simplicity of use and reliability are important. Storing occurrence of alarms and documenting compliance in internal memory in the monitor help the homecare service company and the remote family cope with the situation. The monitor shown in Figure 72.1 stores in digital memory the time, date, and duration of (1) each use of the monitor; (2) occurrence of all equipment alarms; and (3) all physiologic alarms including respiratory waveforms, heart rate, and ECG for up to a 45-day period. These data in the form of a report (see Figure 72.3) can be downloaded to a laptop PC or sent via modem to the homecare service company or directly to the physician.

72.4 Conclusions

Devices that can provide positive patient outcomes with reduced overall cost to the healthcare system while being safe, reliable, and user-friendly will succeed based on pending healthcare changes. Future technology in the areas of sensors, communications, and memory capabilities should continue to increase the potential effectiveness of homecare management programs by using increasingly sophisticated devices. The challenge for the medical device designer is to provide cost-effective, reliable, and easy-to-use solutions that can be readily adopted by the multidisciplinary aspects of homecare medicine while meeting FDA requirements.

Defining Terms

Apnea: Cessation of breathing. Apnea can be classified as **central, obstructive,** or mixed, which is a combination.

Apnea of prematurity: Apnea in which the incidence and severity increases with decreasing gestational age attributable to immaturity of the respiratory control system. The incidence has increased due to improved survival rates for very-low-birth-weight premature infants.

Apparent life-threatening episode (ALTE): An episode characterized by a combination of apnea, color change, muscle tone change, choking, or gagging. To the observer it may appear the infant has died.

Capitated fee: A fixed payment for *total* program services versus the more traditional fee for service in which each individual service is charged.

Cardiac pacemaker: A device that electrically stimulates the heart at a certain rate used in absence of normal function of the heart's sino-atrial node.

Central apnea: Apnea secondary to lack of respiratory or diaphragmatic effort.

Chemotherapy: Treatment of disease by chemical agents. Term popularly used when fighting cancer chemically.

Colostomy: The creation of a surgical hole as an alternative opening of the colon.

CPR (cardiopulmonary resuscitation): Artificially replacing heart and respiration function through rhythmic pressure on the chest.

CPT-code (current procedural terminology code): A code used to describe specific procedures/tests developed by the AMA.

Electrocardiogram (ECG): The electric potential recorded across the chest due to depolarization of the heart muscle with each heartbeat.

Enteral nutrition: Chemical nutrition injected intestinally.

Food and Drug Administration (FDA): Federal agency that oversees and regulates foods, drugs, and medical devices.

Functional electrical stimulation (FES): Electric stimulation of peripheral nerves or muscles to gain functional, purposeful control over partially or fully paralyzed muscles.

Incontinence: Loss of voluntary control of the bowel or bladder.

Obstructive apnea: Apnea in which the effort to breath continues but airflow ceases due to obstruction or collapse of the airway.

Ostomy: Surgical procedure that alters the bladder or bowel to eliminate through an artificial passage.

Parenteral nutrition: Chemical nutrition injected subcutaneously, intramuscular, intrasternally, or intravenously.

Sphincter: A band of muscle fibers that constricts or closes an orifice.

Sudden infant death syndrome (SIDS): The sudden death of an infant which is unexplained by history or postmortem exam.

Transcutaneous electrical nerve stimulation (TENS): Electrical stimulation of sensory nerve fibers resulting in control of pain.

References

AAMI (1988). Association for the Advancement of Medical Instrumentation Technical Information Report. Apnea Monitoring by Means of Thoracic Impedance Pneumography, Arlington, Virg.

FDA (November 1993). Reviewers Guidance for Premarket Notification Submissions (Draft), Anesthesiology and Respiratory Device Branch, Division of Cardiovascular, Respiratory, and Neurological Devices. Food and Drug Administration. Washington, DC.

FIND/SVP (1992). The Market for Home Care Products, a Market Intelligence Report. New York.

Henderson-Smart D.J. 1992. Apnea of prematurity. In R. Beckerman, R. Brouillette, and C. Hunt (Eds.), *Respiratory Control Disorders in Infants and Children*, pp. 161–177, Baltimore, Williams and Wilkins.

Hunt C.E. (1992). Sudden infant death syndrome. In R. Beckerman, R. Brouillette, and C. Hunt (Eds.), *Respiratory Control Disorders in Infants and Children*, pp. 190–211, Baltimore, Williams and Wilkins.

Kahn A., Rebuffat E., Franco P., et al. (1992). Apparent life-threatening events and apnea of infancy. In R. Beckerman, R. Brouillette, and C. Hunt (Eds.), *Respiratory Control Disorders in Infants and Children*, pp 178–189, Baltimore, Williams and Wilkins.

Kelly D.H. (1992). Home monitoring. In R. Beckerman, R. Brouillette, and C. Hunt (Eds.), *Respiratory Control Disorders in Infants and Children,* pp. 400–412, Baltimore, Williams and Wilkins.

NIH (1987). Infantile Apnea and Home Monitoring Report of NIH Consensus Development Conference, US Department of Health and Human Services, NIH publication 87-2905.

Pierson D.J. (1994). Controversies in home respiratory care: Conference summary. *Respir. Care* 39: 294.

Wall L.L., Norton P.A., and Dehancey J.O.L. 1993. *Practical Urology,* Baltimore, Williams and Wilkins.

73

Virtual Instrumentation: Applications in Biomedical Engineering

Eric Rosow
Hartford Hospital
Premise Development Corporation

Joseph Adam
Premise Development Corporation

73.1 Overview

73.1.1 A Revolution — Graphical Programming and Virtual Instrumentation

Over the last decade, the graphical programming revolution has empowered engineers to develop customized systems, the same way the spreadsheet has empowered business managers to analyze financial data. This software technology has resulted in another type of revolution — the virtual instrumentation revolution, which is rapidly changing the instrumentation industry by driving down costs without sacrificing quality.

Virtual Instrumentation can be defined as:

> A layer of software and/or hardware added to a general-purpose computer in such a fashion that users can interact with the computer as though it were their own custom-designed traditional electronic instrument.

Today, computers can serve as the engine for instrumentation. Virtual instruments utilize the open architecture of industry-standard computers to provide the processing, memory, and display capabilities; while the off-the-shelf, inexpensive interface boards plugged into an open bus, standardized communications bus provides the vehicle for the instrument's capabilities. As a result, the open architecture of PCs and workstations allow the functionality of virtual instruments to be user defined. In addition, the processing power of virtual instruments is much greater than stand-alone instruments. This advantage will continue to accelerate due to the rapid technology evolution of PCs and workstations that results from the huge investments made in this industry.

The major benefits of virtual instrumentation include increased performance and reduced costs. In addition, because the user controls the technology through software, the flexibility of virtual instrumentation is unmatched by traditional instrumentation. The modular, hierarchical programming environment of virtual instrumentation is inherently reusable and reconfigurable.

73.2 Virtual Instrumentation and Biomedical Engineering

Virtual Instrumentation applications have encompassed nearly every industry including the telecommunications, automotive, semiconductor, and biomedical industries. In the fields of health care and biomedical engineering, virtual instrumentation has empowered developers and end-users to conceive of, develop, and implement a wide variety of research-based biomedical applications and executive information tools. These applications fall into several categories including: clinical research, equipment testing and quality assurance, data management, and performance improvement.

In a collaborative approach, physicians, researchers, and biomedical and software engineers at Hartford Hospital (Hartford, CT) and Premise Development Corporation (Avon, CT) have developed various data acquisition and analysis systems that successfully integrate virtual instrumentation principles in a wide variety of environments. These include:

- "The EndoTester™," a patented quality assurance system for fiberoptic endoscopes
- A Non-Invasive Pulmonary Diffusion and Cardiac Output Measurement System
- A Cardiovascular Pressure-Dimension Analysis System
- "BioBench™," a powerful turnkey application for physiological data acquisition and analysis
- "PIVIT™," a Performance Indicator Virtual Instrument Toolkit to manage and forecast financial data
- A "Virtual Intelligence Program" to manage the discrete components within the continuum of care "BabySave™," an analysis and feedback system for apnea interruption via vibratory stimulation

This chapter will describe several of these applications and describe how they have allowed clinicians and researchers to gain new insights, discover relationships that may not have been obvious, and test and model hypotheses based on acquired data sets. In some cases, these applications have been developed into commercial products to address test and measurement needs at other healthcare facilities throughout the world.

73.2.1 Example 1: BioBench™ — A Virtual Instrument Application for Data Acquisition and Analysis of Physiological Signals

The biomedical industry is an industry that relies heavily on the ability to acquire, analyze, and display large quantities of data. Whether researching disease mechanisms and treatments by monitoring and storing physiological signals, researching the effects of various drugs interactions, or teaching students in labs where students study physiological signs and symptoms, it was clear that there existed a strong demand for a flexible, easy-to-use, and cost-effective tool. In a collaborative approach, biomedical engineers, software engineers and clinicians, and researchers created a suite of virtual instruments called BioBench™.

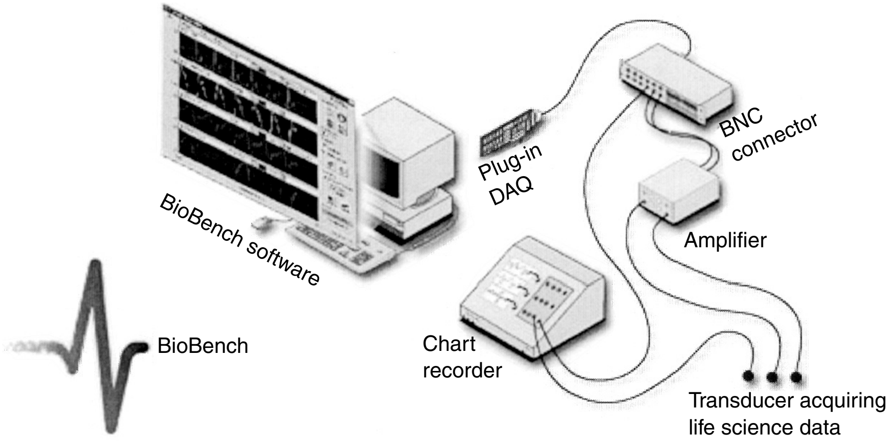


FIGURE 73.1 A typical biomedical application using BioBench. (Courtesy of National Instruments.)

BioBench™ (National Instruments, Austin, TX) is a new software application designed for physiological data acquisition and analysis. It was built with LabVIEW™, the world's leading software development environment for data acquisition, analysis, and presentation.[1] Coupled with National Instruments data acquisition (DAQ) boards, BioBench integrates the PC with data acquisition for the life sciences market.

Many biologists and physiologists have made major investments over time in data acquisition hardware built before the advent of modern PCs. While these scientists cannot afford to throw out their investment in this equipment, they recognize that computers and the concept of virtual instrumentation yield tremendous benefits in terms of data analysis, storage, and presentation. In many cases, traditional medical instrumentation may be too expensive to acquire and maintain. As a result, researchers and scientists are opting to create their own PC-based data monitoring systems in the form of virtual instruments.

Other life scientists, who are just beginning to assemble laboratory equipment, face the daunting task of selecting hardware and software needed for their application. Many manufacturers for the life sciences field focus their efforts on the acquisition of raw signals and converting these signals into measurable linear voltages. They do not concentrate on digitizing signals or the analysis and display of data on the PC. BioBench™ is a low-cost turnkey package that requires no programming. BioBench is compatible with any isolation amplifier or monitoring instrument that provides an analog output signal. The user can acquire and analyze data immediately because BioBench automatically recognizes and controls the National Instruments DAQ hardware, minimizing configuration headaches.

Some of the advantages of PC-Based Data Monitoring include:

- Easy-to-use-software applications
- Large memory and the PCI bus
- Powerful processing capabilities
- Simplified customization and development
- More data storage and faster data transfer
- More efficient data analysis

Figure 73.1 illustrates a typical setup of a data acquisition experiment using BioBench. BioBench also features pull-down menus through which the user can configure devices. Therefore, those who have made large capital investments can easily migrate their existing equipment into the computer age. Integrating a combination of old and new physiological instruments from a variety of manufacturers is an important and straightforward procedure. In fact, within the clinical and research setting, it is a common requirement

[1] BioBench™ was developed for National Instruments (Austin, TX) by Premise Development Corporation (Avon, CT).

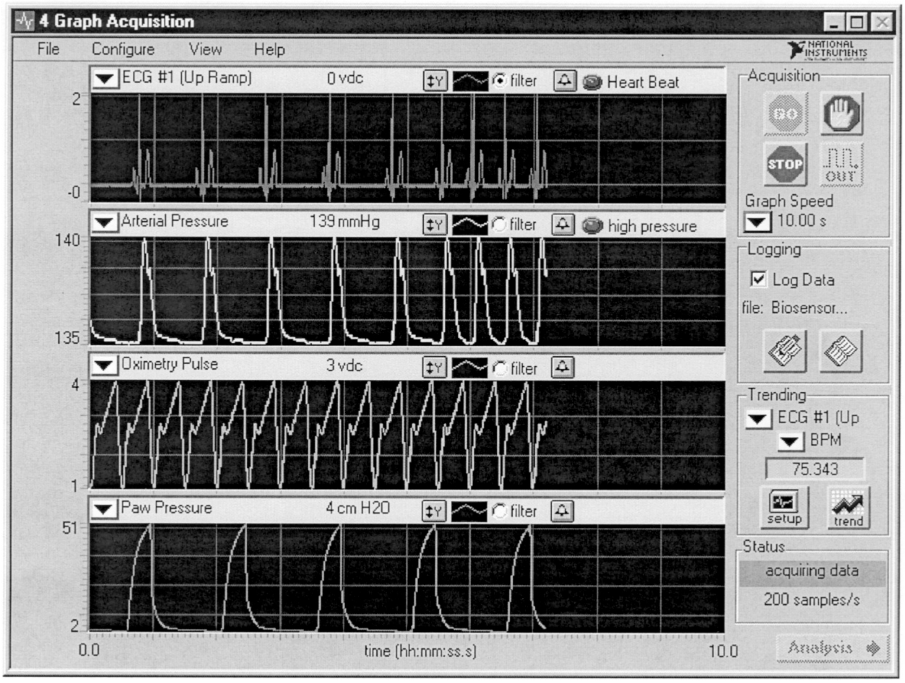
to be able to acquire multiple physiological signals from a variety of medical devices and instruments which do not necessarily communicate with each other. Often times, this situation is compounded by the fact that end-users would like to be able to view and analyze an entire waveform and not just an average value. In order to accomplish this, the end-user must acquire multiple channels of data at a relatively high sampling rate and have the ability to manage many large data files. BioBench can collect up to 16 channels simultaneously at a sampling rate of 1000 Hz per channel. Files are stored in an efficient binary format which significantly reduces the amount of hard disk and memory requirements of the PC. During data acquisition, a number of features are available to the end-user. These features include:

Data Logging: Logging can be enabled prior to or during an acquisition. The application will either prompt the user for a descriptive filename or it can be configured to automatically assign a filename for each acquisition. Turning the data logging option on and off creates a log data event record that can be inspected in any of the analysis views of BioBench.

Event Logging: The capacity to associate and recognize user commands associated with a data file may be of significant value. BioBench has been designed to provide this capability by automatically logging user-defined events, stimulus events, and file logging events. With user-defined events, the user can easily enter and associate date and time-stamped notes with user actions or specific subsets of data. Stimulus events are also data and time-stamped and provide the user information about whether a stimulus has been turned on or off. File logging events note when data has been logged to disk. All of these types of events are stored with the raw data when logging data to file and they can be searched for when analyzing data.

Alarming: To alert the user about specific data values and thresholds, BioBench incorporates user-defined alarms for each signal which is displayed. Alarms appear on the user interface during data acquisition and notify the user that an alarm condition has occurred.

Figure 73.2 is an example of the Data Acquisition mode of BioBench. Once data has been acquired, BioBench can employ a wide array of easy-to-use analysis features. The user has the choice of importing recently acquired data or opening a data file that had been previously acquired for comparison or teaching


ECG #1 (Up Ramp) 0 vdc ● filter Heart Beat

Arterial Pressure 139 mmHg ○ filter high pressure

Oximetry Pulse 3 vdc ○ filter

Paw Pressure 4 cm H20 ○ filter

Graph Speed 10.00 s

☑ Log Data

file: Biosensor...

Trending: ECG #1 (Up BPM 75.343

Status: acquiring data 200 samples/s

FIGURE 73.2 BioBench acquisition mode with alarms enabled.

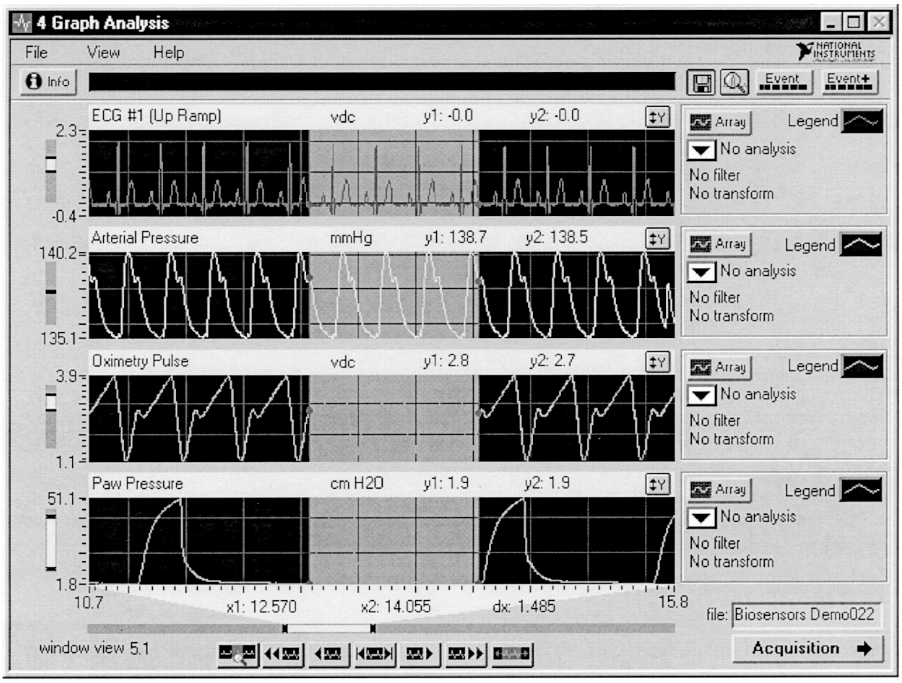


FIGURE 73.3 BioBench analysis mode.

purposes. Once a data set has been selected and opened, BioBench allows the user to simply select and highlight a region of interest and choose the analysis options to perform a specific routine.

BioBench implements a wide array of scalar and array analyses. For example, scalar analysis tools will determine the minimum, maximum, mean, integral, and slope of a selected data set, while the array analysis tools can employ Fast Fourier Transforms (FFTs), peak detection, histograms, and X vs. Y plots.

The ability to compare multiple data files is very important in analysis and BioBench allows the user to open an unlimited number of data files for simultaneous comparison and analysis. All data files can be scanned using BioBench's search tools in which the user can search for particular events that are associated with areas of interest. In addition, BioBench allows the user to employ filters and transformations to their data sets and all logged data can be easily exported to a spreadsheet or database for further analysis. Finally, any signal acquired with BioBench can be played back, thus taking lab experience into the classroom. Figure 73.3 illustrates the analysis features of BioBench.

73.2.2 Example 2: A Cardiovascular Pressure-Dimension Analysis System

73.2.2.1 Introduction

The intrinsic contractility of the heart muscle (myocardium) is the single most important determinant of prognosis in virtually all diseases affecting the heart (e.g., coronary artery disease, valvular heart disease, and cardiomyopathy). Furthermore, it is clinically important to be able to evaluate and track myocardial function in other situations, including chemotherapy (where cardiac dysfunction may be a side effect of treatment) and liver disease (where cardiac dysfunction may complicate the disease).

The most commonly used measure of cardiac performance is the ejection fraction. Although it does provide some measure of intrinsic myocardial performance, it is also heavily influenced by other factors such as heart rate and loading conditions (i.e., the amount of blood returning to the heart and the pressure against which the heart ejects blood).

Better indices of myocardial function based on the relationship between pressure and volume throughout the cardiac cycle (pressure–volume loops) exist. However, these methods have been limited because they require the ability to track ventricular volume continuously during rapidly changing loading conditions. While there are many techniques to measure volume under steady state situations, or at end-diastole and end-systole (the basis of ejection fraction determinations), few have the potential to record volume during changing loading conditions.

Echocardiography can provide online images of the heart with high temporal resolution (typically 30 frames/sec). Since echocardiography is radiation-free and has no identifiable toxicity, it is ideally suited to pressure–volume analyses. Until recently however, its use for this purpose has been limited by the need for manual tracing of the endocardial borders, an extremely tedious and time-consuming endeavor.

73.2.2.2 The System

Biomedical and software engineers at Premise Development Corporation (Avon, CT), in collaboration with physicians and researchers at Hartford Hospital, have developed a sophisticated research application called the "Cardiovascular Pressure-Dimension Analysis (CPDA) System." The CPDA system acquires echocardiographic volume and area information from the acoustic quantification (AQ) port, in conjunction with vetricular pressure(s) and ECG signals to rapidly perform pressure–volume and pressure–area analyses. This fully automated system allows cardiologists and researchers to perform online pressure–dimension and stroke work analyses during routine cardiac catheterizations and open-heart surgery. The system has been designed to work with standard computer hardware. Analog signals for ECG, pressure, and area/volume (AQ) are connected to a standard BNC terminal board. Automated calibration routines ensure that each signal is properly scaled and allows the user to immediately collect and analyze pressure–dimension relationships.

The CPDA can acquire up to 16 channels of data simultaneously. Typically, only three physiological parameters, ECG, pressure, and the AQ signals are collected using standard data acquisition hardware. In addition, the software is capable of running on multiple operating systems including Macintosh, Windows 95/98/NT, and Solaris. The CPDA also takes advantage of the latest hardware developments and form-factors and can be used with either a desktop or a laptop computer.

The development of an automated, online method of tracing endocardial borders (Hewlett–Packard's AQ Technology) (Hewlett–Packard Medical Products Group, Andover, MA) has provided a method for rapid online area and volume determinations. Figure 73.4 illustrates this AQ signal from a Hewlett–Packard

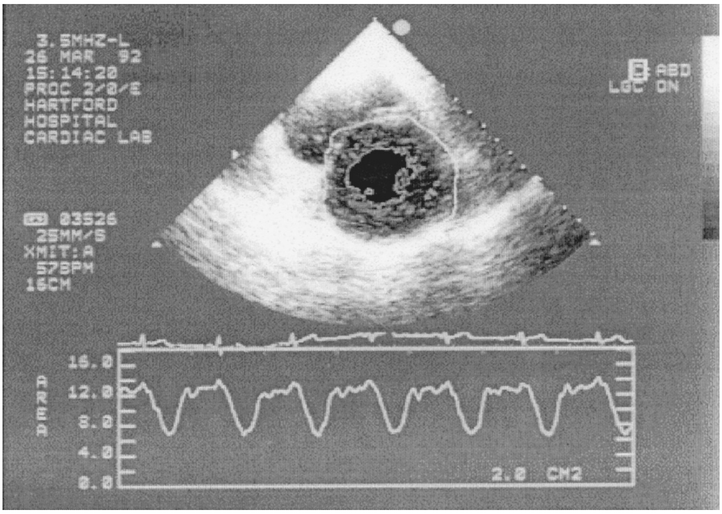
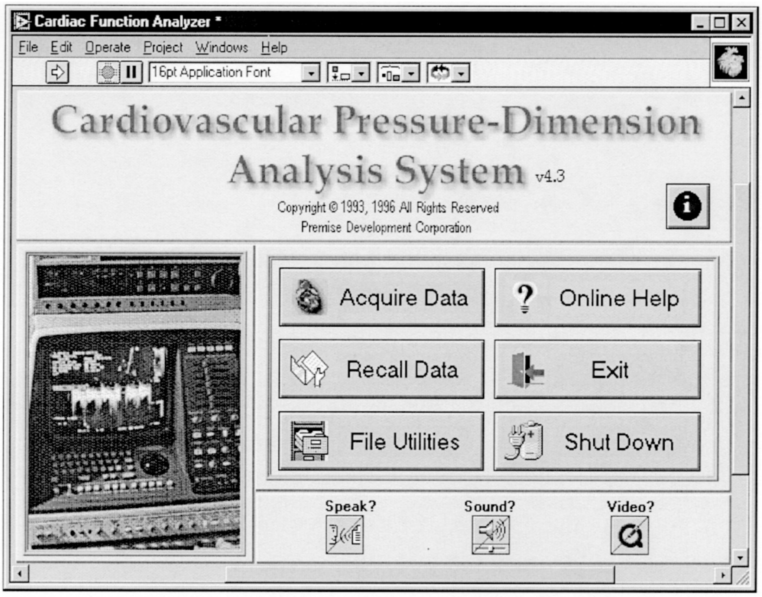


FIGURE 73.4 The Acoustic Quantification (AQ) signal. (Courtesy of Hewlett–Packard.)



Cardiovascular Pressure-Dimension Analysis System v4.3

Copyright © 1993, 1996 All Rights Reserved
Premise Development Corporation

FIGURE 73.5 Cardiovascular pressure-dimension analysis main menu.

Sonos Ultrasound Machine. This signal is available as an analog voltage (-1 to $+1$ V) through the Sonos Dataport option (BNC connector).

73.2.2.3 Data Acquisition and Analysis

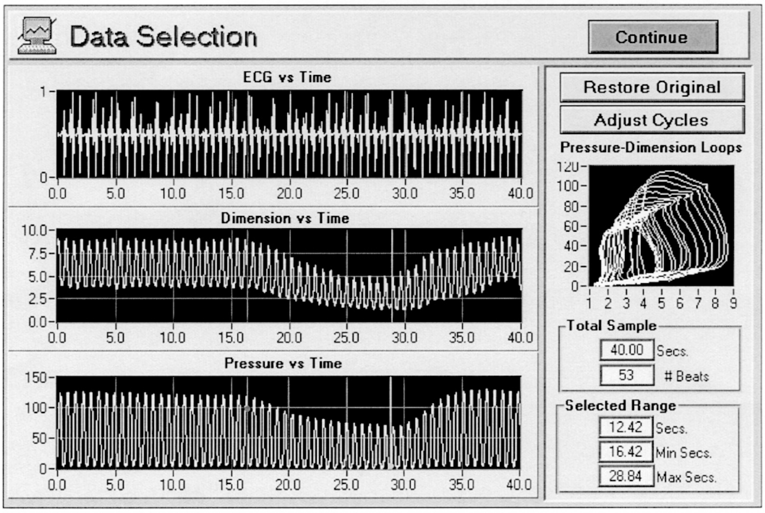
Upon launching this application, the user is presented with a dialog box that reviews the license agreement and limited warranty. Next, the Main Menu is displayed, allowing the user to select from one of six options as shown in Figure 73.5.

73.2.2.4 Clinical Significance

Several important relationships can be derived from this system. Specifically, a parameter called the *End-Systolic Pressure–Volume Relationship* (*ESPVR*) describes the line of best fit through the peak-ratio (maximum pressure with respect to minimum volume) coordinates from a series of pressure–volume loops generated under varying loading conditions. The slope of this line has been shown to be a sensitive index of myocardial contractility that is independent of loading conditions. In addition, several other analyses, including *time varying elastance* ($E_{\max}$) and *stroke work*, are calculated. Time-varying elastance is measured by determining the maximum slope of a regression line through a series of isochronic pressure–volume coordinates. Stroke work is calculated by quantifying the area of each pressure–volume loop. Statistical parameters are also calculated and displayed for each set of data. Figure 73.7 illustrates the pressure–dimension loops and each of the calculated parameters along with the various analysis options. Finally, the user has the ability to export data sets into spreadsheet and database files and export graphs and indicators into third-party presentation software packages such as Microsoft PowerPoint®.

73.3 Summary

Virtual Instrumentation allows the development and implementation of innovative and cost-effective biomedical applications and information management solutions. As the healthcare industry continues to respond to the growing trends of managed care and capitation, it is imperative for clinically useful, cost-effective technologies to be developed and utilized. As application needs will surely continue to change,



Total Sample

40.00 Secs.

53 # Beats

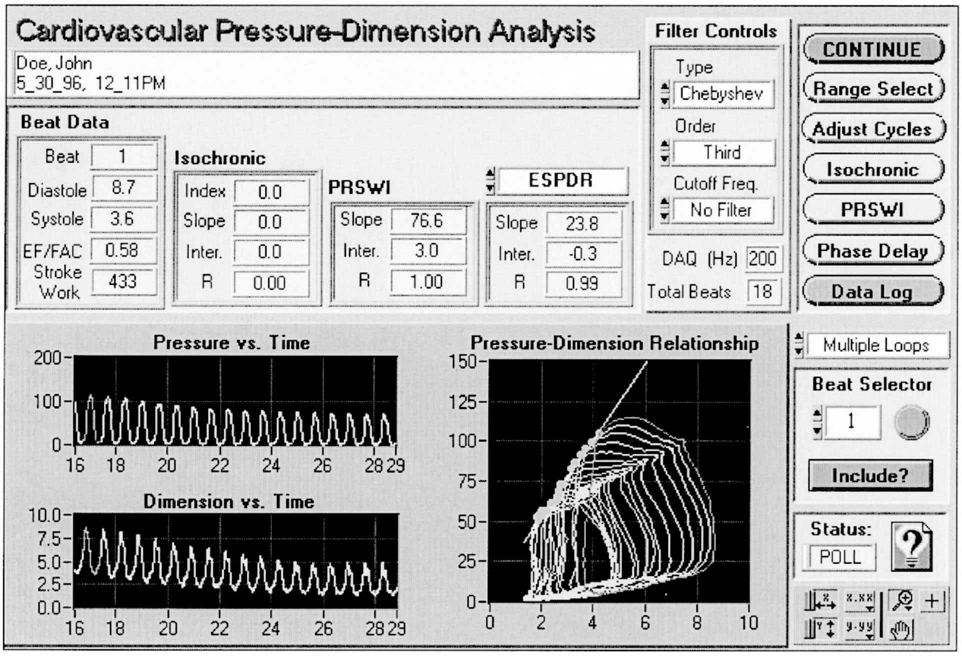
Selected Range

12.42 Secs.

16.42 Min Secs.

28.84 Max Secs.

FIGURE 73.6 The data selection front panel.



Doe, John
5_30_96, 12_11PM

Beat Data

Beat 1

Diastole 8.7

Systole 3.6

EF/FAC 0.58

Stroke Work 433

Isochronic

Index 0.0

Slope 0.0

Inter. 0.0

R 0.00

PRSWI

Slope 76.6

Inter. 3.0

R 1.00

ESPDR

Slope 23.8

Inter. -0.3

R 0.99

Filter Controls

Type Chebyshev

Order Third

Cutoff Freq. No Filter

DAQ (Hz) 200

Total Beats 18

Multiple Loops

Beat Selector 1

Status: POLL

FIGURE 73.7 The cardiac cycle analysis front panel.

virtual instrumentation systems will continue to offer users flexible and powerful solutions without requiring new equipment or traditional instruments.

References

[1] 1. Fisher, J.P., Mikan, J.S., Rosow, E., Nagle, J., Fram, D.B., Maffucci, L.M., McKay, R.G., and Gillam, L.D., "Pressure-Dimension Analysis of Regional Left Ventricular Performance Using

Echocardiographic Automatic Boundary Detection: Validation in an Animal Model of Inotropic Modulation," *J. Am. Coll. Cardiol.* 19, 262A, 1992.

[2] Fisher, J.P., McKay, R.G., Mikan, J.S., Rosow, E., Nagle, J., Mitchel, J.F., Kiernan, F.J., Hirst, J.A., Primiano, C.A., Fram, D.B., and Gillam, L.D., Hartford Hospital and University of Connecticut, Hartford, CT, *"Human Left Ventricular Pressure–Area and Pressure–Volume Analysis Using Echocardiographic Automatic Boundary Detection." 65th Scientific Session of the American Heart Association* (11/92).

[3] Fisher, J.P., Mitchel, J.F., Rosow, E., Mikan, J.S., Nagle, J., Kiernan, F.J., Hirst, J.A., Primiano, and Gillam, L.D., Hartford Hospital and University of Connecticut, Hartford, CT,*"Evaluation of Left Ventricular Diastolic Pressure–Area Relations with Echocardiographic Automatic Boundary Detection," 65th Scientific Session of the American Heart Association* (11/92).

[4] Fisher, J.P., McKay, R.G., Mikan, J.S., Rosow, E., Nagle, J., Mitchel, J.F., Fram, D.B., and Gillam, L.D., Hartford Hospital, "*A Comparison of Echocardiographic Methods of Evaluating Regional LV Systolic Function: Fractional Area Change vs. the End-Systolic Pressure–Area Relation." 65th Scientific Session of the American Heart Association* (11/92).

[5] Fisher, J.P., McKay, R.G., Rosow, E. Mikan, J. Nagle, J. Hirst, J.A., Fram, D.B., and Gillam, L.D., "*On-Line Derivation of Human Left Ventricular Pressure–Volume Loops and Load Independent Indices of Contractility Using Echocardiography with Automatic Boundary Detection-A Clinical Reality." Circulation* 88, I–304, 1993.

[6] Fisher, J.P., Chen, C., Krupowies, N., Li Li, Kiernan, F.J., Fram, D.B., Rosow, E., Adam, J., and Gillam, L.D., *"Comparison of Mechanical and Pharmacologic Methods of Altering Loading Conditions to Determine End-Systolic Indices of Left Ventricle Function." Circulation* 90, 1–494, 1994.

[7] Fisher, J.P., Martin, J., Day, F.P., Rosow, E., Adam, J., Chen, C., and Gillam, L.D. "Validation of a Less Invasive Method for Determining Preload Recruitable Stroke Work Derived with Echocardiographic Automatic Boundary Detection." *Circulation* 92, 1–278, 1995.

[8] Fontes, M.L., Adam, J., Rosow, E., Mathew, J., and DeGraff, A.C. "Non-Invasive Cardiopulmonary Function Assessment System," *J. Clin. Monit.* 13, 413, 1997.

[9] Johnson, G.W., *LabVIEW Graphical Programming: Practical Applications in Instrumentation and Control*, 2nd ed., McGraw-Hill, New York, 1997.

[10] Mathew, J.P., Adam, J., Rosow, E., Fontes, M.L., Davis, L., Barash, P.G., and Gillam, L., "Cardiovascular Pressure-Dimension Analysis System," *J. Clin. Monit.* 13, 423, 1997.

[11] National Instruments. Measurement and Automation Catalog, National Instruments, Austin, TX, 1999.

[12] Rosow, E. "Technology and Sensors," Presented at the *United States Olympic Committee's Sports Equipment and Technology Conference*, Colorado Springs, CO; November 19–23, 1992.

[13] Rosow, E. "Biomedical Applications using LabVIEW," Presented at the *New England Society of Clinical Engineering, Sturbridge*, MA, November 14, 1993.

[14] Rosow, E., Adam, J.S., Nagle, J., Fisher, J.P., and Gillam, L.D. Premise Development Corporation, Avon, CT and Hartford Hospital, Hartford, CT, "A Cardiac Analysis System: LabVIEW and Acoustic Quantification (AQ)," Presented at the *Association for the Advancement of Medical Instrumentation* in Washington, D.C., May 21–25, 1994.

VII

Clinical Engineering

Yadin David
Texas Children's Hospital

OVER THE PAST 100 YEARS, the health care system's dependence on medical technology for the delivery of its services has grown continuously. To some extent, all professional care providers depend on technology, be it in the area of preventive medicine, diagnosis, therapeutic care, rehabilitation, administration, or health-related education and training. Medical technology enables practitioners to intervene through integrated interactions with their patients in a cost-effective, efficient, and safe manner. As a result, the field of clinical engineering has emerged as the discipline of biomedical engineering that fulfills the need to manage the deployment of medical technology and to integrate it appropriately with desired clinical practices.

The healthcare delivery system presents a very complex environment where facilities, equipment, materials, and a full range of human interventions are involved. It is in this clinical environment that patients of various ages and conditions, trained staff, and the wide variety of medical technology converge. This complex mix of interactions may lead to unacceptable risk when programs for monitoring, controlling, improving, and educating all entities involved are not appropriately integrated by qualified professionals.

This section of clinical engineering focuses on the methodology for administering critical engineering services that vary from facilitation of innovation and technology transfer to the performance of technology assessment and operations support and on the management tools with which today's clinical engineer needs to be familiar. With increased awareness of the value obtained by these services, new career opportunities are created for clinical engineers.

In addition to highlighting the important roles that clinical engineers serve in many areas, the section focuses on those areas of the clinical engineering field that enhance the understanding of the "bigger picture." With such an understanding, the participation in and contribution by clinical engineers to this enlarged scope can be fully realized. The adoption of the tools described here will enable clinical engineers to fulfill their new role in the evolving health care delivery system.

All the authors in this section recognize this opportunity and here recognized for volunteering their talent and time so that others can excel as well.

74

Clinical Engineering: Evolution of a Discipline

Joseph D. Bronzino
Trinity College

74.1 Who Is a Clinical Engineer?

As discussed in the introduction to this *Handbook*, biomedical engineers apply the concepts, knowledge, and techniques of virtually all engineering disciplines to solve specific problems in the biosphere, that is, the realm of biology and medicine. When biomedical engineers work within a hospital or clinic, they are more appropriately called *clinical engineers*. But what exactly is the definition of the term *clinical engineer*? For the purposes of this handbook, a *clinical engineer* is defined as an engineer who has graduated from an accredited academic program in engineering or who is licensed as a professional engineer or engineer-in-training and is engaged in the application of scientific and technological knowledge developed through engineering education and subsequent professional experience within the health care environment in support of clinical activities. Furthermore, clinical environment means that portion of the health care system in which patient care is delivered, and clinical activities include direct patient care, research, teaching, and public service activities intended to enhance patient care.

74.2 Evolution of Clinical Engineering

Engineers were first encouraged to enter the clinical scene during the late 1960s in response to concerns about patient safety as well as the rapid proliferation of clinical equipment, especially in academic medical centers. In the process, a new engineering discipline — clinical engineering — evolved to provide the technological support necessary to meet these new needs. During the 1970s, a major expansion of clinical engineering occurred, primarily due to the following events:

- The Veterans' Administration (VA), convinced that clinical engineers were vital to the overall operation of the VA hospital system, divided the country into biomedical engineering districts, with a chief biomedical engineer overseeing all engineering activities in the hospitals in that district.
- Throughout the United States, clinical engineering departments were established in most of the large medical centers and hospitals and in some smaller clinical facilities with at least 300 beds.
- Clinical engineers were hired in increasing numbers to help these facilities use existing technology and incorporate new technology.

Having entered the hospital environment, routine electrical safety inspections exposed the clinical engineer to all types of patient equipment that was not being maintained properly. It soon became obvious that electrical safety failures represented only a small part of the overall problem posed by the presence of medical equipment in the clinical environment. The equipment was neither totally understood nor properly maintained. Simple visual inspections often revealed broken knobs, frayed wires, and even evidence of liquid spills. Investigating further, it was found that many devices did not perform in accordance with manufacturers' specifications and were not maintained in accordance with manufacturers' recommendations. In short, electrical safety problems were only the tip of the iceberg. The entrance of clinical engineers into the hospital environment changed these conditions for the better. By the mid-1970s, complete performance inspections before and after use became the norm, and sensible inspection procedures were developed. In the process, clinical engineering departments became the logical support center for all medical technologies and became responsible for all the biomedical instruments and systems used in hospitals, the training of medical personnel in equipment use and safety, and the design, selection, and use of technology to deliver safe and effective health care.

With increased involvement in many facets of hospital/clinic activities, clinical engineers now play a multifaceted role (Figure 74.1). They must interface successfully with many "clients," including clinical staff, hospital administrators, regulatory agencies, etc., to ensure that the medical equipment within the hospital is used safely and effectively.

Today, hospitals that have established centralized clinical engineering departments to meet these responsibilities use clinical engineers to provide the hospital administration with an objective option of equipment function, purchase, application, overall system analysis, and preventive maintenance policies.

Some hospital administrators have learned that with the in-house availability of such talent and expertise, the hospital is in a far better position to make more effective use of its technological resources [Bronzino, 1992]. By providing health professionals with needed assurance of safety, reliability, and efficiency in using new and innovative equipment, clinical engineers can readily identify poor-quality and ineffective equipment, thereby resulting in faster, more appropriate utilization of new medical equipment. Typical pursuits of clinical engineers, therefore, include:

- Supervision of a hospital clinical engineering department that includes clinical engineers and biomedical equipment technicians (BMETs)
- Prepurchase evaluation and planning for new medical technology
- Design, modification, or repair of sophisticated medical instruments or systems
- Cost-effective management of a medical equipment calibration and repair service
- Supervision of the safety and performance testing of medical equipment performed by BMETs
- Inspection of all incoming equipment (i.e., both new and returning repairs)
- Establishment of performance benchmarks for all equipment
- Medical equipment inventory control
- Coordination of outside engineering and technical services performed by vendors

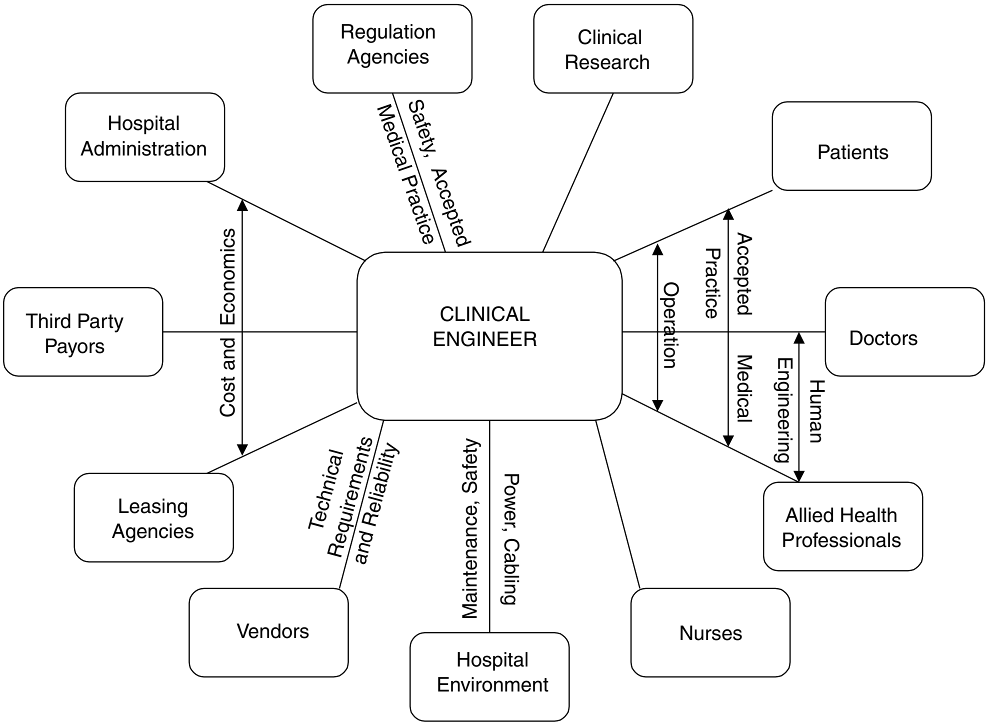


FIGURE 74.1 Diagram illustrating the range of interactions of a clinical engineer.

- Training of medical personnel in the safe and effective use of medical devices and systems
- Clinical applications engineering, such as custom modification of medical devices for clinical research, evaluation of new noninvasive monitoring systems, etc.
- Biomedical computer support
- Input to the design of clinical facilities where medical technology is used, for example, operating rooms (ORs), intensive care units, etc.
- Development and implementation of documentation protocols required by external accreditation and licensing agencies

Clinical engineers thus provide extensive engineering services for the clinical staff and, in recent years, have been increasingly accepted as valuable team members by physicians, nurses, and other clinical professionals. Furthermore, the acceptance of clinical engineers in the hospital setting has led to different types of engineering–medicine interactions, which in turn have improved health care delivery.

74.3 Hospital Organization and the Role of Clinical Engineering

In the hospital, management organization has evolved into a diffuse authority structure that is commonly referred to as the *triad model*. The three primary components are the governing board (trustees), hospital administration (CEO and administrative staff), and the medical staff organization. The role of the governing board and the chief executive officer are briefly discussed below to provide some insight regarding their individual responsibilities and their interrelationship.

74.3.1 Governing Board (Trustees)

The **Joint Commission on the Accreditation of Healthcare Organizations (JCAHO)** summarizes the major duties of the governing board as "adopting by-laws in accordance with its legal accountability and its responsibility to the patient." The governing body, therefore, requires both medical and paramedical departments to monitor and evaluate the quality of patient care, which is a critical success factor in hospitals today. To meet this goal, the governing board essentially is responsible for establishing the mission statement and defining the specific goals and objectives that the institution must satisfy. Therefore, the trustees are involved in the following functions:

- Establishing the policies of the institution
- Providing equipment and facilities to conduct patient care
- Ensuring that proper professional standards are defined and maintained (i.e., providing quality assurance)
- Coordinating professional interests with administrative, financial, and community needs
- Providing adequate financing by securing sufficient income and managing the control of expenditures
- Providing a safe environment
- Selecting qualified administrators, medical staff, and other professionals to manage the hospital

In practice, the trustees select a hospital chief administrator who develops a plan of action that is in concert with the overall goals of the institution.

74.3.2 Hospital Administration

The hospital administrator, the chief executive officer of the medical enterprise, has a function similar to that of the chief executive officer of any corporation. The administrator represents the governing board in carrying out the day-to-day operations to reflect the broad policy formulated by the trustees. The duties of the administrator are summarized as follows:

- Preparing a plan for accomplishing the institutional objectives, as approved by the board.
- Selecting medical chiefs and department directors to set standards in their respective fields.
- Submitting for board approval an annual budget reflecting both expenditures and income projections.
- Maintaining all physical properties (plant and equipment) in safe operating condition.
- Representing the hospital in its relationships with the community and health agencies.
- Submitting to the board annual reports that describe the nature and volume of the services delivered during the past year, including appropriate financial data and any special reports that may be requested by the board.

In addition to these administrative responsibilities, the chief administrator is charged with controlling cost, complying with a multitude of governmental regulations, and ensuring that the hospital conforms to professional norms, which include guidelines for the care and safety of patients.

74.4 Clinical Engineering Programs

In many hospitals, administrators have established clinical engineering departments to manage effectively all the technological resources, especially those relating to medical equipment, that are necessary for providing patient care. The primary objective of these departments is to provide a broad-based engineering program that addresses all aspects of medical instrumentation and systems support.

Figure 74.2 illustrates the organizational chart of the medical support services division of a typical major medical facility. Note that within this organizational structure, the director of clinical engineering reports directly to the vice-president of medical support services. This administrative relationship is extremely important because it recognizes the important role clinical engineering departments play in delivering quality care. It should be noted, however, that in other common organizational structures, clinical engineering services may fall under the category of "facilities," "materials management," or even just

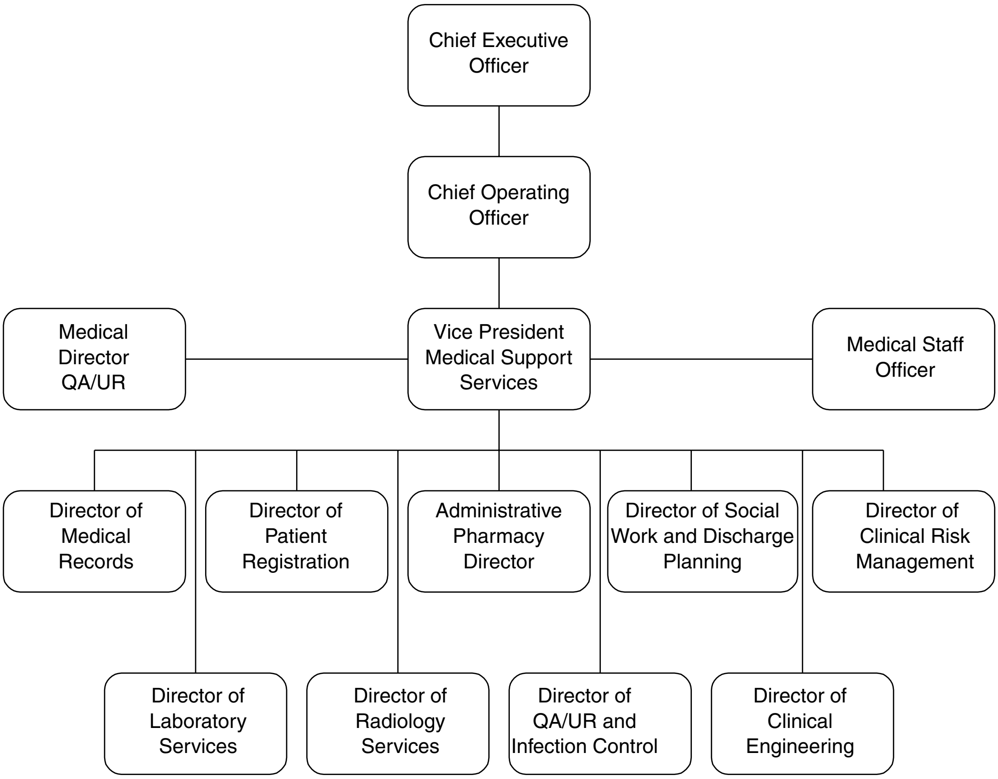

Chief Executive Officer
Chief Operating Officer
Medical Director QA/UR
Vice President Medical Support Services
Medical Staff Officer
Director of Medical Records
Director of Patient Registration
Administrative Pharmacy Director
Director of Social Work and Discharge Planning
Director of Clinical Risk Management
Director of Laboratory Services
Director of Radiology Services
Director of QA/UR and Infection Control
Director of Clinical Engineering

FIGURE 74.2 Organizational chart of medical support services division for a typical major medical facility. This organizational structure points out the critical interrelationship between the clinical engineering department and the other primary services provided by the medical facility.

"support services." Clinical engineers also can work directly with clinical departments, thereby bypassing much of the hospital hierarchy. In this situation, clinical departments can offer the clinical engineer both the chance for intense specialization and, at the same time, the opportunity to develop personal relationships with specific clinicians based on mutual concerns and interests.

Once the hospital administration appoints a qualified individual as director of clinical engineering, the person usually functions at the department-head level in the organizational structure of the institution and is provided with sufficient authority and resources to perform the duties efficiently and in accordance with professional norms. To understand the extent of these duties, consider the job title for "clinical engineering director" as defined by the World Health Organization [Issakov et al., 1990]:

General Statement. The clinical engineering director, by his or her education and experience, acts as a manager and technical director of the clinical engineering department. The individual designs and directs the design of equipment modifications that may correct design deficiencies or enhance the clinical performance of medical equipment. The individual may also supervise the implementation of those design modifications. The education and experience that the director possesses enables him or her to analyze complex medical or laboratory equipment for purposes of defining corrective maintenance and developing appropriate preventive maintenance or performance assurance protocols. The clinical engineering director works with nursing and medical staff to analyze new medical equipment needs and participates in both the prepurchase planning process and the incoming testing process. The individual also participates in the equipment management process through involvement in the system development, implementation, maintenance, and modification processes.

Duties and Responsibilities. The director of clinical engineering has a wide range of duties and responsibilities. For example, this individual:

- Works with medical and nursing staff in the development of technical and performance specifications for equipment requirements in the medical mission.
- Once equipment is specified and the purchase order developed, generates appropriate testing of the new equipment.
- Does complete performance analysis on complex medical or laboratory equipment and summarizes results in brief, concise, easy-to-understand terms for the purposes of recommending corrective action or for developing appropriate preventive maintenance and performance assurance protocols.
- Designs and implements modifications that permit enhanced operational capability. May supervise the maintenance or modification as it is performed by others.
- Must know the relevant codes and standards related to the hospital environment and the performance assurance activities. (Examples in the United States are NFPA 99, UL 544, and JCAHO, and internationally, IEC-TC 62.)
- Is responsible for obtaining the engineering specifications (systems definitions) for systems that are considered unusual or one-of-a-kind and are not commercially available.
- Supervises in-service maintenance technicians as they work on codes and standards and on preventive maintenance, performance assurance, corrective maintenance, and modification of new and existing patient care and laboratory equipment.
- Supervises parts and supply purchase activities and develops program policies and procedures for same.
- Sets departmental goals, develops budgets and policy, prepares and analyzes management reports to monitor department activity, and manages and organizes the department to implement them.
- Teaches measurement, calibration, and standardization techniques that promote optimal performance.
- In equipment-related duties, works closely with maintenance and medical personnel. Communicates orally and in writing with medical, maintenance, and administrative professionals. Develops written procedures and recommendations for administrative and technical personnel.

Minimum Qualifications. A bachelor's degree (4 years) in an electrical or electronics program or its equivalent is required (preferably with a clinical or biomedical adjunct). A master's degree is desirable. A minimum of 3 years' experience as a clinical engineer and 2 years in a progressively responsible supervisory capacity is needed. Additional qualifications are as follows:

- Must have some business knowledge and management skills that enable him or her to participate in budgeting, cost accounting, personnel management, behavioral counseling, job description development, and interviewing for hiring or firing purposes. Knowledge and experience in the use of microcomputers are desirable.
- Must be able to use conventional electronic trouble-shooting instruments such as multimeters, function generators, oscillators, and oscilloscopes. Should be able to use conventional machine shop equipment such as drill presses, grinders, belt sanders, brakes, and standard hand tools.
- Must possess or be able to acquire knowledge of the techniques, theories, and characteristics of materials, drafting, and fabrication techniques in conjunction with chemistry, anatomy, physiology, optics, mechanics, and hospital procedures.
- Clinical engineering certification or professional engineering registration is required.

74.4.1 Major Functions of a Clinical Engineering Department

It should be clear by the preceding job description that clinical engineers are first and foremost engineering professionals. However, as a result of the wide-ranging scope of interrelationships within the medical setting, the duties and responsibilities of clinical engineering directors are extremely diversified. Yet a common thread is provided by the very nature of the technology they manage. Directors of clinical engineering departments are usually involved in the following core functions:

Technology Management: Developing, implementing, and directing equipment management programs. Specific tasks include accepting and installing new equipment, establishing preventive maintenance and repair programs, and managing the inventory of medical instrumentation. Issues such as cost-effective use and quality assurance are integral parts of any *technology management* program. The director advises the hospital administrator of the budgetary, personnel, space, and test equipment requirements necessary to support this equipment management program.

Risk management: Evaluating and taking appropriate action on incidents attributed to equipment malfunctions or misuse. For example, the clinical engineering director is responsible for summarizing the technological significance of each incident and documenting the findings of the investigation. He or she then submits a report to the appropriate hospital authority and, according to the Safe Medical Devices Act of 1990, to the device manufacturer, the Food and Drug Administration (FDA), or both.

Technology assessment: Evaluating and selecting new equipment. The director must be proactive in the evaluation of new requests for capital equipment expenditures, providing hospital administrators and clinical staff with an in-depth appraisal of the benefits/advantages of candidate equipment. Furthermore, the process of **technology assessment** for all equipment used in the hospital should be an ongoing activity.

Facilities design and project management: Assisting in the design of new or renovated clinical facilities that house specific medical technologies. This includes operating rooms, imaging facilities, and radiology treatment centers.

Training: Establish and deliver instructional modules for clinical engineering staff as well as clinical staff on the operation of medical equipment.

In the future, it is anticipated that clinical engineering departments will provide assistance in the application and management of many other technologies that support patient care, including computer support (which includes the development of virtual instrumentation), telecommunications, and facilities operations.

Defining Terms

JCAHO, Joint Commission on the Accreditation of Healthcare Organizations: The accrediting body responsible for checking compliance of a hospital with approved rules and regulations regarding the delivery of health care.

Technology assessment: Involves an evaluation of the safety, efficiency, and cost-effectiveness, as well as consideration of the social, legal, and ethical effects, of medical technology.

References

Bronzino J.D. 1992. *Management of Medical Technology: A Primer for Clinical Engineers*. Boston, Butter-worth.

ICC. 1991. International Certification Commission's Definition of a Clinical Engineer, International Certification Commission Fact Sheet. Arlington, VA, ICC.

Issakov A., Mallouppas A., and McKie J. 1990. Manpower development for a healthcare technical service. Report of the World Health Organization, WHO/SHS/NHP/90.4.

Further Reading

1. Journals: *Journal of Clinical Engineering and Journal of Medical Engineering and Physics, Biomedical Instrumentation and Technology.*

75

Management and Assessment of Medical Technology

Yadin David
Texas Children's Hospital

Thomas M. Judd
Kaiser Permanente

As medical technology continues to evolve, so does its impact on patient outcome, hospital operations, and financial efficiency. The ability to plan for this evolution and its subsequent implications has become a major challenge in most decisions of health care organizations and their related industries. Therefore, there is a need to adequately plan for and apply those management tools which optimize the deployment of medical technology and the facilities that house it. Successful management of the technology and facilities will ensure a good match between the needs and the capabilities of staff and technology, respectively. While different types and sizes of hospitals will consider various strategies of actions, they all share the need to manage efficient utilization of their limited resources and its monitoring. Technology is one of these resources, and while it is frequently cited as the culprit behind cost increases, the well-managed technology program contribute to a significant containment of the cost of providing quality patient care. Clinical engineer's skills and expertise are needed to facilitate the adoption of an objective methodology for implantation of a program that will match the hospital's needs and operational conditions. Whereas both the knowledge and practice patterns of management in general are well organized in today's literature, the management of the health care delivery system and that of medical technology in the clinical environment has not yet reached that same high level. However, as we begin to understand the relationship between the methods and information that guide the decision-making processes regarding the management of medical technology that are being deployed in this highly complex environment, the role of the qualified clinical engineer becomes more valuable. This is achieved by reformulating the technology management process, which starts with the strategic planning process, continues with the **technology assessment** process, leads to the equipment planning and procurement processes, and finally ends with the assets management process. Definition of terms used in this chapter are provided at the end of the chapter.

75.1 The Health Care Delivery System

Societal demands on the health care delivery system revolve around cost, technology, and expectations. To respond effectively, the delivery system must identify its goals, select and define its priorities, and then wisely allocate its limited resources. For most organizations, this means that they must acquire only appropriate technologies and manage what they have already more effectively. To improve performance and reduce costs, the delivery system must recognize and respond to the key dynamics in which it operates, must shape and mold its planing efforts around several existing health care trends and directions, and must respond proactively and positively to the pressures of its environment. These issues and the technology manager's response are outlined here (1) technology's positive impact on care quality and effectiveness, (2) an unacceptable rise in national spending for health care services, (3) a changing mix of how Americans are insured for health care, (4) increases in health insurance premiums for which **appropriate technology** application is a strong limiting factor, (5) a changing mix of health care services and settings in which care is delivered, and (6) growing pressures related to technology for hospital capital spending and budgets.

75.1.1 Major Health Care Trends and Directions

The major trends and directions in health care include (1) changing location and design of treatment areas, (2) evolving benefits, coverages, and choices, (3) extreme pressures to manage costs, (4) treating of more acutely ill older patients and the prematurely born, (5) changing job structures and demand for skilled labor, (6) the need to maintain a strong cash flow to support construction, equipment, and information system developments, (7) increased competition on all sides, (8) requirement for information systems that effectively integrate clinical and business issues, (9) changing reimbursement policies that reduce new purchases and lead to the expectation for extended equipment life cycles, (10) internal **technology planning and management programs** to guide decision making, (11) technology planning teams to coordinate adsorption of new and replacement technologies, as well as to suggest delivery system changes, and (12) equipment maintenance costs that are emerging as a significant expense item under great administrative scrutiny.

75.1.2 System Pressures

System pressures include (1) society's expectations — highest quality care at the lowest reasonable price, where quality is a function of personnel, facilities, technology, and clinical procedures offered; (2) economic conditions — driven often by reimbursement criteria; (3) legal-pressures — resulting primarily from malpractice issues and dealing with rule-intensive "government" clients; (4) regulatory — multistate delivery systems with increased management complexity, or heavily regulated medical device industries facing free-market competition, or hospitals facing the Safe Medical Devices Act reporting requirements and credentialling requirements; (5) ethics — deciding who gets care and when; and (6) technology pressures — organizations having enough capabilities to meet community needs and to compete successfully in their marketplaces.

75.1.3 The Technology Manager's Responsibility

Technology managers should (1) become deeply involved and committed to technology planning and management programs in their system, often involving the need for greater personal responsibilities and expanded credentials, (2) understand how the factors above impact their organization and how technology can be used to improve outcomes, reduce costs, and improve quality of life for patients, (3) educate other health care professionals about how to demonstrate the value of individual technologies through involving financial, engineering, **quality of care**, and management perspective, and (4) assemble a team of caregivers with sufficient broad clinical expertise and administrators with planning and financial expertise to contribute their knowledge to the assessment process [1].

75.2 Strategic Technology Planning

75.2.1 Strategic Planning Process

Leading health care organizations have begun to combine **strategic technology planning** with other technology management activities in program that effectively integrate new technologies with their existingly technology base. This has resulted in high-quality care at a reasonable cost. Among those who have been its leading catalysts, ECRI (formerly the Emergency Care Research Institute) is known for articulating this program [2] and encouraging its proliferation initially among regional health care systems and now for single or multihospital systems as well [3]. Key components of the program include clinical strategic-planning, technology strategic planning, technology assessment, interaction with capital budgeting, acquisition and deployment, resource (or equipment assets) management, and monitoring and evaluation. A proper technology strategic plan is derived from and supports as well-defined clinical strategic plan [4].

75.2.2 Clinical and Technology Strategic Plan

Usually considered long-range and continually evolving, a clinical strategic plan is updated annually. For a given year, the program begins when key hospital participants, through the strategic planning process, assess what clinical services the hospital should be offering in its referral area. They take into account health care trends, demographic and market share data, and space and facilities plans. They analyze their facility's strengths and weaknesses, goals and objectives, competition, and existing technology base. The outcome of this process is a clinical strategic plan that establishes the organization's vision for the year and referral area needs and the hospital's objectives in meeting them.

It is not possible to adequately complete a clinical strategic plan without engaging in the process of strategic technology planning. A key role for technology managers is to assist their organizations throughout the combined clinical and technology strategic planning processes by matching available technical capabilities, both existing and new, with clinical requirements. To accomplish this, technology managers must understand why their institution's values and mission are set as they are, pursue their institution's strategic plans through that knowledge, and plan in a way that effectively allocates limited resources. Although a technology manager may not be assigned to develop an institution's overall strategic plan, he or she must understand and believe it in order to offer good input for hospital management. In providing this input, a technology manager should determine a plan for evaluating the present state of the hospital's technological deployment, assist in providing a review of emerging technological innovations and their possible impact on the hospital, articulate justifications and provisions for adoption of new technologies or enhancement of existing ones, visit research and laboratories and exhibit areas at major medical and scientific meetings to view new technologies, and be familiar with the institution and its equipment users' abilities to assimilate new technology.

The past decade has shown a trend toward increased legislation in support of more federal regulations in health care. These and other pressures will require that additional or replacement medical technology be well anticipated and justified. As a rationale for technology adoption, the Texas Children's Hospital focuses on the issues of clinical necessity, management support, and market preference. Addressing the issue of clinical necessity, the hospital considers the technology's comparison against medical standard of care, its impact on the level of care and quality of life, its improvement on intervention's accuracy and **safety**, its impact on the rate of recovery, the needs or desires of the community, and the change in service volume or focus. On the issue of management support, the hospital estimates if the technology will create a more effective care plan and decision-making process, improve operational efficiency in the current service programs, decrease liability exposure, increase compliance with regulations, reduce workload and dependence on user skill level ameliorate departmental support, or enhance clinical proficiency. Weighting the issue of market preference, the hospital contemplate if it will improve access to care, increase customer convenience and satisfaction, enhance the organization's image and market share, decrease the cost of adoption and ownership, or provide a return on its investment.

75.2.3 Technology Strategic Planning Process

When the annual clinical strategic planning process has started and hospital leaders have begun to analyze or reaffirm what clinical services they want to offer to the community, the hospital can then conduct efficient technology strategic planning. Key elements of this planning involve (1) performing an initial audit of existing technologies, (2) conducting a technology assessment for new and emerging technologies for fit with current or desired clinical services, (3) planning for replacement and selection of new technologies, (4) setting priorities for technology acquisition, and (5) developing processes to implement equipment acquisition and monitor ongoing utilization. "Increasingly, hospitals are designating a senior manager (e.g., an administrator, the director of planning, the director of clinical engineering) to take the responsibility for technology assessment and planning. That person should have the primary responsibility for developing the strategic technology plan with the help of key physicians, department managers, and senior executives" [2].

Hospitals can form a medical technology advisory committee (MTAC), overseen by the designated senior manager and consisting of the types of members mentioned above, to conduct the strategic technology planning process and to annually recommend technology priorities to the hospital strategic planning committee and capital budget committee. It is especially important to involve physicians and nurses in this process.

In the initial technology audit, each major clinical service or product line must be analyzed to determine how well the existing technology base supports it. The audit can be conducted along service lines (radiology, cardiology, surgery) or technology function (e.g., imaging, therapeutic, diagnostic) by a team of designated physicians, department heads, and technology managers. The team should begin by developing a complete hospital-wide assets inventory, including the quantity and quality of equipment. The team should compare the existing technology base against known and evolving standards-of-care information, patient outcome data, and known equipment problems. Next, the team should collect and examine information on technology utilization to assess its appropriate use, the opportunities for improvement, and the risk level. After reviewing the technology users' education needs as they relate to the application and servicing of medical equipment, the team should credential users for competence in the application of new technologies. Also, the auditing team should keep up with published clinical protocols and practice guidelines using available health care **standards** directories and utilize clinical outcome data for quality-assurance and risk-management program feedback [5].

While it is not expected that every hospital has all the required expertise in-house to conduct the initial technology audit or ongoing technology assessment, the execution of this planning process is sufficiently critical for a hospital's success that outside expertise should be obtained when necessary. The audit allows for the gathering of information about the status of the existing technology base and enhances the capability of the medical technology advisory committee to assess the impact of new and emerging technologies on their major clinical services.

All the information collected from the technology audit results and technology assessments is used in developing budget strategies. Budgeting is part of strategic technology planning in that a 2- to 5-year long-range capital spending plan should be created. This is in addition to the annual capital budget preparation that takes into account 1 year at a time. The MTAC, as able and appropriate, provides key information regarding capital budget requests and makes recommendations to the capital budget committee each year. The MTAC recommends priorities for replacement as well as new and emerging technologies that over a period of several years guides that acquisition that provides the desired service developments or enhancements. Priorities are recommended on the basis of need, risk, cost (acquisition, operational and maintenance), utilization, and fit with the clinical strategic plan.

75.3 Technology Assessment

As medical technology continues to evolve, so does its impact on patient outcome, hospital operations, and financial resources. The ability to manage this evolution and its subsequent implications has become a major challenge for all health care organizations. Successful management of technology will ensure a good match between needs and capabilities and between staff and technology. To be successful, an ongoing technology assessment process must be an integral part of an ongoing technology planning and management program at the hospital, addressing the needs of the patient, the user, and the support team. This facilitates better equipment planning and utilization of the hospital's resources. The manager who is knowledgeable about his or her organization's culture, equipment users' needs, the environment within which equipment will be applied, equipment engineering, and emerging technological capabilities will be successful in proficiently implementing and managing technological changes [6].

It is in the technology assessment process that the clinical engineering/technology manager professional needs to wear two hats: that of the manager and that of the engineer. This is a unique position, requiring expertise and detailed preparation, that allows one to be a key leader and contributor to the decision-making process of the medical technology advisory committee (MTAC).

The MTAC uses an ad hoc team approach to conduct technology assessment of selected services and technologies throughout the year. The ad hoc teams may incorporate representatives of equipment users, equipment service providers, physicians, purchasing agents, reimbursement mangers, representatives of administration, and other members from the institution as applicable.

75.3.1 Prerequisites for Technology Assessment

Medical technology is a major strategic factor in positioning and creating a positive community perception of the hospital. Exciting new biomedical devices and systems are continually being introduced. And they are introduced at a time when the pressure on hospitals to contain expenditures is mounting. Therefore, forecasting the deployment of medical technology and the capacity to continually evaluate its impact on the hospital require that the hospital be willing to provide the support for such a program. (*Note*: Many organizations are aware of the principle that an in-house "champion" is needed in order to provide for the leadership that continually and objectively plans ahead. The champion and the program being "championed" may use additional in-house or independent expertise as needed. To get focused attention on the technology assessment function and this program in larger, academically affiliated and government hospitals, the position of a chief technology officer is being created.) Traditionally, executives rely on their staff to produce objective analyses of the hospital's technological needs. Without such analyses, executives may approve purchasing decisions of sophisticated biomedical equipment only to discover later that some needs or expected features were not included with this installation, that those features are not yet approved for delivery, or that the installation has not been adequately planned.

Many hospitals perform technology assessment activities to project needs for new assets and to better manage existing assets. Because the task is complex, an interdisciplinary approach and a cooperative attitude among the assessment team leadership is required. The ability to integrate information from disciplines such as clinical, technical, financial, administrative, and facility in a timely and objective manner is critical to the success of the assessment. This chapter emphasizes how technology assessment fits within a technology planning and management program and recognizes the importance of corporate skills forecasting medical equipment changes and determining the impact of changes on the hospital's market position. Within the technology planning and management program, the focus on capital assets management of medical equipment should not lead to the exclusion of accessories, supplies, and the disposables also required.

Medical equipment has a life cycle that can be identified as (1) the innovation phase, which includes the concept, basic and applies research, and development, and (2) the adoption phase, which begins with the clinical studies, through diffusion, and then widespread use. These phases are different from each other in the scope of professional skills involved, their impact on patient care, compliance with regulatory requirements, and the extent of the required operational support. In evaluating the applicability of a device or a system for use in the hospital, it is important to note in which phase of its life cycle the equipment currently resides.

75.3.2 Technology Assessment Process

More and more hospitals are faced with the difficult phenomenon of a capital equipment requests list that is much larger than the capital budget allocation. The most difficult decision, then, is the one that matches clinical needs with the financial capability. In doing so, the following questions are often raised: How do we avoid costly technology mistakes? How do we wisely target capital dollars for technology? How do we avoid medical staff conflicts as they relate to technology? How do we control equipment-related risks? and How do we maximize the useful life of the equipment or systems while minimizing the cost ownership? A hospital's clinical engineering department can assist in providing the right answers to these questions.

Technology assessment is a component of technology planning that begins with the analysis of the hospital's existing technology base. It is easy to perceive then that technology assessment, rather than an equipment comparison, is a new major function for a clinical engineering department [7]. It is important that clinical engineers be well prepared for the challenge. They must have a full understanding of the mission of their particular hospitals, a familiarity with the health care delivery system, and the cooperation of hospital administrators and the medical staff. To aid in the technology assessment process, clinical engineers need to utilize the following tools (1) access to national database services, directories, and libraries, (2) visits to scientific and clinical exhibits, (3) a network with key industry contacts, and (4) a relationship with peers throughout the country [8].

The need for clinical engineering involvement in the technology assessment process becomes evident when recently purchased equipment or its functions are underutilized, users have ongoing problems with equipment, equipment maintenance costs become excessive, the hospital is unable to comply with standards or guidelines (i.e., JCAHO requirements) for equipment management, a high percentage of equipment is awaiting repair, or training for equipment operators is inefficient due to shortage of allied health professionals. A deeper look at the symptoms behind these problems would likely reveal a lack of a central clearinghouse to collect, index, and monitor all technology-related information for future planning purposes, the absence of procedures for identifying emerging technologies for potential acquisition, the lack of a systematic plan for conducting technology assessment, resulting in an ability to maximize the benefits from deployment of available technology, the inability to benefit from the organization's own previous experience with a particular type of technology, the random replacement of medical technologies rather than a systematic plan based on a set of well-developed criteria, and the lack of integration of technology acquisition into the strategic and capital planning of the hospital.

To address these issues, efforts to develop a technology microassessment process were initiated at one leading private hospital with the following objectives (1) accumulate information on medical equipment, (2) facilitate systematic planning, (3) create an administrative structure supporting the assessment process and its methodology, (4) monitor the replacement of outdated technology, and (5) improve the capital budget process by focusing on long-term needs relative to the acquisition of medical equipment [9].

The process, in general, and the collection of up-to-date pertinent information, in particular, require the expenditure of certain resources and the active participation of designated hospital staff in networks providing technology assessment information. For example, corporate membership in organizations and societies that provide such information needs to be considered, as well as subscriptions to certain computerized database and printed sources [10].

At the example hospital, and MTAC was formed to conduct technology assessment. It was chaired by the director of clinical engineering. Other managers from equipment user departments usually serve as the MTAC's designated technical coordinators for specific task forces. Once the committee accepted a request from an individual user, it identified other users that might have an interest in that equipment or system and authorized the technical coordinator to assemble a task force consisting of users identified by the MTAC. This task force then took responsibility for the establishment of performance criteria that would be used during this particular assessment. The task force also should answer the questions of effectiveness, safety, and **cost-effectiveness** as they relate to the particular assessment. During any specific period, there may be multiple task forces, each focusing on a specific equipment investigation.

The task force technical coordinator cooperates with the material management department in conducting a market survey, in obtaining the specified equipment for evaluation purposes, and in scheduling vendor-provided in-service training. The coordinator also confers with clinical staff to determine if they have experience with the equipment and the maturity level of the equipment under assessment. After establishment of a task force, the MTACs technical coordinator is responsible for analyzing the clinical experiences associated with the use of this equipment, for setting evaluation objectives, and for devising appropriate technical tests in accord with recommendations from the task force. Only equipment that successfully passes the technical tests will proceed to a clinical trial. During the clinical trial, a task force-appointed clinical coordinator collects and reports a summary of experiences gained. The technical coordinator then combines the results from both the technical tests and the clinical trial into a summary report for MTAC review and approval. In this role, the clinical engineer/technical coordinator serves as a multidisciplinary professional, bridging the gap between the clinical and technical needs of the hospital. To complete the process, financial staff representatives review the protocol.

The technology assessment process at this example hospital begins with a department or individual filling out two forms (1) a request for review (RR) form and (2) a capital asset request (CAR) form. These forms are submitted to the hospital's product standards committee, which determines if an assessment process is to be initiated, and the priority for its completion. It also determines if a previously established standard for this equipment already exists (if the hospital is already using such a technology) — if so, an assessment is not needed.

On the RR, the originator delineates the rationale for acquiring the medical device. For example, the originator must tell how the item will improve quality of patient care, who will be its primary user, and how it will improve ease of use. On the CAR, the originator describes the item, estimates its cost, and provides purchase justification. The CAR is then routed to the capital budget office for review. During this process, the optimal financing method for acquisition is determined. If funding is secured, the CAR is routed to the material management department, where, together with the RR, it will be processed. The rationale for having the RR accompany the CAR is to ensure that financial information is included as part of the assessment process. The CAR is the tool by which the purchasing department initiates a market survey and later sends product requests for bid. Any request for evaluation that is received without a CAR or any CAR involving medical equipment that is received without a request for evaluation is returned to the originator without action. Both forms are then sent to the clinical engineering department, where a designated technical coordinator will analyze the requested technology maturity level and results of clinical experience with its use, review trends, and prioritize various manufactures' presentations for MTAC review.

Both forms must be sent to the MTAC if the item requested is not currently used by the hospital or if it does not conform to previously adopted hospital standards. The MTAC has the authority to recommend either acceptance or rejection of any request for review, based on a consensus of its members. A task force consisting of potential equipment users will determine the "must have" equipment functions, review the impact of the various equipment configurations, and plan technical and clinical evaluations.

If the request is approved by the MTAC, the requested technology or equipment will be evaluated using technical and performance standards. Upon completion of the review, a recommendation is returned to the hospital's products standard committee, which reviews the results of the technology assessment, determines whether the particular product is suitable as a hospital standard, and decides if it should be purchased. If approved, the request to purchase will be reviewed by the capital budget committee (CBC) to determine if the required expenditure meets with available financial resources and if or when it may be feasible to make the purchase. To ensure coordination of the technology assessment program, the chairman of the MTAC also serves as a permanent member of the hospital's CBC. In this way, there is a planned integration between technology assessment and budget decisions.

75.4 Equipment Assets Management

An accountable, systemic approach will ensure that cost-effective, efficacious, safe, and appropriate equipment is available to meet the demands of quality patient care. Such an approach requires that existing medical equipment resources be managed and that the resulting management strategies have measurable outputs that are monitored and evaluated. Technology managers/clinical engineers are well positioned to organize and lead this function. It is assumed that cost accounting is managed and monitored by the health care organization's financial group.

75.4.1 Equipment Management Process

Through traditional assets management strategies, medical equipment can be comprehensively managed by clinical engineering personnel. First, the management should consider a full range of strategies for equipment technical support. Plans may include use of a combination of equipment service providers such as manufacturers, third-party service groups, shared services, and hospital-based (in-house) engineers and biomedical equipment technicians (BMETs). All these service providers should be under the general responsibility of the technology manager to ensure optimal equipment performance through comprehensive and ongoing best-value equipment service. After obtaining a complete hospital medical equipment inventory (noting both original manufacturer and typical service provider), the management should conduct a thorough analysis of hospital accounts payable records for at least the past 2 years, compiling all service reports and preventative maintenance-related costs from all possible sources. The manager then should document in-house and external provider equipment service costs, extent of maintenance coverage for each inventory time, equipment-user operating schedule, quality of maintenance coverage for each item, appropriateness of the service provider, and reasonable maintenance costs. Next, he or she should establish an effective equipment technical support process. With an accurate inventory and best-value service providers identified, service agreements/contracts should be negotiated with external providers using prepared terms and conditions, including a log-in system. There should be an in-house clinical engineering staff ensuring ongoing external provider cost control utilizing several tools. By asking the right technical questions and establishing friendly relationships with staff, the manager will be able to handle service purchase orders (POs) by determining if equipment is worth repairing and obtaining exchange prices for parts. The staff should handle service reports to review them for accuracy and proper use of the log-in system. They also should match invoices with the service reports to verify opportunities and review service histories to look for symptoms such as need for user training, repeated problems, run-on calls billed months apart, or evidence of defective or worn-out equipment. The manager should take responsibility for emergency equipment rentals. Finally, the manager should develop, implement, and monitor all the service performance criteria.

To optimize technology management programs, clinical engineers should be willing to assume responsibilities for technology planning and management in all related areas. They should develop policies and procedures for their hospital's management program. With life-cycle costs determined for key high-risk or high-cost devices, they should evaluate methods to provide additional cost savings in equipment operation and maintenance. They should be involved with computer networking systems within the hospital. As computer technology applications increase, the requirements to review technology-related information in a number of hospital locations will increase. They should determine what environmental conditions and facility changes are required to accommodate new technologies or changes in standards and guidelines. Lastly, they should use documentation of equipment performance and maintenance costs along with their knowledge of current clinical practices to assist other hospital personnel in determining the best time and process for planning equipment replacement [11].

75.4.2 Technology Management Activities

A clinical engineering department, through outstanding performance in traditional equipment management, will win its hospital's support and will be asked to be involved in a full range of technology management activities. The department should start an equipment control program that encompasses routine performance testing, inspection, periodic and preventive maintenance, on-demand repair services, incidents investigation, and actions on recalls and hazards. The department should have multidisciplinary involvement in equipment acquisition and replacement decisions, development of new services, and planning of new construction and major renovations, including intensive participation by clinical engineering, materials management, and finance. The department also should initiate programs for training all users of patient care equipment, quality improvement (QI), as it relates to technology use, and technology-related **risk management** [12].

75.4.3 Case Study: A Focus on Medical Imaging

In the mid-1980s, a large private multihospital system contemplated the startup of a corporate clinical engineering program. The directors recognized that involvement in a diagnostic imaging equipment service would be key to the economic success of the program. They further recognized that maintenance cost reductions would have to be balanced with achieving equal or increased quality of care in the utilization of that equipment.

Programs startup was in the summer of 1987 in 3 hospitals that were geographically close. Within the first year, clinical engineering operations began in 11 hospitals in 3 regions over a two-state area. By the fall of 1990, the program included 7 regions and 21 hospitals in a five-state area. The regions were organized, typically, into teams including a regional manager and 10 service providers, serving 3 to 4 hospitals, whose average size was 225 beds. Although the staffs were stationed at the hospitals, some specialists traveled between sites in the region to provide equipment service. Service providers included individuals specializing in the areas of diagnostic imaging (x-ray and computed tomography [CT]), clinical laboratory, general biomedical instrumentation, and respiratory therapy.

At the end of the first 18 months, the program documented over $1 million in savings for the initial 11 hospitals, a 23% reduction from the previous annual service costs. Over 63% of these savings were attributable to "in-house" service x-ray and CT scanner equipment. The mix of equipment maintained by 11 imagining service providers — from a total staff of 30 — included approximately 75% of the radiology systems of any kind found in the hospitals and 5 models of CT scanners from the three different manufacturers.

At the end of 3 years in 1990, program-wide savings had exceeded 30% of previous costs for participating hospitals. Within the imaging areas of the hospitals, savings approached and sometimes exceed 50% of initial service costs. The 30 imaging service providers — out of a total staff of 62 — had increased their coverage of radiology equipment to over 95%, had increased involvement with CT to include nine models from five different manufacturers, and had begun in-house work in other key imaging modalities.

Tracking the financial performance of the initial 11 hospitals over the first 3 years of the program yields the following composite example: a hospital of 225 beds was found to have equipment service costs of $540,000 prior to program startup. Sixty-three percent of these initial costs (or $340,000) was for the maintenance of the hospital's x-ray and CT scanner systems. Three years later, annual service costs for this equipment were cut in half, to approximately $170,000. That represents a 31% reduction in hospital-wide costs due to the imaging service alone.

This corporate clinical engineering operation is, in effect, a large in-house program serving many hospitals that all have common ownership. The multihospital corporation has significant purchasing power in the medical device marketplace and provides central oversight of the larger capital expenditures for its hospitals. The combination of the parent organization's leverage and the program's commitment to serve only hospitals in the corporation facilitated the development of positive relationships with medical device manufacturers. Most of the manufacturers did not see the program as competition but rather as a potentially helpful ally in the future marketing and sales of their equipment and systems. What staff provided these results? All service providers were either medical imaging industry or military trained. All were experienced at troubleshooting electronic subsystems to component level, as necessary. Typically, these individuals had prior experience on the manufacture's models of equipment under their coverage. Most regional managers had prior industry, third party, or in-house imaging service management experience. Each service provider had the test equipment necessary for day-to-day duties. Each individual could expect at least 2 weeks of annual service training to keep appropriate skills current. Desired service training could be acquired in a timely manner from manufactures and third-party organizations. Spare or replacement parts inventory was minimal because of the program's ability to get parts from manufacturers and other sources either locally or shipped in overnight.

As quality indicators for the program, the management measured user satisfaction, equipment downtime, documentation of technical staff service training, types of user equipment errors and their effect on patient outcomes, and regular attention to hospital technology problems. User satisfaction surveys indicated a high degree of confidence in the program service providers by imaging department mangers. Problems relating to technical, management, communication, and financial issues did occur regularly, but the regional manager ensured that they were resolved in a timely manner. Faster response to daily imaging equipment problems, typically by on-site service providers, coupled with regular preventive maintenance (PM) according to established procedures led to reduced equipment downtime. PM and repair service histories were captured in a computer documentation system that also tracked service times, costs, and user errors and their effects. Assisting the safety committee became easier with ability to draw a wide variety of information quickly from the program's documenting system.

Early success in imaging equipment led to the opportunity to do some additional value-added projects such as the moving and reinstallation of x-ray rooms that preserved exiting assets and opened up valuable space for installation of newer equipment and upgrades of CT scanner systems. The parent organization came to realize that these technology management activities could potentially have a greater financial and quality impact on the hospital's health care delivery than equipment management. In the example of one CT upgrade (which was completed over two weekends with no downtime), there was a positive financial impact in excess of $600,000 and improved quality of care by allowing faster off-line diagnosis of patient scans. However, opportunity for this kind of contribution would never have occurred without the strong base of a successful equipment management program staffed with qualified individuals who receive ongoing training.

75.5 Equipment Acquisition and Deployment

75.5.1 Process of Acquiring Technology

Typically, medical device systems will emerge from the strategic technology planning and technology assessment processes as required and budgeted needs. At acquisition time, a needs analysis should be conducted, reaffirming clinical needs and device intended applications. The "request for review" documentation from the assessment process or capital budget request and incremental financial analysis from the planning process may provide appropriate justification information, and a capital asset request (CAR) form should be completed [13]. Materials management and clinical engineering personnel should ensure that this item is a candidate for centralized and coordinated acquisition of similar equipment with other hospital departments. Typical hospital prepurchase evaluation guidelines include an analysis of needs and development of a specification list, formation of a vendor list and requesting proposals, analyzing proposals and site planning, evaluating samples, selecting finalists, making the award, delivery and installation, and acceptance testing. Formal request for proposals (RFPs) from potential equipment vendors are required for intended acquisitions whose initial or life-cycle cost exceeds a certain threshold, that is, $100,000. Finally, the purchase takes place, wherein final equipment negotiations are conducted and purchase documents are prepared, including a purchase order.

75.5.2 Acquisition Process Strategies

The cost-of-ownership concept can be used when considering what factors to include in cost comparisons of competing medical devices. Cost of ownership encompasses all the direct and indirect expenses associated with medical equipment over its lifetime [4]. It expresses the cost factors of medical equipment for both the initial price of the equipment (which typically includes the equipment, its installation, and initial training cost) and over the long term. Long-term costs include ongoing training, equipment service, supplies, connectivity, upgrades, and other costs. Health care organizations are just beginning to account for a full range of cost-of-ownership factors in their technology assessment and acquisition processes, such as acquisition costs, operating costs, and maintenance costs (installation, supplies, downtime, training, spare parts, test equipment and tools, and depreciation). It is estimated that the purchase price represents only 20% of the life-cycle cost of ownership.

When conducting needs analysis, actual utilization information form the organization's existing same or similar devices can be very helpful. One leading private multihospital system has implemented the following approach to measuring and developing relevant management feedback concerning equipment utilization. It is conducting equipment utilization review for replacement planning, for ongoing accountability of equipment use, and to provide input before more equipment is purchased. This private system attempts to match product to its intended function and to measure daily (if necessary) the equipment's actual utilization. The tools they use include knowing their hospital's entire installed base of certain kinds of equipment, that is, imaging systems. Utilization assumptions for each hospital and its clinical procedural mix are made. Equipment functional requirements to meet the demands of the clinical procedures are also taken into account.

Life-cycle cost analysis is a tool used during technology planning, assessment, or acquisition "either to compare high-cost, alternative means for providing a service or to determine whether a single project or technology has a positive or negative economic value. The strength of the life-cycle cost analysis is that it examines the cash flow impact of an alternative over its entire life, instead of focusing solely on initial capital investments" [4].

"Life-cycle cost analysis facilitates comparisons between projects or technologies with large initial cash outlays and those with level outlays and inflows over time. It is most applicable to complex, high-cost choices among alternative technologies, new service, and different means for providing a given service. Life-cycle cost analysis is particularly useful for decisions that are too complex and ambiguous for experience and subjective judgment alone. It also helps decision makers perceive and include costs that often are hidden or ignored, and that may otherwise invalidate results" [11].

"Perhaps the most powerful life-cycle cost technique is net present value (NPV) analysis, which explicitly accounts for inflation and foregone investment opportunities by expressing future cash flows in present dollars" [11].

Examples where LCC and NPV analysis prove very helpful are in deciding whether to replace/rebuild or buy/lease medical imaging equipment. The kinds of costs captured in life-cycle cost analysis, include decision-making costs, planning agency/certificate of need costs (if applicable), financing, initial capital investment costs including facility changes, life-cycle maintenance and repairs costs, personnel costs, and other (reimbursement consequences, resale, etc.).

One of the best strategies to ensure that a desired technology is truly of value to the hospital is to conduct a careful analysis in preparation for its assimilation into hospital operations. The process of equipment prepurchase evaluation provides information that can be used to screen unacceptable performance by either the vendor or the equipment before it becomes a hospital problem.

Once the vendor has responded to informal requests or formal RFPs, the clinical engineering department should be responsible for evaluating the technical response, while the materials management department should devaluate the financial responses.

In translating clinical needs into a specification list, key features or "must have" attributes of the desired device are identified. In practice, clinical engineering and materials management should develop a "must have" list and an extras list. The extras list contains features that may tip the decision in favor of one vendor, all other factors being even. These specification lists are sent to the vendor and are effective in a self-elimination process that results in a time savings for the hospital. Once the "must have" attributes have been satisfied, the remaining candidate devices are evaluated technically, and the extras are considered. This is accomplished by assigning a weighting factor (i.e., 0 to 5) to denote the relative importance of each of the desired attributes. The relative ability of each device to meet the defined requirements is then rated [14].

One strategy that strengthens the acquisition process is the conditions-of-sale document. This multifaceted document integrates equipment specifications, performance, installation requirements, and follow-up services. The conditions-of-sale document ensures that negotiations are completed before a purchase order is delivered and each participant is in agreement about the product to be delivered. As a document of compliance, the conditions-of-sale document specifies the codes and standards having jurisdiction over that equipment. This may include provisions for future modification of the equipment, compliance with standards under development, compliance with national codes, and provision for software upgrades.

Standard purchase orders that include the conditions of sale for medical equipment are usually used to initiate the order. At the time the order is placed, clinical engineering is notified of the order. In addition to current facility conditions, the management must address installation and approval requirements, responsibilities, and timetable; payment, assignment, and cancellation; software requirements and updates; documentation; clinical and technical training; acceptance testing (hospital facility and vendor); warranty, spare parts, and service; and price protection.

All medical equipment must be inspected and tested before it is placed into service regardless of whether it is purchased, leased, rented, or borrowed by the hospital. In any hospital, clinical engineering should receive immediate notification if a very large device or system is delivered directly into another department (e.g., imaging or cardiology) for installation. Clinical engineering should be required to sign off on all purchase orders for devices after installation and validation of satisfactory operation. Ideally, the warranty period on new equipment should not begin until installation and acceptance testing are completed. It is not uncommon for a hospital to lose several months of free parts and service by the manufacturer when new equipment is, for some reason, not installed immediately after delivery.

75.5.3 Clinical Team Requirements

During the technology assessment and acquisition processes, clinical decision makers analyze the following criteria concerning proposed technology acquisitions, specifically as they relate to clinical team requirements: ability of staff to assimilate the technology, medical staff satisfaction (short term and long term), impact on staffing (numbers, functions), projected utilization, ongoing related supplies required, effect on delivery of care and outcomes (convenience, safety, or standard of care), result of what is written in the clinical practice guidelines, credentialling of staff required, clinical staff initial and ongoing training required, and the effect on existing technology in the department or on other services/departments.

Defining Terms

Appropriate technology [14]: A term used initially in developing countries, referring to selecting medical equipment that can "appropriately" satisfy the following constraints: funding shortages, insufficient numbers of trained personnel, lack of technical support, inadequate supplies of consumables/accessories, unreliable water and power utilities/supplies, and lack of operating and maintenance manuals. In the context of this chapter, appropriate technology selection must take into consideration local health needs and disease prevalence, the need for local capability of equipment maintenance, and availability of resources for ongoing operational and technical support.

Clinical engineers/biomedical engineers: As we began describing the issues with the management of medical technology, it became obvious that some of the terms are being used interchangeably in the literature. For example, the terms engineers, clinical engineers, biomedical equipment technicians, equipment managers, and health care engineers are frequently used. For clarification, in this chapter we will refer to clinical engineers and the clinical engineering department as a representative group for all these terms.

Cost-effectiveness [14]: A mixture of quantitative and qualitative considerations. It includes the health priorities of the country or region at the macro assessment level and the community needs at the institution micro assessment level. Product life-cycle cost analysis (which, in turn, includes initial purchase price, shipping, renovations, installation, supplies, associated disposables, cost per use, and similar quantitative measures) is a critical analysis measure. Life-cycle cost also takes into account staff training, ease of use, service, and many other cost factors. But experience and judgement about the relative importance of features and the ability to fulfill the intended purpose also contribute critical information to the cost-effectiveness equation.

Equipment acquisition and deployment: Medical device systems and products typically emerge from the strategic technology planning process as "required and budgeted" needs. The process that follows, which ends with equipment acceptance testing and placement into general use, is known as the equipment acquisition and deployment process.

Health care technology: Health care technology includes the devices, equipment, systems, software, supplies, pharmaceuticals, biotechnologies, and medical and surgical procedures used in the prevention, diagnosis, and treatment of disease in humans, for their rehabilitation, and for assistive purposes. In short, technology is broadly defined as encompassing virtually all the human interventions intended to cope with disease and disabilities, short of spiritual alternatives. This chapter focuses on medical equipment products (devices, systems, and software) rather than pharmaceuticals, biotechnologies, or procedures [14]. The concept of technology also encompasses the facilities that house both patients and products. Facilities cover a wide spectrum — from the modern hospital on one end to the mobile imaging trailer on the other.

Quality of care (QA) and quality of improvement (QI): Quality assurance (QA) and Quality improvement (QI) are formal sets of activities to measure the quality of care provided; these usually include a process for selecting, monitoring, and applying corrective measures. The 1994 Joint Commission on the Accreditation of Healthcare Organizations (JCAHO) standards require hospital QA, programs to focus on patient outcomes as a primary reference. JCAHO standards for plant, technology, and safety management (PTSM), in turn, require certain equipment management practices and QA or QI activities. Identified QI deficiencies may influence equipment planning, and QI audits may increase awareness of technology overuse or under utilization.

Risk management: Risk management is a program that helps the hospital avoid the possibility of risks, minimize liability exposure, and stay compliant with regulatory reporting requirements. JCAHO PTSM standards require minimum technology-based risk-management activities. These include clinical engineering's determination of technology-related incidents with follow-up steps to prevent recurrences and evaluation and documentation of the effectiveness of these steps.

Safety: Safety is the condition of being safe from danger, injury, or damage. It is judgment about the acceptability of risk in a specified situation (e.g., for a given medical problem) by a provider with specified training at a specified type of facility equipment.

Standards [14]: A wide variety of formal standards and guidelines related to health care technology now exists. Some standards apply to design, development, and manufacturing practices for devices, software, and pharmaceuticals; some are related to the construction and operation of a health care facility; some are safety and performance requirements for certain classes of technologies, such as standards related to radiation or electrical safety; and others relate to performance, or even construction specifications, for specific types of technologies. Other standards and guidelines deal with administrative, medical, and surgical procedures and the training of clinical personnel. Standards and guidelines are produced and adopted by government agencies, international organizations, and professional and specialty organizations and societies. ECRI's Healthcare Standards Directory lists over 20,000 individual standards and guidelines produced by over 600 organizations and agencies from North America alone.

Strategic technology planning: Strategic technology planning encompasses both technologies new to the hospital and replacements for existing equipment that are to be acquired over several quarters. Acquisitions can be proposed for reasons related to safety, standard-of-care issues, and age or obsolescence of existing equipment. Acquisitions also can be proposed to consolidate several service area, expand a service area to reduce cost of service, or add a new service area.

Strategic technology planning optimizes the way the hospital's capital resources contribute to its mission. It encourages choosing new technologies that are cost-effective, and it also allows the hospital to be competitive in offering state-of-the-art services. Strategic technology planning works for a single department, product line, or clinical service. It can be limited to one or several high-priority areas. It also can be used for an entire multihospital system or geographic region [2].

Technology assessment: Assessment of medical technology is any process used for examining and reporting properties of medical technology used in health care, such as safety, efficacy, feasibility, and indications for use, cost, and cost-effectiveness, as well as social, economic, and ethical consequences, whether intended or unintended [15]. A primary technology assessment is one that seeks new, previously nonexistent data through research, typically employing long-term clinical studies of the type described below. A secondary technology assessment is usually based on published data, interviews, questionnaires, and other information-gathering methods rather than original research that creates new, basic data.

In technology assessment, there are six basic objectives that the clinical engineering department should have in mind. First, there should be ongoing monitoring of developments concerning new and emerging technologies. For new technologies, there should be an assessment of the clinical efficacy, safety, and cost/benefit ratio, including their effects on established technologies. There should be an evaluation of the short- and long-term costs and benefits of alternate approaches to managing specific clinical conditions. The appropriateness of existing technologies and their clinical uses should be estimated, while outmoded technologies should be identified and eliminated from their duplicative uses. The department should rate specific technology-based interventions in terms of improved overall value (quality and outcomes) to patients, providers, and payers. Finally, the department should facilitate a continuous uniformity between needs, offerings, and capabilities [16].

The locally based (hospital or hospital group) technology assessment described in this chapter is a process of secondary assessment that attempts to judge whether a certain medical equipment/product can be assimilated into the local operational environment.

Technology diffusion [14]: The process by which a technology is spread over time in a social system. The progression of technology diffusion can be described in four stages. The emerging or applied research stage occurs around the time of initial clinical testing. In the new stage, the technology has passed the phase of clinical trials but is not yet in widespread use. During the established stage, the technology is considered by providers to be a standard approach to a particular condition and diffuses into general use. Finally, in the obsolete/outmoded stage, the technology is superseded by another and is demonstrated to be ineffective or harmful.

Technology life cycle: Technology has a life cycle — a process by which technology is created, tested, applied, and replaced or abandoned. Since the life cycle varies from basic research and innovation to obsolescence and abatement, it is critical to know the maturity of a technology prior to making decisions regarding its adoption. Technology forecast assessment of pending technological changes are the investigative tools that support systematic and rational decisions about the utilization of a given institution's technological capabilities.

Technology planning and management [16]: Technology planning and management are an accountable, systematic approach to ensuring that cost-effective, efficacious, appropriate, and safe equipment is available to meet the demands of quality patient care and allow an institution to remain competitive. Elements include in-house service management, management and analysis of equipment external service providers, involvement in the equipment acquisition process, involvement of appropriate hospital personnel in facility planning and design, involvement in reducing technology-related patient and staff incidents, training equipment users, reviewing equipment replacement needs, and ongoing assessment of emerging technologies [2].

References

[1] ECRI. Healthcare Technology Assessment Curriculum. Philadelphia, August 1992.

[2] Banata H.D. Institute of Medicine. Assessing Medical Technologies. Washington, National Academy Press, 1985.

[3] Lumsdon K. Beyond technology assessment: balancing strategy needs, strategy. *Hospitals* 15: 25, 1992.

[4] ECRI. Capital, Competition, and Constraints: Managing Healthcare in the 1990s. A Guide for Hospital Executives. Philadelphia, 1992.

[5] Berkowtiz D.A. and Solomon R.P. Providers may be missing opportunities to improve patient outcomes. *Costs, Outcomes Measure Manage* May–June: 7, 1991.

[6] ECRI. Regional Healthcare Technology Planning and Management Program. Philadelphia, 1990.

[7] Sprague G.R. Managing technology assessment and acquisition. *Health Exec.* 6: 26, 1988.

[8] David Y. Technology-related decision-making issues in hospitals. In *IEEE Engineering in Medicine and Biology Society. Proceedings of the 11th Annual International Conference*, 1989.

[9] Wagner M. Promoting hospitals high-tech equipment. *Mod. Healthcare* 46, 1989.

[10] David Y. *Medical Technology 2001. CPA Healthcare Conference*, 1992.

[11] ECRI. Special Report on Technology Management, Health Technology. Philadelphia, 1989.

[12] ECRI. Special Report on Devices and Dollars, Philadelphia, 1988.

[13] Gullikson M.L., David Y., and Brady M.H. An automated risk management tool. JCAHO, Plant, Technology and Safety Management Review, PTSM Series, no. 2, 1993.

[14] David Y., Judd T., and ECRI. Special Report on Devices and Dollars, Philadelphia, 1988. Medical Technology Management, SpaceLabs Medical, Inc., Redmond, WA, 1993.

[15] Bronzino J.D. (ed). *Management of Medical Technology: A Primer for Clinical Engineers.* Stoneham, MA, Butterworth, 1992.

[16] David Y. *Risk Measurement For Managing Medical Technology. Conference Proceedings*, PERM-IT 1997, Australia.

76

Risk Factors, Safety, and Management of Medical Equipment

Michael L. Gullikson
Texas Children's Hospital

76.1 Risk Management: A Definition

Inherent in the definition of risk management is the implication that the hospital environment cannot be made risk-free. In fact, the nature of medical equipment — to invasively or noninvasively perform diagnostic, therapeutic, corrective, or monitoring intervention on behalf of the patient — implies that risk is present. Therefore, a standard of acceptable risk must be established that defines manageable risk in a real-time economic environment.

Unfortunately, a preexistent, quantitative standard does not exist in terms of, for instance, mean time before failure (MTBF), number of repairs or repair redos per equipment item, or cost of maintenance that provides a universal yardstick for risk management of medical equipment. Sufficient clinical management of risk must be in place that can utilize safeguards, preventive maintenance, and failure analysis information to minimize the occurrence of injury or death to patient or employee or property damage. Therefore, a process must be put in place that will permit analysis of information and modification of the preceding factors to continuously move the medical equipment program to a more stable level of manageable risk.

Risk factors that require management can be illustrated by the example of the "double-edge" sword concept of technology (see Figure 76.1). The front edge of the sword represents the cutting edge of technology and its beneficial characteristics: increased quality, greater availability of technology, timeliness of test results and treatment, and so on. The back edge of the sword represents those liabilities which must be addressed to effectively manage risk: the hidden costs discussed in the next paragraph, our dependence on technology, incompatibility of equipment, and so on [1].

For example, the purchase and installation of a major medical equipment item may only represent 20% of the lifetime cost of the equipment [2]. If the operational budget of a nursing floor does not include

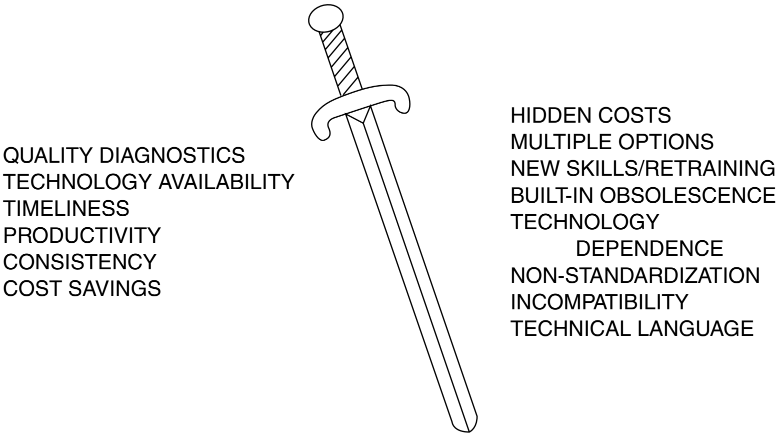


FIGURE 76.1 Double-edged sword concept of risk management.

the other 80% of the equipment costs, the budget constraints may require cutbacks where they appear to minimally affect direct patient care. Preventive maintenance, software upgrades that address "glitches," or overhaul requirements may be seen as unaffordable luxuries. Gradual equipment deterioration without maintenance may bring the safety level below an acceptable level of manageable risk.

Since economic factors as well as those of safety must be considered, a balanced approach to risk management that incorporates all aspects of the medical equipment lifecycle must be considered.

The operational flowchart in Figure 76.2 describe the concept of medical equipment life-cycle management from the clinical engineering department viewpoint. The flowchart includes planning, evaluation, and initial purchase documentation requirements. The condition of sale, for example, ensures that technical manuals, training, replacement parts, etc. are received so that all medical equipment might be fully supported in-house after the warranty period. Introduction to the preventive maintenance program, unscheduled maintenance procedures, and retirement justification must be part of the process. Institutional-wide cooperation with the life-cycle concept requires education and patience to convince health care providers of the team approach to managing medical equipment technology.

This balanced approach requires communication and comprehensive planning by a health care team responsible for evaluation of new and shared technology within the organization. A medical technology evaluation committee (see Figure 76.3), composed of representatives from administration, medical staff, nursing, safety department, biomedical engineering, and various services, can be an effective platform for the integration of technology and health care. Risk containment is practiced as the committee reviews not only the benefits of new technology but also the technical and clinical liabilities and provides a 6-month followup study to measure the effectiveness of the selection process. The history of risk management in medical equipment management provides helpful insight into its current status and future direction.

76.2 Risk Management: Historical Perspective

Historically, risk management of medical equipment was the responsibility of the clinical engineer (Figure 76.4). The engineer selected medical equipment based on individual clinical department consultations and established preventive maintenance (PM) programs based on manufacturer's recommendation and clinical experience. The clinical engineer reviewed the documentation and "spot-checked" equipment used in the hospital. The clinical engineer met with biomedical supervisors and technicians to discuss PM completion and to resolve repair problems. The clinical engineer then attempted to analyze failure information to avoid repeat failure.

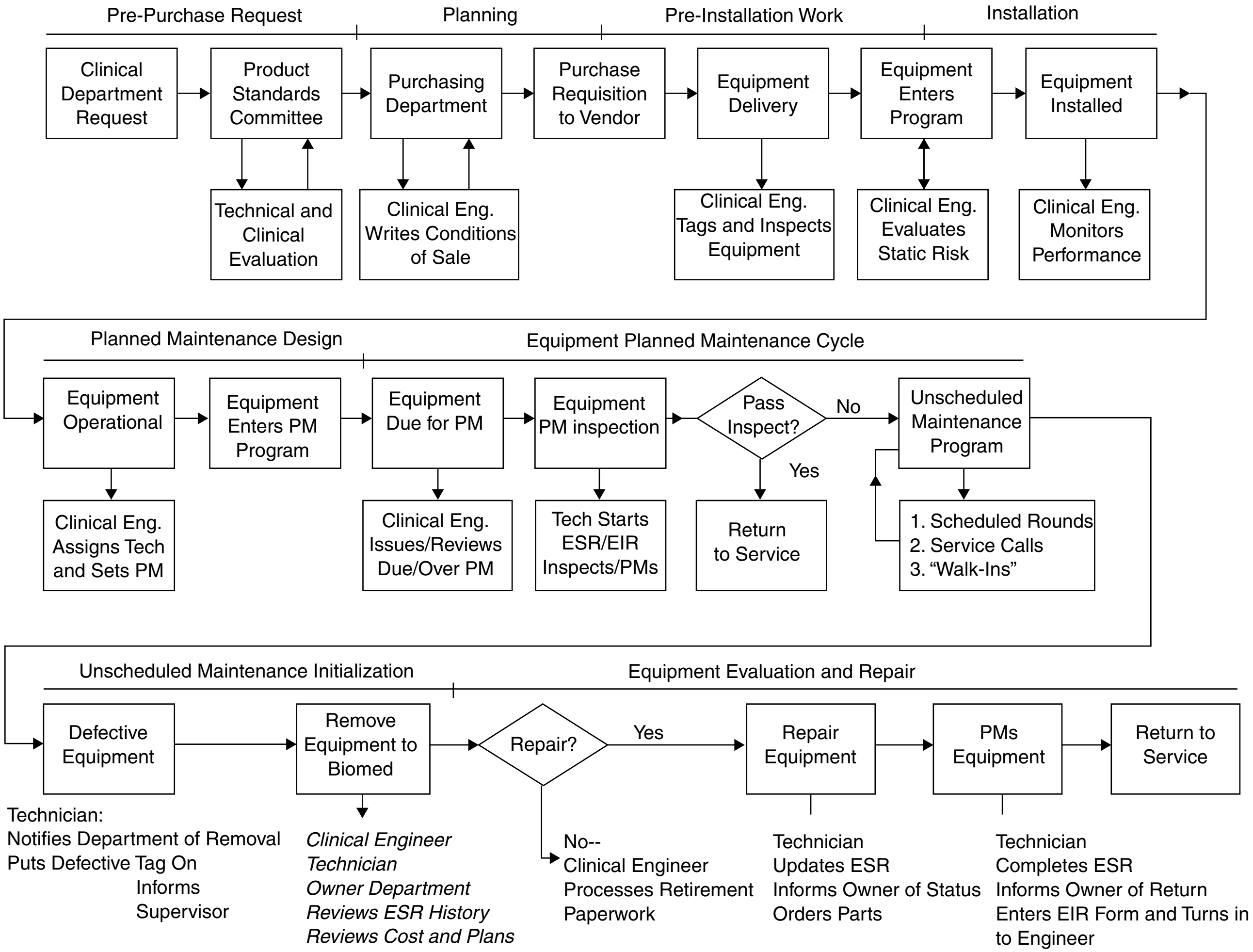


FIGURE 76.2 Biomedical engineering equipment management system (BEEMS).

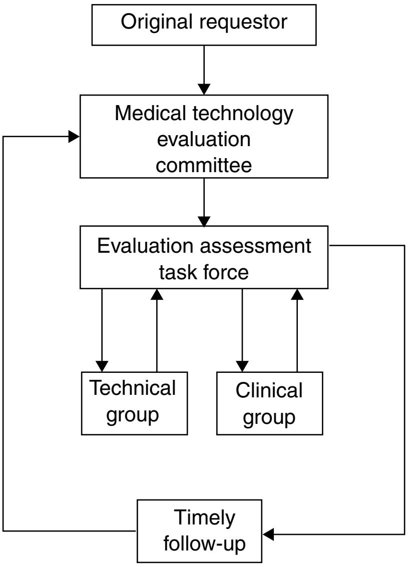

Original requestor
Medical technology evaluation committee
Evaluation assessment task force
Technical group
Clinical group
Timely follow-up

FIGURE 76.3 Medical technology evaluation committee.

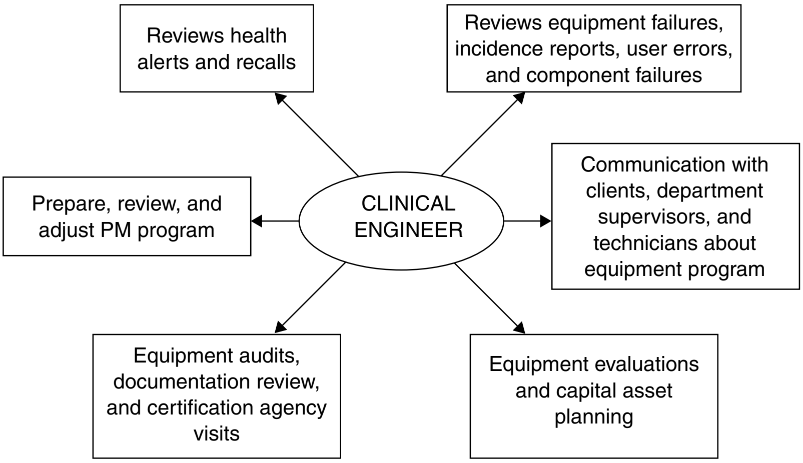

Reviews health alerts and recalls
Reviews equipment failures, incidence reports, user errors, and component failures
Prepare, review, and adjust PM program
CLINICAL ENGINEER
Communication with clients, department supervisors, and technicians about equipment program
Equipment audits, documentation review, and certification agency visits
Equipment evaluations and capital asset planning

FIGURE 76.4 Operational flowchart.

However, greater public awareness of safety issues, increasing equipment density at the bed-side, more sophisticated software-driven medical equipment, and financial considerations have made it more difficult for the clinical engineer to singularly handle risk issues. In addition, the synergistic interactions of various medical systems operating in proximity to one another have added another dimension to the risk formula. It is not only necessary for health care institutions to manage risk using a team approach, but it is also becoming apparent that the clinical engineer requires more technology-intensive tools to effectively contribute to the team effort [3].

76.3 Risk Management: Strategies

Reactive risk management is an outgrowth of the historical attitude in medical equipment management that risk is an anomaly that surfaces in the form of a failure. If the failure is analyzed and proper operational

100	Medical Equipment Operator Error
101	Medical Equipment Failure
102	Medical Equipment Physical Damage
103	Reported Patient Injury
104	Reported Employee Injury
105	Medical Equipment Failed PM
108	Medical Equipment MBA

FIGURE 76.5 Failure codes.

procedures, user in-services, and increased maintenance are supplied, the problem will disappear and personnel can return to their normal work. When the next failure occurs, the algorithm is repeated. If the same equipment fails, the algorithm is applied more intensely. This is a useful but not comprehensive component of risk management in the hospital. In fact, the traditional methods of predicting the reliability of electronic equipment from field failure data have not been very effective [4]. The health care environment, as previously mentioned, inherently contains risk that must be maintained at a manageable level. A reactive tool cannot provide direction to a risk-management program, but it can provide feedback as to its efficiency.

The engine of the reactive risk-management tool is a set of failure codes (see Figure 76.5) that flag certain anomalous conditions in the medical equipment management program. If operator training needs are able to be identified, then codes 100, 102, 104, and 108 (MBA equipment returned within 9 days for a subsequent repair) may be useful. If technician difficulties in handling equipment problems are of concern, then 108 may be of interest. The key is to develop failure codes not in an attempt to define all possible anomaly modalities but for those which can clearly be defined and provide unambiguous direction for the correction process. Also, the failure codes should be linked to equipment type, manufacturer/model, technician service group, hospital, and clinical department. Again, when the data are analyzed, will the result be provided to an administrator, engineer, clinical departmental director, or safety department? This should determine the format in which the failure codes are presented.

A report intended for the clinical engineer might be formatted as in Figure 76.6. It would consist of two parts, sorted by equipment type and clinical department (not shown). The engineer's report shows the failure code activity for various types of equipment and the distribution of those failure codes in clinical departments.

Additionally, fast data-analysis techniques introduced by NASA permit the survey of large quantities of information in a three-dimensional display [5] (Figure 76.7). This approach permits viewing time-variable changes from month to month and failure concentration in specific departments and equipment types.

The importance of the format for failure modality presentation is critical to its usefulness and acceptance by health care professionals. For instance, a safety director requests the clinical engineer to provide a list of equipment that, having failed, could have potentially harmed a patient or employee. The safety director is asking the clinical engineer for a clinical judgment based on clinical as well as technical factors. This is beyond the scope of responsibility and expertise of the clinical engineer. However, the request can be addressed indirectly. The safety director's request can be addressed in two steps first, providing a list of high-risk equipment (assessed when the medical equipment is entered into the equipment inventory) and, second, a clinical judgment based on equipment failure mode, patient condition, and so on. The flowchart in Figure 76.8 provides the safety director with useful information but does not require the clinical engineer to make an unqualified clinical judgment. If the "failed PM" failure code were selected from the list of high-risk medical equipment requiring repair, the failure would be identified by the technician during routine preventive maintenance and most likely the clinician still would find the equipment clinically efficacious. This condition is a "high risk, soft failure" or a high-risk equipment item whose failure is least likely to cause injury. If the "failed PM" code were not used, the clinician would question the clinical

Source	Failed PM	Fall in Items	% Fail	Reported Fail–OK	Physical Damage	Patient Injury	Employee Injury	Back Again	Equip. Fail	Equip. Count	% Fail
10514	PULMONARY INSTR TEXAS CHILDREN'S HOSP										
1 NON-TAGGED EQUIPMENT	0		0.00	1	5				14		0.00
1280 THERMOMETER, ELECTRONIC	2		0.00						1	99	1.01
1292 RADIANT WARMER, INFANT	1		0.00						3	63	4.76
1306 INCUBATOR, NEONATAL	0	2	0.00		7				9	56	16.07
1307 INCUBATOR, TRANSPORT, NEONATAL	9	3	33.33		1			1 *	4	9	44.44
1320 PHOTOTHERAPHY UNIT, NEONATAL	4		0.00						2	28	7.14
1321 INFUSION PUMP	35		0.00	15	9			3 *	34	514	6.61
1357 SUCTION, VAC POWERED, BODY FLUID	0		0.00						4	358	1.12
1384 EXAMINATION LIGHT, AC-POWERED	11		0.00						1	47	2.13
1447 CARDIAC MONITOR W/ RATE ALARM	0		0.00	1	1			1 *	1		0.00
1567 SURGICAL NERVE STIMULATOR /LOC	0		0.00						2	15	13.33
1624 OTOSCOPE	73		0.00						1	101	0.99
1675 OXYGEN GAS ANALYZER	0		0.00						3	44	6.82
1681 SPIROMETER DIAGNOSTIC	0		0.00						1	8	12.50
1703 AIRWAY PRESSURE MONITOR	1		0.00					1 *	7	9	77.78
1735 BREATHING GAS MIXER	13		0.00					1 *	4	38	10.53
1749 HYPO/HYPERTHERMIA DEVICE	1		0.00						1	3	33.33
1762 NEBULIZER	25		0.00						1	56	1.79
1787 VENTILATOR CONTINUOUS	96	7	7.29		1			1 *	11	99	11.11
1788 VENTILATOR NONCONTINUOUS	1		0.00		1				3	27	11.11
2014 HEMODIALYSIS SYSTEM ACCESSORIE	0		0.00						3	1	300.00
2051 PERITONEAL DIALYSIS SYS & ACC	4		0.00						1	5	20.00
2484 SPECTROPHOTOMETER, MASS	0		0.00						2	3	66.67
2695 POWERED SUCTION PUMP	16		0.00		1			1 *	1	32	3.13
5028 PH METER	0		0.00		1				1	4	25.00
5035 COMPUTER & PERIPHERALS	0		0.00						3	18	16.67
5081 OXYGEN MONITOR	0		0.00	3	1				15	102	14.71
5082 RESPIRATION ANALYZER	1		0.00						1	5	20.00
5097 EXAM TABLE	75		0.00						1	86	1.16
5113 PRINTER	2		0.00						1	12	8.33
5126 ADDRESSOGRAPH	2		0.00						4	1	400.00
9102 STADIOMETER	0		0.00						1	8	12.50
17211 ANESTHESIA MONITOR	23		0.00						2	23	8.70
90063 POWER SUPPLY, PORTABLE	20		0.00						1	25	4.00
Total for TEXAS CHILDREN'S HOSP	415	12		20	28			9	144	1899	

FIGURE 76.6 Engineer's failure analysis report.

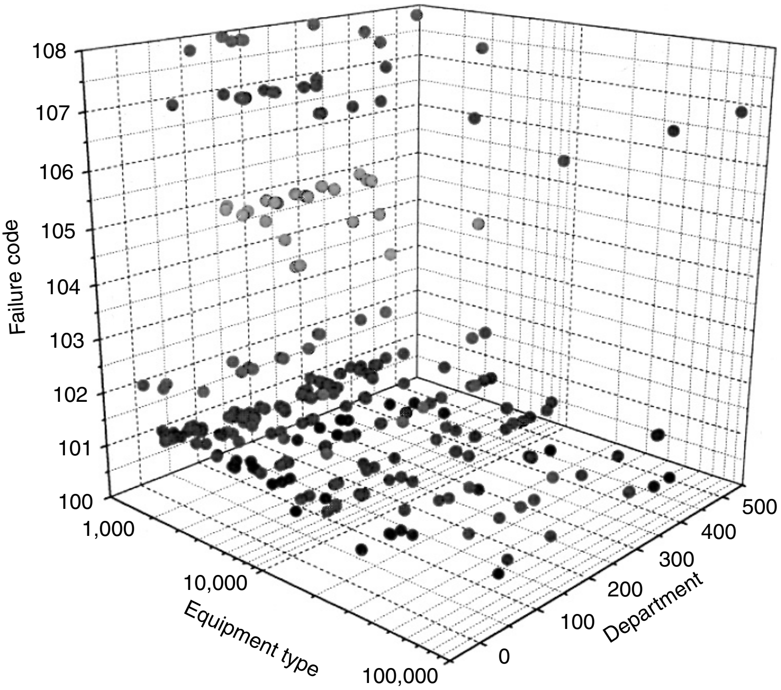


FIGURE 76.7 Failure code analysis using a 3D display.

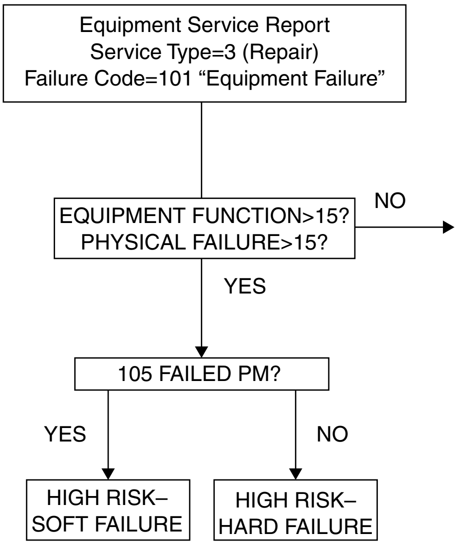


FIGURE 76.8 High-risk medical equipment failures.

efficacy of the medical equipment item, and the greater potential for injury would be identified by "high risk, hard failure." Monitoring the distribution of high-risk equipment in these two categories assists the safety director in managing risk.

Obviously, a more forward-looking tool is needed to take advantage of the failure codes and the plethora of equipment information available in a clinical engineering department. This proactive tool should use failure codes, historical information, the "expert" knowledge of the clinical engineer, and the baseline of an established "manageable risk" environment (perhaps not optimal but stable).

The overall components and process flow for a proactive risk-management tool [6] are presented in Figure 76.9. It consists of a two-component static risk factor, a two-component dynamic risk factor, and two "shaping" or feedback loops.

The static risk factor classifies new equipment by a generic equipment type: defibrilator, electrocardiograph, pulse oximeter, etc. When equipment is introduced into the equipment database, it is assigned to two different static risk (Figure 76.10) categories [7]. The first is the equipment function that defines the application and environment in which the equipment item will operate. The degree of interaction with the patient is also taken into account. For example, a therapeutic device would have a higher risk assignment than a monitoring or diagnostic device. The second component of the static risk factor is the physical risk category. It defines the worst-cases scenario in the event of equipment malfunction. The correlation between equipment function and physical risk on many items might make the two categories appear redundant. However, there are sufficient equipment types where there is not the case. A scale of 1–25 is assigned to each risk category. The larger number is assigned to devices demonstrating greater risk because of their function or the consequences of device failure. The 1–25 scale is an arbitrary assignment, since a validated scale of risk factors for medical equipment, as previously described, is nonexistent. The risk points assigned to the equipment from these two categories are algebraically summed and designated the static risk factor. This value remains with the equipment type and the individual items within that equipment type permanently. Only if the equipment is used in a clinically variant way or relocated to a functionally different environment would this assignment be reviewed and changed.

The dynamic component (Figure 76.11) of the risk-management tool consists or two parts. The first is a maintenance requirement category that is divided into 25 equally spaced divisions, ranked by least (1) to greatest (25) average manhours per device per year. These divisions are scaled by the maintenance hours for the equipment type requiring the greatest amount of maintenance attention. The amount of nonplanned (repair) manhours from the previous 12 months of service reports is totaled for each equipment type. Since this is maintenance work on failed equipment items, it correlates with the risk associated with that equipment type.

If the maintenance hours of an equipment type are observed to change to the point of placing it in a different maintenance category, a flag notifies the clinical engineer to review the equipment-type category. The engineer may increase the PM schedule to compensate for the higher unplanned maintenance hours. If the engineer believes the system "overacted," a "no" decision adjusts a scaling factor by a −5%. Progressively, the algorithm is "shaped" for the equipment maintenance program in that particular institution. However, to ensure that critical changes in the average manhours per device for each equipment type is not missed during the shaping period, the system is initialized. This is accomplished by increasing the average manhours per device for each equipment type to within 5% of the next higher maintenance requirement division. Thus the system is sensitized to variations in maintenance requirements.

The baseline is now established for evaluating individual device risk. Variations in the maintenance requirement hours for any particular equipment type will, for the most part, only occur over a substantial period of time. For this reason, the maintenance requirement category is designated a "slow" dynamic risk element.

The second dynamic element assigns weighted risk points to individual equipment items for each unique risk occurrence. An occurrence is defined as when the device:

- Exceeds the American Hospital Association Useful Life Table for Medical Equipment or exceeds the historical MTBF for that manufacturer and model
- Injures a patient or employee
- Functionally fails or fails to pass a PM inspection
- Is returned for repair or returned for rerepair within 9 days of a previous repair occurrence
- Misses a planned maintenance inspection

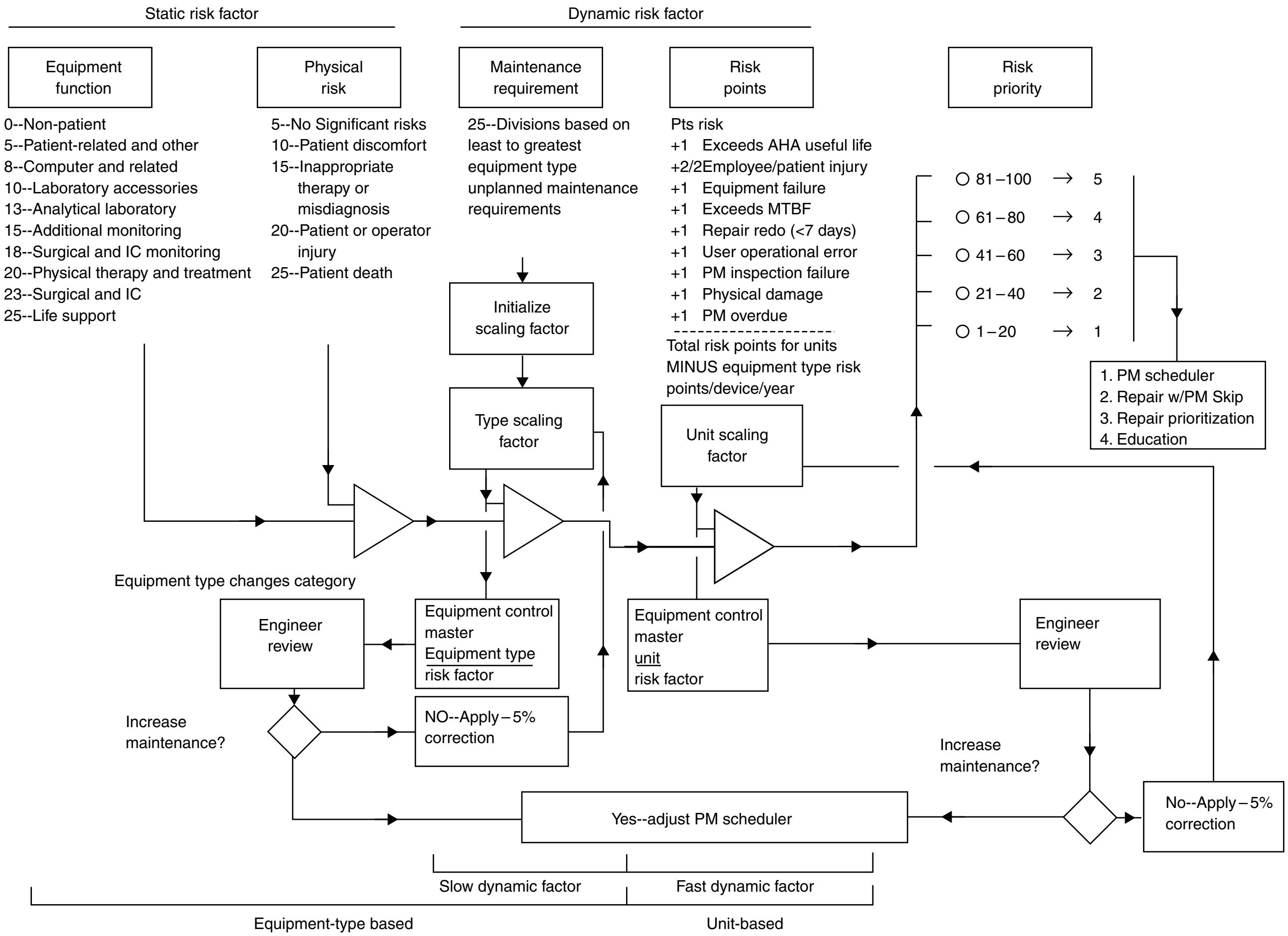


FIGURE 76.9 Biomedical engineering risk-management tool.

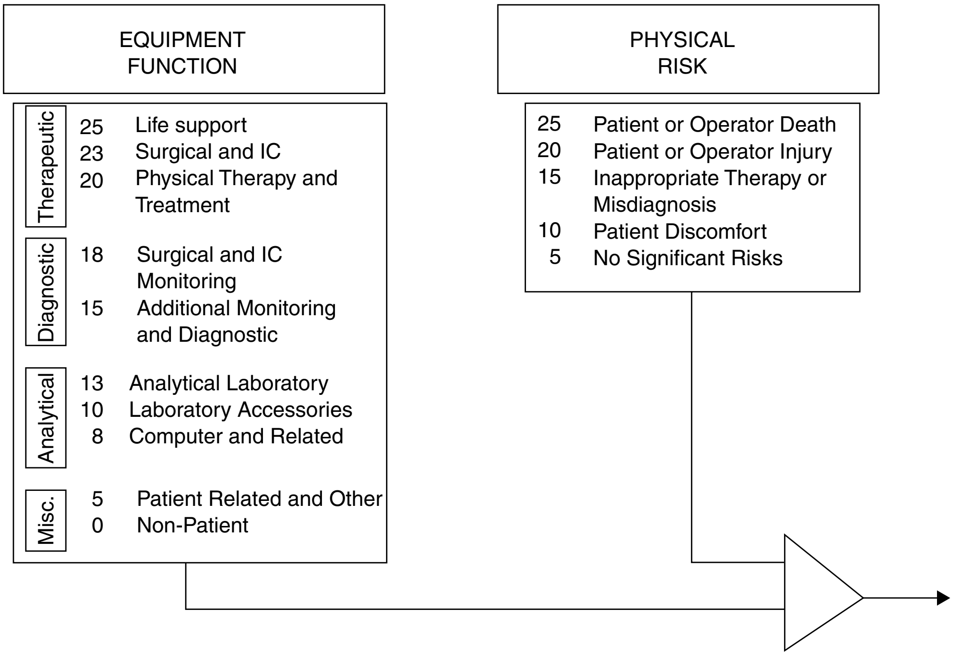

EQUIPMENT FUNCTION
Therapeutic
25 Life support
23 Surgical and IC
20 Physical Therapy and Treatment
Diagnostic
18 Surgical and IC Monitoring
15 Additional Monitoring and Diagnostic
Analytical
13 Analytical Laboratory
10 Laboratory Accessories
8 Computer and Related
Misc.
5 Patient Related and Other
0 Non-Patient
PHYSICAL RISK
25 Patient or Operator Death
20 Patient or Operator Injury
15 Inappropriate Therapy or Misdiagnosis
10 Patient Discomfort
5 No Significant Risks

FIGURE 76.10 Static risk components.

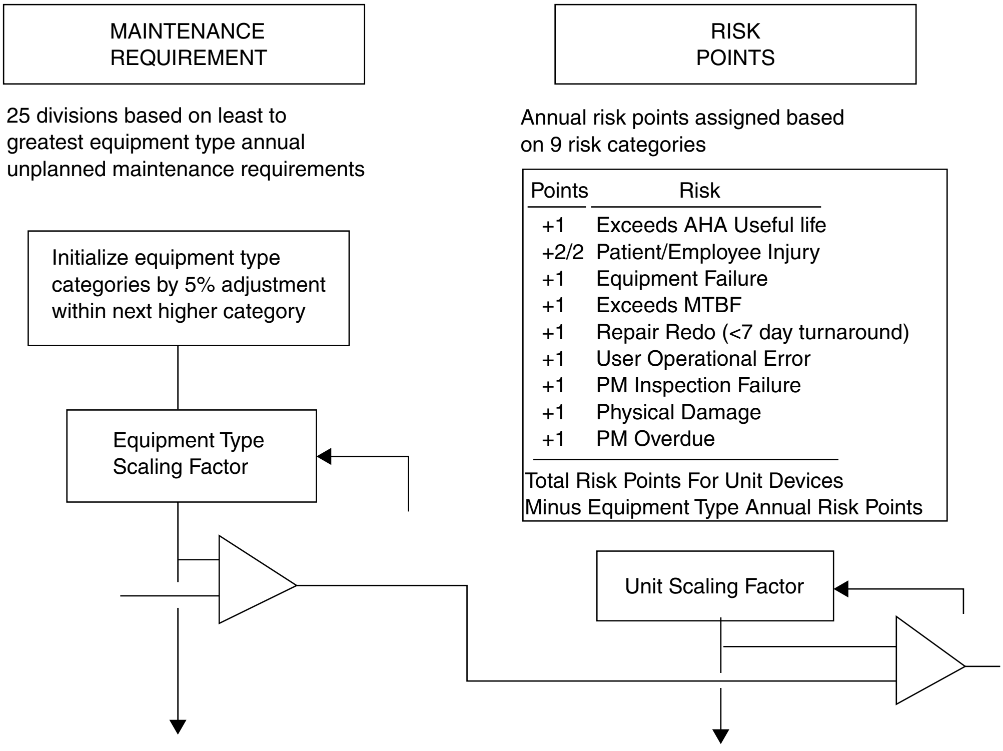

MAINTENANCE REQUIREMENT
25 divisions based on least to greatest equipment type annual unplanned maintenance requirements
Initialize equipment type categories by 5% adjustment within next higher category
Equipment Type Scaling Factor
RISK POINTS
Annual risk points assigned based on 9 risk categories
Points Risk
+1 Exceeds AHA Useful life
+2/2 Patient/Employee Injury
+1 Equipment Failure
+1 Exceeds MTBF
+1 Repair Redo (<7 day turnaround)
+1 User Operational Error
+1 PM Inspection Failure
+1 Physical Damage
+1 PM Overdue
Total Risk Points For Unit Devices
Minus Equipment Type Annual Risk Points
Unit Scaling Factor

FIGURE 76.11 Dynamic risk components.

- Is subjected to physical damage
- Was reported to have failed but the problem was determined to be a user operational error

Their risk occurrences include the failure codes previously described. Although many other risk occurrences could be defined, these nine occurrences have been historically effective in managing equipment risk. The risk points for each piece or equipment are algebraically summed over the previous year. Since the yearly total is a moving window, the risk points will not continue to accumulate but will reflect a recent historical average risk. The risk points for each equipment type are also calculated. This provides a baseline to measure the relative risk of devices within an equipment type. The average risk points for the equipment type are subtracted from those for each piece of equipment within the equipment type. If the device has a negative risk point value, the device's risk is less than the average device in the equipment type. If positive, then the device has higher risk than the average device. This positive or negative factor is algebraically summed to the risk values from the equipment function, physical risk, and maintenance requirements. The annual risk points for an individual piece of equipment might change quickly over several months. For this reason, this is the "fast" component of the dynamic risk factor.

The concept of risk has now been quantified in term of equipment function, physical risk, maintenance requirements, and risk occurrences. The total risk count for each device then places it in one of five risk priority groups that are based on the sum of risk points. These groups are then applied in various ways to determine repair triage, PM triage, educational and in-service requirements and test equipment/parts, etc. in the equipment management program.

Correlation between the placement of individual devices in each risk priority group and the levels of planned maintenance previously assigned by the clinical engineer have shown that the proactive risk-management tool calculates a similar maintenance schedule as manually planned by the clinical engineer. In other words, the proactive risk-management tool algorithm places equipment items in a risk priority group commensurate with the greater or lesser maintenance as currently applied in the equipment maintenance program.

As previously mentioned, the four categories and the 1 to 25 risk levels within each category are arbitrary because a "gold standard" for risk management is nonexistent. Therefore, the clinical engineer is given input into the dynamic components making up the risk factor to "shape the system" based on the equipment's maintenance history and the clinical engineer's experience. Since the idea of a safe medical equipment program involves "judgment about the acceptability of risk in a specified situation" [8], this experience is a necessary component of the risk-assessment tool for a specific health care setting.

In the same manner, the system tracks the unit device's assigned risk priority group. If the risk points for a device change sufficiently to place it in a different group, it is flagged for review. Again, the clinical engineer reviews the particular equipment item and decides if corrective action is prudent. Otherwise, the system reduces the scaling factor by 5%. Over a period of time, the system will be "formed" to what is acceptable risk and what deserves closer scrutiny.

76.4 Risk Management: Application

The information can be made available to the clinical engineer in the form of a risk assessment report (see Figure 76.12). The report lists individual devices by property tag number (equipment control number), manufacturer, model, and equipment type. Assigned values for equipment function and physical risk are constant for each equipment type. The maintenance sensitizing factor enables the clinical engineer to control the algorithm's response to the maintenance level of an entire equipment type. These factors combine to produce the slow risk factor (equipment function + physical risk + maintenance requirements). The unit risk points are multiplied for the unit scaling factor, which allows the clinical engineer to control the algorithm's response to static and dynamic risk components on individual pieces of equipment. This number is then added to the slow risk factor to determine the risk factor for each item. The last two columns are the risk priority that the automated system has assigned and the PM level set by the clinical

Equip. Control Number	Manuf.	Model	Equipment Type	Equip. Func.	Phys. Risk	Maint. Avg. Requir	Hours	Maint. Sensitiz. Factor	Slow Risk Factor	Unit Risk Points	Unit Scaling Factor	Risk Factor	Risk Priority	Equip. Type Priority
Manager:		PHYSIOLOGICAL GROUP												
17407	322	4000A	NIBP SYSTEM	18	15	1	1.66	1.78	38	6.62	1.00	41	3	2
17412	322	4000A	NIBP SYSTEM	18	15	1	1.66	1.78	38	6.62	1.00	41	3	2
17424	322	4000A	NIBP SYSTEM	18	15	1	1.66	1.78	38	6.62	1.00	41	3	2
17431	322	4000A	NIBP SYSTEM	18	15	1	1.66	1.78	38	6.62	1.00	41	3	2
15609	65	BW5	BLOOD & PLMA WARMING DEVICE	5	5	1	2.51	1.17	14	10.64	1.00	22	2	1
3538	167	7370000	HR/RESP MONITOR	18	15	1	0.10	29.47	35	8.69	1.00	43	3	2
3543	167	7370000	HR/RESP MONITOR	18	15	1	0.10	29.47	35	7.69	1.00	42	3	2
15315	167	7370000	HR/RESP MONITOR	18	15	1	0.10	29.47	35	7.69	1.00	42	3	2
17761	167	7370000	HR/RESP MONITOR	18	15	1	0.10	29.47	35	6.69	1.00	41	3	2
18382	574	N100C	PULSE OXIMETER	18	15	1	0.70	4.21	35	7.54	1.00	42	3	2
180476	574	N100C	PULSE OXIMETER	18	15	1	0.70	4.21	35	7.54	1.00	42	3	2
16685	167	7275217	2 CHAN CHART REC	18	15	1	0.42	7.02	37	6.83	1.00	41	3	2

FIGURE 76.12 Engineer's risk-assessment report.

engineer. This report provides the clinical engineer with information about medical equipment that reflects a higher than normal risk factor for the equipment type to which it belongs.

The proactive risk management tool can be used to individually schedule medical equipment devices for PM based on risk assessment. For example, why should newer patient monitors be maintained at the same maintenance level as older units if the risk can be demonstrated to be less? The tool is used as well to prioritize the planned maintenance program. For instance, assume a PM cycle every 17 weeks is started on January 1 for a duration of 1 week. Equipment not currently available for PM can be inspected at a later time as a function of the risk priority group for that device. In other words, an equipment item with a risk priority of 2, which is moderately low, would not be overdue for 2/5 of the time between the current and the next PM start date or until the thirteenth week after the start of a PM cycle of 17 weeks. The technicians can complete more critical overdue equipment first and move on to less critical equipment later.

Additionally, since PM is performed with every equipment repair, is it always necessary to perform the following planned PM? Assume for a moment that unscheduled maintenance was performed 10 weeks into the 17 weeks between the two PM periods discussed above. IF the equipment has a higher risk priority of the three, four, or five, the equipment is PMed as scheduled in April. However, if a lower equipment risk priority of one or two is indicated, the planned maintenance is skipped in April and resumed in July. The intent of this application is to reduce maintenance costs, preserve departmental resources, and minimize the war and tear on equipment during testing.

Historically, equipment awaiting service has been placed in the equipment holding area and inspected on a first in, first out (FIFO) basis when a technician is available. A client's request to expedite the equipment repair was the singular reason for changing the work priority schedule. The proactive risk-management tool can prioritize the equipment awaiting repair, putting the critical equipment back into service more quickly, subject to the clinical engineer's review.

76.5 Case Studies

Several examples are presented of the proactive risk-assessment tool used to evaluate the performance of medical equipment within a program.

The ventilators in Figure 76.13 show a decreasing unit risk factor for higher equipment tag numbers. Since devices are put into service with ascending tag numbers and these devices are known to have

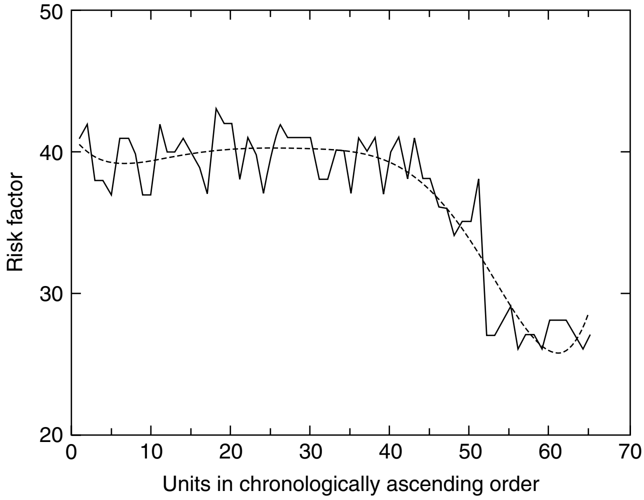


FIGURE 76.13 Ventilator with time-dependent risk characteristics.

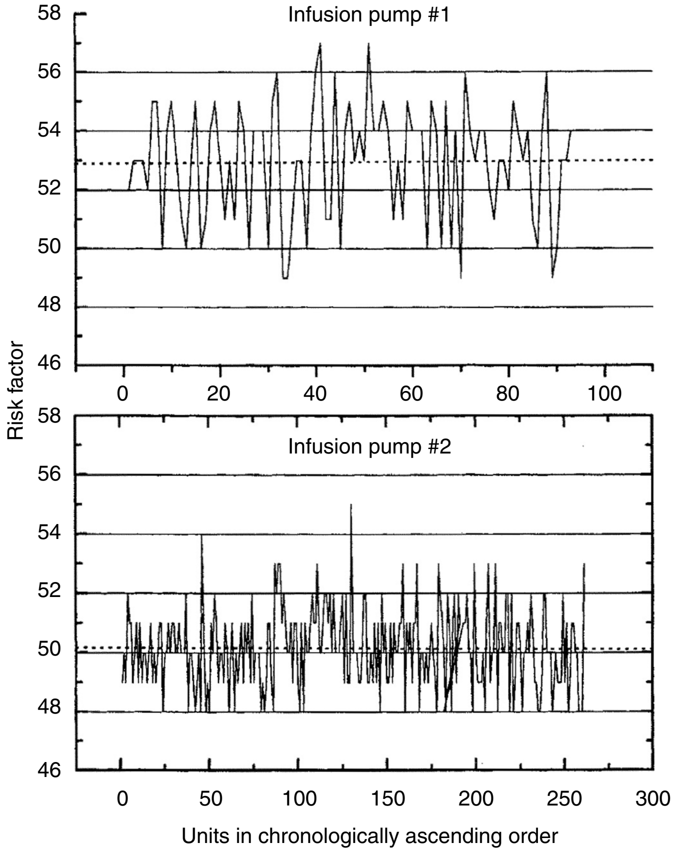


FIGURE 76.14 Time-independent risk characteristics infusion pump #1.

been purchased over a period of time, the x-axis represents a chronological progression. The ventilator risk factor is decreasing for newer units and could be attributable to better maintenance technique or manufacturer design improvements. This device is said to have a time-dependent risk factor.

A final illustration uses two generations of infusion pumps from the same manufacturer. Figure 76.14 shows the older vintage pump as Infusion Pump 1 and the newer version as Infusion Pump 2. A linear regression line for the first pump establishes the average risk factor as 53 with a standard deviation of 2.02 for the 93 pumps in the analysis. The second pump, a newer version of the first, had an average risk factor of 50 with a standard deviation of 1.38 for 261 pumps. Both pumps have relatively time-independent risk factors. The proactive risk-management tool reveals that this particular brand of infusion pump in the present maintenance program is stable over time and the newer pump has reduced risk and variability of risk between individual units. Again, this could be attributable to tighter manufacturing control or improvements in the maintenance program.

76.6 Conclusions

In summary, superior risk assessment within a medical equipment management program requires better communication, teamwork, and information analysis and distribution among all health care providers. Individually, the clinical engineer cannot provide all the necessary components for managing risk in the health care environment. Using historical information to only address equipment-related problems, after an incident, is not sufficient. The use of a proactive risk-management tool is necessary.

The clinical engineer can use this tool to deploy technical resources in a cost-effective manner. In addition to the direct economic benefits, safety is enhanced as problem equipment is identified and monitored more frequently. The integration of a proactive risk-assessment tool into the equipment management program can more accurately bring to focus technical resources in the health care environment.

References

[1] Gullikson M.L. (1994). Biotechnology Procurement and Maintenance II: Technology Risk Management. *Proceedings of the 3rd Annual International Pediatric Colloquium*, Houston, TX.

[2] David Y. (1992). Medical Technology 2001. *Health Care Conference*, Texas Society of Certified Public Accountants, San Antonio, TX.

[3] Gullikson M.L. (1993). An Automated Risk Management Tool. Plant, Technology, and Safety Management Series, Joint Commission on the Accreditation of Healthcare Facilities (JCAHO) Monograph 2.

[4] Pecht M.L. and Nash F.R. (1994). Predicting the reliability of electronic equipment. *Proceedings of the IEEE* 82: 990.

[5] Gullikson M.L. and David Y. (1993). Risk-Based Equipment Management Systems. *Proceedings of the 9th National Conference on Medical Technology Management*, American Society for Hospital Engineering of the American Hospital Association (AHA), New York, Orleans, LA.

[6] Gullikson M.L. (1992). Biomedical Equipment Maintenance System. *Proceedings of the 27th Annual Meeting and Exposition, Hospital and Medical Industry Computerized Maintenance Systems*, Association for the Advancement of Medical Instrumentation (AAMI), Anaheim, CA.

[7] Fennigkoh L. (1989). Clinical Equipment Management. Plant, Technology, and Safety Management Monograph 2.

[8] David Y. and Judd T. (1993). *Medical Technology Management*, Biophysical Measurement Series. Spacelabs Medical, Inc., Redmond, WA.

77

Clinical Engineering Program Indicators

Dennis D. Autio
Robert L. Morris
Dybonics, Inc.

The role, organization, and structure of clinical engineering departments in the modern health care environment continue to evolve. During the past 10 years, the rate of change has increased considerably faster than mere evolution due to fundamental changes in the management and organization of health care. Rapid, significant changes in the health care sector are occurring in the United States and in nearly every country. The underlying drive is primarily economic, the recognition that resources are finite.

Indicators are essential for survival of organizations and are absolutely necessary for effective management of change. Clinical engineering departments are not exceptions to this rule. In the past, most clinical engineering departments were task-driven and their existence justified by the tasks performed. Perhaps the most significant change occurring in clinical engineering practice today is the philosophical shift to a more business-oriented, cost-justified, bottom-line-focused approach than has been generally the case in the past.

Changes in the health care delivery system will dictate that clinical engineering departments justify their performance and existence on the same basis as any business, the performance of specific functions at a high quality level and at a competitive cost. Clinical engineering management philosophy must change from a purely task-driven methodology to one that includes the economics of department performance. Indicators need to be developed to measure this performance. Indicator data will need to be collected and analyzed. The data and indicators must be objective and defensible. If it cannot be measured, it cannot be managed effectively.

Indicators are used to measure performance and function in three major areas. Indicators should be used as internal measurements and monitors of the performance provided by individuals, teams, and the department. These essentially measure what was done and how it was done. Indicators are essential during quality improvement and are used to monitor and improve a process. A third important type of program indicator is the benchmark. It is common knowledge that successful businesses will continue to use benchmarks, even though differing terminology will be used. A business cannot improve its competitive position unless it knows where it stands compared with similar organizations and businesses.

Different indicators may be necessary depending on the end purpose. Some indicators may be able to measure internal operations, quality improvement, and external benchmarks. Others will have a more restricted application.

It is important to realize that a single indicator is insufficient to provide the information on which to base significant decisions. Multiple indicators are necessary to provide cross-checks and verification. An example might be to look at the profit margin of a business. Even if the profit margin per sale is 100%, the business will not be successful if there are few sales. Looking at single indicators of gross or net profit will correct this deficiency but will not provide sufficient information to point the way to improvements in operations.

77.1 Department Philosophy

A successful clinical engineering department must define its mission, vision, and goals as related to the facility's mission. A mission statement should identify what the clinical engineering department does for the organization. A vision statement identifies the direction and future of the department and must incorporate the vision statement of the parent organization. Department goals are then identified and developed to meet the mission and vision statements for the department and organization. The goals must be specific and attainable. The identification of goals will be incomplete without at least implied indicators. Integrating the mission statement, vision statement, and goals together provides the clinical engineering department management with the direction and constraints necessary for effective planning.

Clinical engineering managers must carefully integrate mission, vision, and goal information to develop a strategic plan for the department. Since available means are always limited, the manager must carefully assess the needs of the organization and available resources, set appropriate priorities, and determine available options. The scope of specific clinical engineering services to be provided can include maintenance, equipment management, and technology management activities. Once the scope of services is defined, strategies can be developed for implementation. Appropriate program indicators must then be developed to document, monitor, and manage the services to be provided. Once effective indicators are implemented, they can be used to monitor internal operations and quality-improvement processes and complete comparisons with external organizations.

77.1.1 Monitoring Internal Operations

Indicators may be used to provide an objective, accurate measurement of the different services provided in the department. These can measure specific individual, team, and departmental performance parameters. Typical indicators might include simple tallies of the quantity or level of effort for each activity, productivity (quantify/effort), percentage of time spent performing each activity, percentage of scheduled IPMs (inspection and preventive maintenance procedures) completed within the scheduled period, mean time per job by activity, repair jobs not completed within 30 days, parts order for greater than 60 days, etc.

77.1.2 Process for Quality Improvement

When program indicators are used in a quality-improvement process, an additional step is required. Expectations must be quantified in terms of the indicators used. Quantified expectations result in the establishment of a threshold value for the indicator that will precipitate further analysis of the process. Indicators combined with expectations (threshold values of the indicators) identify the opportunities for program improvement. Periodic monitoring to determine if a program indicator is below (or above, depending on whether you are measuring successes or failures) the established threshold will provide a flag to whether the process or performance is within acceptable limits. If it is outside acceptable limits for the indicator, a problem has been identified. Further analysis may be required to better define the problem. Possible program indicators for quality improvement might include the number of repairs completed within 24 or 48 h, the number of callbacks for repairs, the number of repair problems caused by user error, the percentage of hazard notifications reviewed and acted on within a given time frame, meeting time targets for generating specification, evaluation or acceptance of new equipment, etc.

An example might be a weekly status update of the percentage of scheduled IPMs completed. Assume that the department has implemented a process in which a group of scheduled IPMs must be completed within 8 weeks. The expectation is that 12% of the scheduled IPMs will be completed each week. The indicator is the percentage of IPMs completed. The threshold value of the indicator is 12% per week increase in the percentage of IPMs completed. To monitor this, the number of IPMs that were completed must be tallied, divided by the total number scheduled, and multiplied by 100 to determine the percentage completed. If the number of completed IPMs is less than projected, then further analysis would be required to identify the source of the problem and determine solutions to correct it. If the percentage of completed IPMs were equal to or greater than the threshold or target, then no action would be required.

77.1.3 External Comparisons

Much important and useful information can be obtained by carefully comparing one clinical engineering program with others. This type of comparison is highly valued by most hospital administrators. It can be helpful in determining performance relative to competitors. External indicators or benchmarks can identify specific areas of activity in need of improvement. They offer insights when consideration is being given to expanding into new areas of support. Great care must be taken when comparing services provided by clinical engineering departments located in different facilities. There are number of factors that must be included in making such comparisons; otherwise, the results can be misleading or misinterpreted. It is important that the definition of the specific indicators used be well understood, and great care must be taken to ensure that the comparison utilizes comparable information before interpreting the comparisons. Failure to understand the details and nature of the comparison and just using the numbers directly will likely result in inappropriate actions by managers and administrators. The process of analysis and explanation of differences in benchmark values between a clinical engineering department and a competitor (often referred to as gap analysis) can lead to increased insight into department operations and target areas for improvements.

Possible external indicators could be the labor cost per hour, the labor cost per repair, the total cost per repair, the cost per bed supported, the number of devices per bed supported, percentage of time devoted to repairs vs. IPMs vs. consultation, cost of support as a percentage of the acquisition value of capital inventory, etc.

77.2 Standard Database

> In God we trust . . . all others bring data!
>
> —*Florida Power and Light*

Evaluation of indicators requires the collection, storage, and analysis of data from which the indicators can be derived. A standard set of data elements must be defined. Fortunately, one only has to look at commercially available equipment management systems to determine the most common data elements used. Indeed, most of the high-end software systems have more data elements than many clinical engineering departments are willing to collect. These standard data elements must be carefully defined and understood. This is especially important if the data will later be used for comparisons with other organizations. Different departments often have different definitions for the same data element. It is crucial that the data collected be accurate and complete. The members of the clinical engineering department must be trained to properly gather, document, and enter the data into the database. It makes no conceptual difference if the database is maintained on paper or using computers. Computers and their databases are ubiquitous and so much easier to use that usually more data elements are collected when computerized systems are used. The effort required for analysis is less and the level of sophistication of the analytical tools that can be used is higher with computerized systems.

The clinical engineering department must consistently gather and enter data into the database. The database becomes the practical definition of the services and work performed by the department. This standardized database allows rapid, retrospective analysis of the data to determine specific indicators identifying problems and assist in developing solutions for implementation. A minimum database should allow the gathering and storage of the following data:

In-House Labor. This consists of three elements: the number of hours spent providing a particular service, the associated labor rate, and the identity of the individual providing the service. The labor cost is not the hourly rate the technician is paid multiplied by the number of hours spent performing the service. It should include the associated indirect costs, such as benefits, space, utilities, test equipment, and tools, along with training, administrative overhead, and many other hidden costs. A simple, straightforward approach to determine an hourly labor rate for a department is to take the total budget of the department and subtract parts' costs, service contract costs, and amounts paid to outside vendors. Divide the resulting amount by the total hours spent providing services as determined from the database. This will provide an average hourly rate for the department.

Vendor Labor. This should include hours spent and rate, travel, and zone charges and any perdiem costs associated with the vendor supplied service.

Parts. Complete information on parts is important for any retrospective study of services provided. This information is similar for both in-house and vendor-provided service. It should include the part number, a description of the part, and its cost, including any shipping.

Timeless. It is important to include a number of time stamps in the data. These should include the date the request was received, data assigned, and date completed.

Problem Identification. Both a code for rapid computer searching and classification and a free text comment identifying the nature of the problem and description of service provided are important. The number of codes should be kept to as few as possible. Detailed classification schemes usually end up with significant inaccuracies due to differing interpretations of the fine gradations in classifications.

Equipment Identification. Developing an accurate equipment history depends on reliable means of identifying the equipment. This usually includes a department- and/or facility-assigned unique identification number as well as the manufacturer, vendor, model, and serial number. Identification numbers provided by asset management are often inadequate to allow tracking of interchangeable modules or important items with a value less than a given amount. Acquisition cost is a useful data element.

Service Requester. The database should include elements allowing identification of the department, person, telephone number, cost center, and location of the service requester.

77.3 Measurement Indicators

Clinical engineering departments must gather objective, quantifiable data in order to assess ongoing performance, identify new quality-improvement opportunities, and monitor the effect of improvement action plans. Since resources are limited and everything cannot be measured, certain selection criteria must be implemented to identify the most significant opportunities for indicators. High-volume, high-risk, or problem-prone processes require frequent monitoring of indicators. A new indicator may be developed after analysis of ongoing measurements or feedback from other processes. Customer feedback and surveys often can provide information leading to the development of new indicators. Department management, in consultation with the quality-management department, typically determines what indicators will be monitored on an ongoing basis. The indicators and resulting analysis are fed back to individuals and work teams for review and improvement of their daily work activities. Teams may develop new indicators during their analysis and implementation of solutions to quality-improvement opportunities.

An indicator is an objective, quantitative measurement of an outcome or process that relates to performance quality. The event being assessed can be either desirable or undesirable. It is objective in that the same measurement can be obtained by different observers. This indicator represents quantitative, measured data that are gathered for further analysis. Indicators can assess many different aspects of quality, including accessibility, appropriateness, continuity, customer satisfaction, effectiveness, efficacy, efficiency, safety, and timeliness.

A program indicator has attributes that determine its utility as a performance measure. The reliability and variability of the indicator are distinct but related characteristics. An indicator is reliable if the same measurement can be obtained by different observers. A valid indicator is one that can identify opportunities for quality improvement. As indicators evolve, their reliability and validity should improve to the highest level possible.

An indicator can specify a part of a process to be measured or the outcome of that process. An outcome indicator assesses the results of a process. Examples include the percentage of uncompleted, scheduled IPMs, or the number of uncompleted equipment repairs not completed within 30 days. A process indicator assesses an important and discrete activity that is carried out during the process. An example would be the number of anesthesia machines in which the scheduled IPM failed or the number of equipment repairs awaiting parts that are uncompleted within 30 days.

Indicators also can be classified as sentinel event indicators and aggregate data indicators. A performance measurement of an individual event that triggers further analysis is called a sentinel-event indicator. These are often undesirable events that do not occur often. These are often related to safety issues and do not lend themselves easily to quality-improvement opportunities. An example may include equipment failures that result in a patient injury.

An aggregate data indicator is a performance measurement based on collecting data involving many events. These events occur frequently and can be presented as a continuous variable indicator or as rate-based indicators. A continuous variable indicator is a measurement where the value can fall anywhere along a continuous scale. Examples could be the number of IPMs scheduled during a particular month or the number of repair requests received during a week. A rate-based variable indicator is the value of a measurement that is expressed as a proportion or a ratio. Examples could be the percentage of IPMs completed each month or the percentage of repairs completed within one workday.

General indicators should be developed to provide a baseline monitoring of the department's performance. They also should provide a cross-check for other indicators. These indicators can be developed to respond to a perceived need within a department or to solve a specific problem.

77.4 Indicator Management Process

The process to develop, monitor, analyze, and manage indicators is shown in Figure 77.1. The different steps in this process include defining the indicator, establishing the threshold, monitoring the indicator, evaluating the indicator, identifying quality-improvement opportunities, and implementing action plans.

Define Indicator. The definition of the indicator to be monitored must be carefully developed. This process includes at least five steps. The event or outcome to be measured must be described. Define any specific terms that are used. Categorize the indicator (sentinel event or rate-based, process or outcome, desirable or undesirable). The purpose for this indicator must be defined, as well as how it is used in specifying and assessing the particular process or outcome.

Establish Threshold. A threshold is a specific data point that identifies the need for the department to respond to the indicator to determine why the threshold was reached. Sentinel-event indicator thresholds are set at zero. Rate indicator thresholds are more complex to define because they may require expert

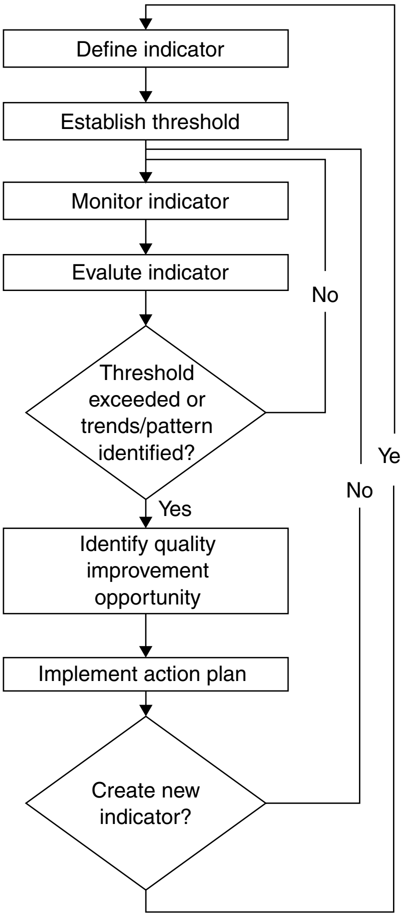


FIGURE 77.1 Indicator management process.

consensus or definition of the department's objectives. Thresholds must be identified, including the process used to set the specific level.

Monitor Indicator. Once the indicator is defined, the data-acquisition process identifies the data sources and data elements. As these data are gathered, they must be validated for accuracy and completeness. Multiple indicators can be used for data validation and cross-checking. The use of a computerized database allows rapid access to the data. A database management tool allows quick sorting and organization of the data. Once gathered, the data must be presented in a format suitable for evaluation. Graphic presentation of data allows rapid visual analysis for thresholds, trends, and patterns.

Evaluate Indicator. The evaluation process analyze and reports the information. This process includes comparing the information with established thresholds and analyzing for any trends or patterns. A trend is the general direction the indicator measurement takes over a period of time and may be desirable or undesirable. A pattern is a grouping or distribution of indicator measurements. A pattern analysis is often triggered when thresholds are crossed or trends identified. Additional indicator information is often required. If an indictor threshold has not been reached, no further action may be necessary, other than continuing to monitor this indicator. The department also may decide to improve its performance level by changing the threshold.

Factors may be present leading to variation of the indicator data. These factors may include failure of the technology to perform properly, failure of the operators to use the technology properly, and failure of the organization to provide the necessary resources to implement this technology properly. Further analysis of these factors may lead to quality-improvement activities later.

Identify Quality-Improvement Opportunity. A quality-improvement opportunity may present itself if an indicator threshold is reached, a trend is identified, or a pattern is recognized. Additional information is then needed to further define the process and improvement opportunities. The first step in the process is to identify a team. This team must be given the necessary resources to complete this project, a timetable to be followed, and an opportunity to periodically update management on the status of the project. The initial phase of the project will analyze the process and establish the scope and definition of the problem. Once the problem is defined, possible solutions can be identified and analyzed for potential implementation. A specific solution to the problem is then selected. The solution may include modifying existing indictors or thresholds to more appropriate values, modifying steps to improve existing processes, or establishing new goals for the department.

Implement Action Plan. An action plan is necessary to identify how the quality-improvement solution will be implemented. This includes defining the different tasks to be performed, the order in which they will be addressed, who will perform each task, and how this improvement will be monitored. Appropriate resources must again be identified and a timetable developed prior to implementation. Once the action plan is implemented, the indicators are monitored and evaluated to verify appropriate changes in the process. New indicators and thresholds may need to be developed to monitor the solution.

77.5 Indicator Example 1: Productivity Monitors

Define Indicator. Monitor the productivity of technical personnel, teams, and the department. Productivity is defined as the total number of documented service support hours compared with the total number of hours available. This is a desirable rate-based outcome indicator. Provide feedback to technical staff and hospital administration regarding utilization of available time for department support activities.

Establish Threshold. At least 50% of available technician time will be spent providing equipment maintenance support services (revolving equipment problems and scheduled IPMs). At least 25% of available technician time will be spent providing equipment management support services (installations, acceptance testing, incoming inspections, equipment inventory database management, hazard notification review).

Monitor Indicator. Data will be gathered every 4 weeks from the equipment work order history database. A trend analysis will be performed with data available from previously monitored 4-week intervals. These data will consist of hours worked on completed and uncompleted jobs during the past 4-week interval.

Technical staff available hours is calculated for the 4-week interval. The base time available is 160 h (40 h/week × 4 week) per individual. Add to this any overtime worked during the interval. Then subtract any holidays, sick days, and vacation days within the interval.

CJHOURS: Hours worked on completed jobs during the interval
UJHOURS: Hours worked on uncompleted jobs during the interval
AHOURS: Total hours available during the 4-week interval

$$\text{Productivity} = (\text{CJHOURS} + \text{UJHOURS})/\text{AHOURS}$$

Evaluate Indicator. The indicator will be compared with the threshold, and the information will be provided to the individual. The individual team member data can be summed for team review. The data from multiple teams can be summed and reviewed by the department. Historical indicator information will be utilized to determine trends and patterns.

Quality-Improvement Process. If the threshold is not met, a trend is identified, or a pattern is observed, a quality-improvement opportunity exists. A team could be formed to review the indicator, examine the process that the indicator measured, define the problem encountered, identify ways to solve the problem, and select a solution. An action plan will then be developed to implement this solution.

Implement Action Plan. During implementation of the action plan, appropriate indicators will be used to monitor the effectiveness of the action plan.

77.6 Indicator Example 2: Patient Monitors IPM

77.6.1 Completion Time

Define Indicator. Compare the mean to complete an IPM for different models of patient monitors. Different manufacturers of patient monitors have different IPM requirements. Identify the most timely process to support this equipment.

Establish Threshold. The difference between the mean time to complete an IPM for different models of patient monitors will not be greater than 30% of the lesser time.

Monitor Indicator. Determine the mean time to complete an IPM for each model of patient monitor. Calculate the percentage difference between the mean time for each model and the model with the least mean time.

Evaluate Indicator. The mean time to complete IPMs was compared between the patient monitors, and the maximum difference noted was 46%. A pattern also was identified in which all IPMs for that one particular monitor averaged 15 min longer than those of other vendors.

Quality-Improvement Process. A team was formed to address this problem. Analysis of individual IPM procedures revealed that manufacturer X requires the case to be removed to access internal filters. Performing an IPM for each monitor required moving and replacing 15 screws for each of the 46 monitors. The team evaluated this process and identified that 5 min could be saved from each IPM if an electric screwdriver was utilized.

Implement Action Plan. Electric screwdrivers were purchased and provided for use by the technician. The completion of one IPM cycle for the 46 monitors would pay for two electric screwdrivers and provide 4 h of productive time for additional work. Actual savings were greater because this equipment could be used in the course of daily work.

77.7 Summary

In the ever-changing world of health care, clinical engineering departments are frequently being evaluated based on their contribution to the corporate bottom line. For many departments, this will require difficult and painful changes in management philosophy. Administrators are demanding quantitative measures of performance and value. To provide the appropriate quantitative documentation required by corporate managers, a clinical engineering a manager must collect available data that are reliable and accurate. Without such data, analysis is valueless. Indicators are the first step in reducing the data to meaningful information that can be easily monitored and analyzed. The indicators can then be used to determine department performance and identify opportunities for quality improvement.

Program indicators have been used for many years. What must change for clinical engineering departments is a conscious evaluation and systematic use of indicators. One traditional indicator of clinical engineering department success is whether the department's budget is approved or not. Unfortunately, approval of the budget as an indicator, while valuable, does not address the issue of predicting long-term survival, measuring program and quality improvements, or allowing frequent evaluation and changes.

There should be monitored indicators for every significant operational aspect of the department. Common areas where program indicators can be applied include monitoring interval department activities, quality-improvement processes, and benchmarking. Initially, simple indicators should be developed. The complexity and number of indicators should change as experience and needs demand.

The use of program indicators is absolutely essential if a clinical engineering departments is to survive. Program and survival are now determined by the contribution of the department to the bottom line of the parent organization. Indicators must be developed and utilized to determine the current contribution of the clinical engineering department to the organization. Effective utilization and management of program indicators will ensure future department contributions.

References

AAMI. 1993. Management Information Report MIR 1: Design of Clinical Engineering Quality Assurance Risk Management Programs. Arlington, VA, Association for the Advancement of Medical Instrumentation.

AAMI. 1993. Management Information Report MIR 2: Guideline for Establishing and Administering Medical Instrumentation Maintenance Programs. Arlington, VA, Association for the Advancement of Medical Instrumentation.

AAMI. 1994. Management Information Report MIR 3: Computerized Maintenance Management Systems for Clinical Engineering. Arlington, VA, Association for the Advancement of Medical Instrumentation.

Bauld T.J. 1987. Productivity: standard terminology and definitions. *J. Clin. Eng.* 12: 139.

Betts W.F. 1989. Using productivity measures in clinical engineering departments. *Biomed. Instrum. Technol.* 23: 120.

Bronzino J.D. 1992. *Management of Medical Technology: A Primer for Clinical Engineers.* Stoneham, MA, Butterworth-Heinemann.

Coopers and Lybrand International, AFSM. 1994. Benchmarking Impacting the Boston Line. Fort Myers, FL, Association for Services Management International.

David Y. and Judd T.M. 1993. Risk management and quality improvement. In *Medical Technology Management*, pp. 72–75. Redmond, WA, SpaceLab Medical.

David Y. and Rohe D. 1986. Clinical engineering program productivity and measurement. *J. Clin. Eng.* 11: 435.

Downs K.J. and McKinney W.D. 1991. Clinical engineering workload analysis: a proposal for standardization. *Biomed. Instrum. Technol.* 25: 101.

Fennigkoh L. 1986. ASHE Technical Document No 055880: medical equipment maintenance performance measures. Chicago, American Society for Hospital Engineers.

Furst E. 1986. Productivity and cost-effectiveness of clinical engineering. *J. Clin. Eng.* 11: 105.

Gordon G.J. 1995. Breakthrough management — a new model for hospital technical services. Arlington, VA, Association for the Advancement of Medical Instrumentation.

Hertz E. 1990. Developing quality indicators for a clinical engineering department. In *Plant, Technology and Safety Management Series: Measuring Quality in PTSM.* Chicago, Joint Commission on Accreditation of Healthcare Organizations.

JCAHO. 1990. Primer on Indicator Development and Application, Measuring Quality in Health Care. Oakbrook, IL, Joint Commission on Accreditation of Healthcare Organizations.

JCAHO. 1994. Framework for improving performance. Oakbrook, IL, Joint Commission on Accreditation of Healthcare Organizations.

Keil O.R. 1989. The challenge of building quality into clinical engineering programs. *Biomed. Instrum. Technol.* 23: 354.

Lodge D.A. 1991. Productivity, efficiency, and effectiveness in the management of healthcare technology: an incentive pay proposal. *J. Clin. Eng.* 16: 29.

Mahachek A.R. 1987. Management and control of clinical engineering productivity: a case study. *J. Clin. Eng.* 12: 127.

Mahachek A.R. 1989. Productivity measurement. Taking the first steps. *Biomed. Instrum. Technol.* 23: 16.

Selsky D.B. et al. 1991. Biomedical equipment information management for the next generation. *Biomed. Instrum. Technol.* 25: 24.

Sherwood M.K. 1991. Quality assurance in biomedical or clinical engineering. *J. Clin. Eng.* 16: 479.

Stiefel R.H. 1991. Creating a quality measurement system for clinical engineering. *Biomed. Instrum. Technol.* 25: 17.

78

Quality of Improvement and Team Building

Joseph P. McClain
Walter Reed Army Medical Center

In today's complex health care environment, quality improvement and team building must go hand in hand. This is especially true for Clinical Engineers and Biomedical Equipment Technicians as the diversity of the field increases and technology moves so rapidly that no one can know all that needs to be known without the help of others. Therefore, it is important that we work together to ensure quality improvement. Ken Blachard, the author of the One Minute Manager series, has made the statement that "all of us are smarter than any one of us" — a synergy that evolves from working together.

Throughout this chapter we will look closely at defining quality and the methods for continuously improving quality, such as collecting data, interpreting indicators, and team building. All this will be put together, enabling us to make decisions based on scientific deciphering of indicators.

Quality is defined as conformance to customer or user requirements. If a product or service does what it is supposed to do, it is said to have high quality. If the product or service fails its mission, it is said to be low quality. Dr. W. Edward Demings, who is known to many as the "father of quality," defined it as surpassing customer needs and expectations throughout the life of the product or service.

Dr. Demings, a trained statistician by profession, formed his theories on quality during World War II while teaching industry how to use statistical methods to improve the quality of military production. After the war, he focused on meeting customer or consumer needs and acted as a consultant to Japanese organizations to change consumers' perceptions that "Made in Japan" meant junk. Dr. Demings predicted that people would be demanding Japanese products in just 5 years, if they used his methods. However, it only took 4, and the rest is history.

78.1 Deming's 14 Points

1. Create constancy of purpose toward improvement of product and service, with an aim to become competitive and to stay in business and provide jobs
2. Adopt the new philosophy. We are in a new economic age. Western management must awaken and lead for change
3. Cease dependence on inspection to achieve quality. Eliminate the needs for mass inspection by first building in quality
4. Improve constantly and forever the system of production and service to improve quality and productivity and thus constantly decrease costs
5. Institute training on the job
6. Institute leadership: The goal is to help people, machines, and gadgets to do a better job
7. Drive out fear so that everyone may work effectively for the organization
8. Break down barriers between departments
9. Eliminate slogans, exhortations, and targets for the workforce
10. Eliminate work standards (quota) on the factory floor
11. Substitute leadership: Eliminate management by objective, by numbers, and numerical goals
12. Remove barriers that rob the hourly worker of the right to pride of workmanship
13. Institute a vigorous program of education and self-improvement
14. Encourage everyone in the company to work toward accomplishing transformation. Transformation is everyone's job

78.2 Zero Defects

Another well-known quality theory, called zero defects (ZD), was established by Philip Crosby. It got results for a variety of reasons. The main reasons are as follows:

1. *A strict and specific management standard.* Management, including the supervisory staff, do not use vague phrases to explain what it wants. It made the quality standard very clear: Do it the right way from the start. As Philip Crosby said, "What standard would you set on how many babies nurses are allowed to drop?"
2. *Complete commitment of everyone.* Interestingly, Crosby denies that ZD was a motivational program. But ZD worked because everyone got deeply into the act. Everyone was encouraged to spot problems, detect errors, and prescribe ways and means for their removal. This commitment is best illustrated by the ZD pledge: "I freely pledge myself to make a constant, conscious effort to do my job right the first time, recognizing that my individual contribution is a vital part of the overall effort."
3. *Removal of actions and conditions that cause errors.* Philip Crosby claimed that at ITT, where he was vice-president for quality, 90% of all error causes could be acted on and fully removed by

first-line supervision. In other words, top management must do its part to improve conditions, but supervisors and employees should handle problems directly. Errors, malfunctions, and/or variances can best be corrected where the rubber hits the road — at the source.

78.3 TQM (Total Quality Management)

The most recent quality theory that has found fame is called TQM (Total Quality Management). It is a strategic, integrated management system for achieving customer satisfaction which involves all managers and employees and uses quantitative methods to continuously improve an organization's processes. Total Quality Management is a term coined in 1985 by the Naval Air Systems Command to describe its management approach to quality improvement. Simply put, TQM is a management approach to long-term success through customer satisfaction. TQM includes the following three principles (1) achieving customer satisfaction, (2) making continuous improvement, and (3) giving everyone responsibility. TQM includes eight practices. These practices are (1) focus on the customer, (2) effective and renewed communications, (3) reliance on standards and measures, (4) commitment to training, (5) top management support and direction, (6) employee involvement, (7) rewards and recognition, and (8) long-term commitment.

78.4 CQI (Continuous Quality Improvement)

Step 8 of the total quality management practices leads us to the quality concept coined by the Joint Commission On Accreditation of Healthcare Organizations and widely used by most health care agencies. It is called CQI (Continuous Quality Management). The principles of CQI are as follows:

Unity of Purpose

- Unity is established throughout the organization with a clear and widely understood vision
- Environment nurtures total commitment from all employees
- Rewards go beyond benefits and salaries to the belief that "We are family" and "We do excellent work"

Looking for Faults in the Systems

- Eighty percent of an organization's failures are the fault of management-controlled systems
- Workers can control fewer than 20% of the problems
- Focus on rigorous improvement of every system, and cease blaming individuals for problems (the 80/20 rule of J.M. Juran and the nineteenth-century economist Vilfredo Pareto)

Customer Focus

- Start with the customer
- The goal is to meet or exceed customer needs and give lasting value to the customer
- Positive returns will follow as customers boast of the company's quality and service

Obsession with Quality

- Everyone's job
- Quality is relentlessly pursued through products and services that delight the customer
- Efficient and effective methods of execution

Recognizing the Structure in Work

- All work has structure
- Structure may be hidden behind workflow inefficiency
- Structure can be studied, measured, analyzed, and improved

Freedom Through Control

- There is control, yet freedom exists by eliminating micromanagement
- Employees standardize processes and communicate the benefits of standardization
- Employees reduce variation in the way work is done
- Freedom comes as changes occur resulting in time to spend on developing improved processes, discovering new markets, and adding other methods to increase productivity

Continued Education and Training

- Everyone is constantly learning
- Educational opportunities are made available to employees
- Greater job mastery is gained and capabilities are broadened

Philosophical Issues on Training

- Training must stay tuned to current technology
- Funding must be made available to ensure that proper training can be attained
- Test, measurement, and diagnostic equipment germane to the mission must be procured and technicians trained on its proper use, calibration, and service
- Creativity must be used to obtain training when funding is scarce
  - Include training in equipment procurement process
  - Contact manufacturer or education facility to bring training to the institution
  - Use local facilities to acquire training, thus eliminating travel cost
  - Allow employees to attend professional seminars where a multitude of training is available

Teamwork

- Old rivalries and distrust are eliminated
- Barriers are overcome
- Teamwork, commitment to the team concept, and partnerships are the focus
- Employee empowerment is critical in the CQI philosophy and means that employees have the authority to make well-reasoned, data-based decisions. In essence, they are entrusted with the legal power to change processes through a rational, scientific approach

Continuous quality improvement is a means for tapping knowledge and creativity, applying participative problem solving, finding and eliminating problems that prevent quality, eliminating waste, instilling pride, and increasing teamwork. Further it is a means for creating an atmosphere of innovation for continued and permanent quality improvement. Continuous quality improvement as outlined by the Joint Commission on Accreditation of Healthcare Organizations is designed to improve the work processes within and across organizations.

78.5 Tools Used for Quality Improvement

The tools listed on the following pages will assist in developing quality programs, collecting data, and assessing performance indicators within the organization. These tools include several of the most frequently used and most of the seven tools of quality. The seven tools of quality are tools that help health care organizations understand their processes in order to improve them. The tools are the cause-and-effect diagram, check sheet, control chart, flowchart, histogram, Pareto chart, and scatter diagram. Additional tools shown are the Shewhart cycle (PDCA process) and the bar chart. The Clinical Engineering Manager must access the situation and determine which tool will work best for his/her situational needs.

Two of the seven tools of quality discussed above are not illustrated. These are the scatter diagram and the check sheet. The scatter diagram is a graphic technique to analyze the relationship between two variations and the check sheet is simple data-recording device. The check sheet is custom designed by the

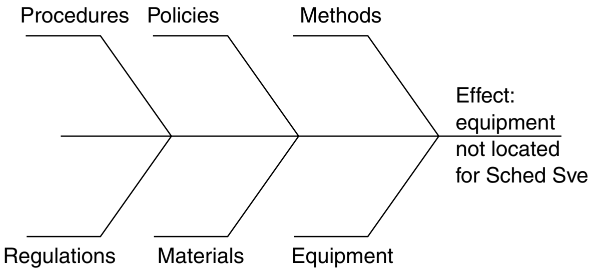


FIGURE 78.1 Cause-and-effect or Ishikawa chart.

user, which facilitates interpretation of the results. Most Biomedical Equipment Technicians use the check sheet on a daily basis when performing preventive maintenance, calibration, or electrical safety checks.

78.5.1 Cause-and-Effect or Ishikawa Chart

This is a tool for analyzing process dispersion (Figure 78.1). The process was developed by Dr. Karou Ishikawa and is also known as the fishbone diagram because the diagram resembles a fish skeleton. The diagram illustrates the main causes and subcauses leading to an effect. The cause-and-effect diagram is one of the seven tools of quality.

The following is an overview of the process:

1. Used in group problem solving as a brainstorming tool to explore and display the possible causes of a particular problem
2. The effect (problem, concern, or opportunity) that is being investigated is stated on the right side, while the contributing causes are grouped in component categories through group brainstorming on the left side
3. This is an extremely effective tool for focusing a group brainstorming session
4. Basic components include environment, methods (measurement), people, money information, materials, supplies, capital equipment, and intangibles

78.5.2 Control Chart

A control chart is a graphic representation of a characteristic of a process showing plotted values of some statistic gathered from that characteristic and one or two control limits (Figure 78.2). It has two basic uses:

1. As a judgment to determine if the process is in control
2. As an aid in achieving and maintaining statistical control

(This chart was used by Dr. W.A. Shewhart for a continuing test of statistical significance.) A control chart is a chart with a baseline, frequently in time order, on which measurement or counts are represented by points that are connected by a straight line with an upper and lower limit. The control chart is one of the seven tools of quality.

78.5.3 Flowchart

A flowchart is a pictorial representation showing all the steps of a process (Figure 78.3). Flowcharts provide excellent documentation of a program and can be a useful tool for examining how various steps in a process are related to each other. Flowcharting uses easily recognizable symbols to represent the type of processing performed. The flowchart is one of the seven tools of quality.

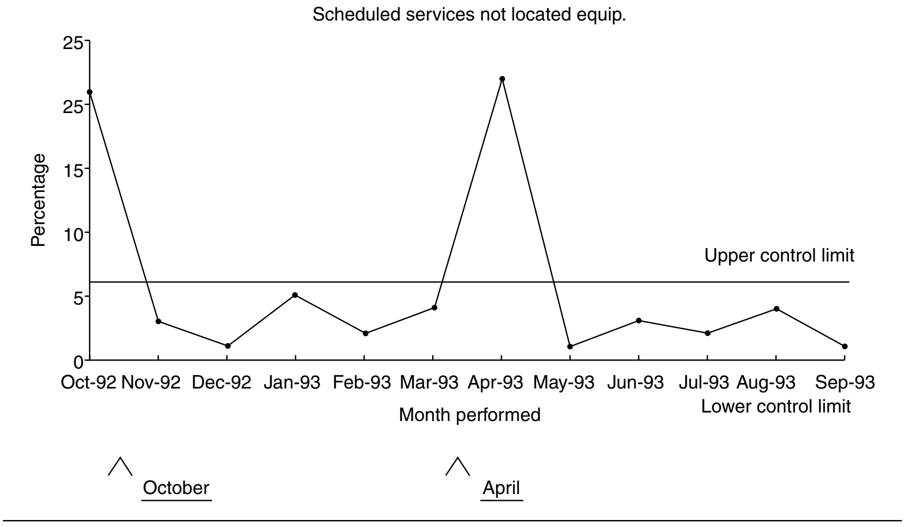


FIGURE 78.2 Control chart.

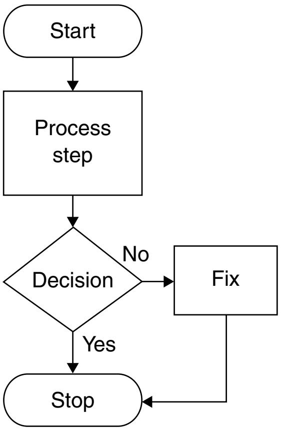


FIGURE 78.3 Flowchart.

78.5.4 Histogram

A graphic summary of variation in a set of data is a histogram (Figure 78.4). The pictorial nature of the histogram lets people see patterns that are difficult to see in a simple table of numbers. The histogram is one of the seven tools of quality.

78.5.5 Pareto Chart

A Pareto chart is a special form of vertical bar graph that helps us to determine which problems to solve and in what order (Figure 78.5). It is based on the Pareto principle, which was first developed by J.M. Juran in 1950. The principle, named after the nineteenth-century economist Vilfredo Pareto, suggests that most effects come from relatively few causes; that is, 80% of the effects come from 20% of the possible causes.

Doing a Pareto chart, based on either check sheets or other forms of data collection, helps us direct our attention and efforts to truly important problems. We will generally gain more by working on the tallest bar than tackling the smaller bars. The Pareto chart is one of the seven tools of quality.

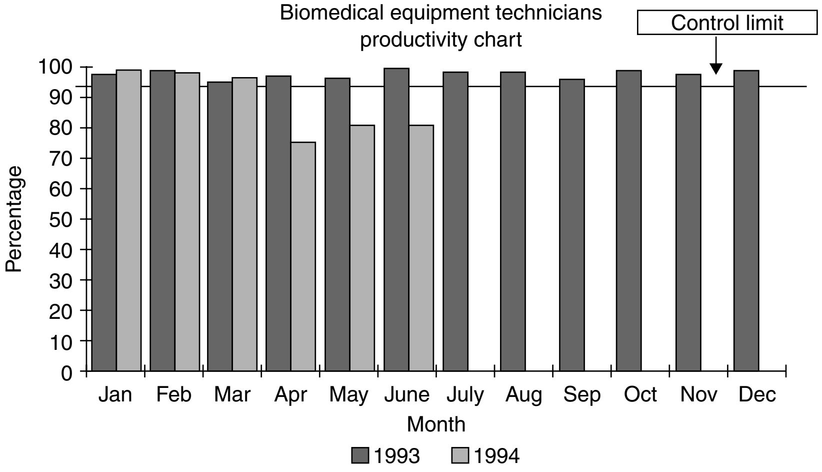


FIGURE 78.4 Histogram.

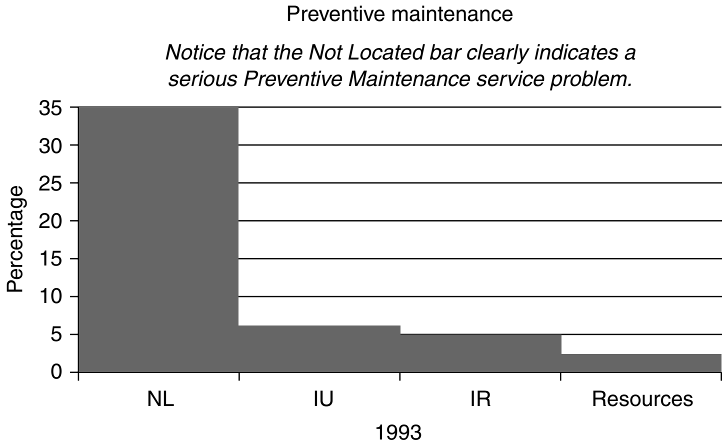


FIGURE 78.5 Pareto chart.

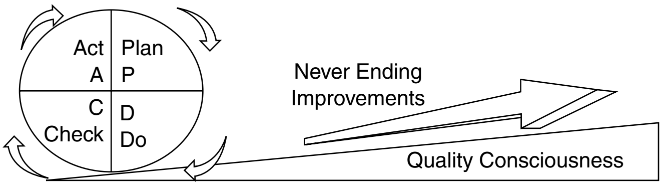


FIGURE 78.6 The Shewhart cycle.

78.5.6 The Plan-Do-Check-Act or Shewhart Cycle

This is a four-step process for quality improvement that is sometimes referred to as the Deming cycle (Figure 78.6). One of the consistent requirements of the cycle is the long-term commitment required. The Shewhart cycle or PDCA cycle is outlined here and has had overwhelming success when used properly. It is also a very handy tool to use in understanding the quality cycle process. The results of the cycle are studied to determine what was learned, what can be predicted, and appropriate changes to be implemented.

78.6 Quality Performance Indicators (QPI)

An indicator is something that suggests the existence of a fact, condition, or quality — an omen (a sign of future good or evil). It can be considered as evidence of a manifestation or symptom of an incipient failure or problem. Therefore, quality performance indicators are measurements that can be used to ensure that quality performance is continuous and will allow us to know when incipient failures are starting so that we may take corrective and preventive actions.

QPI analysis is a five-step process:

Step 1: Decide what performance we need to track
Step 2: Decide the data that need to be collected to track this performance
Step 3: Collect the data
Step 4: Establish limits, a parameter, or control points
Step 5: Utilize BME (management by exception) — where a performance exceeds the established control limits, it is indicating a quality performance failure, and corrective action must be taken to correct the problem

In the preceding section, there were several examples of QPIs. In the Pareto chart, the NL = not located, IU = in use, IR = in repair. The chart indicates that during the year 1994, 35% of the equipment could not be located to perform preventive maintenance services. This indicator tells us that we could eventually have a serious safety problem that could impact on patient care, and if not corrected, it could prevent the health care facility from meeting accreditation requirements. In the control chart example, an upper control limit of 6% "not located equipment" is established as acceptable in any one month. However, this upper control limit is exceeded during the months of April and October. This QPI could assist the clinical and Biomedical Equipment Manager in narrowing the problem down to a 2-month period. The histogram example established a lower control limit for productivity at 93%. However, productivity started to drop off in May, June, and July. This QPI tells the manager that something has happened that is jeopardizing the performance of his or her organization. Other performance indicators have been established graphically in Figure 78.7 and Figure 78.8. See if you can determine what the indicators are and what the possible cause might be. You may wish to use these tools to establish QPI tracking germane to your own organization.

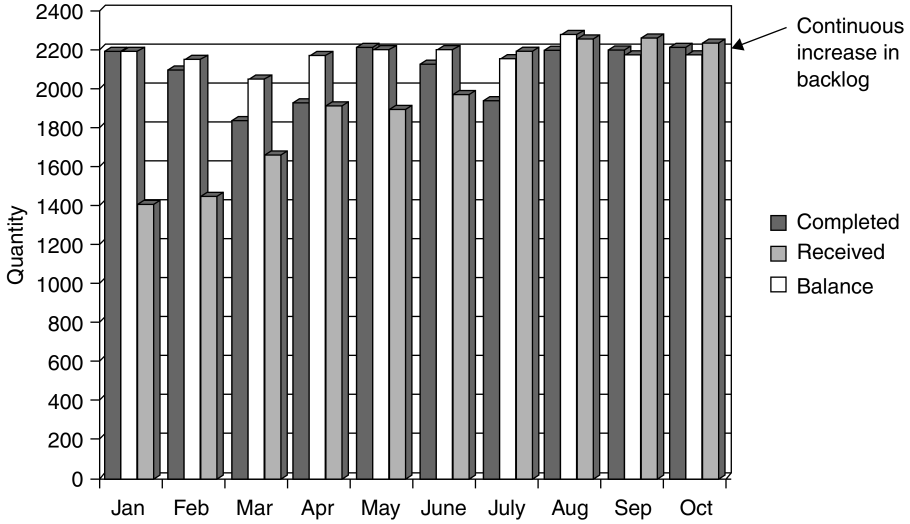


FIGURE 78.7 Sample repair service report.

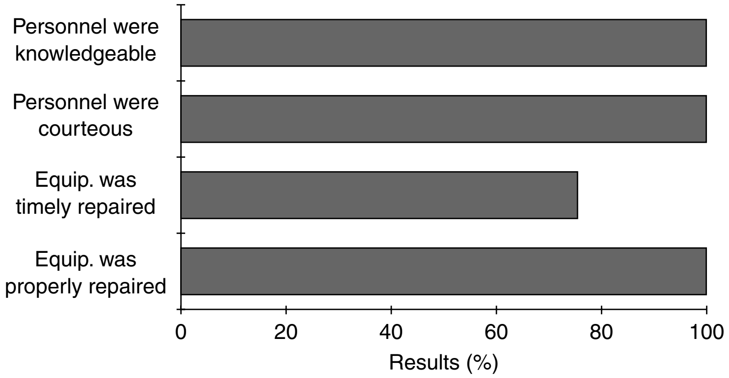


FIGURE 78.8 Customer satisfaction survey. August–September 1994.

78.7 Teams

A team is a formal group of persons organized by the company to work together to accomplish certain goals and objectives. Normally, when teams are used in quality improvement programs, they are designed to achieve the organization's vision. The organization's vision is a statement of the desired end state of the organization articulated and deployed by the executive leadership. Organizational visions are inspiring, clear, challenging, reasonable, and empowering. Effective visions honor the past, while they prepare for the future. The following are types of teams that are being used in health care facilities today. Some of the names may not be common, but their definitions are very similar if not commensurate.

78.7.1 Process Action Teams (PAT)

Process action teams are composed of those who are involved in the process being investigated. The members of a PAT are often chosen by their respective managers. The primary consideration for PAT membership is knowledge about the operations of the organization and consequently the process being studied. The main function of a PAT is the performance of an improvement project. Hence customers are often invited to participate on the team. PATs use basic statistical and other tools to analyze a process and identify potential areas for improvement. PATs report their findings to an Executive Steering Committee or some other type of quality management improving group. ("A problem well defined is half solved." John Dewey, American philosopher and educator; 1859–1952.)

78.7.2 Transition Management Team (TMT)

The transition management team (see *Harvard Business Review*, November–December 1993, pp. 109–118) is normally used for a major organizational change such as restructuring or reengineering. The TMT can be initiated due to the findings of a PAT, where it has been indicated that the process is severely broken and unsalvageable. The TMT is not a new layer of bureaucracy or a job for fading executives. The TMT oversees the large-scale corporate change effort. It makes sure that all change initiatives fit together. It is made up of 8 to 12 highly talented leaders who commit all their time making the transition a reality. The team members and what they are trying to accomplish must be accepted by the power structure of the organization. For the duration of the change process, they are the CEO's version of the National Guard. The CEO should be able to say, "I can sleep well tonight, the TMT is managing this." In setting up a TMT, organizations should adopt a fail–safe approach: Create a position to oversee the emotional and behavioral issues unless you can prove with confidence that you do not need one.

78.7.3 Quality Improvement Project Team (QIPT)

A quality improvement project team can be initiated due to the findings of a PAT, where it has been indicated that the process is broken. The main agenda of the QIPT is to improve the work process that managers have identified as important to change. The team studies this process methodically to find permanent solutions to problems. To do this, members can use many of the tools described in this chapter and in many other publications on quality and quality improvement available from schools, bookstores, and private organizations.

78.7.4 Executive Steering Committee (ESC)

This is an executive-level team composed of the Chief Executive Officer (CEO) of the organization and the executive staff that reports directly to the CEO. Whereas an organization may have numerous QMBs, PATs, and QIPTs, it has only one ESC. The ESC identifies strategic goals for organizational quality improvement efforts. It obtains information from customers to identify major product and service requirements. It is through the identification of these major requirements that quality goals for the organization are defined. Using this information, the ESC lists, prioritizes, and determines how to measure the organization's goals for quality improvement. The ESC develops the organization's improvement plan and manages the execution of that plan to ensure that improvement goals are achieved.

78.7.5 Quality Management Board (QMB)

This is a permanent cross-functional team made up of top and midlevel managers who are jointly responsible for a specific product, service, or process. The structure of the board intended to improve communication and cooperation by providing vertical and horizontal "links" throughout the organization.

78.8 Process Improvement Model

This process is following the Joint Commission on the Accreditation of Healthcare Organizations' Quality Cube, a method of assessing the quality of the organization.

Plan:

1. Identify the process to be monitored
2. Select important functions and dimensions of performance applicable to the process identified
3. Design a tool for collection of data

Measure (Under this heading, you will document how, when, and where data was collected):

1. Collect data
2. Select the appropriate tool to deliver your data (charts, graphs, tables, etc.)

Assess (Document findings under this heading):

1. Interpret data collected
2. Design and implement change
    - Redesign the process or tool if necessary
    - If no changes are necessary, then you have successfully used the Process Improvement pathway

Improvement (Document details here):

1. Set in place the process to gain and continue the improvement

Outcome (Document all changes here):

1. Positive changes made to improve quality of care based on Performance Improvement Activity

78.8.1 Problem-Solving Model

The FOCUS-PDCA/PMAIO Process Improvement Model is a statistical-based quality control method for improving processes. This approach to problem-solving could be used by all Process Action Teams to ensure uniformity within an organization. FOCUS-PDCA/PMAIO is as follows:

F — Find a process to improve.
O — Organize a team that knows the process.
C — Clarify current knowledge of the process.
U — Understand the cause or variations.
S — Select the process to improve.
P — Plan the improvement. P — Plan
D — Do the improvement (pilot test). M — Measure
C — Check the results of the improvement. A — Assess
A — Act to hold the gain. I — Improve
O — Outcome

78.9 Summary

Although quality can be simply defined as conformity to customer or user requirements, it has many dimensions. Seven of them are described here (1) performance, (2) aesthetics, (3) reliability (how dependably it performs), (4) availability (there when you need it), (5) durability (how long it lasts), (6) extras or features (supplementary items), and (7) serviceability (how easy it is to get serviced). The word PARADES can help you remember this.

78.9.1 PARADES — Seven Dimensions of Quality

1. *Performance*: A product or service that performs its intended function well scores high on this dimension of quality
2. *Aesthetics*: A product or service that has a favorable appearance, sound, taste, or smell is perceived to be of good quality
3. *Reliability*: Reliability, or dependability, is such an important part of product quality that quality-control engineers are sometimes referred to as reliability engineers
4. *Availability*: A product or service that is there when you need it
5. *Durability*: Durability can be defined as the amount of use one gets from a product before it no longer functions properly and replacement seems more feasible than constant repair
6. *Extras*: Feature or characteristics about a product or service that supplements its basic functioning (i.e., remote control dialing on a television)
7. *Serviceability*: Speed, courtesy, competence, and ease of repair are all important quality factors

78.9.2 Quality Has a Monetary Value!

Good quality often pays for itself, while poor quality is expensive in both measurable costs and hidden costs. The hidden costs include loss of goodwill, including loss of repeat business and badmouthing of the firm. High quality goods and services often carry a higher selling price than do those of low quality. This information is evidenced by several reports in the Wall Street Journal, Forbes Magazine, Money Magazine, Business Week Magazine, etc. A good example is the turnaround of Japanese product sales using quality methodologies outlined in The Deming Guide to Quality and Competitive Position, by Howard S. and Shelly J. Gitlow. As Dr. Demings has stated, quality improvement must be continuous!

> Quality is never an accident; it has always the result of intelligent energy.
>
> *John Ruskin, 1819–1900, English art critic and historian*
> *Seven Lamps of Architecture*

References

Bittel L.R. 1985. *Every Supervisor Should Know*, 5th ed., pp. 455–456. New York, Gregg Division/ McGraw-Hill.

DuBrin A.J. 1994. *Essentials of Management*, 3rd ed. Cleveland, South-Western Publishing Co.

Duck J.D. 1993. Managing change — the art of balancing. *Harvard Business Review*, November–December 1993, pp. 109–118.

Gitlow H.S. and Gitlow S.J. 1987. *The Deming Guide to Quality and Competitive Position.* Englewood Cliffs, NJ, Prentice-Hall.

Goal/QPC. 1988. *The Memory Jogger. A Pocket Guide of Tools for Continuous Improvement.* Massachusetts, Goal/QPC.

Ishikawa K. 1991. *Guide to Quality Control.* New York, Quality Resources (Asian Productivity Organization, Tokyo, Japan).

Joint Commission on Accreditation of Healthcare Organizations — Comprehensive Accreditation Manual for Hospitals — Official Handbook, Library of Congress Number 96-076721, 1998, Oakbrook Terrace, Illinois 60181.

Juran J.M. 1979. *Quality Control Handbook*, 3rd ed. New York, McGraw-Hill.

Katzenbach J.R. and Smith D.K. 1994. *The Wisdom of Teams.* New York, Harper Business. A Division of Harper Collins Publishers.

Mizuno S. 1988. *Management for Quality Improvement. The Seven New QC Tools.* Boston, Productivity Press.

Sholters P.R. 1993. *The Team Handbook. How to Use Teams to Improve Quality.* Madison, WI, Joiner Associates, Inc.

Walton M. 1986. *The Deming Management Method.* New York, Putnam Publishing Group.

79

A Standards Primer for Clinical Engineers

Alvin Wald
Columbia University

79.1 Introduction

The development, understanding, and use of standards is an important component of a clinical engineer's activities. Whether involved in industry, a health care facility, governmental affairs, or commercial enterprise, one way or another, the clinical engineer will find that standards are a significant aspect of professional activities. With the increasing emphasis on health care cost containment and efficiency, coupled with the continued emphasis on patient outcome, standards must be viewed both as a mechanism to reduce expenses and as another mechanism to provide quality patient care. In any case, standards must be addressed in their own right, in terms of technical, economic, and legal implications.

It is important for the clinical engineer to understand fully how standards are developed, how they are used, and most importantly, how they affect the entire spectrum of health related matters. Standards exist that address systems (protection of the electrical power distribution system from faults), individuals (means to reduce potential electric shock hazards), and protection of the environment (disposal of deleterious waste substances).

From a larger perspective, standards have existed since biblical times. In the Book of Genesis (Chap. 6, ver. 14), Noah is given a construction standard by God, "Make thee an ark of gopher wood; rooms shalt thou make in the ark, and shalt pitch it within and without with pitch." Standards for weights and measures have played an important role in bringing together human societies through trade and commerce. The earliest record of a standard for length comes from ancient Egypt, in Dynasty IV (ca. 3000 BC). This length was the royal cubit, 20.620 in. (52.379 cm), as used in construction of the Great Pyramid.

The importance of standards to society is illustrated in the Magna Carta, presented by the English barons to King John in 1215 on the field at Runnymede. Article 35 states:

> There shall be standard measures of wine, beer, and corn — the London quarter — throughout the whole of our kingdom, and a standard width of dyed, russet and halberject cloth — two ells within the selvedges; and there shall be standard weights also.

The principles of this article appear in the English Tower system for weight and capacity, set in 1266 by the assize of Bread and Ale Act:

> An English penny called a sterling, round and without any clipping, shall weigh thirty-two wheatcorns in the midst of the ear; and twenty ounces a pound: and eight pounds do make a gallon of wine, and eight gallons of wine do make a bushell, which is the eighth part of a quarter.

In the United States, a noteworthy use of standards occurred after the Boston fire of 1689. With the aim of rapid rebuilding of the city, the town fathers specified that all bricks used in construction were to be 9 × 4 × 4 in. An example of standardization to promote uniformity in manufacturing practices was the contract for 10,000 muskets awarded to Eli Whitney by President Thomas Jefferson in 1800. The apocryphal story is that Eli Whitney (better known to generations of grammar school children for his invention of the cotton gin) assembled a large number of each musket part, had one of each part randomly selected, and then assembled a complete working musket. This method of production, the complete interchangeability of assembly parts, came to be known as the "armory method," replacing hand crafting, which at that time had been the prevailing method of manufacturing throughout the world.

79.2 Definitions

A most general definition of a standard is given by Rowe (1983). "A standard is a multi-party agreement for establishing an arbitrary criterion for reference." Each word used in the definition by Rowe corresponds to a specific characteristic that helps to define the concept of a standard. Multi means more than one party, organization, group, government, agency, or individual. Agreement means that the concerned parties have come to some mutually agreed upon understanding of the issues involved and of ways to resolve them. This understanding has been confirmed via some mechanism such as unanimity, consensus, ballot, or other means that has been specified. Establishing defines the purpose of the agreement — to create the standard and carry forth its provisions.

Arbitrary emphasizes an understanding by the parties that there are no absolute criteria in creating the standard. Rather, the conditions and values chosen are based on the most appropriate knowledge and conditions available at the time the standard was established. Criterions are those features and conditions that the parties to the agreement have chosen as the basis for the standard. Not all issues may be addressed, but only those deemed, for whatever reasons, suitable for inclusion.

A different type of definition of a standard is given in The United States Office of Management and Budget Circular A-119:

> . . . a prescribed set of rules, conditions, or requirements concerned with the definition of terms; classification of components; delineation of procedures; specifications of materials, performance, design, or operations; or measurement of quality and quantity in describing materials, products, systems, services, or practices.

A code is a compilation of standards relating to a particular area of concern, that is, a collection of standards. For example, local government health codes contain standards relating to providing of health care to members of the community. A regulation is an organization's way of specifying that some particular standard must be adhered to. Standards, codes, and regulations may or may not have legal implications, depending on whether the promulgating organization is governmental or private.

79.3 Standards for Clinical Engineering

There is a continually growing body of standards that affect health care facilities, and hence clinical engineering. The practitioner of health care technology must constantly search out, evaluate, and apply appropriate standards. The means to reconcile the conflicts of technology, cost considerations, the different jurisdictions involved, and the implementation of the various standards is not necessarily apparent. One technique that addresses these concerns and has proven to yield a consistent practical approach is a structured framework of the various levels of standards. This hierarchy of standards is a conceptual model that the clinical engineer can use to evaluate and apply to the various requirements that exist in the procurement and use of health care technology.

Standards have different purposes, depending on their particular applications. A hierarchy of standards can be used to delineate those conditions for which a particular standard applies. There are four basic categories, any one or all of which may be in simultaneous operation:

1. Local or proprietary standards (perhaps more properly called regulations) are developed to meet the internal needs of a particular organization
2. Common interest standards serve to provide uniformity of product or service throughout an industry or profession
3. Consensus standards are agreements amongst interested participants to address an area of mutual concern
4. Regulatory standards are mandated by an authority having jurisdiction to define a particular aspect of concern

In addition, there are two categories of standards adherence (1) voluntary standards, which carry no inherent power of enforcement, but provide a reference point of mutual understanding, and (2) mandatory standards, which are incumbent upon those to whom the standard is addressed, and enforceable by the authority having jurisdiction.

The hierarchy of standards model can aid the clinical engineer in the efficient and proper use of standards. More importantly, it can provide standards developers, users, and the authorities having jurisdiction in these matters with a structure by which standards can be effectively developed, recognized, and used to the mutual benefit of all.

79.4 A Hierarchy of Standards

Local, or proprietary standards, are developed for what might be called internal use. An organization that wishes to regulate and control certain of its own activities issues its own standards. Thus, the standard is local in the sense that it is applied in a specific venue, and it is proprietary in that it is the creation of a completely independent administration. For example, an organization may standardize on a single type of an electrocardiograph monitor. This standardization can refer to a specific brand or model, or to specific functional or operational features. In a more formal sense, a local standard may often be referred to as an institutional Policy and Procedure. The policy portion is the why of it; the procedure portion is the how. It must be kept in mind that standards of this type that are too restrictive will limit innovation and progress, in that they cannot readily adapt to novel conditions. On the other hand, good local standards contribute to lower costs, operational efficiency, and a sense of coherence within the organization.

Sometimes, local standards may originate from requirements of a higher level of regulation. For example, the Joint Commission for Accreditation of Healthcare Organizations (JCAHO) (formerly the Joint Commission for Hospital Accreditation (JCAH), a voluntary organization (but an organization that hospitals belong to for various reasons, for example, accreditation, reimbursement, approval of training programs), does not set standards for what or how equipment should be used. Rather, the JCAHO requires that each hospital set its own standards on how equipment is selected, used, and maintained. To monitor compliance with this requirement, the JCAHO inspects whether the hospital follows its own standards. In one sense, the most damaging evidence that can be adduced against an organization (or an individual) is that it (he) did not follow its (his) own standards.

Common interest standards are based on a need recognized by a group of interested parties, which will further their own interests, individually or collectively. Such standards are generally accepted by affected interests without being made mandatory by an authority; hence they are one type of voluntary standard. These standards are often developed by trade or professional organizations to promote uniformity in a product or process. This type of standard may have no inducement to adherence except for the benefits to the individual participants. For example, if you manufacture a kitchen cabinet that is not of standard size, it will not fit into the majority of kitchens and thus it will not sell. Uniformity of screw threads is another example of how a product can be manufactured and used by diverse parties, and yet be absolutely interchangeable. More recently, various information transfer standards allow the interchange of computer-based information amongst different types of instruments and computers.

Consensus standards are those that have been developed and accepted in accordance with certain well defined criteria so as to assure that all points of view have been considered. Sometimes, the adjective "consensus" is used as a modifier for a "voluntary standard." Used in this context, consensus implies that all interested parties have been consulted and have come to a general agreement on the provisions of the standard. The development of a consensus standard follows an almost ritualistic procedure to insure that fairness and due process are maintained. There are various independent voluntary and professional organizations that sponsor and develop standards on a consensus basis (see below). Each such organization has its own particular rules and procedures to make sure that there is a true consensus in developing a standard.

In the medical products field, standards are sometimes difficult to implement because of the independent nature of manufacturers and their high level of competition. A somewhat successful standards story is the adoption of the DIN configuration for ECG lead-cable connection by the Association for the Advancement of Medical Instrumentation (AAMI). The impetus for this standard was the accidental electrocution of several children brought about by use of the previous industry standard lead connection (a bare metal pin, as opposed to the new recessed socket). Most (but not all) manufacturers of ECG leads and cables now adhere to this standard. Agreement on this matter is in sharp contrast to the inability of the health care manufacturing industry to implement a standard for ECG cable connectors. Even though a standard was written, the physical configuration of the connector is not necessarily used by manufacturers in production, nor is it demanded by medical users in purchasing. Each manufacturer uses a different connector, leading to numerous problems in supply and incompatibility for users. This is an example of a voluntary standard, which for whatever reasons, is effectively ignored by all interested parties.

However, even though there have been some failures in standardization of product features, there has also been significant progress in generating performance and test standards for medical devices. A number of independent organizations sponsor development of standards for medical devices. For example, the American Society for Testing and Materials (ASTM) has developed, "Standard Specification for Minimum Performance and Safety Requirements for Components and Systems of Anesthesia Gas Machines (F1161-88)." Even though there is no statutory law that requires it, manufacturers no longer produce, and thus hospitals can no longer purchase anesthesia machines without the built-in safety features specified in this standard. AAMI has sponsored numerous standards that relate to performance of specific medical devices, such as defibrillators, electrosurgical instruments, and electronic sphygmomanometers. These standards are compiled in the AAMI publication, "Essential Standards for Biomedical Equipment Safety and Performance." The National Fire Protection Association (NFPA) publishes "Standard for Health Care Facilities (NFPA 99)," which covers a wide range of safety issues relating to facilities. Included are sections that deal with electricity and electrical systems, central gas and vacuum supplies, and environmental conditions. Special areas such as anesthetizing locations, laboratories, and hyperbaric facilities are addressed separately. Mandatory standards have the force of law or other authority having jurisdiction.

Mandatory standards imply that some authority has made them obligatory. Mandatory standards can be written by the authority having jurisdiction, or they can be adapted from documents prepared by others as proprietary or consensus standards. The authority having jurisdiction can be a local hospital or even a department within the hospital, a professional society, a municipal or state government, or an agency of the federal government that has regulatory powers.

In the United States, hospitals are generally regulated by a local city or county authority, and/or by the state. These authorities set standards in the form of health codes or regulations, which have the force of law. Often, these local bodies consider the requirements of a voluntary group, the Joint Commission for Accreditation of Healthcare Organizations, in their accreditation and regulatory processes.

American National Standards. The tradition in the United States is that of voluntary standards. However, once a standard is adopted by an organization, it can be taken one step further. The American National Standards Institute (ANSI) is a private, nongovernment, voluntary organization that acts as a coordinating body for standards development and recognition in the United States. If the development process for a standard meets the ANSI criteria of open deliberation of legitimate concerns, with all interested parties coming to a voluntary consensus, then the developers can apply (but are not required) to have their standard designated as an American National Standard. Such a designation does not make a standard any more legitimate, but it does offer some recognition as to the process by which it has been developed. ANSI also acts as a clearing house for standards development, so as to avoid duplication of effort by various groups that might be concerned with the same issues. ANSI is also involved as a U.S. coordinating body for many international standards activities.

An excellent source that lists existing standards and standards generating organizations, both nationally and internationally, along with some of the workings of the FDA (see below), is the "Medical Device Industry Fact Book" [Allen, 1996].

79.5 Medical Devices

On the national level, oversight is generally restricted to medical devices, and not on operational matters. Federal jurisdiction of medical devices falls under the purview of the Department of Health and Human Services, Public Health Service, Food and Drug Administration (FDA), Center for Devices and Radiological Health. Under federal law, medical devices are regulated under the "Medical Device Amendments of 1976" and the "Radiation Control for Health and Safety Act of 1968." Additional regulatory authorization is provided by the "Safe Medical Devices Act of 1990," the "Medical Device Amendments of 1992," the "FDA Reform and Enhancement Act of 1996," and the "Food and Drug Administration Modernization Act of 1997."

A medical device is defined by Section 201 of the Federal Food, Drug, and Cosmetic Act (as amended), as an:

> instrument, apparatus, implement, machine, contrivance, implant, *in vitro* reagent, or other similar or related article including any component, part, or accessory which is:
> recognized in the official National Formulary, or the United States Pharmacopeia, or any supplement to them;
> intended for use in the diagnosis of disease or other conditions, or in the care, mitigation, treatment, or prevention of disease, in man or other animals, or
> intended to affect the structure of any function of the body of man or other animals; and which does not achieve its primary intended purposes through chemical action within or on the body of man . . . and which is not dependent upon being metabolized for the achievement of its primary intended purposes.

The major thrust of the FDA has been in the oversight of the manufacture of medical devices, with specific requirements based on categories of perceived risks. The 1976 Act (Section 513) establishes three classes of medical devices intended for human use:

Class I. General controls regulate devices for which controls other than performance standards or premarket approvals are sufficient to assure safety and effectiveness. Such controls include regulations that (1) prohibit adulterated or misbranded devices; (2) require domestic device manufacturers and initial distributors to register their establishments and list their devices; (3) grant FDA authority to ban certain devices; (4) provide for notification of risks and of repair, replacement, or refund; (5) restrict the sale, distribution, or use of certain devices; and (6) govern Good Manufacturing Practices, records, and reports, and inspections. These minimum requirements apply also to Class II and Class III devices.

Class II. Performance Standards apply to devices for which general controls alone do not provide reasonable assurance of safety and efficacy, and for which existing information is sufficient to establish a performance standard that provides this assurance. Class II devices must comply not only with general controls, but also with an applicable standard developed under Section 514 of the Act. Until performance standards are developed by regulation, only general controls apply.

Class III. Premarket Approval applies to devices for which general controls do not suffice or for which insufficient information is available to write a performance standard to provide reasonable assurance of safety and effectiveness. Also, devices which are used to support or sustain human life or to prevent impairment of human health, devices implanted in the body, and devices which present a potentially unreasonable risk of illness or injury. New Class III devices, those not "substantially equivalent" to a device on the market prior to enactment (May 28, 1976), must have approved Premarket Approval Applications (Section 510 k).

Exact specifications for General Controls and Good Manufacturing Practices (GMP) are defined in various FDA documents. Aspects of General Controls include yearly manufacturer registration, device listing, and premarket approval. General Controls are also used to regulate adulteration, misbranding and labeling, banned devices, and restricted devices. Good Manufacturing Practices include concerns of organization and personnel; buildings and equipment; controls for components, processes, packaging, and labeling; device holding, distribution, and installation; manufacturing records; product evaluation; complaint handling; and a quality assurance program. Design controls for GMP were introduced in 1996. They were motivated by the FDA's desire to harmonize its requirements with those of a proposed international standard (ISO 13485). Factors that need to be addressed include planning, input and output requirements, review, verification and validation, transfer to production, and change procedures, all contained in a history file for each device. Device tracking is typically required for Class III life-sustaining and implant devices, as well as postmarket surveillance for products introduced starting in 1991.

Other categories of medical devices include combination devices, in which a device may incorporate drugs or biologicals. Combination devices are controlled via intercenter arrangements implemented by the FDA.

Transitional devices refer to devices that were regulated as drugs, prior to the enactment of the Medical Device Amendments Act of 1976. These devices were automatically placed into Class III, but may be transferred to Class I or II.

A custom device may be ordered by a physician for his/her own use or for a specific patient. These devices are not generally available, and cannot be labeled or advertised for commercial distribution.

An investigational device is one that is undergoing clinical trials prior to premarket clearance. If the device presents a significant risk to the patient, an Investigational Device Exemption must be approved by the FDA. Information must be provided regarding the device description and intended use, the origins of the device, the investigational protocol, and proof of oversight by an Institutional Review Board to insure informed patient consent. Special compassionate or emergency use for a nonapproved device or for a nonapproved use can be obtained from the FDA under special circumstances, such as when there is no other hope for the patient.

Adverse Events. The Safe Medical Devices Act of 1990 included a provision by which both users and manufacturers (and distributors) of medical devices are required to report adverse patient events that may be related to a medical device. Manufacturers must report to the FDA if a device (a) may have caused or contributed to a death or serious injury, or (b) malfunctioned in such a way as would be likely to cause or contribute to a death or serious injury if the malfunction were to reoccur. Device users are required to notify the device manufacturer of reportable incidents, and must also notify the FDA in case of a device-related death. In addition, the FDA established a voluntary program for reporting device problems that may not have caused an untoward patient event, but which may have the potential for such an occurrence under altered circumstances.

New devices. As part of the General Controls requirements, the FDA must be notified prior to marketing any new (or modifying an existing) device for patient use. This premarket notification, called the 510(k) process after the relevant section in the Medical Device Amendments Act, allows the FDA to review the device for safety and efficacy.

There are two broad categories that a device can fall into. A device that was marketed prior to May 28, 1976 (the date that the Medical Device Amendments became effective) can continue to be sold. Also, a product that is "substantially equivalent" to a preamendment device can likewise be marketed. However, the FDA may require a premarket approval application for any Class III device (see below). Thus, these preamendment devices and their equivalents are approved by "grandfathering." (Premarket notification to the FDA is still required to assure safety and efficacy). Of course, the question of substantial equivalency is open to an infinite number of interpretations. From the manufacturer's perspective, such a designation allows marketing the device without a much more laborious and expensive premarket approval process.

A new device that the FDA finds is not substantially equivalent to a premarket device is automatically placed into Class III. This category includes devices that provide functions or work through principles not present in preamendment devices. Before marketing, this type of device requires a Premarket Approval Application by the manufacturer, followed by an extensive review by the FDA. (However, the FDA can reclassify such devices into Class I or II, obviating the need for premarket approval.) The review includes scientific and clinical evaluation of the application by the FDA and by a Medical Advisory Committee (composed of outside consultants). In addition, the FDA looks at the manufacturing and control processes to assure that all appropriate regulatory requirements are being adhered to. Clinical (use of real patients) trials are often required for Class III devices in order to provide evidence of safety and efficacy. To carry out such trials, an Investigational Device Exemption must be issued by the FDA.

The Food and Drug Administration Modernization Act of 1997, which amends section 514 of the Food, Drug, and Cosmetic Act, has made significant changes in the above regulations. These changes greatly simplify and accelerate the entire regulatory process. For example, the law exempts from premarket notification Class I devices that are not intended for a use that is of substantial importance in preventing impairment of human health, or that do not present a potential unreasonable risk of illness or injury. Almost 600 Class I generic devices have been so classified by the agency. In addition, the FDA will specify those Class II devices for which a 510(k) submission will also not be required.

Several other regulatory changes have been introduced by the FDA to simplify and speed up the approval process. So-called "third party" experts will be allowed to conduct the initial review of all Class I and low-to-intermediate risk Class II devices. Previously, the FDA was authorized to create standards for medical devices. The new legislation allows the FDA to recognize and use all or parts of various appropriate domestic and internationally recognized consensus standards that address aspects of safety and effectiveness relevant to medical devices.

79.6 International Standards

Most sovereign nations have their own internal agencies to establish and enforce standards. However, in our present world of international cooperation and trade, standards are tending towards uniformity across national boundaries. This internationalization of standards is especially true since formation of the European Common Market. The aim here is to harmonize the standards of individual nations by promulgating directives for medical devices that address "Essential Requirements" [Freeman, 1993] (see below). Standards in other areas of the world (Asia, Eastern Europe) are much more fragmented, with each country specifying regulations for its own manufactured and imported medical devices.

There are two major international standards generating organizations, both based in Europe, the International Electrotechnical Commission (IEC) and the International Organization for Standardization (ISO). Nations throughout the world participate in the activities of these organizations.

The International Electrotechnical Commission (IEC), founded in 1906, oversees, on an international level, all matters relating to standards for electrical and electronic items. Membership in the IEC is held by a National Committee for each nation. The United States National Committee (USNC) for IEC was founded in 1907, and since 1931 has been affiliated with ANSI. USNC has its members representatives from professional societies, trade associations, testing laboratories, government entities, other organizations, and individual experts. The USNC appoints a technical advisor and a technical advisory group for each IEC Committee and Subcommittee to help develop a unified United States position. These advisory groups are drawn from groups that are involved in the development of related U.S. national standards.

Standards are developed by Technical Committees (TC), Subcommittees (SC), and Working Groups (WG). IEC TC 62, "Electrical Equipment in Medical Practice," is of particular interest here. One of the basic standards of this Technical Committee is document 601-1, "Safety of Medical Electrical Equipment, Part 1: General Requirements for Safety," 2nd Edition (1988) and its Amendment 1 (1991), along with Document 601-1-1, "Safety Requirements for Medical Electrical Systems" (1992).

The International Organization for Standardization (ISO) oversees aspects of device standards other than those related to electrotechnology. This organization was formed in 1946 with a membership comprised of the national standards organizations of 26 countries. There are currently some 90 nations as members. The purpose of the ISO is to "facilitate international exchange of goods and services and to develop mutual cooperation in intellectual, scientific, technological, and economic ability." ISO addresses all aspects of standards except for electrical and electronic issues, which are the purview of the International Electrotechnical Commission. ANSI has been the official United States representative to ISO since its inception. For each Committee or Subcommittee of the ISO in which ANSI participates, a U.S. Technical Advisory Group (TAG) is formed. The administrator of the TAG is, typically, that same U.S. organization that is developing the parallel U.S. standard.

Technical Committees (TC) of the ISO concentrate on specific areas of interest. There are Technical Committees, Subcommittees, Working Groups, and Study Groups. One of the member national standards organizations serves as the Secretariat for each of these technical bodies.

One standard of particular relevancy to manufacturers throughout the world is ISO 9000. This standard was specifically developed to assure a total quality management program that can be both universally recognized and applied to any manufacturing process. It does not address any particular product or process, but is concerned with structure and oversight of how processes are developed, implemented, monitored, and documented. An independent audit must be passed by any organization to obtain ISO 9000 registration. Many individual nations and manufacturers have adopted this standard and require that any product that they purchase be from a source that is ISO 9000 compliant.

The European Union was, in effect, created by the Single Europe Act (EC-92), as a region "without internal frontiers in which the free movement of goods, persons, and capital is ensured." For various products and classes of products, the European Commission issues directives with regard to safety and other requirements, along with the means for assessing conformity to these directives. Products that comply with the appropriate directives can then carry the CE mark. EU member states ratify these directives into national law.

Two directives related to medical devices are the Medical Devices Directive (MDD), enacted in 1993 (mandatory as of June 15,1998), and the Active Implanted Medical Devices Directive (AIMDD), effective since 1995. Safety is the primary concern of this system, and as in the United States, there are three classes of risk. These risks are based on what and for how long the device touches, and its effects. Safety issues include electrical, mechanical, thermal, radiation, and labeling. Voluntary standards that address these issues are formulated by the European Committee for Standardization (CEN) and the European Committee for Electrotechnical Standardization (CENELEC).

79.7 Compliance with Standards

Standards that were originally developed on a voluntary basis may take on mandatory aspects. Standards that were developed to meet one particular need may be used to satisfy other needs as well. Standards will be enforced and adhered to if they meet the needs of those who are affected by them. For example, consider a standard for safety and performance for a defibrillator. For the manufacturer, acceptance and sales are a major consideration in both the domestic and international markets. People responsible for specifying, selecting, and purchasing equipment may insist on adherence to the standard so as to guarantee safety and performance. The user, physician, or other health care professional, will expect the instrument to have certain operational and performance characteristics to meet medical needs. Hospital personnel want a certain minimum degree of equipment uniformity for ease of training and maintenance. The hospital's insurance company and risk manager want equipment that meets or exceeds recognized safety standards. Third party payers, that is private insurance companies or government agencies, insist on equipment that is safe, efficacious, and cost effective. Accreditation agencies, such as local health agencies or professional societies, often require equipment to meet certain standards. More basically, patients, workers, and society as a whole have an inherent right to fundamental safety. Finally, in our litigatious society, there is always the threat of civil action in the case of an untoward event in which a "non-standard," albeit "safe," instrument was involved. Thus, even though no one has stated "this standard must be followed," it is highly unlikely that any person or organization will have the temerity to manufacture, specify, or buy an instrument that does not "meet the standard."

Another example of how standards become compulsory is via accreditation organizations. The Joint Commission for Accreditation of Healthcare Organizations has various standards (requirements). This organization is a private body that hospitals voluntarily accept as an accrediting agent. However, various health insurance organizations, governmental organizations, and physician specialty boards for resident education use accreditation by the JCAHO as a touchstone for quality of activities. Thus, an insurance company might not pay for care in a hospital that is not accredited, or a specialty board might not recognize resident training in such an institution. Thus, the requirements of the JCAHO, in effect, become mandatory standards for health care organizations.

A third means by which voluntary standards can become mandatory is by incorporation. Existing standards can be incorporated into a higher level of standards or codes. For example, various state and local governments incorporate standards developed by voluntary organizations, such as the National Fire Protection Association, into their own building and health codes. These standards then become, in effect, mandatory government regulations, and have the force of (civil) law. In addition, as discussed above, the FDA will now recognize voluntary standards developed by recognized organizations.

79.8 Limitations of Standards

Standards are generated to meet the expectations of society. They are developed by organizations and individuals to meet a variety of specific needs, with the general goals of promoting safety and efficiency. However, as with all human activities, problems with the interpretation and use of standards do occur. Engineering judgment is often required to help provide answers. Thus, the clinical engineer must consider the limits of standards, a boundary that is not clear and is constantly shifting. Yet clinical engineers must always employ the highest levels of engineering principles and practices. Some of the limitations and questions of standards and their use will be discussed below.

79.8.1 Noncompliance with a Standard

Sooner or later, it is likely that a clinical engineer will either be directly involved with or become aware of deviation from an accepted standard. The violation may be trivial, with no noticeable effect, or there may be serious consequences. In the former case, either the whole incident may be ignored, or nothing more may be necessary than a report that is filed away, or the incident can trigger some sort of corrective action. In the latter case, there may be major repercussions involving investigation, censure, tort issues, or legal actions. In any event, lack of knowledge about the standard is not a convincing defense. Anyone who is in a position that requires knowledge about a standard should be fully cognizant of all aspects of that standard. In particular, one should know the provisions of the standard, how they are to be enforced, and the potential risks of noncompliance. Nonetheless, noncompliance with a standard, in whole or in part, may be necessary to prevent a greater risk or to increase a potential benefit to the patient. For example, when no other recourse is available, it would be defensible to use an electromagnet condemned for irreparable excessive leakage current to locate a foreign body in the eye of an injured person, and thus save the patient's vision. Even if the use of this device resulted in a physical injury or equipment damage, the potential benefit to the patient is a compelling argument for use of the noncompliant device. In such a case, one should be aware of and prepared to act on the possible hazard (excessive electrical current, here). A general disclaimer making allowance for emergency situations is often included in policy statements relating to use of a standard. Drastic conditions require drastic methods.

79.8.2 Standards and the Law

Standards mandated by a government body are not what is called "black letter law," that is a law actually entered into a criminal or civil code. Standards are typically not adopted in the same manner as laws, that is, they are not approved by a legislative body, ratified by an elected executive, and sanctioned by the courts. The usual course for a mandated standard is via a legislative body enacting a law that establishes or assigns to an executive agency the authority to regulate the concerned activities. This agency, under the control of the executive branch of government, then issues standards that follow the mandate of its enabling legislation. If conflicts arise, in addition to purely legal considerations, the judiciary must interpret the intent of the legislation in comparison with its execution. This type of law falls under civil rather than criminal application.

The penalty for noncompliance with a standard may not be criminal or even civil prosecution. Instead, there are administrative methods of enforcement, as well as more subtle yet powerful methods of coercion. The state has the power (and the duty) to regulate matters of public interest. Thus, the state can withhold or withdraw permits for construction, occupancy, or use. Possibly more effective, the state can withhold means of finance or payments to violators of its regulations. Individuals injured by failure to abide by a standard may sue for damages in civil proceedings. However, it must be recognized that criminal prosecution is possible when the violations are most egregious, leading to human injury or large financial losses.

79.8.3 Incorporation and Revision

Because of advances in technology and increases in societal expectations, standards are typically revised periodically. For example, the National Fire Protection Association revises and reissues its "Standard for Health Care Facilities" (NFPA 99) every three years. Other organizations follow a five year cycle of review, revision, and reissue of standards. These voluntary standards, developed in good faith, may be adapted by governmental agencies and made mandatory, as discussed above. When a standard is incorporated into a legislative code, it is generally referenced as to a particular version and date. It is not always the case that a newer version of the standard is more restrictive. For example, ever since 1984, the National Fire Protection Association "Standard for Health Care Facilities" (NFPA 99) does not require the installation of isolated power systems (isolation transformers and line isolation monitors) in anesthetizing locations that do not use flammable anesthetic agents, or in areas that are not classified as wet locations. A previous version of this standard, "Standard for the Use of Inhalation Anesthetics (Flammable and Nonflammable)," (NFPA 56A-1978) did require isolated power. However, many State Hospital Codes have incorporated, by name and date, the provisions of the older standard, NFPA 56A. Thus, isolated power may still be required, by code, in new construction of all anesthetizing locations, despite the absence of this requirement in the latest version of the standard that addresses this issue. In such a case, the organization having jurisdiction in the matter must be petitioned to remedy this conflict between new and old versions of the standard.

79.8.4 Safety

The primary purpose of standards in clinical practice is to assure the safety of patient, operator, and bystanders. However, it must be fully appreciated that there is no such thing as absolute safety. The more safety features and regulations attached to a device, the less useful and the more cumbersome and costly may be its actual use. In the development, interpretation, and use of a standard, there are questions that must be asked: What is possible? What is acceptable? What is reasonable? Who will benefit? What is the cost? Who will pay?

No one can deny that medical devices should be made as safe as possible, but some risk will always remain. In our practical world, absolute safety is a myth. Many medical procedures involve risk to the patient. The prudent physician or medical technologist will recognize the possible dangers of the equipment and take appropriate measures to reduce the risk to a minimum. Some instruments and procedures are inherently more dangerous than others. The physician must make a judgment, based on his/her own professional knowledge and experience, as well as on the expectations of society, whether using a particular device is less of a risk than using an alternative device or doing nothing. Standards will help — but they do not guarantee complete safety, a cure, or legal and societal approval.

79.8.5 Liability

Individuals who serve on committees that develop standards, as well as organizations involved in such activities, are justifiably concerned with their legal position in the event that a lawsuit is instituted as a result of a standard that they helped to bring forth. Issues involved in such a suit may include restraint of trade, in case of commercial matters, or to liability for injury due to acts of commission or of omission. Organizations that sponsor standards or that appoint representatives to standards developing groups often have insurance for such activities. Independent standards committees and individual members of any standards committees may or may not be covered by insurance for participation in these activities. Although in recent times only one organization and no individual has been found liable for damages caused by improper use of standards (see following paragraph), even the possibility of being named in a lawsuit can intimidate even the most self-confident “expert.” Thus, it is not at all unusual for an individual who is asked to serve on a standards development committee first to inquire as to liability insurance coverage. Organizations that develop standards or appoint representatives also take pains to insure that all of their procedures are carefully followed and documented so as to demonstrate fairness and prudence.

The dark side of standards is the implication that individuals or groups may unduly influence a standard to meet a personal objective, for example, to dominate sales in a particular market. If standards are developed or interpreted unfairly, or if they give an unfair advantage to one segment, then restraint of trade charges can be made. This is why standards to be deemed consensus must be developed in a completely open and fair manner. Organizations that sponsor standards that violate this precept can be held responsible. In 1982, the United States Supreme Court, in the Hydrolevel Case [Perry, 1982], ruled that the American Society of Mechanical Engineers was guilty of antitrust activities because of the way some of its members, acting as a committee to interpret one of its standards, issued an opinion that limited competition in sales so as to unfairly benefit their own employers. This case remains a singular reminder that standards development and use must be inherently fair.

79.8.6 Inhibition

Another charge against standards is that they inhibit innovation and limit progress [Flink, 1984]. Ideally, standards should be written to satisfy minimum, yet sufficient, requirements for safety, performance, and efficacy. Improvements or innovations would still be permitted so long as the basic standard is followed. From a device users point of view, a standard that is excessively restrictive may limit the scope of permissible professional activities. If it is necessary to abrogate a standard in order to accommodate a new idea or to extend an existing situation, then the choice is to try to have the standard changed, which may be very time consuming, or to act in violation of the standard and accept the accompanying risks and censure.

79.8.7 *Ex Post Facto*

A question continually arises as to what to do about old equipment (procedures, policies, facilities, etc.) when a new standard is issued or an old standard is revised so that existing items become obsolete. One approach, perhaps the simplest, is to do nothing. The philosophy here being that the old equipment was acquired in good faith and conformed to the then existing standards. As long as that equipment is usable and safe, there is no necessity to replace it. Another approach is to upgrade the existing equipment to meet the new standard. However, such modification may be technically impractical or financially prohibitive. Finally, one can simply throw out all of the existing equipment (or sell it to a second-hand dealer, or use the parts for maintenance) and buy everything new. This approach would bring a smile of delight from the manufacturer and a scream of outrage from the hospital administrator. Usually what is done is a compromise, incorporating various aspects of these different approaches.

79.8.8 Costs

Standards cost both time and money to propose, develop, promulgate, and maintain. Perhaps the greatest hindrance to more participation in standards activities by interested individuals is the lack of funds to attend meetings where the issues are discussed and decisions are made. Unfortunately, but nonetheless true, organizations that can afford to sponsor individuals to attend such meetings have considerable influence in the development of that standard. On the other hand, those organizations that do have a vital interest in a standard should have an appropriate say in its development. A consensus of all interested parties tempers the undue influence of any single participant. From another viewpoint, standards increase the costs of manufacturing devices, carrying out procedures, and administering policies. This incremental cost is, in turn, passed on to the purchaser of the goods or services. Whether or not the increased cost justifies the benefits of the standard is not always apparent. It is impossible to realistically quantify the costs of accidents that did not happen or the confusion that was avoided by adhering to a particular standard. However, it cannot be denied that standards have made a valuable contribution to progress, in the broadest sense of that word.

79.9 Conclusions

Standards are just like any other human activity, they can be well used or a burden. The danger of standards is that they will take on a life of their own; and rather than serve a genuine need will exist only as a justification of their own importance. This view is expressed in the provocative and iconoclastic book by Bruner and Leonard [1989], and in particular in their Chapter 9, *"Codes and Standards: Who Makes the Rules?"* However, the raison d'être of standards is to do good. It is incumbent upon clinical engineers, not only to understand how to apply standards properly, but also how to introduce, modify, and retire standards as conditions change. Furthermore, the limitations of standards must be recognized in order to realize their maximum benefit. No standard can replace diligence, knowledge, and a genuine concern for doing the right thing.

References

Allen, A. (ed.) 1996. *Medical Device Industry Fact Book*, 3rd ed. Canon Communications, Santa Monica, CA.

American Society for Testing and Materials (ASTM). 1916. Race Street, Philadelphia, PA 19103.

Association for the Advancement of Medical Instrumentation (AAMI) 3330 Washington Boulevard, Suite 400, Arlington, VA 22201.

Bruner, J.M.R. and Leonard, P.F. 1989. *Electricity, Safety and the Patient*, Year Book Medical Publishers, Chicago.

Flink, R. 1984. Standards: Resource or Constraint? *IEEE Eng. Med. Biol. Mag.*, 3: 14–16.

Food and Drug Administration, Center for Devices and Radiological Health, 5600 Fishers Lane, Rockville, MD 20857. URL: http://www.fda.gov/.

Freeman, M. 1993. The EC Medical Devices Directives, *IEEE Eng. Med. Biol. Mag.*, 12: 79–80.

International Organization for Standardization (ISO), Central Secretariat, 1 rue de Varembe, Case postale 56, CH 1211, Geneva 20, Switzerland. URL: http://www.iso.ch/index.html.

International Electrotechnical Commission (IEC), Central Office, 3 rue de Varembé, P. O. Box 131, CH-1211, Geneva 20, Switzerland. URL: http://www.iec.ch/.

Joint Commission for Accreditation of Healthcare organizations (JCAHO) 1 Renaissance Boulevard, Oakbrook, IL 60181.

National Fire Protection Association (NFPA) Batterymarch Park, Quincy, MA 02269.

Perry, T.S. 1982. Antirust Ruling Chills Standards Setting, *IEEE Spectrum*, 19: 52–54.

Rowe, W.D. 1983. Design and Performance Standards. In *Medical Devices: Measurements, Quality Assurance, and Standards*, C.A. Caceres, H.T. Yolken, R.J. Jones, and H.R. Piehler (eds.), pp. 29–40. American Society for Testing and Materials, Philadelphia, PA.

80

Regulatory and Assessment Agencies

Mark E. Bruley
Vivian H. Coates
ECRI

Effective management and development of clinical and biomedical engineering departments (hereafter called clinical engineering departments) in hospitals requires a basic knowledge of relevant regulatory and technology assessment agencies. Regulatory agencies set standards of performance and record keeping for the departments and the technology for which they are responsible. Technology assessment agencies are information resources for what should be an ever expanding role of the clinical engineer in the technology decision-making processes of the hospital's administration.

This chapter presents an overview of regulatory and technology assessment agencies in the United States, Canada, Europe, and Australia that are germane to clinical engineering. Due to the extremely large number of such agencies and information resources, we have chosen to focus on those of greatest relevance and/or informational value. The reader is directed to the references and sources of further information presented at the end of the chapter.

80.1 Regulatory Agencies

Within the healthcare field, there are over 38,000 applicable standards, clinical practice guidelines, laws, and regulations [ECRI, 1999]. Voluntary standards are promulgated by more than 800 organizations; mandatory standards by more than 300 state and federal agencies. Many of these organizations and agencies issue guidelines that are relevant to the vast range of healthcare technologies within the responsibility of clinical engineering departments. Although many of these agencies also regulate the manufacture and clinical use of healthcare technology, such regulations are not directly germane to the management of a clinical department and are not presented.

For the clinical engineer, many agencies promulgate regulations and standards in the areas of, for example, electrical safety, fire safety, technology management, occupational safety, radiology and nuclear medicine, clinical laboratories, infection control, anesthesia and respiratory equipment, power distribution, and medical gas systems. In the United States medical device problem reporting is also regulated by many state agencies and by the U.S. Food and Drug Administration (FDA) via its MEDWATCH program. It is important to note that, at present, the only direct regulatory authority that the FDA has over U.S. hospitals is in the reporting of medical device related accidents that result in serious injury or death.

Chapter 80 discusses in detail many of the specific agency citations. Presented below are the names and addresses of the primary agencies whose codes, standards, and regulations have the most direct bearing on clinical engineering and technology management:

American Hospital Association
1 North Franklin
Chicago, IL 60606
(312) 422-3000
Website: www.aha.org

American College of Radiology
1891 Preston White Drive
Reston, VA 22091
(703) 648-8900
Website: www.acr.org

American National Standards Institute
11 West 42nd Street
13th Floor, New York, NY 10036
(212) 642-4900
Website: www.ansi.org

American Society for Hospital Engineering
840 North Lake Shore Drive
Chicago, IL 60611
(312) 280 5223
Website: www.ashe.org

American Society for Testing and Materials
1916 Race Street
Philadelphia, PA 19103
(215) 299-5400
Website: www.astm.org

Association for the Advancement of
Medical Instrumentation
3330 Washington Boulevard
Suite 400, Arlington, VA 22201
(703) 525-4890
Website: www.aami.org

Australian Institute of Health and Welfare
GPO Box 570
Canberra, ACT 2601
Australia, (61) 06-243-5092
Website: www.aihw.gov.au

British Standards Institution
2 Park Street
London, W1A 2BS
United Kingdom
(44) 071-629-9000
Website: www.bsi.org.uk

Canadian Healthcare Association
17 York Street
Ottawa, ON K1N 9J6
Canada, (613) 241-8005
Website: www.canadian-healthcare.org

CSA International
178 Rexdale Boulevard
Etobicoke, ON M9W 1R3
Canada, (416) 747-4000
Website: www.csa-international.org

Center for Devices and Radiological Health
Food and Drug Administration
9200 Corporate Boulevard
Rockville, MD 20850
(301) 443-4690
Website: www.fda.gov/cdrh

Compressed Gas Association, Inc.
1725 Jefferson Davis Highway
Suite 1004, Arlington, VA 22202
(703) 412-0900

ECRI
5200 Butler Pike
Plymouth Meeting, PA 19462
(610) 825-6000; (610) 834-1275 (fax)
Websites: www.ecri.org; www.ecriy2k.org
www.mdsr.ecri.org

Environmental Health Directorate
Health Protection Branch
Health Canada
Environmental Health Centre
19th Floor, Jeanne Mance Building
Tunney's Pasture
Ottawa, ON K1A 0L2 Canada
(613) 957-3143
Website: www.hc-sc.gc.ca/hpb/index_e.html

Therapeutic Products Programme
Health Canada
Holland Cross, Tower B
2nd Floor, 1600 Scott Street
Address Locator #3102D1
Ottawa, ON K1A 1B6
(613) 954-0288
Website: www.hc-sc.gc.ca/hpb-dgps/therapeut

Food and Drug Administration
MEDWATCH, FDA Medical Products
Reporting Program
5600 Fishers Lane
Rockville, MD 20857-9787
(800) 332-1088
Website: www.fda.gov/cdrh/mdr.html

Institute of Electrical and Electronics Engineers
445 Hoes Lane
P.O. Box 1331
Piscataway, NJ 08850-1331
(732) 562-3800
Website: www.standards.ieee.org

International Electrotechnical Commission
Box 131
3 rue de Varembe, CH 1211
Geneva 20, Switzerland
(41) 022-919-0211
Website: www.iec.ch

International Organization for
Standardization
1 rue de Varembe
Case postale 56, CH 1211
Geneva 20
Switzerland
(41) 022-749-0111
Website: www.iso.ch

Joint Commission on Accreditation
of Healthcare Organizations
1 Renaissance Boulevard
Oakbrook Terrace, IL 60181
(630) 792-5600
Website: www.jcaho.org

Medical Devices Agency
Department of Health
Room 1209, Hannibal House
Elephant and Castle
London, SE1 6TQ
United Kingdom
(44) 171-972-8143
Website: www.medical-devices.gov.uk

National Council on Radiation
Protection and Measurements
7910 Woodmont Avenue, Suite 800
Bethesda, MD 20814
(310) 657-2652
Website: www.ncrp.com

National Fire Protection Association
1 Batterymarch Park
PO Box 9101
Quincy, MA 02269-9101
(617) 770-3000
Website: www.nfpa.org

Nuclear Regulatory Commission
11555 Rockville Pike, Rockville
MD 20852, (301) 492-7000
Website: www.nrc.gov

Occupational Safety and Health Administration
US Department of Labor
Office of Information and Consumer Affairs
200 Constitution Avenue, NW
Room N3647, Washington, DC 20210
(202) 219-8151
Website: www.osha.gov

ORKI
National Institute for Hospital and
Medical Engineering
Budapest dios arok 3, H-1125
Hungary, (33) 1-156-1522

Radiation Protection Branch
Environmental Health Directorate
Health Canada, 775 Brookfield Road
Ottawa, ON K1A 1C1
Website: www.hc-sc.gc.ca/ehp/ehd/rpb

Russian Scientific and Research Institute
Russian Public Health Ministry
EKRAN, 3 Kasatkina Street
Moscow, Russia 129301
(44) 071-405-3474

Society of Nuclear Medicine, Inc.
1850 Samuel Morse Drive
Reston, VA 20190-5316, (703) 708-9000
Website: www.snm.org

Standards Association of Australia
PO Box 1055, Strathfield
NSW 2135, Australia
(61) 02-9746-4700
Website: www.standards.org.au

Therapeutic Goods Administration
PO Box 100, Wooden, ACT 2606
Australia, (61) 2-6232-8610
Website: www.health.gov.au/tga

Underwriters Laboratories, Inc.
333 Pfingsten Road
Northbrook, IL 60062-2096
(847) 272-8800
Website: www.ul.com

VTT, Technical Research Center of Finland
Postbox 316
SF-33101 Tampere 10
Finland, (358) 31-163300
Website: www.vti.fi

80.2 Technology Assessment Agencies

Technology assessment is the practical process of determining the value of a new or emerging technology in and of itself or against existing or competing technologies using safety, efficacy, effectiveness, outcome, risk management, strategic, financial, and competitive criteria. Technology assessment also considers ethics and law as well as health priorities and cost-effectiveness compared to competing technologies. A "technology" is defined as devices, equipment, related software, drugs, biotechnologies, procedures, and therapies; and systems used to diagnose or treat patients. The processes of technology assessment are discussed in detail in Chapter 76.

Technology assessment is not the same as technology acquisition/procurement or technology planning. The latter two are processes for determining equipment vendors, soliciting bids, and systematically determining a hospital's technology related needs based on strategic, financial, risk management, and clinical criteria. The informational needs differ greatly between technology assessment and the acquisition/procurement or planning processes. This section focuses on the resources applicable to technology assessment.

Worldwide, there are nearly 400 organizations (private, academic, and governmental), providing technology assessment information, databases, or consulting services. Some are strictly information clearing houses, some perform technology assessment, and some do both. For those that perform assessments, the quality of the information generated varies greatly from superficial studies to in-depth, well referenced analytical reports. In 1997, the U.S. Agency for Health Care Policy and Research (AHCPR) designated 12 "Evidence-Based Practice Centers" (EPC) to undertake major technology assessment studies on a contract basis. Each of these EPCs are noted in the list below and general descriptions of each center may be viewed on the internet at the AHCPR Website http://www.ahcpr.gov/clinic/epc/.

Language limitations are a significant issue. In the ultimate analysis, the ability to undertake technology assessment requires assimilating vast amounts of information, most of which exists only in the English language. Technology assessment studies published by the International Society for Technology Assessment in Health Care (ISTAHC), by the World Health Organization, and other umbrella organizations are generally in English. The new International Health Technology Assessment database being developed by ECRI in conjunction with the U.S. National Library of Medicine contains more than 30,000 citations to technology assessments and related documents.

Below are the names, mailing addresses, and Internet Website addresses of some of the most prominent organizations undertaking technology assessment studies:

Agence Nationale pour le Develeppement
de l'Evaluation Medicale
159 Rue Nationale
Paris 75013
France
(33) 42-16-7272
Website: www.upml.fr/andem/andem.htm

Agencia de Evaluacion de
Technologias Sanitarias
Ministerio de Sanidad y Consumo
Instituto de Salud Carlos III, AETS
Sinesio Delgado 6, 28029 Madrid
Spain, (34) 1-323-4359
Website: www.isciii.es/aets

Agence Nationale pour
le Develeppement
de l'Evaluation Medicale
159 Rue Nationale
Paris 75013
France
(33) 42-16-7272
Website: www.upml.fr/andem/andem.htm

Alberta Heritage Foundation for
Medical Research
125 Manulife Place
10180-101 Street
Edmonton, AB T5J 345
(403) 423-5727
Website: www.ahfmr.ab.ca

American Association of Preferred
Provider Organizations
601 13th Street, NW
Suite 370 South
Washington, DC 20005
(202) 347-7600

American Academy of Neurology
1080 Montreal Avenue
St. Paul, MN 55116-2791
(612) 695-2716
Website: www.aan.com

American College of Obstetricians
and Gynecologists
409 12th Street, SW
Washington, DC 20024
(202) 863-2518
Website: www.acog.org

Australian Institute of
Health and Welfare
GPO Box 570
Canberra, ACT 2601
Australia
(61) 06-243-5092
Website: www.aihw.gov.au

Battelle Medical Technology
Assessment and Policy
Research Center (MEDTAP)
901 D Street, SW
Washington, DC 20024
(202) 479-0500
Website: www.battelle.org

Blue Cross and Blue Shield Association
Technology Evaluation Center
225 N Michigan Avenue
Chicago, IL 60601-7680
(312) 297-5530
(312) 297-6080 (publications)
Website: www.bluecares.com/new/clinical
(An EPC of AHCPR)

British Columbia Office of Health
Technology Assessment
Centre for Health Services & Policy Research,
University of British Columbia
429-2194 Health Sciences Mall
Vancouver, BC V6T 1Z3
Canada, (604) 822-7049
Website: www.chspr.ubc.ca

British Institute of Radiology
36 Portland Place
London, W1N 4AT
United Kingdom
(44) 171-580-4085
Website: www.bir.org.uk

Canadian Coordinating Office for
Health Technology Assessment
110-955 Green Valley Crescent
Ottawa ON K2C 3V4
Canada, (613) 226-2553
Website: www.ccohta.ca

Canadian Healthcare Association
17 York Street
Ottawa, ON K1N 9J6
Canada, (613) 241-8005
Website: www.canadian-healthcare.org

Catalan Agency for Health
Technology Assessment
Travessera de les Corts 131-159
Pavello Avenue
Maria, 08028 Barcelona
Spain, (34) 93-227-29-00
Website: www.aatm.es

Centre for Health Economics
University of York
York Y01 5DD
United Kingdom
(44) 01904-433718
Website: www.york.ac.uk

Center for Medical
Technology Assessment
Linköping University
5183 Linköping, Box 1026 (551-11)
Sweden, (46) 13-281-000

Center for Practice and Technology
Assessment Agency for Health
Care Policy and Research (AHCPR)
6010 Executive Boulevard, Suite 300
Rockville, MD 20852
(301) 594-4015
Website: www.ahcpr.gov

Committee for Evaluation and Diffusion
of Innovative Technologies
3 Avenue Victoria
Paris 75004, France
(33) 1-40-273-109

Conseil d'evaluation des technologies
de la sante du Quebec
201 Cremazie Boulevard East
Bur 1.01, Montreal
PQ H2M 1L2, Canada
(514) 873-2563
Website: www.msss.gouv.qc.ca

Danish Hospital Institute
Landermaerket 10
Copenhagen K
Denmark DK1119
(45) 33-11-5777

Danish Medical Research Council
Bredgade 43
1260 Copenhagen
Denmark
(45) 33-92-9700

Danish National Board of Health
Amaliegade 13, PO Box 2020
Copenhagen K, Denmark DK1012
(45) 35-26-5400

Duke Center for Clinical Health
Policy Research
Duke University Medical Center
2200 West Main Street, Suite 320
Durham, NC 27705
(919) 286-3399
Website: www.clinipol.mc.duke.edu
(An EPC of AHCPR)

ECRI
5200 Butler Pike
Plymouth Meeting, PA 19462
(610) 825-6000
(610) 834-1275 fax
Websites: www.ecri.org
www.ecriy2k.org
www.mdsr.ecri.org
(An EPC of AHCPR)

Finnish Office for Health Care
Technology Assessment
PO Box 220
FIN-00531 Helsinki
Finland, (35) 89-3967-2296
Website: www.stakes.fi/finohta

Frost and Sullivan, Inc.
106 Fulton Street
New York, NY 10038-2786
(212) 233-1080
Website: www.frost.com

Health Council of
the Netherlands
PO Box 1236
2280 CE, Rijswijk
The Netherlands
(31) 70-340-7520

Health Services Directorate
Strategies and Systems for Health
Health Promotion
Health Promotion and Programs Branch
Health Canada
1915B Tunney's Pasture
Ottawa, ON K1A 1B4
Canada, (613) 954-8629
Website: www.hc-sc.gc.ca/hppb/hpol

Health Technology Advisory Committee
121 East 7th Place, Suite 400
PO Box 64975
St. Paul, MN 55164-6358
(612) 282-6358

Hong Kong Institute of Engineers
9/F Island Centre
No. 1 Great George Street
Causeway Bay
Hong Kong

Institute for Clinical PET
7100-A Manchester Boulevard
Suite 300
Alexandria, VA 22310
(703) 924-6650
Website: www.icpet.org

Institute for Clinical
Systems Integration
8009 34th Avenue South
Minneapolis, MN 55425
(612) 883-7999
Website: www.icsi.org

Institute for Health Policy Analysis
8401 Colesville Road, Suite 500
Silver Spring, MD 20910
(301) 565-4216

Institute of Medicine (U.S.)
National Academy of Sciences
2101 Constitution Avenue, NW
Washington, DC 20418
(202) 334-2352
Website: www.nas.edu/iom

International Network of Agencies for
Health Technology Assessment
c/o SBU, Box 16158
S-103 24 Stockholm
Sweden, (46) 08-611-1913
Website: www.sbu.se/sbu-site/links/inahta

Johns Hopkins Evidence-based
Practice Center
The Johns Hopkins
Medical Institutions
2020 E Monument Street, Suite 2-600
Baltimore, MD 21205-2223
(410) 955-6953
Website: www.jhsph.edu/Departments/Epi/
(An EPC of AHCPR)

McMaster University Evidence-based
Practice Center
1200 Main Street West, Room 3H7
Hamilton, ON L8N 3Z5
Canada
(905) 525-9140 ext. 22520
Website: http://hiru.mcmaster.ca.epc/
(An EPC of AHCPR)

Medical Alley
1550 Utica Avenue, South
Suite 725
Minneapolis, MN 55416
(612) 542-3077
Website: www.medicalalley.org

Medical Devices Agency
Department of Health
Room 1209
Hannibal House
Elephant and Castle
London, SE1 6TQ
United Kingdom
(44) 171-972-8143
Website: www.medical-devices.gov.uk

Medical Technology Practice
Patterns Institute
4733 Bethesda Avenue, Suite 510
Bethesda, MD 20814
(301) 652-4005
Website: www.mtppi.org

MEDTAP International
7101 Wisconsin Avenue, Suite 600
Bethesda MD 20814
(301) 654-9729
Website: www.medtap.com

MetaWorks, Inc.
470 Atlantic Avenue
Boston, MA 02210
(617) 368-3573 ext. 206
Website: www.metawork.com
(An EPC of AHCPR)

National Institute of
Nursing Research, NIH
31 Center Drive
Room 5B10. MSC 2178
Bethesda, MD 20892-2178
(301) 496-0207
Website: www.nih.gov/ninr

National Commission on
Quality Assurance
2000 L Street NW, Suite 500
Washington. DC 20036
(202) 955-3500
Website: www.ncqa.org

National Committee of Clinical
Laboratory Standards (NCCLS)
940 West Valley Road, Suite 1400
Wayne, PA 19087-1898
(610) 688-0100
Website: www.nccls.org

National Coordinating Center for
Health Technology Assessment
Boldrewood (Mailpoint 728)
Univ of Southampton SO16 7PX
United Kingdom, (44) 170-359-5642
Website: www.soton.ac.uk/~hta/address.htm

National Health and Medical Research Council
GPO Box 9848
Canberra, ACT Australia
(61) 06-289-7019

New England Medical Center
Center for Clinical Evidence Synthesis
Division of Clinical Research
750 Washington Street, Box 63
Boston, MA 02111
(617) 636-5133
Website: www.nemc.org/medicine/ccr/cces.htm
(An EPC of AHCPR)

New York State Department of Health
Tower Building, Empire State Plaza
Albany, NY 12237
(518) 474-7354
Website: www.health.state.ny.us

NHS Centre for Reviews and Dissemination
University of York
York Y01 5DD, United Kingdom
(44) 01-904-433634
Website: www.york.ac.uk

Office of Medical Applications of Research
NIH Consensus Program Information Service
PO Box 2577, Kensington
MD 20891, (301) 231-8083
Website: odp.od.nih.gov/consensus

Ontario Ministry of Health
Hepburn Block
80 Grosvenor Street
10th Floor
Toronto, ON M7A 2C4
(416) 327-4377

Oregon Health Sciences University
Division of Medical Informatics
and Outcomes Research
3181 SW Sam Jackson Park Road
Portland, OR 97201-3098
(503) 494-4277
Website: www.ohsu.edu/epc
(An EPC of AHCPR)

Pan American Health Organization
525 23rd Street NW
Washington, DC 20037-2895
(202) 974-3222
Website: www.paho.org

Physician Payment Review
Commission (PPRC)
2120 L Street NW, Suite 510
Washington, DC 20037
(202) 653-7220

Prudential Insurance Company
of America Health Care Operations and
Research Division
56 N Livingston Avenue
Roseland, NJ 07068
(201) 716-3870

Research Triangle Institute
3040 Cornwallis Road
PO Box 12194
Research Triangle Park, NC 27709-2194
(919) 541-6512
(919) 541-7480
Website: www.rti.org/epc/
(An EPC of AHCPR)

San Antonio Evidence-based Practice Center
University of Texas Health Sciences Center
Department of Medicine
7703 Floyd Curl Drive
San Antonio, TX 78284-7879
(210) 617-5190
Website: www.uthscsa.edu/
(An EPC of AHCPR)

Saskatchewan Health
Acute and Emergency
Services Branch
3475 Albert Street
Regina, SK S4S 6X6
(306) 787-3656

Scottish Health Purchasing
Information Centre
Summerfield House
2 Eday Road
Aberdeen AB15 6RE
Scotland
United Kingdom
(44) 0-1224-663-456 ext. 75246
Website: www.nahat.net/shpic

Servicio de Evaluacion de
Technologias Sanitarias
Duque de Wellington 2
E01010 Vitoria-Gasteiz
Spain, (94) 518-9250
E-mail: osteba-san@ej-gv.es

Society of Critical Care Medicine
8101 E Kaiser Boulevard
Suite 300
Anaheim, CA 92808-2259
(714) 282-6000
Website: www.sccm.org

Swedish Council on Technology
Assessment in Health Care
Box 16158
S-103 24 Stockholm
Sweden
(46) 08-611-1913
Website: www.sbu.se

Southern California EPC-RAND
1700 Main Street
Santa Monica, CA 90401
(310) 393-0411 ext. 6669
Website: www.rand.org/organization/health/epc/
(An EPC of AHCPR)

Swiss Institute for Public
Health Technology Programme
Pfrundweg 14
CH-5001 Aarau
Switzerland
(41) 064-247-161

TNO Prevention and Health
PO Box 2215
2301 CE Leiden
The Netherlands
(31) 71-518-1818
Website: www.tno.n1/instit/pg/index.html

University HealthSystem Consortium
2001 Spring Road, Suite 700
Oak Brook, IL 60523
(630) 954-1700
Website: www.uhc.edu

University of Leeds
School of Public Health
30 Hyde Terrace
Leeds L52 9LN
United Kingdom
Website: www.leeds.ac.uk

USCF-Stanford University EPC
University of California
San Francisco
505 Parnassus Avenue
Room M-1490, Box 0132
San Francisco, CA 94143-0132
(415) 476-2564
Website: www.stanford.edu/group/epc/
(An EPC of AHCPR)

U.S. Office of Technology
Assessment (former address)
600 Pennsylvania Avenue SE
Washington, DC 20003
Note: OTA closed on 29 Sep, 1995.
However, documents can be
accessed via the internet at
www.wws.princeton.edu/§ota/html2/cong.html
Also a complete set of
OTA publications is
available on CD-ROM; contact
the U.S. Government Printing Office
(www.gpo.gov) for more information

Veterans Administration
Technology Assessment Program
VA Medical Center (152M)
150 S Huntington Avenue
Building 4
Boston, MA 02130
(617) 278-4469
Website: www.va.gov/resdev

Voluntary Hospitals of America, Inc.
220 East Boulevard
Irving, TX 75014
(214) 830-0000

Wessex Institute of Health Research
and Development
Boldrewood Medical School
Bassett Crescent East
Highfield, Southampton SO16 7PX
United Kingdom
(44) 01-703-595-661
Website: www.soton.ac.uk/∼wi/index.html

World Health Organization
Distribution Sales, CH 1211
Geneva 27, Switzerland 2476
(41) 22-791-2111
Website: www.who.ch
Note: Publications are also available from the
WHO Publications Center, USA,
at (518) 436-9686.

References

ECRI. *Healthcare Standards Official Directory*. ECRI, Plymouth Meeting, PA, 1999.

Eddy D.M. *A Manual for Assessing Health Practices & Designing Practice Policies: The Explicit Approach.* American College of Physicians, Philadelphia, PA, 1992.

Goodman C., Ed. *Medical Technology Assessment Directory.* National Academy Press, Washington, DC, 1988.

Marcaccio K.Y., ed. *Gale Directory of Databases. Volume 1*: *Online Databases.* Gale Research International, London, 1993.

van Nimwegen Chr., ed. *International List of Reports on Comparative Evaluations of Medical Devices.* TNO Centre for Medical Technology, Leiden, the Netherlands, 1993.

Further Information

A comprehensive listing of healthcare standards and the issuing organizations is presented in the Healthcare Standards Directory published by ECRI. The Directory is well organized by keywords, organizations and their standards, federal and state laws, legislation and regulations, and contains a complete index of names and addresses.

The International Health Technology Assessment database is produced by ECRI. A portion of the database is also available in the U.S. National Library of Medicine's new database called HealthSTAR. Internet access to HealthSTAR is through Website address http://igm.nlm.nih.gov. A description of the database may be found at http://www.nlm.nih.gov/pubs/factsheets/healthstar.html.

81

Applications of Virtual Instruments in Health Care

Eric Rosow
Hartford Hospital
Premise Development Corporation

Joseph Adam
Premise Development Corporation

81.1 Applications of Virtual Instruments in Health Care

Virtual Instrumentation (which was previously defined in Chapter 73, "*Virtual Instrumentation: Applications in Biomedical Engineering*") allows organizations to effectively harness the power of the PC to access, analyze, and share information throughout the organization. With vast amounts of data available from increasingly sophisticated enterprise-level data sources, potentially useful information is often left hidden due to a lack of useful tools. Virtual instruments can employ a wide array of technologies such as multidimensional analyses and Statistical Process Control (SPC) tools to detect patterns, trends, causalities, and discontinuities to derive knowledge and make informed decisions.

Today's enterprises create vast amounts of raw data and recent advances in storage technology, coupled with the desire to use this data competitively, has caused a data glut in many organizations. The healthcare industry in particular is one that generates a tremendous amount of data. Tools such as databases and spreadsheets certainly help manage and analyze this data; however databases, while ideal for extracting data are generally not suited for graphing and analysis. Spreadsheets, on the other hand, are ideal for analyzing and graphing data, but this can often be a cumbersome process when working with multiple data files. Virtual instruments empower the user to leverage the best of both worlds by creating a suite of user-defined applications which allow the end-user to convert vast amounts of data into information which is ultimately transformed into knowledge to enable better decision making.

This chapter will discuss several virtual instrument applications and tools that have been developed to meet the specific needs of healthcare organizations. Particular attention will be placed on the use of quality control and "performance indicators" which provide the ability to trend and forecast various metrics. The use of SPC within virtual instruments will also be demonstrated. Finally, a nontraditional application of virtual instrumentation will be presented in which a "peer review" application has been developed to allow members of an organization to actively participate in the Employee Performance Review process.

81.1.1 Example Application #1: The EndoTester™ — A Virtual Instrument-Based Quality Control and Technology Assessment System for Surgical Video Systems

The use of endoscopic surgery is growing, in large part because it is generally safer and less expensive than conventional surgery, and patients tend to require less time in a hospital after endoscopic surgery. Industry experts conservatively estimate that about 4 million minimally invasive procedures were performed in 1996. As endoscopic surgery becomes more common, there is an increasing need to accurately evaluate the performance characteristics of endoscopes and their peripheral components.

The assessment of the optical performance of laparoscopes and video systems is often difficult in the clinical setting. The surgeon depends on a high quality image to perform minimally invasive surgery, yet assurance of proper function of the equipment by biomedical engineering staff is not always straightforward. Many variables in both patient and equipment may result in a poor image. Equipment variables, which may degrade image quality, include problems with the endoscope, either with optics or light transmission. The light cable is another source of uncertainty as a result of optical loss from damaged fibers. Malfunctions of the charge coupled device (CCD) video camera are yet another source of poor image quality. Cleanliness of the equipment, especially lens surfaces on the endoscope (both proximal and distal ends) are particularly common problems. Patient variables make the objective assessment of image quality more difficult. Large operative fields and bleeding at the operative site are just two examples of patient factors that may affect image quality.

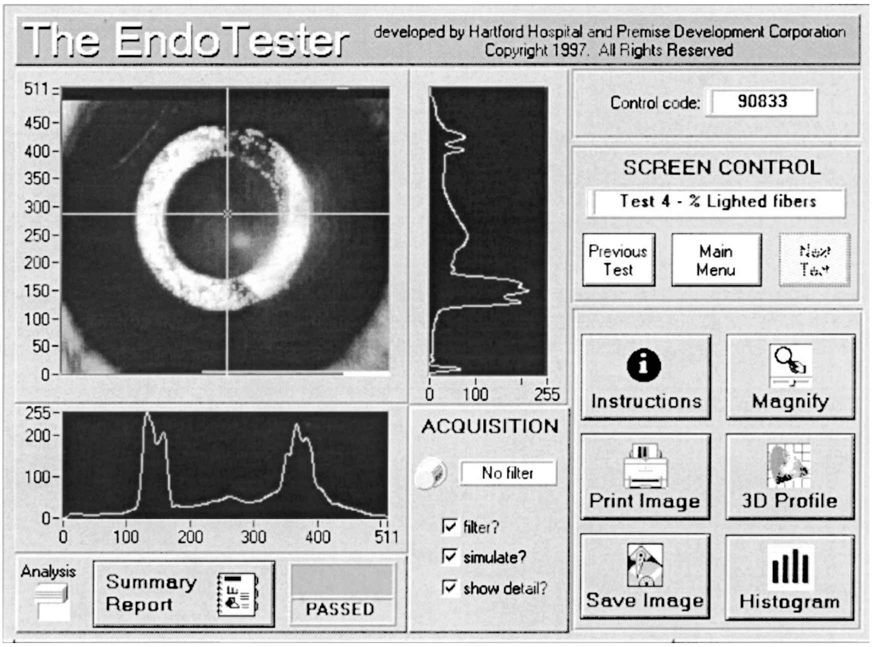
The evaluation of new video endoscopic equipment is also difficult because of the lack of objective standards for performance. Purchasers of equipment are forced to make an essentially subjective decision about image quality. By employing virtual instrumentation, a collaborative team of biomedical engineers, software engineers, physicians, nurses, and technicians at Hartford Hospital (Hartford, CT) and Premise Development Corporation (Avon, CT) have developed an instrument, the EndoTester™, with integrated software to quantify the optical properties of both rigid and flexible fiberoptic endoscopes. This easy-to-use optical evaluation system allows objective measurement of endoscopic performance prior to equipment purchase and in routine clinical use as part of a program of prospective maintenance.

The EndoTester™ was designed and fabricated to perform a wide array of quantitative tests and measurements. Some of these tests include (1) Relative light loss, (2) Reflective symmetry, (3) Lighted (good) fibers, (4) Geometric distortion, and (5) Modulation transfer function (MTF). Each series of tests is associated with a specific endoscope to allow for trending and easy comparison of successive measurements.

Specific information about each endoscope (i.e., manufacturer, diameter, length, tip angle, department/unit, control number, and operator), the reason for the test (i.e., quality control, pre/post repair, etc.), and any problems associated with the scope are also documented through the electronic record. In addition, all the quantitative measurements from each test are automatically appended to the electronic record for life-cycle performance analysis.

Figure 81.1 and Figure 81.2 illustrate how information about the fiberoptic bundle of an endoscope can be displayed and measured. This provides a record of the pattern of lighted optical fibers for the endoscope under test. The number of lighted pixels will depend on the endoscope's dimensions, the distal end geometry, and the number of failed optical fibers. New fiber damage to an endoscope will be apparent by comparison of the lighted fiber pictures (and histogram profiles) from successive tests. Statistical data is also available to calculate the percentage of working fibers in a given endoscope.

In addition to the two-dimensional profile of lighted fibers, this pattern (and all other image patterns) can also be displayed in the form of a three-dimensional contour plot. This interactive graph may be viewed from a variety of viewpoints in that the user can vary the elevation, rotation, size, and perspective controls.



The EndoTester developed by Hartford Hospital and Premise Development Corporation Copyright 1997. All Rights Reserved

Control code: 90833

SCREEN CONTROL

Test 4 - % Lighted fibers

ACQUISITION

No filter

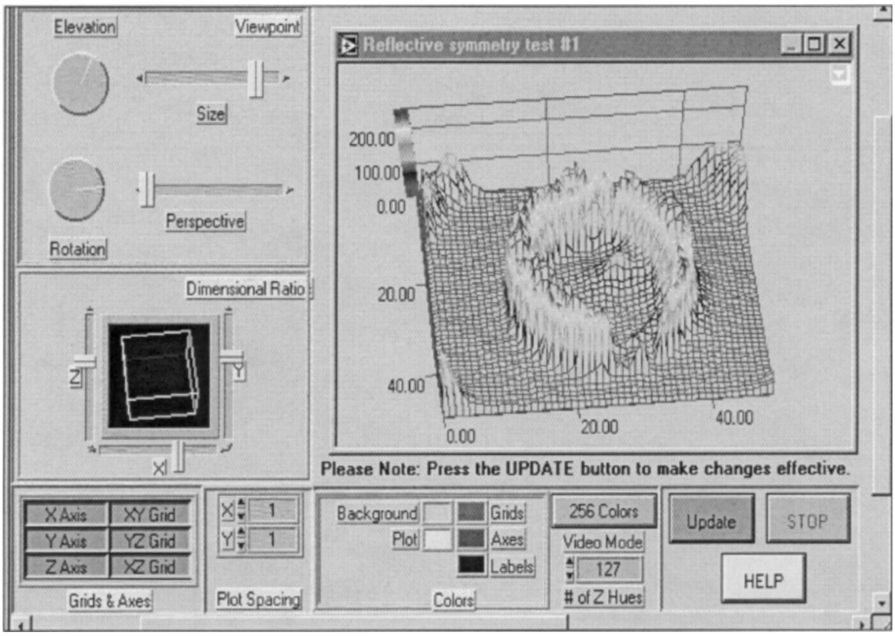
☑ filter?

☑ simulate?

☑ show detail?

PASSED

FIGURE 81.1 Endoscope tip reflection.



Reflective symmetry test #1

Please Note: Press the UPDATE button to make changes effective.

X 1

Y 1

256 Colors

Video Mode

127

# of Z Hues

FIGURE 81.2 Endoscope profiling module.

Figure 81.2 illustrates how test images for a specific scope can be profiled over time (i.e., days, months, years) to identify degrading performance. This profile is also useful to validate repair procedures by comparing test images before and after the repair.

The EndoTester™ has many applications. In general, the most useful application is the ability to objectively measure an endoscope's performance prior to purchase, and in routine clinical use as part of a program of prospective maintenance. Measuring parameters of scope performance can facilitate equipment purchase. Vendor claims of instrument capabilities can be validated as a part of the negotiation process. Commercially available evaluation systems (for original equipment manufacturers) can cost upward of $50,000, yet by employing the benefits of virtual instrumentation and a standard PC, an affordable, yet highly accurate test system for rigid and flexible fiberoptic endoscopes can now be obtained by clinical institutions.

In addition to technology assessment applications, the adoption of disposable endoscopes raises another potential use for the EndoTester™. Disposable scopes are estimated to have a life of 20 to 30 procedures. However, there is no easy way to determine exactly when a scope should be "thrown away." The EndoTester™ could be used to define this end-point.

The greatest potential for this system is as part of a program of preventive maintenance. Currently, in most operating rooms, endoscopes are removed from service and sent for repair when they fail in clinical use. This causes operative delay with attendant risk to the patient and an increase in cost to the institution. The problem is difficult because an endoscope may be adequate in one procedure but fail in the next which is more exacting due to clinical variables such as large patient size or bleeding. Objective assessment of endoscope function with the EndoTester™ may eliminate some of these problems.

Equally as important, an endoscope evaluation system will also allow institutions to ensure value from providers of repair services. The need for repair can be better defined and the adequacy of the repair verified when service is completed. This ability becomes especially important as the explosive growth of minimally invasive surgery has resulted in the creation of a significant market for endoscope repairs and service. Endoscope repair costs vary widely throughout the industry with costs ranging from $500 to 1500 or more per repair. Inappropriate or incomplete repairs can result in extending surgical time by requiring the surgeon to "switch scopes" (in some cases several times) during a surgical procedure.

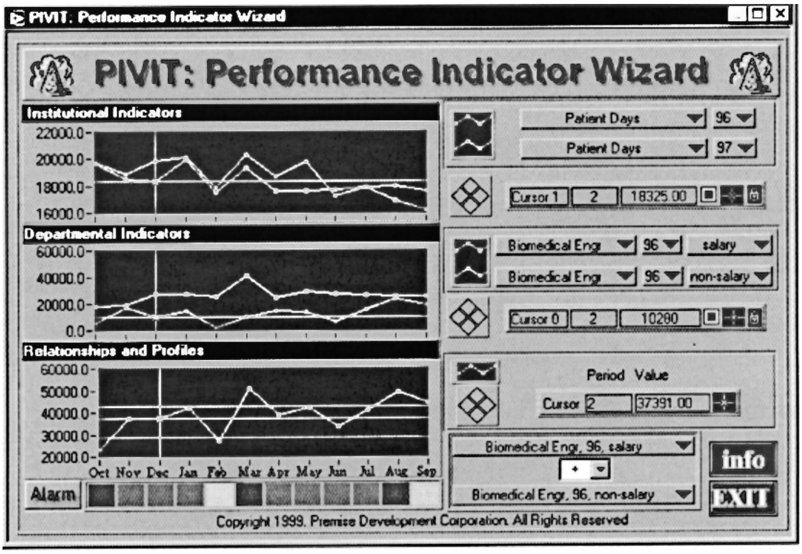
Given these applications, we believe that the EndoTester™ can play an important role in reducing unnecessary costs, while at the same time improving the quality of the endoscopic equipment and the outcome of its utilization. It is the sincere hope of the authors that this technology will help to provide accurate, affordable and easy-to-acquire data on endoscope performance characteristics which clearly are to the benefit of the healthcare provider, the ethical service providers, manufacturers of quality products, the payers, and, of course, the patient.

81.1.2 Example Application #2: PIVIT™ — Performance Indicator Virtual Instrument Toolkit

Most of the information management examples presented in this chapter are part of an application suite called PIVIT™. PIVIT is an acronym for "Performance Indicator Virtual Instrument Toolkit" and is an easy-to-use data acquisition and analysis product. PIVIT was developed specifically in response to the wide array of information and analysis needs throughout the healthcare setting.

The PIVIT applies virtual instrument technology to assess, analyze, and forecast clinical, operational, and financial performance indicators. Some examples include applications which profile institutional indicators (i.e., patient days, discharges, percent occupancy, ALOS, revenues, expenses, etc.), and departmental indicators (i.e., salary, nonsalary, total expenses, expense per equivalent discharge, DRGs, etc.). Other applications of PIVIT include 360° Peer Review, Customer Satisfaction Profiling, and Medical Equipment Risk Assessment.

The PIVIT can access data from multiple data sources. Virtually any parameter can be easily accessed and displayed from standard spreadsheet and database applications (i.e., Microsoft Access, Excel, Sybase, Oracle, etc.) using Microsoft's Open Database Connectivity (ODBC) technology. Furthermore, multiple parameters can be profiled and compared in real-time with any other parameter via interactive polar plots and three-dimensional displays. In addition to real-time profiling, other analyses such as SPC can be employed to view large data sets in a graphical format. SPC has been applied successfully for decades to help companies reduce variability in manufacturing processes. These SPC tools range from Pareto graphs



Patient Days 96

Patient Days 97

Cursor 1 | 2 | 18325.00

Biomedical Engr 96 salary

Biomedical Engr 96 non-salary

Cursor 0 | 2 | 10280

Cursor 2 | 37391.00

Biomedical Engr, 96, salary

+

Biomedical Engr, 96, non-salary

Copyright 1999. Premise Development Corporation. All Rights Reserved

FIGURE 81.3 PIVIT™ — Performance Indicator Wizard displays institutional and departmental indicators.

to Run and Control charts. Although it will not be possible to describe all of these applications, several examples are provided below to illustrate the power of PIVIT.

81.1.3 Trending, Relationships, and Interactive Alarms

Figure 81.3 illustrates a virtual instrument that interactively accesses institutional and department specific indicators and profiles them for comparison. Data sets can be acquired directly from standard spreadsheet and database applications (i.e., Microsoft Access®, Excel®, Sybase®, Oracle®, etc.). This capability has proven to be quite valuable with respect to quickly accessing and viewing large sets of data. Typically, multiple data sets contained within a spreadsheet or database had to be selected and then a new chart of this data had to be created. Using PIVIT, the user simply selects the desired parameter from any one of the pull-down menus and this data set is instantly graphed and compared to any other data set.

Interactive "threshold cursors" dynamically highlight when a parameter is over and/or under a specific target. Displayed parameters can also be ratios of any measured value, for example, "Expense per Equivalent Discharge" or "Revenue to Expense Ratio." The indicator color will change based on how far the data value exceeds the threshold value (i.e., from green to yellow to red). If multiple thresholds are exceeded, then the entire background of the screen (normally gray) will change to red to alert the user of an extreme condition.

Finally, multimedia has been employed by PIVIT to alert designated personnel with an audio message from the personal computer or by sending an automated message via e-mail, fax, pager, or mobile phone.

The PIVIT also has the ability to profile historical trends and project future values. Forecasts can be based on user-defined history (i.e., "Months for Regression"), the type of regression (i.e., linear, exponential, or polynomial), the number of days, months, or years to forecast, and if any offset should be applied to the forecast. These features allow the user to create an unlimited number of "what if" scenarios and allow only the desired range of data to be applied to a forecast. In addition to the graphical display of data values, historical and projected tables are also provided. These embedded tables look and function very much like a standard spreadsheet.

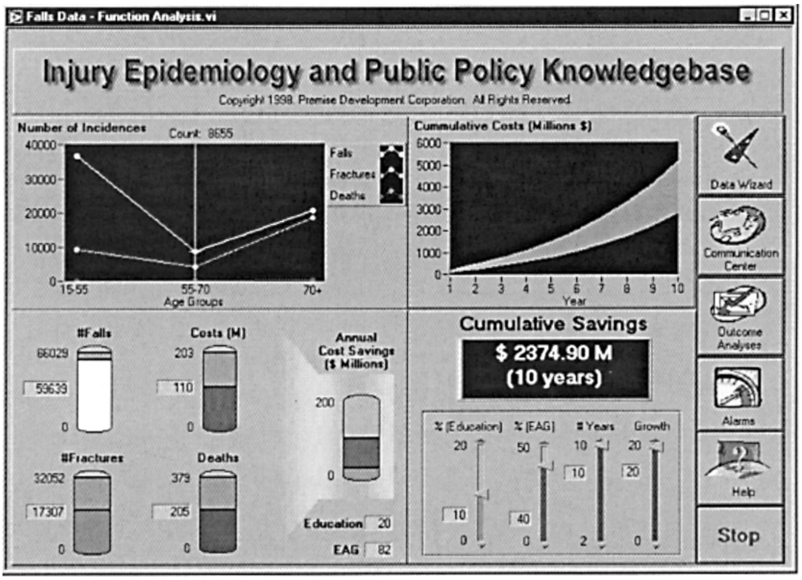


FIGURE 81.4 Injury epidemiology and public policy knowledgebase.

81.1.4 Data Modeling

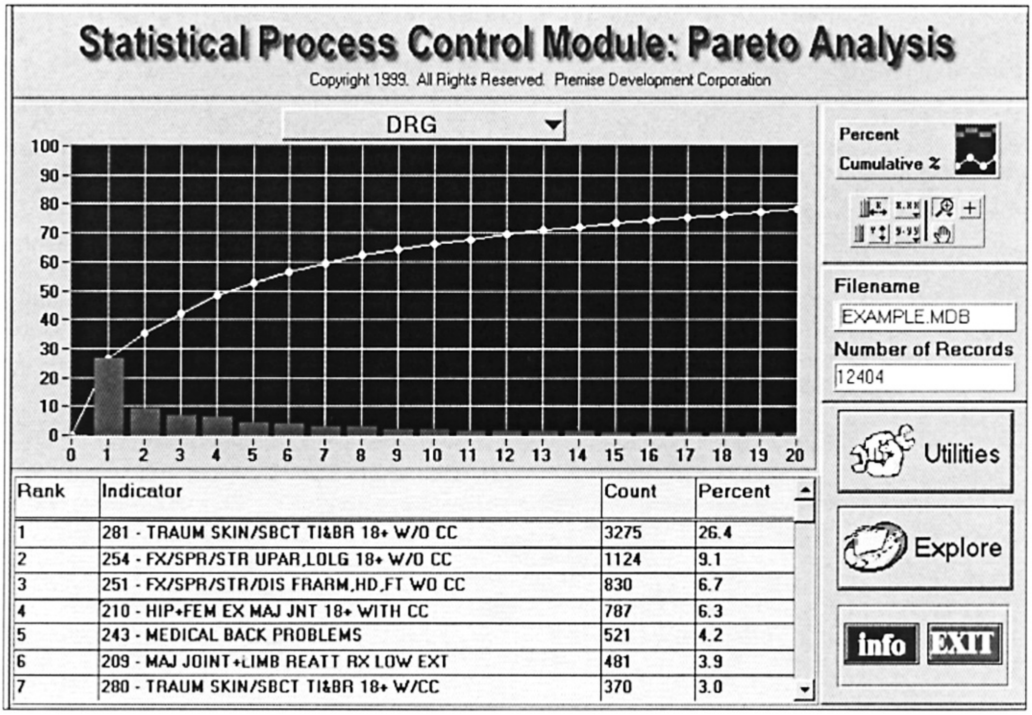
Figure 81.4 illustrates another example of how virtual instrumentation can be applied to financial modeling and forecasting. This example graphically profiles the annual morbidity, mortality, and cost associated with falls within the state of Connecticut. Such an instrument has proved to be an extremely effective modeling tool due to its ability to interactively highlight relationships and assumptions, and to project the cost and/or savings of employing educational and other interventional programs.

Virtual instruments such as these are not only useful with respect to modeling and forecasting, but perhaps more importantly, they become a "knowledgebase" in which interventions and the efficacy of these interventions can be statistically proven. In addition, virtual instruments can employ standard technologies such as Dynamic Data Exchange (DDE), ActiveX, or TCP/IP to transfer data to commonly used software applications such as Microsoft Access® or Microsoft Excel®. In this way, virtual instruments can measure and graph multiple signals while at the same time send this data to another application which could reside on the network or across the Internet.

Another module of the PIVIT application is called the "Communications Center." This module can be used to simply create and print a report or it can be used to send e-mail, faxes, messages to a pager, or even leave voice-mail messages. This is a powerful feature in that information can be easily and efficiently distributed to both individuals and groups in real-time.

Additionally, Microsoft Agent® technology can be used to pop-up an animated help tool to communicate a message, indicate an alarm condition, or can be used to help the user solve a problem or point out a discrepancy that may have otherwise gone unnoticed. Agents employ a "text-to-speech" algorithm to actually "speak" an analysis or alarm directly to the user or recipient of the message. In this way, on-line help and user support can also be provided in multiple languages.

In addition to real-time profiling of various parameters, more advanced analyses such as SPC can be employed to view large data sets in a graphical format. SPC has been applied successfully for decades to help companies reduce variability in manufacturing processes. It is the opinion of this author that SPC has enormous applications throughout healthcare. For example, Figure 81.5 shows how Pareto analysis can be applied to a sample trauma database of over 12,000 records. The Pareto chart may be frequency or percentage depending on front panel selection and the user can select from a variety of different parameters by clicking on the "pull-down" menu. This menu can be configured to automatically display each database



Filename: EXAMPLE.MDB

Number of Records: 12404

Rank	Indicator	Count	Percent
1	281 - TRAUM SKIN/SBCT TI&BR 18+ W/O CC	3275	26.4
2	254 - FX/SPR/STR UPAR,LOLG 18+ W/O CC	1124	9.1
3	251 - FX/SPR/STR/DIS FRARM,HD,FT WO CC	830	6.7
4	210 - HIP+FEM EX MAJ JNT 18+ WITH CC	787	6.3
5	243 - MEDICAL BACK PROBLEMS	521	4.2
6	209 - MAJ JOINT+LIMB REATT RX LOW EXT	481	3.9
7	280 - TRAUM SKIN/SBCT TI&BR 18+ W/CC	370	3.0

FIGURE 81.5 Statistical process control — Pareto analysis of a sample trauma registry.

field directly from the database. In this example, various database fields (i.e., DRG, Principal Diagnosis, Town, Payer, etc.) can be selected for Pareto analysis. Other SPC tools include run charts, control charts, and process capability distributions.

81.1.5 Medical Equipment Risk Criteria

Figure 81.6 illustrates a virtual instrument application which demonstrates how four "static" risk categories (and their corresponding values) are used to determine the inclusion of clinical equipment in the Medical Equipment Management Program at Hartford Hospital. Each risk category includes specific sub-categories that are assigned points, which when added together according to the formula listed below, yield a total score which ranges from 4 to 25.

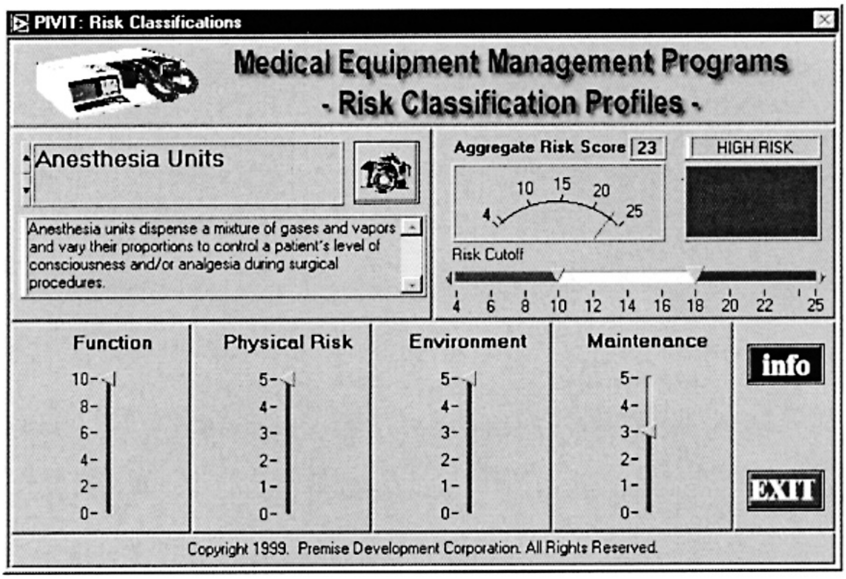
Considering these scores, the equipment is categorized into five priority levels (High, Medium, Low, Grey List, and Non-Inclusion into the Medical Equipment Management Program). The four static risk categories are:

Equipment function (EF): Stratifies the various functional categories (i.e., therapeutic, diagnostic, analytical, and miscellaneous) of equipment. This category has "point scores" which range from 1 (miscellaneous, non-patient related devices) to 10 (therapeutic, life support devices)

Physical risk (PR): Lists the "worst case scenario" of physical risk potential to either the patient or the operator of the equipment. This category has "point scores" which range from 1 (no significant identified risk) to 5 (potential for patient and/or operator death)

Environmental use classification (EC): Lists the primary equipment area in which the equipment is used and has "point scores" which range from 1 (non-patient care areas) to 5 (anesthetizing locations)

Preventive maintenance requirements (MR): Describes the level and frequency of required maintenance and has "point scores" which range from 1 (not required) to 5 (monthly maintenance)


PIVIT: Risk Classifications

Medical Equipment Management Programs
- Risk Classification Profiles -

Anesthesia Units

Anesthesia units dispense a mixture of gases and vapors and vary their proportions to control a patient's level of consciousness and/or analgesia during surgical procedures.

Aggregate Risk Score 23

HIGH RISK

Copyright 1999. Premise Development Corporation. All Rights Reserved.

FIGURE 81.6 Medical equipment risk classification profiler.

The aggregate static risk score is calculated as follows:

$$\text{Aggregate Risk Score} = \text{EF} + \text{PR} + \text{EC} + \text{MR} \tag{81.1}$$

Using the criteria's system described above, clinical equipment is categorized according to the following priority of testing and degree of risk:

High risk: Equipment that scores between and including 18 to 25 points on the criteria's evaluation system. This equipment is assigned the highest risk for testing, calibration, and repair

Medium risk: Equipment that scores between and including 15 to 17 points on the criteria's evaluation system

Low risk: Equipment that scores between and including 12 to 14 points on the criteria's evaluation system

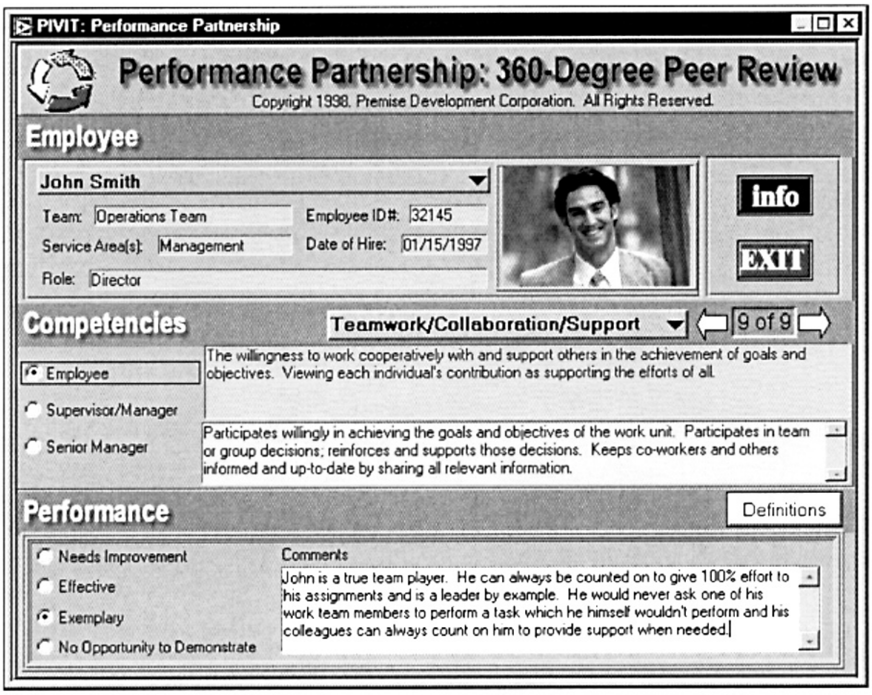
Hazard surveillance (gray): Equipment that scores between and including 6 and 11 points on the criteria's evaluation system is visually inspected on an annual basis during the hospital hazard surveillance rounds

Medical equipment management program deletion: Medical equipment and devices that pose little risk and scores less than 6 points may be deleted from the management program as well as the clinical equipment inventory

Future versions of this application will also consider "dynamic" risk factors such as: user error, mean-time-between failure (MTBF), device failure within 30 days of a preventive maintenance or repair, and the number of years beyond the American Hospital Association's recommended useful life

81.1.6 Peer Performance Reviews

The virtual instrument shown in Figure 81.7 has been designed to easily acquire and compile performance information with respect to institution-wide competencies. It has been created to allow every member of a team or department to participate in the evaluation of a co-worker (360° peer review). Upon running the application, the user is presented with a "Sign-In" screen where he or she enters their username and password. The application is divided into three components. The first (top section) profiles the employee



Performance Partnership: 360-Degree Peer Review

Copyright 1998. Premise Development Corporation. All Rights Reserved.

Employee

John Smith

Team: Operations Team

Employee ID#: 32145

Service Area(s): Management

Date of Hire: 01/15/1997

Role: Director

Competencies

Teamwork/Collaboration/Support

9 of 9

● Employee

○ Supervisor/Manager

○ Senior Manager

The willingness to work cooperatively with and support others in the achievement of goals and objectives. Viewing each individual's contribution as supporting the efforts of all.

Participates willingly in achieving the goals and objectives of the work unit. Participates in team or group decisions; reinforces and supports those decisions. Keeps co-workers and others informed and up-to-date by sharing all relevant information.

Performance

○ Needs Improvement

○ Effective

● Exemplary

○ No Opportunity to Demonstrate

Comments

John is a true team player. He can always be counted on to give 100% effort to his assignments and is a leader by example. He would never ask one of his work team members to perform a task which he himself wouldn't perform and his colleagues can always count on him to provide support when needed.

FIGURE 81.7 Performance reviews using virtual instrumentation.

and relevant service information. The second (middle section) indicates each competency as defined for employees, managers, and senior managers. The last (bottom) section allows the reviewer to evaluate performance by selecting one of four "radio buttons" and also provide specific comments related to each competency. This information is then compiled (with other reviewers) as real-time feedback.

References

[1] American Society for Quality Control. American National Standard. Definitions, Symbols, Forumulas, and Tables for Control Charts, 1987.

[2] Breyfogle, F.W. *Statistical Methods for Testing, Development and Manufacturing*, John Wiley & Sons, New York, 1982.

[3] Carey, R.G. and Lloyd, R.C. *Measuring Quality Improvement in Healthcare: A Guide to Statistical Process Control Applications*, 1995.

[4] Fennigkow, L. and Lagerman, B. Medical Equipment Management. 1997 EC/PTSM Series/No. 1; Joint Commission on Accreditation of Hospital Organizations, 1997, pp. 47–54.

[5] Frost & Sullivan Market Intelligence, file 765, The Dialog Corporation, Worldwide Headquarters, 2440 W. El Camino Real, Mountain View, CA 94040.

[6] Inglis, A. *Video Engineering*, McGraw Hill, New York, 1993.

[7] Kutzner J., Hightower L., and Pruitt C. Measurement and Testing of CCD Sensors and Cameras, *SMPTE Journal*, 325–327, 1992.

[8] Measurement of Resolution of Camera Systems, *IEEE Standard* 208, 1995.

[9] Medical Device Register 1997, Volume 2, Montvale NJ, Medical Economics Data Production Company, 1997.

[10] Montgomery, D.C. *Introduction to Statistical Quality Control*, 2nd ed., John Wiley & Sons, New York, 1992.

[11] Rosow, E. Virtual Instrumentation Applications with BioSensors, presented at the *Biomedical Engineering Consortium for Connecticut (BEACON) Biosensor Symposium*, Trinity College, Hartford, CT, October 2, 1998.

[12] Rosow, E., Adam, J., and Beatrice, F. The EndoTester': A Virtual Instrument Endoscope Evaluation System for Fiberoptic Endoscopes, *Biomedical Instrumentation and Technology*, 480–487, September/October 1998.

[13] Surgical Video Systems, *Health Devices*, 24, 428–457, 1995.

[14] Walker, B. *Optical Engineering Fundamentals*, McGraw Hill, New York, 1995.

[15] Wheeler, D.J. and Chambers, D.S. *Understanding Statistical Process Control*, 2nd ed., SPC Press, 1992.

VIII

Ethical Issues Associated with the Use of Medical Technology

Subrata Saha
Clemson University

Joseph D. Bronzino
Trinity College/Biomedical Engineering Alliance for Connecticut (BEACON)

BIOMEDICAL ENGINEERING IS RESPONSIBLE for many of the recent advances in modern medicine. These development have led to new treatment modalities that have significantly improved not only medical care, but the quality of life for many patients in our society. However, along with such positive outcomes new ethical dilemmas and challenges have also emerged. These include (1) involvement of humans in clinical research, (2) definition of death and the issue of euthanasia, (3) animal experimentation and human trials for new medical devices, (4) patient access to sophisticated and high cost medical technology, (5) regulation of new biomaterials and devices. With these issues in mind, this section discusses some of these topics. The first chapter focuses on the concept of professional ethics and its importance to the practicing biomedical engineer. The second chapter deals with the role medical technology has played in the definition of death and the dilemmas posed by advocates of euthanasia. The third chapter focuses on the use of animals and humans in research and clinical experimentation. The final chapter addresses the issue of regulating the use of devices, materials, etc. in the care of patients.

Since the space allocated in the handbook is limited, a complete discussion of many ethical dilemmas encountered by practicing biomedical engineers is beyond the scope of this section. Therefore, it is our sincere hope that the readers of this handbook will further explore these ideas from other texts and articles, some of which are references at the end of the chapter. Clearly, a course on biomedical ethics should be an essential component of any bioengineering curriculum.

With new developments in biotechnology and genetic engineering, we need to ask ourselves not only if we can do it, but also "should it be done?" As professional engineers we also have an obligation to educate the public and other regulatory agencies regarding the social implications of such new developments. It is our hope that the topics covered in this can provide an impetus for further discussion of the ethical issues and challenges faced by the bioengineer during the course of his/her professional life.

82

Beneficence, Nonmaleficence, and Medical Technology

Joseph D. Bronzino
Trinity College

Two moral norms have remained relatively *constant across* the various moral codes and oaths that have been formulated for health-care deliverers since the beginning of Western medicine in classical Greek civilization, namely beneficence — the provision of benefits — and nonmaleficence — the avoidance of doing harm. These norms are traced back to a body of writings from classical antiquity known as the *Hippocratic Corpus*. Although these writings are associated with the name of Hippocrates, the acknowledged founder of Western medicine, medical historians remain uncertain whether any, including the *Hippocratic Oath*, were actually his work. Although portions of the Corpus are believed to have been authored during the 6th century B.C., other portions are believed to have been written as late as the beginning of the Christian Era. Medical historians agree, though, that many of the specific moral directives of the *Corpus* represent neither the actual practices nor the moral ideals of the majority of physicians of ancient Greece and Rome.

Nonetheless, the general injunction, *"As to disease, make a habit of two things — to help or, at least, to do no harm,"* was accepted as a fundamental medical ethical norm by at least some ancient physicians. With the decline of Hellenistic civilization and the rise of Christianity, beneficence and nonmaleficence became increasingly accepted as the fundamental principles of morally sound medical practice. Although beneficence and nonmaleficence were regarded merely as concomitant to the craft of medicine in classical Greece and Rome, the emphasis upon compassion and the brotherhood of humankind, central to Christianity, *increasingly made* these norms the only acceptable motives for medical practice. Even today the provision of benefits and the avoidance of doing harm are stressed just as much in virtually all contemporary Western codes of conduct for health professionals as they were in the oaths and codes that guided the health-care providers of past centuries.

Traditionally, the ethics of medical care have given greater prominence to nonmaleficence than to beneficence. This priority was grounded in the fact that, historically, medicine's capacity to do harm far exceeded its capacity to protect and restore health. Providers of health care possessed many treatments that posed clear and genuine risks to patients but that offered little prospect of benefit. Truly effective therapies were all too rare. In this context, it is surely rational to give substantially higher priority to avoiding harm than to providing benefits.

The advent of modern science changed matters dramatically. Knowledge acquired in laboratories, tested in clinics, and verified by statistical methods has increasingly dictated the practices of medicine. This *ongoing alliance* between medicine and science became a critical source of the plethora of technologies that now pervades medical care. The impressive increases in therapeutic, preventive, and rehabilitative capabilities that these technologies have provided have pushed beneficence to the forefront of medical morality. Some have even gone so far as to hold that the old medical ethic of *"Above all, do no harm"* should be superseded by the new ethic that *"The patient deserves the best."* However, the rapid advances in medical technology capabilities have also produced great uncertainty as to what is most beneficial or least harmful for the patient. In other words, along with increases in ability to be beneficent, medicine's technology has generated much debate about what actually counts as beneficent or nonmaleficent treatment. To illustrate this point, let us turn to several specific moral issues posed by the use of medical technology [Bronzino, 1992, 1999].

82.1 Defining Death: A Moral Dilemma Posed by Medical Technology

Supportive and resuscitative devices, such as the respirator, found in the typical modern intensive care unit provide a useful starting point for illustrating how technology has rendered medical morality more complex and problematic. Devices of this kind allow clinicians to sustain respiration and circulation in patients who have suffered massive brain damage and total permanent loss of brain *function.* These technologies force us to ask: precisely when does a human life end? When is a human being indeed dead? This is not the straightforward factual matter it may appear to be. All of the relevant facts may show that the patient's brain has suffered injury grave enough to destroy its functioning forever. The facts may show that such an individual's circulation and respiration would permanently cease without artificial support. Yet these facts do not determine whether treating such an individual as a corpse is morally appropriate. To know this, it is necessary to know or perhaps to decide on those features of living persons that are essential to their status as "living persons." It is necessary to know or decide which human qualities, if irreparably lost, make an individual identical in all morally relevant respects to a corpse. Once those qualities have been specified, deciding whether total and irreparable loss of brain function constitutes death becomes a straightforward factual matter. Then, it would simply have to be determined if such loss itself deprives the individual of those qualities. If it does, the individual is morally identical to a corpse. If not, then the individual must be regarded and treated as a living person.

The traditional criterion of death has been irreparable cessation of heart beat, respiration, and blood pressure. This criterion would have been quickly met by anyone suffering massive trauma to the brain prior to the development of modern supportive technology. Such technology allows indefinite artificial maintenance of circulation and respiration and, thus, forestalls what once was an inevitable consequence of severe brain injury. The existence and use of such technology therefore challenges the traditional criterion of death and forces us to consider whether continued respiration and circulation are in themselves sufficient to distinguish a living individual from a corpse. Indeed, total and irreparable loss of brain function, referred to as "brainstem death," "whole brain death," and, simply, "brain death," has been widely accepted as the legal standard for death. By this standard, an individual in a state of brain death is legally indistinguishable from a corpse and may be legally treated as one even though respiratory and circulatory functions may be sustained through the intervention of technology. Many take this legal standard to be the morally appropriate one, noting that once destruction of the brain stem has occurred, the brain *cannot function* at all, and the body's regulatory mechanisms will fail unless artificially sustained. Thus, mechanical sustenance of an individual in a state of brain death is merely postponement of the inevitable and sustains nothing of the personality, character, or consciousness of the individual. It is merely the mechanical intervention that differentiates such an individual from a corpse and a mechanically ventilated corpse is a corpse nonetheless.

Even with a consensus that brainstem death is death and thus that an individual in such a state is indeed a corpse, hard cases remain. Consider the case of an individual in a persistent vegetative state, the condition known as "neocortical death." Although severe brain injury has been suffered, enough brain function remains to make mechanical *sustenance* of respiration and circulation unnecessary. In a persistent vegetative state, an individual exhibits no purposeful response to external stimuli and no evidence of self-awareness. The eyes may open periodically and the individual may exhibit sleep–wake cycles. Some patients even yawn, make chewing motions, or swallow spontaneously. Unlike the complete unresponsiveness of individuals in a state of brainstem death, a variety of simple and complex responses can be elicited from an individual in a persistent vegetative state. Nonetheless, the chances that such an individual will regain consciousness virtually do not exist. Artificial feeding, kidney dialysis, and the like make it possible to sustain an individual in a state of neocortical death for decades. This sort of condition and the issues it raises were exemplified by the famous case of Karen Ann Quinlan. James Rachels [1986] provided the following description of the situation created by Quinlan's condition:

> In April 1975, this young woman ceased breathing for at least two 15-min periods, for reasons that were never made clear. As a result, she suffered severe brain damage, and, in the words of the attending physicians, was reduced to a "chronic vegetative state" in which she "no longer had any cognitive function." Accepting the doctors' judgment that there was no hope of recovery, her parents sought permission from the courts to disconnect the respirator that was keeping her alive in the intensive care unit of a New Jersey hospital.
>
> The trial court, and then the Supreme Court of New Jersey, agreed that Karen's respirator could be removed. So it was disconnected. However, the nurse in charge of her care in the Catholic hospital opposed this decision and, anticipating it, had begun to wean her from the respirator so that by the time it was disconnected she could remain alive without it. So Karen did not die. Karen remained alive for ten additional years. In June 1985, she finally died of acute pneumonia. Antibiotics, which would have fought the pneumonia, were not given.

If brainstem death is death, is neocortical death also death? Again, the issue is not a straightforward factual matter. For, it too, is a matter of specifying which features of living individuals distinguish them from corpses and so make treatment of them as corpses morally impermissible. Irreparable cessation of respiration and circulation, the classical criterion for death, would entail that an individual in a persistent vegetative state is not a corpse and so, morally speaking, must not be treated as one. The brainstem death criterion for death would also entail that a person in a state of neocortical death is not yet a corpse. On this criterion, what is crucial is that brain damage be severe enough to cause failure of the body's regulatory mechanisms.

Is an individual in a state of neocortical death any less in possession of the characteristics that distinguish the living from cadavers than one whose respiration and circulation are mechanically maintained? Of course, it is a matter of what the relevant characteristics are, and it is a matter that society must decide. It is not one that can be settled by greater medical information or more powerful medical devices. Until society decides, it will not be clear what would count as beneficent or nonmaleficent treatment of an individual in a state of neocortical death.

82.2 Euthanasia

A long-standing issue in medical ethics, which has been made more pressing by medical technology, is euthanasia, the deliberate termination of an individual's life for the individual's own good. Is such an act ever a permissible use of medical resources? Consider an individual in a persistent vegetative state. On the assumption that such a state is not death, withdrawing life support would be a deliberate termination of a human life. Here a critical issue is whether the quality of a human life can be so low or so great a liability to the individual that deliberately taking action to hasten death or at least not to postpone death is morally defensible. Can the quality of a human life be so low that the value of extending its quantity is totally negated? If so, then Western medicine's traditional commitment to providing benefits and avoiding harm would seem to make cessation of life support a moral requirement in such a case.

Consider the following hypothetical version of the kind of case that actually confronts contemporary patients, their families, health-care workers, and society as a whole. Suppose a middle-aged man suffers a brain hemorrhage and loses consciousness as a result of a ruptured aneurysm. Suppose that he never regains consciousness and is hospitalized in a state of neocortical death, a chronic vegetative state. He is maintained by a surgically implanted gastronomy tube that drips liquid nourishment from a plastic bag directly into his stomach. The care of this individual takes $7\frac{1}{2}$ of nursing time daily and includes (1) shaving, (2) oral hygiene, (3) grooming, (4) attending to his bowels and bladder, and so forth.

Suppose further that his wife undertakes legal action to force his care givers to end all medical treatment, including nutrition and hydration, so that complete bodily death of her husband will occur, She presents a preponderance of evidence to the court to show that her husband would have wanted this result in these circumstances.

The central moral issue raised by this sort of case is whether the quality of the individual's life is sufficiently compromised by neocortical death to make intentioned termination of that life morally permissible. While alive, he made it clear to both family and friends that he would prefer to be allowed to die rather than be mechanically maintained in a condition of irretrievable loss of consciousness. Deciding whether the judgment in such a case should be allowed requires deciding which capacities and qualities make life worth living, which qualities are sufficient to endow it with value worth sustaining, and whether their absence justifies deliberate termination of a life, at least when this would be the wish of the individual in question. Without this decision, the traditional norms of medical ethics, beneficence and nonmaleficence, provide no guidance. Without this decision, it cannot be determined whether termination of life support is a benefit or a harm to the patient.

An even more difficult type of case was provided by the case of Elizabeth Bouvia. Bouvia, who had been a lifelong quadriplegic sufferer of cerebral palsy, was often in pain, completely dependent upon others, and spent all of her time bedridden. Bouvia, after deciding that she did not wish to continue such a life, entered Riverside General Hospital in California. She desired to be kept comfortable while starving to death. Although she remained adamant during her hospitalization, Bouvia's requests were denied by hospital officials with the legal sanction of the courts.

Many who might believe that neocortical death renders the quality of life sufficiently low to justify termination of life support, especially when this agrees with the individual's desires, would not arrive at this conclusion in a case like Bouvia's. Whereas neocortical death completely destroys consciousness and makes purposive interaction with the individual's environment impossible, Bouvia was fully aware and mentally alert. She had previously been married and had even acquired a college education. Televised interviews with her portrayed a very intelligent person who had great skill in presenting persuasive arguments to support her wish not to have her life continued by artificial means of nutrition. Nonetheless, she judged her life to be of such low quality that she should be allowed to choose to deliberately starve to death. Before the existence of life support technology, maintenance of her life against her will might not have been possible at all and at least would have been far more difficult.

Should Elizabeth Bouvia's judgment have been accepted? Her case is more difficult than the care of a patient in a chronic vegetative state because, unlike such an *individual, she* was able to engage in meaningful interaction with her environment. Regarding an individual who cannot speak or otherwise meaningfully interact with others as nothing more than living matter, as a "human vegetable," is not especially difficult. Seeing Bouvia this way is not easy. Her awareness, intelligence, mental acuity, and ability to interact with others means that although her life is one of discomfort, indignity, and complete dependence, she is not a mere "human vegetable."

Despite the differences between Bouvia's situation and that of someone in a state of neocortical death, the same issue is posed. Can the quality of an individual's life be so low that deliberate termination is morally justifiable? How that question is answered is a matter of what level of quality of life, if any, is taken to be sufficiently low to justify deliberately acting to end it or deliberately failing to end it. If there is such a level, the conclusion that it is not always beneficent or even nonmaleficent to use life-support technology must be accepted.

Another important issue here is the respect for individual autonomy. For the cases of Bouvia and the hypothetical instance of neocortical death discussed earlier, both concern voluntary euthanasia, that is, euthanasia voluntarily requested by the patient. A long-standing commitment, vigorously defended by various schools of thought in Western moral philosophy, is the notion that competent adults should be free to conduct their lives as they please as long as they do not impose undeserved harm on others. Does this commitment entail a right to die? Some clearly believe that it does. If one owns anything at all, surely one owns one's life. In the two cases discussed earlier, neither individual sought to impose undeserved harm on anyone else, nor would satisfaction of their wish to die do so. What justification can there be then for not allowing their desires to be fulfilled?

One plausible answer is based upon the very respect of individual autonomy at issue here. A necessary condition, in some views, of respect for autonomy is the willingness to take whatever measures are necessary to protect it, including measures that restrict autonomy. An autonomy-respecting reason offered against laws that prevent even competent adults from voluntarily entering lifelong slavery is that such an exercise of autonomy is self-defeating and has the consequence of undermining autonomy altogether. Same token, an individual who acts to end his own life thereby exercises his autonomy in a manner that places it in jeopardy of permanent loss. Many would regard this as justification for using the coercive force of the law to prevent suicide. This line of thought does not fit the case of an individual in a persistent vegetative state because his/her autonomy has been destroyed by the circumstances that rendered him/her neocortically dead. It does fit Bouvia's case though. Her actions indicate that she is fully competent and her efforts to use medical care to prevent the otherwise inevitable pain of starvation is itself an exercise of her autonomy. Yet, if allowed to succeed, those very efforts would destroy her autonomy as they destroy her. On this reasoning, her case is a perfect instance of limitation of autonomy being justified by respect for autonomy and of one where, even against the wishes of a competent patient, the life-saving power of medical technology should be used.

82.2.1 Active vs. Passive Euthanasia

Discussions of the morality of euthanasia often distinguish active from passive euthanasia in light of the distinction made between killing a person and letting a person die, a distinction that rests upon the difference between an act of commission and an act of omission. When failure to take steps that could effectively forestall death results in an individual's demise, the resultant death is an act of omission and a case of letting a person die. When a death is the result of doing something to hasten the end of a person's life (e.g., giving a lethal injection), that death is caused by an act of commission and is a case of killing a person. When a person is allowed to die, death is a result of an act of omission, and the motive is the person's own good, the omission is an instance of passive euthanasia. When a person is killed, death is the result of an act of commission, and the motive is the person's own good, the commission is an instance of active euthanasia.

Does the difference between passive and active euthanasia, which reduces to a difference in how death comes about, make any moral difference? It does in the view of the American Medical Association. In a statement adopted on December 4, 1973, the House of Delegates of the American Medical Association asserted the following [Rachels, 1978]:

The intentional termination of the life of one human being by another — mercy killing — is contrary to that for which the medical profession stands and is contrary to the policy of the American Medical Association (AMA).

The cessation of extraordinary means to prolong the life of the body where there is irrefutable evidence that biological death is imminent is the decision of the patient and immediate family. The advice of the physician would be freely available to the patient and immediate family.

In response to this position, Rachels [1978, 1986] answered with the following:

> The AMA policy statement isolates the crucial issue very well, the crucial issue is "intentional termination of the life of one human being by another." But after identifying this issue and forbidding "mercy killing," the statement goes on to deny that the cessation of treatment is the intentional termination of a life. This is where the mistake comes in, for what is the cessation of treatment in those circumstances (where the intention is to release the patient from continued suffering), if it is not "the intentional termination of the life of one human being by another?"

As Rachels correctly argued, when steps that could keep an individual alive are omitted for the person's own good, this omission is as much the intentional termination of life as taking active measures to cause death. Not placing a patient on a respirator due to a desire not to prolong suffering is an act intended to end life as much as the administration of a lethal injection. In many instances the main difference between the two cases is that the latter would release the individual from his pain and suffering more quickly than the former. Dying can take time and involve considerable pain even if nothing is done to prolong life. Active killing can be done in a manner that causes death painlessly and instantly. This difference certainly does not render killing, in this context, morally worse than letting a person die. In so far as the motivation is merciful (as it must be if the case is to be a genuine instance of euthanasia) because the individual is released more quickly from a life that is disvalued than otherwise, the difference between killing and letting one die may provide support for active euthanasia. According to Rachels, the common rejoinder to this argument is the following:

> The important difference between active and passive euthanasia is that in passive euthanasia the doctor does not do anything to bring about the patient's death. The doctor does nothing and the patient dies of whatever ills already afflict him. In active euthanasia, however, the doctor does something to bring about the patient's death: he kills the person. The doctor who gives the patient with cancer a lethal injection has himself caused his patient's death; whereas if he merely ceases treatment, the cancer is the cause of death.

According to this rejoinder, in active euthanasia someone must do something to bring about the patient's death, and in passive euthanasia the patient's death is caused by illness rather than by anyone's conduct. Surely this is mistaken. Suppose a physician deliberately decides not to treat a patient who has a routinely curable ailment and the patient dies. Suppose further that the physician were to attempt to exonerate himself by saying, "I did nothing. The patient's death was the result of illness. I was not the cause of death." Under current legal and moral norms, such a response would have no credibility. As Rachels noted, *"it would be no defense at all for him to insist that he didn't do anything. He would have done something very serious indeed, for he let his patient die."*

The physician would be blameworthy for the patient's death as surely as if he had actively killed him. If causing death is justifiable under a given set of circumstances, whether it is done by allowing death to occur or by actively causing death is morally irrelevant. If causing someone to die is not justifiable under a given set of circumstances, whether it is done by allowing death to occur or by actively causing death is also morally irrelevant. Accordingly, if voluntary passive euthanasia is morally justifiable in the light of the duty of beneficence, so is voluntary active euthanasia. Indeed, given that the benefit to be achieved is more quickly realized by means of active euthanasia, it may be preferable to passive euthanasia in some cases.

82.2.2 Involuntary and Nonvoluntary Euthanasia

An act of euthanasia is involuntary if it hastens the individual's death for his own good but against his wishes. To take such a course would be to destroy a life that is valued by its possessor. Therefore, it is no different in any morally relevant way from unjustifiable homicide. There are only two legitimate reasons for hastening an innocent person's death against his will: self-defense and saving the lives of a larger number of other innocent persons. Involuntary euthanasia does not fit either of these justifications. By definition, it is done for the good of the person who is euthanized and for self-defense or saving innocent others.

No act that qualifies as involuntary euthanasia can be morally justifiable. Hastening a person's death for his own good is an instance of non-voluntary euthanasia when the individual is incapable of agreeing or disagreeing. Suppose it is clear that a particular person is sufficiently self-conscious to be regarded a person but cannot make his wishes known. Suppose also that he is suffering from the kind of ailment that, in the eyes of many persons, makes one's life unendurable. Would hastening his death be permissible? It would be if there were substantial evidence that he has given prior consent. This person may have told friends and relatives that under certain circumstances efforts to prolong his life should not be undertaken or continued. He might have recorded his wishes in the form of a Living Will (as shown in the box) or on audio or videotape. Where this kind of substantial evidence of prior consent exists, the decision to hasten death would be morally justified. A case of this scenario would be virtually a case of voluntary euthanasia.

But what about an instance in which such evidence is not available? Suppose the person at issue has never had the capacity for competent consent or dissent from decisions concerning his life. It simply cannot be known what value the individual would place on his life in his present condition of illness. What should be done is a matter of what is taken to be the greater evil — mistakenly ending the life of an innocent person for whom that life has value or mistakenly forcing him to endure a life that he radically disvalues.

To My Family, My Physician, My Clergyman, and My Lawyer:

If the time comes when I can no longer take part in decisions about my own future, let this statement stand as testament of my wishes: If there is no reasonable expectation of my recovery from physical or mental disability. I, ________, request that I be allowed to die and not be kept alive by artificial means or heroic measures. Death is as much a reality as birth, growth, maturity, and old age — it is the one certainty. I do not fear death as much as I fear the indiginity of deterioration, dependence, and hopeless pain. I ask that drugs be mercifully administered to me for the terminal suffering even if they hasten the moment of death.

This request is made after careful consideration. Although this document is not legally binding, you who care for me will, I hope, feel morally bound to follow its mandate. I recognize that it places a heavy burden of responsibility upon you, and it is with the intention of sharing that responsibility and of mitigating any feelings of guild that this statement is made.

Signed:______________
Date: ______________

Witnessed by:
____
____

Living Will statutes have been passed in at least 35 states and the District of Columbia. For a Living Will to be a legally binding document, the person signing it must be of sound mind at the time the will is made and shown not to have altered his opinion in the interim between the signing and his illness. The witnesses must not be able to benefit from the individual's death.

82.2.3 Should Voluntary Euthanasia Be Legalized?

Recent events have raised the question: "Should voluntary euthanasia be legalized?" Some argue that even if voluntary euthanasia is morally justifiable, it should be prohibited by social policy nonetheless. According to this position, the problem with voluntary euthanasia is its impact on society as a whole. In other words, the overall disutility of allowing voluntary euthanasia outweighs the good it could do for its beneficiaries. The central moral concern is that legalized euthanasia would eventually erode respect for human life and ultimately become a policy under which "socially undesirable" persons would have their deaths hastened (by omission or commission). The experience of Nazi Germany is often cited in support of this fear. What began there as a policy of euthanasia soon became one of eliminating individuals deemed racially inferior or otherwise undesirable. The worry, of course, is that what happened there can happen here as well. If social policy encompasses efforts to hasten the deaths of people, respect for human life in general is eroded and all sorts of abuses become socially acceptable, or so the argument goes.

No one can provide an absolute guarantee that the experience of Nazi Germany would not be repeated, but there is reason to believe that its likelihood is negligible. The medical moral duty of beneficence justifies only voluntary euthanasia. It justifies hastening an individual's death only for the individual's benefit and only with the individual's consent. To kill or refuse to save people judged socially undesirable is not to engage in euthanasia at all and violates the medical moral duty of nonmaleficence. As long as only voluntary euthanasia is legalized, and it is clear that involuntary euthanasia is not and should never be, no degeneration of the policy need occur. Furthermore, such degeneration is not likely to occur if the beneficent nature of voluntary euthanasia is clearly distinguished from the maleficent nature of involuntary euthanasia and any policy of exterminating the socially undesirable. Euthanasia decisions must be scrutinized carefully and regulated strictly to ensure that only voluntary cases occur, and severe penalties must be established to deter abuse.

References

Bronzino, J.D. Chapter 190. Beneficence, Nonmaleficence and Technological Progress. *The Biomedical Engineering Handbook*. CRC Press, Boca Raton, Fl, 1995; 2000.

Bronzino, J.D. Chapter 10. Medical and Ethical Issues in Clinical Engineering Practice. *Management of Medical Technology*. Butterworth, 1992.

Bronzino, J.D. Chapter 20. Moral and Ethical Issues Associated with Medical Technology. *Introduction to Biomedical Engineering*. Academic Press, New York, 1999.

Rachels, J. "Active and Passive Euthanasia," In *Moral Problems,* 3rd ed., Rachels, J. (Ed.), Harper and Row, New York, 1978.

Rachels, J. *Ethics at the End of Life: Euthanasia and Morality,* Oxford University Press, Oxford, 1986.

Further Information

Dubler, N.N. and Nimmons D. *Ethics on Call.* Harmony Books, New York, 1992.

Jonsen, A.R. *The New Medicine and the Old Ethics*, Harvard University Press, Cambridge, MA, 1990.

Seebauer, E.G. and Barry R.L. *Fundamentals of Ethics for Scientists and Engineers*, Oxford University Press, Oxford, 2001.

83

Ethical Issues Related to Clinical Research

Joseph D. Bronzino
Trinity College

83.1 Introduction

The Medical Device Amendment of 1976, and its updated 1990 version, requires approval from the Food and Drug Administration (FDA) before new devices are marketed and imposes requirements for the clinical investigation of new medical devices on human subjects. Although the statute makes interstate commerce of an unapproved new medical device generally unlawful, it provides an exception to allow interstate distribution of unapproved devices in order to conduct clinical research on human subjects. This investigational device exemption (IDE) can be obtained by submitting to the FDA *"a protocol for the proposed clinical testing of the device, reports of prior investigations of the device, certification that the study has been approved by a local institutional review board, and an assurance that informed consent will be obtained from each human subject"* [Bronzino et al., 1990a,b; Bronzino 1992; 1995; 1999; 2000].

With respect to clinical research on humans, the FDA distinguishes devices into two categories: devices that pose significant risk and those that involve insignificant risk. Examples of the former included orthopedic implants, artificial hearts, and infusion pumps. Examples of the latter include various dental devices and daily-wear contact lenses. Clinical research involving a significant risk device cannot begin until an institutional review board (IRB) has approved both the protocol and the informed consent form and the FDA itself has given permission. This requirement to submit an IDE application to the FDA is waived in the case of clinical research where the risk posed is insignificant. In this case, the FDA requires only that approval from an IRB be obtained certifying that the device in question poses only insignificant risk. In deciding whether to approve a proposed clinical investigation of a new device, the IRB and the FDA must determine the following:

1. Risks to subjects are minimized
2. Risks to subjects are reasonable in relation to the anticipated benefit and knowledge to be gained

3. Subject selection is equitable
4. Informed consent materials and procedures are adequate
5. Provisions for monitoring the study and protecting patient information are acceptable

The FDA allows unapproved medical devices to be used without an IDE in three types of situations: emergency use, treatment use, and feasibility studies. However, in each instance there are specific ethical issues.

83.2 Ethical Issues in Feasibility Studies

Manufacturers seeking more flexibility in conducting investigations in the early developmental stages of a device have submitted a petition to the FDA, requesting that certain limited investigations of significant risk devices be subject to abbreviated IDE requirements. In a feasibility study, or "limited investigation," human research on a new device would take place at a single institution and involve no more than ten human subjects. *The sponsor of a limited investigation would be required to submit to the FDA a "Notice of Limited Investigation," which would include a description of the device,* a *summary of the purpose of the investigation, the protocol, a sample of the informed consent form, and a certification of approval by the responsible IRB. In certain circumstances, the FDA could require additional information, or require the submission of a full IDE application, or suspend the investigation* [Bronzino et al., 1990a,b].

Investigations of this kind would be limited to certain circumstances (1) investigations of new uses of existing devices, (2) investigations involving temporary or permanent implants during the early developmental stages, and (3) investigations involving modification of an existing device.

To comprehend adequately the ethical issues posed by clinical use of unapproved medical devices outside the context of an IDE, it is necessary to utilize the distinctions between practice, nonvalidated practice, and research elaborated in the previous pages. How do those definitions apply to feasibility studies?

Clearly, the goal of this sort of study, that is, generalizable knowledge, makes it an issue of research rather than practice. Manufacturers seek to determine the performance of a device with respect to a particular patient population in an effort to gain information about its efficacy and safety. Such information would be important in determining whether further studies (animal or human) need to be conducted, whether the device needs modification before further use, and the like. The main difference between use of an unapproved device in a feasibility study and use under the terms of an IDE is that the former would be subject to significantly less intensive FDA review than the latter. This, in turn, means that the responsibility for ensuring that use of the device is ethically sound would fall primarily to the IRB of the institution conducting the study.

The ethical concerns posed here are best comprehended with a clear understanding of what justifies research. Ultimately, no matter how much basic research and animal experimentation has been conducted on a given device, the risks and benefits it poses for humans cannot be adequately determined until it is actually used on humans.

The benefits of research on humans lie primarily in the knowledge that is yielded and the generalizable information that is provided. This information is crucial to medical science's ability to generate new modes and instrumentalities of medical treatment that are both efficacious and safe. Accordingly, for necessary but insufficient condition for experimentation to be ethically sound, it must be scientifically sound [Capron, 1978, 1986].

Although scientific soundness is a necessary condition of ethically acceptable research on humans, it is not of and by itself sufficient. Indeed, it is widely recognized that the primary ethical concern posed by such investigation is the use of one person by another to gather knowledge or other benefits where these benefits may only partly or not at all accrue to the first person. In other words, the human subjects of such research are at risk of being mere research resources, as having value only for the ends of the research. Research upon human beings runs the risk of failing to respect them as people. The notion that human beings are not mere things but entities whose value is inherent rather than wholly instrumental is one of the most widely held norms of contemporary Western society. That is, human beings are not valuable wholly or solely for the uses to which they can be put. They are valuable simply by being the kinds of entities they are. To treat them as such is to respect them as people.

Respecting individuals as people is generally agreed to entail two requirements in the context of biomedical experimentation. First, since what is most generally taken to make human beings people is their autonomy — their ability to make rational choices for themselves — treating individuals as people means respecting that autonomy. This requirement is met by ensuring that no competent person is subjected to any clinical intervention without first giving voluntary and informed consent. Second, respect for people means that the physician will not subject a human to unnecessary risks and will minimize the risks to patients in required procedures.

Much of the ethical importance of the scrutiny that the FDA imposes upon use of unapproved medical devices in the context of an IDE derives from these two conditions of ethically sound research. The central ethical concern posed by use of medical devices in a feasibility study is that the decreased degree of FDA scrutiny will increase the likelihood that either or both of these conditions will not be met. This possibility may be especially great because many manufacturers of medical devices are, after all, commercial enterprises, companies that are motivated to generate profit and thus to get their devices to market as soon as possible with as little delay and cost as possible. These self-interested motives are likely, at times, to conflict with the requirements of ethically sound research and thus to induce manufacturers to fail (often unwittingly) to meet these requirements. Note that profit is not the only motive that might induce manufacturers to contravene the requirements of ethically sound research on humans. A manufacturer may sincerely believe that its product offers great benefit to many people or to a population of especially needy people and so from this utterly altruistic motive may be prompted to take shortcuts that compromise the quality of the research. Whether the consequences being sought by the research are desired for reasons of self-interest, altruism, or both, the ethical issue is the same. Research subjects may be placed at risk of being treated as mere objects rather than as people.

What about the circumstances under which feasibility studies would take place? Are these not sufficiently different from the "normal" circumstances of research to warrant reduced FDA scrutiny? As noted earlier, manufacturers seek to be allowed to engage in feasibility studies in order to investigate new uses of existing devices, to investigate temporary or permanent implants during the early developmental stages, and to investigate modifications to an existing device. As also noted earlier, a feasibility study would take place at only one institution and would involve no more than ten human subjects. Given these circumstances, is the sort of research that is likely to occur in a feasibility study less likely to be scientifically unsound or to fail to respect people in the way that normal research upon humans does in "normal" circumstances?

Such research would be done on a very small subject pool, and the harm of any ethical lapses would likely affect fewer people than if such lapses occurred under more usual research circumstances. Yet even if the harm done is limited to a failure to respect the ten or fewer subjects in a single feasibility study, the harm would still be ethically wrong. To wrong ten or fewer people is not as bad as to wrong in the same way more than ten people but it is to engage in wrongdoing nonetheless. In either case, individuals are reduced to the status of mere research resources and their dignity as people is not properly respected.

Are ethical lapses more likely to occur in feasibility studies than in studies that take place within the requirements of an IDE? Although nothing in the preceding discussion provides a definitive answer to this question, it is a question to which the FDA should give high priority in deciding whether to allow this type of exception to IDE use of unapproved medical devices. The answer to this question might be quite different when the device at issue is a temporary or permanent implant than when it is an already approved device being put to new uses or modified in some way. Whatever the contemplated use under the feasibility studies mechanism, the FDA would be ethically advised not to allow this kind of exception to IDE use of an unapproved device without a reasonably high level of certainty that research subjects would not be placed in greater jeopardy than in "normal" research circumstances.

83.3 Ethical Issues in Emergency Use

What about the mechanism for avoiding the rigors of an IDE for emergency use?

The FDA has authorized emergency use where an unapproved device offers the only alternative for saving the life of a dying patient, but an IDE has not yet been approved for the device or its use, or an IDE has been approved but the physician who wishes to use the device is not an investigator under the IDE [Bronzino et al., 1990a,b].

Because the purpose of emergency use of an unapproved device is to attempt to save a dying patient's life under circumstances where no other alternative is at hand, this sort of use constitutes practice rather than research. Its aim is primarily benefit to the patient rather than provision of new and generalizable information. Because this sort of use occurs prior to the completion of clinical investigation of the device, it constitutes a nonvalidated practice. What does this mean?

First, it means that while the aim of the use is to save the life of the patient, the nature and likelihood of the potential benefits and risks engendered by use of the device are far more speculative than in the sort of clinical intervention that constitutes validated practice. In validated practice, thorough investigation, including preclinical studies, animal studies, and studies on human subjects of a device has established its efficacy and safety. The clinician thus has a well-founded basis upon which to judge the benefits and risks such an intervention poses for his patients.

It is precisely this basis that is lacking in the case of a nonvalidated practice. Does this mean that emergency use of an unapproved device should be regarded as immoral? This conclusion would follow only if there were no basis upon which to make an assessment of the risks and benefits of the use of the device. The FDA requires that a physician who engages in emergency use of an unapproved device must *"have substantial reason to believe that benefits will exist. This means that there should be a body of preclinical and animal tests allowing a prediction of the benefit to a human patient."*

Thus, although the benefits and risks posed by use of the device are highly speculative, they are not entirely speculative. Although the only way to validate a new technology is to engage in research on humans at some point, not all nonvalidated technologies are equal. Some will be largely uninvestigated, and assessment of their risks and benefits will be wholly or almost wholly speculative. Others will at least have the support of preclinical and animal tests. Although this is not sufficient support for incorporating use of a device into regular clinical practice, it may however represent sufficient support to justify use in the desperate circumstances at issue in emergency situations. Desperate circumstances can justify desperate actions, but desperate actions are not the same as reckless actions, hence the ethical soundness of the FDA's requirement that emergency use be supported by solid results from preclinical and animal tests of the unapproved device.

A second requirement that the FDA imposes on emergency use of unapproved devices is the expectation that physicians *"exercise reasonable foresight with respect to potential emergencies and make appropriate arrangements under the IDE procedures. Thus, a physician should not 'create' an emergency in order to circumvent IRB review and avoid requesting the sponsor's authorization of the unapproved use of a device."*

From a Kantian point of view, which is concerned with protecting the dignity of people, it is a particularly important requirement to create an emergency in order to avoid FDA regulations, which prevent the patient being treated as a mere resource whose value is reducible to a service of the clinician's goals. Hence, the FDA is quite correct to insist that emergencies are circumstances that reasonable foresight would not anticipate.

Also especially important here is the nature of the patient's consent. Individuals facing death are especially vulnerable to exploitation and deserve greater measures for their protection than might otherwise be necessary. One such measure would be to ensure that the patient, or his legitimate proxy, knows the highly speculative nature of the intervention being offered. That is, to ensure that it is clearly understood that the clinician's estimation of the intervention's risks and benefits is far less solidly grounded than in the case of validated practices. The patient's consent must be based upon an awareness that the particular device has not undergone complete and rigorous testing on humans and that estimations of its potential are based wholly upon preclinical and animal studies. Above all the patient must not be led to believe that there is complete understanding of the risks and benefits of the intervention. Another important point here is to ensure that the patient is aware that the options he is facing are not simply life or death but may include life of a severely impaired quality, and therefore that even if his life is saved, it may be a life of significant impairment. Although desperate circumstance may legitimize desperate actions, the decision to take such actions must rest upon the informed and voluntary consent of the patient, especially when he/she is an especially vulnerable patient.

It is important here for a clinician involved in emergency use of an unapproved device to recognize that these activities constitute a form of nonvalidated practice and not research. Hence, the primary obligation is to the well being of the patient. The patient enters into the relationship with the clinician with the same trust that accompanies any normal clinical situation. To treat this sort of intervention as if it were an instance of research and hence justified by its benefits to science and society would be to abuse this trust.

83.4 Ethical Issues in Treatment Use

The FDA has adopted regulations authorizing the use of investigational new drugs in certain circumstances — where a patient has not responded to approved therapies. This "treatment use" of unapproved new drugs is not limited to life-threatening emergency situations, but rather is also available to treat "serious" diseases or conditions.

The FDA has not approved treatment use of unapproved medical devices, but it is possible that a manufacturer could obtain such approval by establishing a specific protocol for this kind of use within the context of an IDE.

The criteria for treatment use of unapproved medical devices would be similar to criteria for treatment use of investigational drugs (1) the device is intended to treat a serious or life-threatening disease or condition, (2) there is no comparable or satisfactory alternative product available to treat that condition, (3) the device is under an IDE, or has received an IDE exemption, or all clinical trials have been completed and the device is awaiting approval, and (4) the sponsor is actively pursuing marketing approval of the investigational device. The treatment use protocol would be submitted as part of the IDE, and would describe the intended use of the device, the rationale for use of the device, the available alternatives and why the investigational product is preferred, the criteria for patient selection, the measures to monitor the use of the device and to minimize risk, and technical information that is relevant to the safety and effectiveness of the device for the intended treatment purpose.

Were the FDA to approve treatment use of unapproved medical devices, what ethical issues would be posed? First, because such use is premised on the failure of validated interventions to improve the patient's condition adequately, it is a form of practice rather than research. Second, since the device involved in an instance of treatment use is unapproved, such use would constitute nonvalidated practice. As such, like emergency use, it should be subject to the FDA's requirement that prior preclinical tests and animal studies have been conducted that provide substantial reason to believe that patient benefit will result. As with emergency use, although this does not prevent assessment of the intervention's benefits and risks from being highly speculative, it does prevent assessment from being totally speculative. Here too, although desperate circumstances can justify desperate action, they do not justify reckless action. Unlike emergency use, the circumstances of treatment use involve serious impairment of health rather than the threat of premature death. Hence, an issue that must be considered is how serious such impairment must be to justify resorting to an intervention whose risks and benefits have not been solidly established.

In cases of emergency use, the FDA requires that physicians not use this exception to an IDE to avoid requirements that would otherwise be in place. This particular requirement would be obviated in instances of treatment use by the requirement that a protocol for such use be previously addressed within an IDE.

As with emergency use of unapproved devices, the patients involved in treatment use would be particularly vulnerable patients. Although they are not dying, they are facing serious medical conditions and are thereby likely to be less able to avoid exploitation than patients under less desperate circumstances.

Consequently, it is especially important that patients be informed of the speculative nature of the intervention and of the possibility that treatment may result in little or no benefit to them.

83.5 The Safe Medical Devices Act

On November 28, 1991, the Safe Medical Devices Act of 1990 (Public Law 101-629) went into effect. This regulation requires a wide range of healthcare institutions, including hospitals, ambulatory-surgical facilities, nursing homes, and outpatient treatment facilities, to report information that "reasonably suggests" the likelihood that the death, serious injury, or serious illness of a patient at that facility has been caused or contributed to by a medical device. When a death is devicerelated, a report must be made directly to the FDA *and* to the manufacturer of the device. When a serious illness or injury is device related, a report must be made to the manufacturer *or to* the FDA in cases where the manufacturer is not known. In addition, summaries of previously submitted reports must be submitted to the FDA on a semiannual basis. Prior to this regulation, such reporting was voluntary. This new regulation was designed to enhance the FDA's ability to quickly learn about problems related to medical devices. It also supplements the medical device reporting (MDR) regulations promulgated in 1984. MDR regulations require that reports of device-related deaths and serious injuries be submitted to the FDA by manufacturers and importers. The new law extends this requirement to users of medical devices along with manufacturers and importers. This act represents a significant step forward in protecting patients exposed to medical devices.

References

Bronzino, J.D. "Medical and Ethical Issues in Clinical Engineering Practice," *Management of Medical Technology.* Butterworth, Chapter 10, 1992.

Bronzino, J.D. "Moral and Ethical Issues Associated with Medical Technology," *Introduction to Biomedical Engineering. Academic Press*, Chapter 20, 1999.

Bronzino, J.D. "Regulation of Medical Device Innovation," *The Biomedical Engineering Handbook.* CRC Press, Chapter 192, 1995; 2000.

Bronzino, J.D., Flannery, E.J., and Wade, M.L. "Legal and Ethical Issues in the Regulation and Development of Engineering Achievements in Medical Technology," Part I *IEEE Engineering in Medicine and Biology,* 1990a.

Bronzino, J.D., Flannery, E.J., and Wade, M.L. "Legal and Ethical Issues in the Regulation and Development of Engineering Achievements in Medical Technology," Part II *IEEE Engineering in Medicine and Biology,* 1990b.

Capron, A. "Human Experimentation: Basic Issues," *The Encyclopedia of Bioethics* Vol. II. The Free Press, Glencoe, II. 1978.

Capron, A. "Human Experimentation," (J.P. Childress et al., eds.) University Publications of America, 1986.

Further Reading

1. Dubler, N.N. and Nimmons, D. *Ethics on Call.* Harmony Books, New York, 1992.
2. Jonsen, A.R. *The New Medicine and the Old Ethics.* Harvard University Press, Cambridge, MA, 1990.