Effectiveness and Outcomes in Health Care: Proceedings of an Invitational Conference

Division of Health Care Services
ISBN: 0-309-54363-0, 246 pages, 6 x 9, (1990)

Effectiveness and Outcomes in Health Care

Proceedings of an Invitational Conference By the Institute of Medicine Division of Health Care Services

Kim A. Heithoff and Kathleen N. Lohr, editors

National Academy Press
Washington, D.C. 1990

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for this report were chosen for their special competencies and with regard for appropriate balance.

The Institute of Medicine was chartered in 1970 by the National Academy of Sciences to enlist distinguished members of the appropriate professions in the examination of policy matters pertaining to the health of the public. In this, the Institute acts under both the Academy's 1863 congressional charter responsibility to be an adviser to the federal government and its own initiative in identifying issues of medical care, research, and education.

This conference was supported by the Health Care Financing Administration, U.S. Department of Health and Human Services, under Basic Ordering Agreement Contract No. 500-89-0008.

Library of Congress Catalog Card No. 90-63194
International Standard Book Number 0-309-04342-5

Publication 90-003

Additional copies of this report are available from: National Academy Press 2101 Constitution Avenue, NW Washington, DC 20418

S225
Printed in the United States of America

INSTITUTE OF MEDICINE

Division of Health Care Services HCFA Effectiveness Initiative

CORE COMMITTEE

KENNETH I. SHINE, *Chair,* Dean, School of Medicine, University of California, Los Angeles

MAUREEN M. HENDERSON, Head, Cancer Prevention Research Center, Fred Hutchinson Cancer Research Center, Seattle, Washington

EMMETT B. KEELER, Senior Mathematician, Economics Department, The RAND Corporation, Santa Monica, California

BARBARA J. MCNEIL, Head, Department of Health Care Policy, Harvard Medical School, Boston, Massachusetts

DAVID G. MURRAY, Professor of Orthopedic Surgery, Health Services Center, State University of New York, Syracuse

ALAN R. NELSON, Associate, Memorial Medical Center, Salt Lake City, Utah

J. SANFORD SCHWARTZ, Professor, General Internal Medicine, Hospital of the University of Pennsylvania, Philadelphia

G. RICHARD SMITH, Associate Professor of Psychiatry, Department of Psychiatry, University of Arkansas for the Medical Sciences, Little Rock

HAROLD G. SOX, Professor and Chairman, Department of Medicine, Dartmouth Medical School, Hanover, New Hampshire

BREAST CANCER COMMITTEE

MARTIN D. ABELOFF, Professor and Associate Director, Johns Hopkins Oncology Center, Baltimore, Maryland

BARBARA DANOFF FOWBLE, Professor, Department of Radiologic Oncology, Hospital of the University of Pennsylvania, Philadelphia

SHELDON GREENFIELD, Senior Scientist, Institute for the Improvement of Medical Care and Health, New England Medical Center, Boston, Massachusetts

VALERIE P. JACKSON, Associate Professor, Department of Radiology, Wishard Memorial Hospital, Indiana University School of Medicine, Indianapolis

LUELLA KLEIN, Professor and Chair, Department of Obstetrics and Gynecology, Emory University, Atlanta, Georgia

MARY K. KNOBS, Section on Medical Oncology, Yale Medical Center, New Haven, Connecticut

JOHN S. MEYER, Pathology Department, St. Luke Hospital, Chesterfield, Missouri

MONICA MORROW, Associate Professor of Surgery, University of Chicago, Chicago, Illinois

WILLIAM C. WOOD, Medical Director, Cancer Center, Massachusetts General Hospital, Boston

ACUTE MYOCARDIAL INFARCTION COMMITTEE

HOOSHANG BOLOOKI, Jackson Memorial Hospital, Miami, Florida

WILLIAM H. CARTER, The Charleston Cardiology Group, Charleston, West Virginia

KATHLEEN DRACUP, Professor of Nursing, School of Nursing, University of California, Los Angeles

KENNETH M. KENT, Director, Cardiac Catheterization Laboratory, Georgetown University School of Medicine, Washington, D.C.

BRUCE C. PATON, Arapahoe Cardiovascular Surgeons, Denver, Colorado

GERALD M. POHOST, Director, Division of Cardiovascular Disease, University of Alabama, Birmingham

JOHN V. RUSSO, John Russo & Shellee Nolan, Washington, D.C.

THOMAS J. RYAN, Chief of Cardiology, Boston University School of Medicine, Boston, Massachusetts

HARRY P. SELKER, Director of Health Services Research Unit, New England Medical Center Hospital, Boston, Massachusetts

GEORGE T. THIBAULT, Chief, Medical Service, VA Medical Center, West Roxbury, Massachusetts

W. DOUGLAS WEAVER, Director, Division of Cardiology, University of Washington, Seattle

MYRON L. WEISFELDT, Director of Cardiology, Johns Hopkins Hospital, Baltimore, Maryland

NANETTE K. WENGER, Director, Cardiac Clinic at Grady Memorial Hospital, Emory University, Atlanta, Georgia

HIP FRACTURE COMMITTEE

CHRISTINE K. CASSEL, Chief, Section of General Internal Medicine, University of Chicago, Chicago, Illinois

JOHN F. FITZGERALD, Assistant Professor of Medicine, Indiana University School of Medicine, Indianapolis

HOWARD S. FRAZIER, Professor of Medicine, Harvard Medical School, Boston, Massachusetts

JOHN J. GARTLAND, Director, Center for Research, Medical Education and Health Care, Jefferson Medical College, Philadelphia, Pennsylvania

CAROL CLARKE HOGUE, Associate Professor, Center for Study of Aging and Human Development, University of North Carolina, Chapel Hill

C. CONRAD JOHNSON, Chief of Endocrinology, Indiana University School of Medicine, Indianapolis

ROSALIE A. KANE, Professor, School of Public Health, University of Minnesota, Minneapolis

ROBERT B. KELLER, Executive Director, Maine Medical Assessment Foundation, Belfast

ROBERT J. LLOYD, Arthritis Rehabilitation Center, Washington, D.C.

JOHN L. MELVIN, Professor and Chair, Department of Physical Medicine and Rehabilitation, Medical College of Wisconsin, Milwaukee

JANA M. MOSSEY, Professor, Department of Psychiatry, Medical College of Pennsylvania, Philadelphia

RAYMOND J. RABIDOUX, President, Henry Ford Continuing Care Corporation, Detroit, Michigan

WAYNE A. RAY, Associate Professor of Biostatistics and Director, Division of Pharmacoepidemiology, Vanderbilt University School of Medicine, Nashville, Tennessee

STUDY STAFF

Division of Health Care Services
KARL D. YORDY, Director
KATHLEEN N. LOHR, Deputy Director
RICHARD A. RETTIG, Senior Staff Officer
KIM A. HEITHOFF, Research Assistant
PATRICK A. MATTINGLY, Consultant
H. DONALD TILLER, Administrative Assistant
THELMA L. COX, Senior Secretary
THERESA H. NALLY, Senior Secretary
Division of Health Promotion and Disease Prevention
MARIA ELENA LARA, Program Officer

ACKNOWLEDGMENTS

The contributions of several members of the Institute of Medicine staff deserve special mention. Among them are H. Donald Tiller, administrative assistant, and Theresa Nally and Thelma Cox, senior secretaries. Kim Heithoff contributed greatly to the smooth logistics of the committee's workshop meetings and review of the first drafts of these proceedings. Richard Rettig, Kathleen Lohr, and Karl Yordy provided steady support and leadership throughout the entire project.

The committee is particularly indebted to Blair Potter for the first draft edit of this report.

Support for this study was provided by the U.S. Department of Health and Human Services, Health Care Financing Administration. We particularly wish to acknowledge the unflagging assistance and guidance of the government's project officer, John Spiegel, of the Health Standards and Quality Bureau.

PREFACE

This conference represented an important step in the Institute of Medicine's (IOM) effort to provide consultation to the Health Care Financing Administration (HCFA) in regard to its research on effectiveness. It provided an opportunity for the IOM's core committee to communicate its initial thoughts about the opportunities and challenges in this type of inquiry. It brought together many of the most knowledgeable individuals in the country whose insights were useful to government and to the research community in understanding the issues raised by effectiveness research.

The IOM convened its committee of experts from clinical medicine, health services research, epidemiology, nursing, and a number of other areas in order to identify clinical conditions for effectiveness research. The decision to identify conditions rather than procedures was based on the committee's desire to focus on the practice of medicine in a clinical setting, rather than upon technology assessment alone. At the same time, it was understood that all of the conditions which we identified can and will be included as part of the inquiries, procedures, and technology.

Initially, five conditions—hip fracture, breast cancer, angina pectoris, congestive heart failure, and acute myocardial infarction—were identified by the committee for possible study. HCFA identified three of these conditions—breast cancer, hip fracture, and acute myocardial infarction—as its priorities. Three workshops were conducted to examine each at greater length. The membership of each workshop was drawn approximately equally from the committee and from individuals who are experts on the particular condition to be studied.

Each workshop framed questions that might be asked in regard to the particular condition being explored, and attempted to identify the research strategies, approaches, and methodologies that might be used. In doing so,

the committee learned a great deal about the currently available HCFA databases and about other issues that must be confronted if effectiveness research is to fulfill its promise. The committee learned about the anxieties that conscientious clinicians and scientists have about effectiveness research.

After each of the workshops, a report was generated for that particular condition. Subsequently, the committee drafted a statement about recurrent concerns which it believed required attention in order that the overall effort in effectiveness research be pursued successfully. That summary document was distributed prior to this conference.

EFFECTIVENESS VERSUS EFFICACY

Among the most important principles to be understood by clinicians and scientists is the difference between effectiveness and efficacy. The randomized clinical trial provides important information with regard to therapeutic alternatives. However, the vast majority of such trials have traditionally excluded the elderly. Without further evidence for each condition, the extrapolation of the results of such trials to the elderly may or may not be justified.

Moreover, there is evidence that physicians themselves have made judgments about the nature of disease and about the treatment of elderly patients that are sometimes inconsistent with the results of clinical trials. For example, some physicians behave as if breast cancer in an elderly woman is a different disease than in a younger patient. As a result, the same range of therapy may not be offered.

We must develop data that will allow judgments to be made rationally. This requires an understanding not only of what may be efficacious, but also of what is actually done and what the outcomes are in the real world of medical care.

METHODOLOGIES

Some have expressed anxiety that effectiveness studies may be performed to the exclusion of other kinds of scientific inquiry. This is both unrealistic and undesirable. It is essential that initial observations regardless of their source—whether the HCFA databases or an investigator's imagination—be addressed by the most appropriate kinds of inquiry. This may require more detailed epidemiological study, demonstration projects, or randomized clinical trials. A few of these studies may be satisfactorily conducted by intramural programs of the federal government, but the vast majority are likely to require funding of research activities conducted by extramural investigators who work in a variety of settings.

IMPACT UPON COSTS

In general, the participants at this conference agreed that effectiveness research is an important endeavor that should be undertaken. There was a general assumption that the results of such research could help physicians make clinical decisions, but there was considerable uncertainty as to the best way of altering physician behavior. Results of effectiveness research might be expected to influence the management and organization of our health care system, but there was general consensus that the results of such research, in and of themselves, are not likely to alter the rate of growth of health care expenditures. However, it was strongly felt that information is absolutely essential if policymakers are to make rational decisions about management, organization, and reimbursement.

RISK ADJUSTMENTS AND OUTCOMES

While HCFA has regularly reported mortality data in a variety of formats, there was a strong consensus that such data were of limited value unless issues of morbidity, disability, function, and cost were also more satisfactorily considered. Comorbidity, risk stratification, and function assessment are major challenges.

Recent studies of transurethral versus transabdominal prostatectomy, for example, suggest that small differences in a mortality rate associated with each procedure, which might range from 1 to 1.5 percent, could be invisible to an active urologist. However, the impact on the Medicare population as a whole might be several thousand deaths a year. At the same time, in the absence of a randomized clinical trial, we are limited in our ability to stratify for risk satisfactorily. Although significant progress is being made in this area, it is not uncommon to find that only 50 to 65 percent of outcomes can be predicted by currently available risk stratification. Clinicians will continue to be skeptical of retrospective results when only a relatively small proportion of risk can be accurately adjusted.

A particular highlight of the discussions was the potential role of the patient in providing assessments of morbidity, disability, and function. Patients' assessment may be of comparable accuracy with that of physicians, and, in some cases, much more easily obtainable.

CONCLUSIONS

As an educator, I am concerned that medical students and house staff understand issues of effectiveness and appropriateness. This will become more important in the future as health maintenance organizations, indepen

dent practice associations, insurance companies, government, corporations, employers, and employees seek to restrain health care costs. Without reliable data, decisions about the provision of health care, the role of prevention, and rehabilitation will continue to be unduly influenced by economics and politics rather than by reason. Although the IOM core committee has identified many concerns, I believe it is important to move forward in effectiveness research, not only with HCFA, but also though linkages with state programs such as Medi-Cal and the private insurance sector. Developments in government and industry will make this possible and imperative.

KENNETH I. SHINE

CHAIR, COMMITTEE ON THE HCFA EFFECTIVENESS INITIATIVE

CONTENTS

PART I

INTRODUCTION

1

Genesis of the Effectiveness Initiative and IOM's Role

Kim A. Heithoff, Kathleen N. Lohr, and Richard A. Rettig

In 1988, the Health Care Financing Administration (HCFA) of the U.S. Department of Health and Human Services (DHHS) proposed a research program called the Effectiveness Initiative to bring the resources of Medicare to bear on the question of what works in the practice of medicine. This initiative, in HCFA's view, was intended to help it fulfill its responsibilities for ensuring the quality of care of some 30 million Medicare beneficiaries.

The initial objectives of the Effectiveness Initiative were (a) to assess the merits of alternative health care interventions; (b) to provide information that would help clinicians in the management of their patients; (c) to assist and improve the Medicare program's quality assurance efforts; and (d) to aid policymakers in allocating Medicare resources. The subsequent evolution of the DHHS effectiveness and outcomes research programs has made it clear that improving patient outcomes is a unifying, primary objective and that identifying additional issues for further research is also important.

HCFA originally identified the following activities as elements of the Effectiveness Initiative: (a) monitoring time trends in the use of services by the Medicare population; (b) analyzing geographic (population-based) variations in the use of services and in outcomes of care; (c) assessing interventions by clinical demonstrations, observational studies, and randomized controlled trials (RCTs) in addition to monitoring and analyses of variations; and (d) feeding information back to clinicians.

THE BROADER CONTEXT

The Effectiveness Initiative did not occur in a vacuum. Within HCFA, it represented another step in the evolution of its responsibilities in quality assurance. Its other responsibilities include the Peer Review Organizations

(PROs), which followed the Professional Standards Review Organizations (PSROs), and the periodic release of hospital mortality data, a highly controversial step that has provided a powerful stimulus for clarifying the usefulness of mortality data as a measure of quality. In a different vein, HCFA has acted, through its Bureau of Data Management and Strategy, to facilitate the research community's access to its major data bases. In general terms, then, the Effectiveness Initiative both extended these earlier efforts and brought them into a more coherent framework.

Elsewhere in DHHS an Outcomes Research Program had been authorized by Congress in 1987 and was being administered by the National Center for Health Services Research (NCHSR). This program, inspired largely by the work of John E. Wennberg and associates in small-area variations in utilization and outcomes of medical interventions, invited research proposals in late 1988 and announced the first four awards in September 1989, a few weeks before the conference held by the Institute of Medicine (IOM). The program intends to make a number of additional awards on a regular cycle.

Conceptually, outcomes and effectiveness research are very similar; differences lie in legislative, administrative, and funding histories. Consequently, when the DHHS, through Secretary Louis Sullivan, announced in mid-1989 that the department was including the HCFA Effectiveness Initiative in a comprehensive outcomes and effectiveness research program, the announcement was greeted with approval by many in the health services research community.

Another strong general influence in the evolution of DHHS efforts has been the emphasis on appropriateness of care. The issue here deals with whether effective medical interventions—effective, that is, in the context of normal practice situations—are being used appropriately or inappropriately, given the indications for use and the characteristics of the particular patient. Robert Brook and his colleagues have been strongly associated with this research emphasis and have published several important papers in the past few years. Appropriateness research, then, constitutes one more converging stream of influence in the broader developments leading to the DHHS effort in effectiveness and outcomes research.

Ideas in good currency influence Congress, as well as the executive branch. Various legislative proposals in 1988 and 1989, therefore, culminated in the Omnibus Budget Reconciliation Act of 1989 and a reorganization of part of the Public Health Service (PHS). The NCHSR was disestablished, and the Agency for Health Care Policy and Research was authorized in its place. The new agency absorbed from the prior organization the functions of health services research (including outcomes research) and technology assessment, especially the PHS advisory function to Medicare.

In addition, a new function was added, namely, responsibility for developing medical practice guidelines. This responsibility represents congres

sional thinking that such guidelines represent the practical application of outcomes and effectiveness research to the practice of medicine.

This set of developments—evolution of the HCFA responsibility in quality assurance, emergence of outcomes research and its incorporation into federal research programs, articulation of the general concern for appropriateness and of specific research approaches in this area, and creation of a new federal agency with responsibility for the development of practice guidelines—provides the nutrient bath in which the IOM contribution has grown.

The Institute of Medicine's Contribution

In planning the Effectiveness Initiative, HCFA consulted widely in 1987 and 1988 with representatives of medicine, health financing, and health services and policy research. It also coordinated its efforts closely with other agencies of DHHS. In August 1988, William L. Roper, then Administrator of HCFA, asked Samuel O. Thief, president of the Institute of Medicine, to convene a group of clinicians to advise the agency on the Effectiveness Initiative. HCFA specifically asked the IOM for advice concerning what clinical conditions ought to receive priority in the initial period of the new program. Clinical conditions, rather than specific procedures or technologies, were chosen as the unit of analysis because they permitted examination of the full range of patient care opportunities, including prevention and follow-up care.

The IOM Clinical Workshop

The IOM hosted a meeting of clinicians for the above purpose in October 1988. This "clinical workshop committee" recommended that five conditions receive highest priority: angina (stable and unstable); acute myocardial infarction; carcinoma of the breast; congestive heart failure; and hip fracture. These conditions were selected because of their high prevalence, the substantial burden they impose on elderly persons, appreciable variations in use of services and in outcomes, high costs, and the existence of alternative ways of managing patient care that reflect professional and clinical disagreement or uncertainty. The committee also recommended a second tier of clinical conditions for later attention: cataracts, depressive disorders, prostatic hypertrophy, and transient ischemic attacks with or without occlusion.[1]

[1] The report of this study was published as *Effectiveness Initiative: Setting Priorities for Clinical Conditions* in April 1989; it is available from the National Academy Press (Report No. IOM-89-04).

The Condition-Specific Workshops

After the clinical workshop, HCFA asked the IOM to organize three additional workshops, one on each of the three clinical conditions of highest priority to the agency—namely breast cancer, acute myocardial infarction, and hip fracture. These condition-specific workshops were held in March, May, and July 1989, respectively. Each had three objectives: (a) to identify key research questions in more detail than had occurred at the October workshop; (b) to identify critical patient care topics deserving further investigation; and (c) to propose appropriate research strategies and methods.[2]

The IOM appointed a core committee to oversee the entire series of workshops. This core group included: Kenneth I. Shine (chair), Maureen M. Henderson, Emmett B. Keeler, Barbara J. McNeil, David G. Murray, Alan R. Nelson, J. Sanford Schwartz, G. Richard Smith, and Harold C. Sox. (Drs. Shine, Murray, Nelson, Smith, and Sox had also been on the clinical workshop committee.) For each meeting, IOM augmented the core group with additional experts in the condition under consideration. The names of all participants can be found in the committee rosters in the front of this monograph.

Effectiveness and Outcomes Conference

The September 1989 IOM conference, "Effectiveness and Outcomes in Health Care," concluded this series of activities. It had four main objectives: to explore the social, clinical, and legislative environment for research on the topic of "what works in the practice of medicine"; to review the conclusions and recommendations of the IOM workshops on breast cancer, acute myocardial infarction, and hip fracture; to highlight four important methodological issues in effectiveness studies, specifically, use of administrative data bases, collection of primary data, development and use of outcome measures, and applications in clinical practice; and to examine the question of where we go from here.

Proceedings

This book represents the proceedings of that conference and is divided into five parts. The first is an introduction consisting of the foregoing description of how the Effectiveness Initiative and IOM's role in it evolved

[2] The reports of the three research workshops have been or are being published as a series: *Breast Cancer; Hip Fracture; Acute Myocardial Infarction,* with the subtitle *Setting Priorities for Effectiveness Research.* All are available from the National Academy Press.

and the summary statement of the IOM core committee. Although most of the issues in the summary are addressed in the individual workshop reports, the committee believed it would be helpful to draw them together in a single statement. That statement was distributed in advance to participants in the September conference to stimulate discussion and is published here as Chapter 2.

The second, third, fourth, and fifth parts comprise the conference proceedings papers. Each section focuses on one of the four main objectives of the conference described above.

2

Promise and Limitations of Effectiveness and Outcomes Research

Summary Statement of the IOM Core Committee

The committee first acknowledges the major contribution of the Health Care Financing Administration (HCFA) in advancing the conceptual and practical ideas behind effectiveness research. Studies of the actual delivery of health care and "what works in the practice of medicine" are a legitimate and important priority for health scientists. We therefore applaud the imagination and efforts of William Roper, former Administrator of HCFA, Acting Administrator Louis Hays, and the HCFA staff for their leadership and energy in stimulating interest and focusing attention upon effectiveness research.

IMPORTANCE OF EFFECTIVENESS RESEARCH

The emerging emphasis on effectiveness is welcome, if not overdue, as a complement to the National Institutes of Health's (NIH) emphasis on efficacy. The distinction between effectiveness and efficacy is an important one. Efficacy is typically defined as the outcome of an intervention when it is applied in "ideal," well-controlled circumstances, such as those inherent in prospective randomized controlled trials (RCTs). By contrast, effectiveness means the outcome of that intervention when it is applied in everyday or average circumstances (such as the daily practice of medicine); the latter may include patient groups that differ marginally or considerably from those studied in RCTs.

The desirability of high-quality effectiveness research was clearly demonstrated in the various committee deliberations. The committees identified numerous areas in which the efficacy of a particular therapy has been documented through RCTs but in which the effectiveness of that therapy is not necessarily predicted by the results of the efficacy studies.

Two examples make this point. First, almost all RCTs in the treatment

of acute myocardial infarction (AMI) exclude individuals over age 65. As a result, it is impossible to extrapolate the results of such studies directly to the Medicare population. Second, although excellent clinical trials have revealed the most efficacious treatment for breast cancer, other research makes it clear that physicians offer different options to their older and their younger patients.

Medicare Data Bases

We also believe that the existing Medicare administrative data bases (known as the Medicare-Medicaid Decision Support System) contain much potentially useful data. Over 31 million elderly individuals are currently in the Medicare program, and they are covered for virtually all inpatient hospital care and a considerable portion of their outpatient care. Beginning in 1990, outpatient prescription drugs and various screening tests will also be covered.[1] Thus, the size and scope of the HCFA data files offer remarkable opportunities for effectiveness studies based on monitoring and surveillance of large populations.

This data base allows us, with considerable accuracy and for specific diagnoses, to (a) track the use of services, the patterns of care, and the costs of those services and care over time; (b) monitor trends in care received and to measure the variations (i.e., patterns of care) by geographical region, institutional providers, type of practitioners, and patient demography; and (c) track what happens to Medicare beneficiaries over time (for example, to learn certain rates of death and utilization-related events such as rehospitalization or use of home health services following a hospital admission).

KEY ASPECTS OF EFFECTIVENESS RESEARCH

Reliability and Validity of Data

All the IOM workshops were concerned with the reliability and validity of data. These problems center on the adequacy of information in the Medicare files about diagnosis, procedures, coding (in general as well as for new technologies), and timing of patient management events.

Data on AMIs, for example, raise the difficulty of separating the hospitalization for the infarction from a hospitalization for cardiac catheterization two weeks later, thus calling into question up to 20 percent of these diagnoses. As another example: when the count of surgical procedures for hip

[1] Editors' Note: Although true at the time this statement was drawn up, these benefits ultimately were not covered because of the repeal of the Medicare Catastrophic Coverage Act late in 1989.

fracture is compared to the overall annual incidence of hip fracture, nearly 20 percent of patients cannot be matched with a procedure.

Moreover, the long lag between the introduction of a discrete new technology and its designation with a unique code means that an intervention such as tissue plasminogen activator cannot yet be identified in the Medicare data bases. Many questions regarding the role and the effectiveness of mammography, biopsy, and surgical therapy for breast cancer require that the temporal relationships of those interventions to outcomes be known.

Efforts by HCFA to improve coding and dating of data and to find methods for validation of data are impressive. The current Medicare files are superior to any similar set of insurance claims files that might be tapped today for national effectiveness research. Nevertheless, this area will continue to present major problems. For one thing, data generated for reimbursement may be inadequate for research purposes. In addition, as long as coding is driven mainly by the need to develop charges for care, biases derived from efforts to maximize reimbursement may be introduced. Assessment of procedures may be limited, for example, because the Medicare files from hospital discharge abstracts code for only three procedures. Moreover, in the case of procedures, the major Medicare data bases (Part A and Part B) are not consistent in the coding systems used. Clinical vagaries also present problems with regard to initial diagnoses and treatments; accurate description of the type of hip fracture is one example.

In short, effectiveness research requires, as does any area of scientific inquiry, confirmation, reconciliation, and validation of data on a continuing basis. Efforts to validate Medicare, Medicaid, and other important data banks must be maintained or even expanded. In some cases, critical data will be made available through the Medicare Peer Review Organizations (PROs), but periodic independent validation of PRO data will also be essential. We conclude that continuing validation of the Medicare data base is essential to the success of an effectiveness research program that relies heavily on those files. Effectiveness research should ensure such validation.

Longitudinal Studies

The three IOM clinical conditions committees noted the need to follow patients over time and across settings of care. The limitations of unsupplemented hospital outcomes data were particularly striking. The core committee recognized the clear need for an episode-of-care approach to analyzing the outcomes of care, which among other things calls for appreciable efforts to collect ambulatory and other out-of-hospital information, including posthospitalization outcomes data.

Again, some examples may be useful. Because of varying lengths of stay after inpatient hospital treatment for AMI, mortality rates from the

acute event may need to be calculated at some specified, minimum period of time following the hospital episodes (for example, 30 days after admission or after discharge) rather than from in-hospital deaths. In-hospital mortality associated with initial treatment of both breast cancer and hip fracture is very low; only longitudinal studies can identify the effectiveness of a particular treatment combination. Finally, appropriate care for hip fracture requires that the earliest stages of rehabilitation occur during acute hospitalization but that rehabilitation be fully pursued in whatever settings are appropriate and available —a rehabilitation hospital, skilled nursing facility, home health care, or family care.

In short, before the effectiveness of a particular therapeutic modality can truly be determined, information about care in a variety of sites and from a variety of practitioners will be needed. The diversity of sites of care, coupled with the requirement for longitudinal studies, presents an exceptional challenge to effectiveness research.

Tracking the Patient

Effectiveness studies must follow the patient across all levels and sites of care. Efforts by HCFA to link their inpatient (Part A) and outpatient (Part B) files will help. Mechanisms for obtaining information about ambulatory care, including care given in offices and clinics, will also be essential.

Adequate information about drug therapy is particularly important. This concern arose for all three conditions addressed in this project. The assessment of treatment for breast cancer depends on knowledge of the chemotherapeutic agents administered, including their timing and schedule. The opportunity to obtain these data for Medicare recipients should be strongly supported, and every attempt should be made to enter such information (for example, at the point of sale) in a manner that facilitates matching drugs with patients and diagnoses. In addition to the challenges offered by outpatient use of medications, data on hospital drugs, including timing of administration and coding of new agents, continue to be needed.

Several instruments intended to capture health status and clinical data are being developed for use at the time of hospital admission and discharge or in conjunction with nursing home or home health care. Such instruments should be simple and comprehensive, yet sensitive and economical in terms of time and money. These attributes would be enhanced by appending to generic instruments some disease-specific risk stratification questions. This requires coordination of instrument development and application. Assessing changes in the health of patients tracked throughout the health care system is easier if the content of the data sets derived from these instruments is consistent.

Health Status Assessment

Because health care aims at more than simply extending life, effectiveness research must consider the range of outcomes—that is, the diversity of health states—appropriate to the patient with the condition under study. Desired outcome may differ with stage of patient management (from screening and prevention through therapy and rehabilitation, to control of symptoms and palliation in terminally ill patients). Each workshop committee recommended that attention and care be given to the definition, design, and development of appropriate measures of outcome as part of the effectiveness research program, and each recognized the need to go beyond administrative files and medical records data to obtain outcome information directly from patients or their families.

The use of a sensitive, yet simple, patient health status instrument (or set of instruments) in longitudinal studies is highly desirable. This might be based on a questionnaire completed by the Medicare patient and/or the patient's family, or both. When that is not feasible, or when multiple views of the patient's health status are required, the relevant information might be provided by the health care practitioner.

Such an instrument should provide information about activities of daily living as part of a functional assessment. Further, it should yield information about emotional aspects of health status and, if possible, some insight into cognitive function.

The core committee endorsed the proposal that such an instrument be applied to 5 percent of all Medicare patients upon their enrollment into the program. Even better, it could also be applied to, say, 5 percent of those enrollees every five years thereafter, to provide a rich longitudinal data base. For those beneficiaries who seek care for an acute illness, this health status document could be updated at the onset of care for the acute episode and at periodic intervals thereafter; a different schedule might be devised for beneficiaries who seek care for chronic illness. In this way, the data from a generic health status instrument would contribute to risk stratification for this 5 percent cohort and would provide insight into the outcomes of efforts at prevention as well as management of acute illness. The instrument might be applied more frequently to the 5 percent cohort as it grows older and the incidence of illness becomes more frequent.

A health status instrument could be used for all Medicare patients as part of any long-term effectiveness studies of treatment for an acute illness. For example, it could be filled out at the time of hospitalization after hip fracture, at discharge, and periodically thereafter to ascertain functional recovery after the episode of hip fracture. The committee believes that such a health status instrument could be used effectively in risk adjustment, as an outcome measurement, and to assess prevention efforts. We recognize that for stratification of patients in mortality analysis, clinical information on

the acute crisis will be more important than overall health status. The main use of health status information is in charting the patient's progress in terms of functioning and long-term recovery, not in acute episodes.

Many reliable and valid generic instruments for measuring health status are available and could be synthesized effectively for this purpose. The committee believes that the Karnofsky index (a physician-completed, cancer-specific index to measure physical status and activities) will not be satisfactory for this purpose and recommends that attention be given to developing a unique health status instrument. We also acknowledge that more specialized instruments for evaluating health status may be required for specific studies and note that many disease-specific instruments are already available. Examples include measures of pain and of psychological factors in illness. The committee believes, however, that a simple, generic document for use at the time of acute illness as well as for follow-up of a 5 percent cohort is critical to the success of effectiveness studies.

Risk Stratification

Little can be concluded about effectiveness when variations in patients' clinical status are unknown and uncontrolled. Risk stratification adjusts for differences among patients, making use of such concepts as case mix and disease stage. By whatever term it is described, the adjustment issue was central for the committees that examined breast cancer, acute myocardial infarction, and hip fracture. The acceptability of effectiveness research results to the clinical community will depend in large part upon convincing evidence that risk stratification and stage of disease have been accounted for.

In breast cancer, for instance, accurately determining the stage of disease at the time of initial diagnosis often determines what therapeutic options are considered or offered to the patient (and thus are at issue in any effectiveness investigation). Disease stage also has a profound impact on outcomes, regardless of treatment. Thus, any comparison that neglects risk adjustment is unavoidably biased. Likewise, understanding the effectiveness of treating hip fracture requires the careful identification and assessment of comorbidity, cognitive function, and previous functional capacity, which are strong predictors of outcomes. HCFA has recognized the importance of this concept; in fact, risk stratification provided crucial insights into HCFA's studies of coronary angioplasty versus coronary bypass surgery as a treatment for AMI.

Developing a good health status instrument should greatly improve risk stratification and determination of disease stage. Efforts to identify morbidity, comorbidity, and acute severity should be encouraged. Identifying the stage of concurrent disease is necessary but may not be sufficient to determine health status, and additional clinical data may be required; these

data might be gathered through the PRO mechanism, demonstration projects, or other kinds of trials.

Newer, innovative methods for risk stratification should also be explored. Use of the Predictive Instrument for Acute Ischemic Heart Disease, developed by Michael Pozen and Harry Selker, to determine an emergency room patient's true probability of having acute cardiac ischemia, including AMI and unstable angina (the group most physicians would consider appropriate for coronary care unit admission), is one example of stratification in studies of AMI. Demographic data, including socioeconomic information and data about prior hospitalization and other use of the acute care system, will also contribute to appropriate risk stratification.

Prevention

The prevention of disease is a long-standing ideal of health care, but it suffers from many theoretical and practical difficulties. The importance of early diagnosis is exemplified by the decision to include screening mammography as a benefit under Medicare. However, the need for research on interventions that *prevent* illness in the Medicare population continues. Prevention involves not only primary prevention (for example, of the original ailment or catastrophic event) but also secondary prevention (of a second hip fracture, another myocardial infarction, or a recurrence of malignancy). Early detection of malignancy using mammography; prevention of atherosclerosis by adequate control of blood pressure, diet, and smoking cessation; and the role of estrogen therapy or certain diuretics in prevention of osteoporosis in hip fracture are all examples of primary or secondary prevention. We need to develop data bases that can identify risk factors and preventive measures that are provided outside the hospital setting, and perhaps without regard to the specific diagnosis at issue.

Because prevention is a long-term intervention, it clearly must begin before age 65 for the Medicare population. Thus, our society has a major stake in preventing disease before persons become eligible for Medicare, if only to limit the burden of disease during Medicare coverage. At the same time, third-party payers for individuals under age 65 have an interest in effective prevention in the younger population. As the powerful HCFA data base grows and improves, the capacity to connect it to data bases for populations under age 65 should be carefully explored and expanded. HCFA has improved Medicaid data in some states, which enhances its ability to develop information about a special segment of the population, namely, the elderly poor. Expanding these connections to other states and exploiting the opportunity to link them with information from private insurers, prepaid group practice systems, and other organizations will be essential if the effectiveness of care before age 65 is to be related to events after age 65.

Despite substantial problems of public-private relationships, accessibility,

confidentiality, economic competitiveness, and similar factors, private projects of this kind should be investigated. Developing such connections at a relatively early stage might facilitate more cost-effective results for the private sector as well as enhance the validity of Medicare's results. The sharing of well-developed approaches to effectiveness research between public and private sectors will benefit all.

The Aging Process

Effectiveness research will provide valuable insights into the aging process in our society, independent of its contributions to our understanding of the effects of health care. Tracking the health status of a cohort of patients and understanding the impact of health status on acute illnesses should aid in decision making about prevention, screening, diagnosis, therapy, and rehabilitation of a cohort of aged patients.

PRINCIPLES TO GUIDE FEDERAL EFFORTS

As federal efforts in effectiveness research evolve, certain principles deserve consideration. The challenges offered by the need for data validation, longitudinal surveillance, risk stratification, health status assessment—together with the need for pursuing a diverse but coordinated approach to effectiveness research—led us to the following observations.

Range of Study Designs

The core committee believes that a diversity of approaches is needed in this program; proper effectiveness research and outcome studies will require research and demonstration projects, case-control studies, and the like to be conducted by a wide variety of investigators. Such studies will be important not only to test instruments and hypotheses, but also to validate further data from HCFA or other sources. Given this need for diversity, and the complexity of funding mechanisms and research methods required, the committee strongly endorses a coordinated, comprehensive, and balanced DHHS approach to effectiveness and outcome research.

Funding

The importance of research funding that emphasizes extramural investigation and investigator-initiated projects without excluding intramural work or research contracts was reiterated by every study in this project. The recommended use of many research methods—including randomized trials, various quasi-experimental efforts such as case-control studies and demon

strations projects, natural history studies, and other approaches—requires a commensurate level of resources for effectiveness projects.

Balance of Approaches

Although Medicare (and possibly other) data bases provide powerful tools for biostatistical and econometric analysis, the committee believes that clinical input and participation are critical in effectiveness and outcomes studies. In this regard we believe that HCFA's decision to involve the IOM in the early stages of its activities has been important. Indeed, all groups working in effectiveness and outcomes research must have a good balance of statistical, economic, and clinical perspectives. Without such balance, the risks of misinterpretation, underinterpretation, or overinterpretation of data are significant.

Coordination

We endorse the idea of appointing a high-level advisory committee of individuals with clinical, economic, statistical, and organizational expertise for the effectiveness research program, as suggested by Assistant Secretary for Health James Mason. We further urge that this body have an experienced staff and adequate support so that it can function effectively.

The DHHS unit responsible for effectiveness research will also be responsible for coordinating the activities of such diverse agencies as the National Center for Health Services Research,[2] HCFA, and others. This coordination will need to (a) support the development of new instruments, data bases, and research methodology that can be shared by investigators, (b) prevent redundancy in funding between agencies, and (c) foster appropriate translation and dissemination of results obtained in effectiveness research to help health care providers, policymakers, and the public.

Both investigators and providers would benefit from regular updates on data bases. Dissemination of information will be a considerable responsibility as the field of effectiveness research evolves. Several workshop participants congratulated HCFA on its willingness to provide investigators with access to HCFA data bases; such accessibility and cooperation are strongly encouraged and supported.

Technology Assessment

The effectiveness research organization should also consider some practical aspects of technology assessment. NIH studies do not provide funds

[2] Now the Agency for Health Care Policy and Research.

for the use of a procedure or technology that is under investigation. Moreover, Medicare and other third-party payers will not pay for such technologies. To obtain meaningful data on effectiveness, HCFA may have to provide financial support for as-yet-unapproved technologies used in an approved study. This difficult issue needs examination.

Influencing Provider Behavior

The effectiveness and outcomes programs of DHHS also need to support studies on how to translate research results into practice. The committee is impressed with how little we know about the factors that influence the behavior of health care providers. High-quality data published in peer-reviewed medical literature and supplemented by direct feedback to physicians (for example, through new computer systems) are important mechanisms. Certainly economic incentives and constraints have a major impact. If economic limitations are relied upon as a principal method to influence behavior, however, they may affect some patients in an undesirable way. We need to understand more about how payment mechanisms affect the behavior of health care providers.

The Confluence of Biomedical, Health Services, and Effectiveness Research

Our society (indeed, the world) has benefited substantially from scientific advances made by biomedical research. The scientific substrate of clinical medicine is maintained and extended by such research. What works in medical practice derives, in a fundamental way, from the success of public and private investments in biomedical research. These efforts deserve continued support.

Clinical epidemiological research, with its emphasis on the incidence and prevalence of disease, provides a necessary reminder of the magnitude and importance of clinical problems across the nation. It thus provides indispensable guidance to policymakers about research priorities. Epidemiological research using claims data has been the principal means of demonstrating the variations in use and outcome of medical interventions.

Health services research has added important new dimensions to our understanding of the mechanisms by which health care is organized, financed, and delivered. It has laid the foundation for connecting clinical data on use with expenditure data on resource consumption; it has nurtured the development of assessments of functional and health status; it has promoted concern for outcomes research; and it has been responsible for numerous advances in measurement, methodology, and data base development.

Effectiveness research adds another dimension to these activities, one that can be extremely valuable in guiding physicians, patients, the public,

and policymakers. It does not supplant existing efforts in biomedical, epidemiological, or health services research. On the contrary, it draws pertinent data from all of these sources and integrates them in an effort to advance the assessment of clinical practice.

The IOM committees for the three workshops were repeatedly confronted with evidence that short-term, quick answers to effectiveness will be rare and of limited value. To describe hospital mortality after a particular surgical procedure for hip fracture provides little meaningful insight into the first-year overall mortality rate of 25 percent among patients experiencing a hip fracture; nor does it convey any important information about the level of function of such a patient six months or a year later. Because level of function has important health and economic consequences, the effectiveness of treatment for this condition requires the longer view.

The core committee reflects all these perspectives in its emphasis on the need for access to data about drugs and procedures (and outpatient care in general), information on risk stratification, the development of appropriate tools for measuring health status, and longitudinal studies of a cohort of Medicare beneficiaries. Developing a consistent, comprehensive federal approach that involves many agencies, adequate dissemination of information, support of diverse analytical approaches, vigorous efforts at validation, and development of effective tools for communicating results to providers of care will do much to advance effectiveness studies.

PART II

OVERVIEW

3

Research on the Effectiveness of Medical Treatment: New Challenges and Opportunities

J. Jarrett Clinton

In this chapter I present the perspective of the Department of Health and Human Services (DHHS) on today's environment for a new component of health services research: medical treatment effectiveness research.

In a November 1988 issue of the *New England Journal of Medicine*, William Roper, then Administrator of the Health Care Financing Administration (HCFA), presented a bold plan to evaluate and improve medical practice in the United States. True, others had also called for and were engaged in measuring the outcomes of medical care, but Dr. Roper's article jolted many into realizing the considerable potential of these measures (particularly as they apply to quality of life) for improving the quality of medical care by using population-based data to indicate which practices are most effective.

Dr. Roper made these activities essential components of HCFA's Effectiveness Initiative. He invited substantive collaborative efforts between the public and private sectors to:

- share data bases among public and private payers,
- create greater uniformity in the collection of information to determine and measure medical outcomes,
- establish a critical role for practicing medical professionals in planning and carrying out research on medical treatment outcomes,
- accelerate the training of health professionals in evaluation sciences such as decision analysis and clinical epidemiology, and
- refine medical practice guidelines (parameters and standards of medical practice created by practitioners and their professional organizations).

Arnold Relman, editor of the *New England Journal of Medicine*, stated in an accompanying editorial that "... no one should underestimate the size or difficulty of the task. However, the logical necessity of this seems clear.

We can no longer afford to provide health care without knowing more about its successes and failings. The Era of Assessment and Accountability is dawning at last."

I expect that most of us endorse and enthusiastically support the positions taken by Dr. Roper and Dr. Relman. For some, this thesis is already a guiding principle; for others, it is still a dream. For some, it may be a source of considerable anxiety. The concept, indeed, is challenging, provocative, and fraught with difficulties. Yet it gives us extraordinary potential for advancing the practice of health care.

Medical Treatment Effectiveness Program

In fiscal year (FY) 1990 DHHS will expand the original Effectiveness Initiative enunciated by Dr. Roper and others into a more formal Medical Treatment Effectiveness Program. The increased visibility of effectiveness research, and professional assimilation of that research, reflects the DHHS belief that years of careful scientific studies in this area have produced strong and credible results. We intend to use these advances in knowledge, and the further questions they raise, to catalyze DHHS support for and participation in this dramatic effort.

Secretary of Health and Human Services Louis Sullivan has assigned primary responsibility for this new program to the Public Health Service (PHS). Consistent with the Dr. Sullivan's desire that the Medical Treatment Effectiveness Program be a cohesive, department-wide effort, PHS is collaborating closely with HCFA to develop sound, fresh, and forward-thinking strategies.

The long-term goal of the program is to change the assessment of health care services, research and financing from a focus on processes, that is, procedures and interventions, to a focus on patient outcomes of these processes. The central questions thus become: Has the patient improved? Has the quality of his or her life improved? By how much?

The specific purpose of the Medical Treatment Effectiveness Program is to improve the effectiveness and appropriateness of health care services by enhancing our understanding of which health care practices are most effective—what works best. Four components, or sets of activities, form the basis of the program.

1. *Collection and development of data* This will be undertaken to expand the data bases available for analysis and to improve the ability to link Medicare files and other data bases on additional populations.
2. *Research on patient outcomes and clinical effectiveness* Specific treatment will be assessed through studies such as small area analysis and multidisciplinary epidemiological research. For example, in FY 1989 the

National Center for Health Services Research (NCHSR)[1] of the PHS awarded four major research grants to assess alternative means of managing of myocardial infarction, different procedures for treatment of cataracts, management of prostatic hyperplasia, and nonsurgical interventions for lower back pain. In addition, NCHSR awarded planning grants for assessments in several areas, including total hip replacement, colon polyps, peripheral vascular disease, and ischemic heart disease.

3. *Dissemination and assimilation of findings* As outcomes research is completed, results will be widely disseminated through journal articles, information networks, and conferences sponsored by HCFA and NCHSR. We will also make use of the resources and expertise of the National Library of Medicine and the Health Resources and Services Administration. As a component of the latter agency, the Bureau of Health Professions will convey appropriate information to geriatric education centers, family medicine departments, general internal medicine departments, and the network of area health education centers which, in some states, are powerful continuing education networks. We also intend to explore new approaches to medical education to ensure that research findings are incorporated in academic curricula, continuing education, and other professional education programs.
4. *Practice guidelines* The fourth and most challenging component of the Medical Treatment Effectiveness Program is the development of practice guidelines, that is, parameters and standards of care. These guidelines must be created by practicing physicians, be based on science, and be practical, explicit, and subject to revisions as needed. The research findings generated by this program and by others will facilitate the development of these guidelines. We expect this process to involve the full participation of the following:

    - professional organizations, such as the American Medical Association, and specialty organizations, such as the American College of Physicians;
    - scientific bodies, including the Institute of Medicine (IOM);
    - academic medical centers;
    - standard-setting organizations, for example, the Joint Commission on Accreditation of Healthcare Organizations;
    - quality measurement organizations, such as Peer Review Organizations (PROs);
    - research-based organizations, for example, the American Medical Review Research Center and the Association of Health Services Research.

[1] As of December 1989, NCHSR became the Agency for Health Care Policy and Research. The agency is the main source of federal support for research on problems related to the quality, delivery, and costs of health services.

In the not-too-distant future, nursing professionals must also be engaged to develop nursing care guidelines. Patient advocacy groups must be incorporated to ensure that the programs, processes, guidelines, and measures are relevant and understandable from the patient's perspective.

Implementation of the Program

To accomplish the goals and objectives of the Medical Treatment Effectiveness Program, Dr. Sullivan intends to implement it from a departmental perspective. Much of the research activity and budget notations will be assigned to NCHSR. Much of the data development work, however, will be done by HCFA. In addition, all components of the PHS, including the National Institutes of Health, the Health Resources and Services Administration, the Food and Drug Administration, the Centers for Disease Control, and the Alcohol, Drug Abuse, and Mental Health Administration, will participate in program development, implementation, and review.

The President's FY 1990 budget request for the program is $52 million. We plan to use this money to support a broad array of activities in each of the four program components—data development, research, dissemination of information, and development of guidelines. Congress is in the process of making final decisions on the President's budget. Collectively, we must ensure that this program has adequate resources to accomplish its far-reaching goals. It has a preliminary House mark of $20 million and a Senate mark of $35 million. I do not need to explain how essential it is that this program be adequately funded right from the start. Certainly, we will implement our agenda, even with reduced funding, but our progress will be slower and our goals more elusive if smaller budgets are appropriated.

I want to reemphasize that the Medical Treatment Effectiveness Program has been incorporated into key objectives enunciated by Dr. Sullivan as goals for his administration at DHHS. We therefore have the full support of the DHHS in facing the greatest of challenges in health care.

Priorities for Research

There has been substantial debate as to who would establish research topic priorities and by what criteria. HCFA has obviously asked the IOM to assist in setting priorities, and this volume contains summary judgments regarding the three clinical conditions that have received substantial review (1,2,3).

We expect to respond to these, as well as to recommendations from appropriate advisory councils, Institutes in the National Institutes of Health, and the Alcohol, Drug Abuse, and Mental Health Administration

Congress, too, powerfully affects us. The Senate Appropriations Committee suggested the following topics for full consideration:

- effectiveness of prevention services,
- effectiveness of alcohol, drug abuse, and mental health treatment programs,
- effectiveness of nonphysician health providers, such as nurse practitioners and physician assistants.

Final judgments on broad research topics will consider all of these perspectives. Through an intradepartmental committee incorporating PHS, HCFA, and other policy offices of DHHS, we are certain that a research agenda can be agreed upon.

New Patterns of Collaboration

The work planned cannot be accomplished within the traditional patterns of relatively distinct and separate research undertakings. New teams of investigators, across disciplines, from the different institutions, and transcending traditional academic and geographic barriers, are essential for program progress. We need new constellations of researchers and data base managers. To facilitate the assimilation of findings, we need the practitioners and the specialty societies.

Yet, collaboration based on mutual cooperation and trust does not occur spontaneously. It requires each person, organization, and institution to reaffirm that quality of care in America transcends the traditional precepts held by each of these entities. It requires strong and visionary leadership.

Finally, as one might expect, we have heard criticisms of this program. These include arguments that:

- There are not sufficient researchers to undertake a large research program of this nature;
- The health services research community is not well organized or comfortable with collaborative efforts;
- Social scientists and physicians have yet to demonstrate large-scale collaborative efforts;
- The data bases for population-based research are inadequate;
- The PRO data bases and processes are too dissimilar across states to create a unified national approach;
- Organized medicine resists guidelines and parameters;
- "True science" is found only in randomized controlled trials.

We believe these arguments are not based on fact and that they ignore potential. Each is a challenge to the private sector and to government—a

challenge to commit ourselves to cooperative efforts to strengthen the scientific foundation on which clinical judgments rest.

I have outlined here an exciting new program reflecting DHHS's determination to ensure that medical care is of the highest quality. We know that there are no quick or easy answers to many of the questions surrounding effectiveness of medical treatment. Because there are none, we are positioning the Medical Treatment Effectiveness Program for the long haul. Uncertainties will give way to scientifically sound research, and answers will come. By going forward with each component of our program—data development, outcomes research, dissemination and assimilation of findings, and development of practice guidelines—we will add more knowledge to the physician's armamentarium. A higher quality of care will become the new standard.

We ask that the private sector join us in this effort. The synergism of federal-private sector collaboration will be the force moving us closer to our mutual goals.

References

1. Jackson, V.P. Breast Cancer. Pp. 53-60 in *Effectiveness and Outcomes in Health Care*. Heithoff, K.A. and Lohr, K.N., eds. Washington, D.C.: National Academy Press, 1990.
2. Murray, D.G. Hip Fracture. Pp. 61-64 in *Effectiveness and Outcomes in Health Care*. Heithoff, K.A. and Lohr, K.N., eds. Washington, D.C.: National Academy Press, 1990.
3. McNeil, B.J. Claims Data and Effectiveness: Acute Myocardial Infarction and Other Examples. Pp. 65-70 in *Effectiveness and Outcomes in Health Care*. Heithoff, K.A. and Lohr, K.N., eds. Washington, D.C.: National Academy Press, 1990.

4

The Health Care Financing Administration and the Effectiveness Initiative

Louis B. Hays

My objective in this chapter is to discuss the basis of the Health Care Financing Administration's (HCFA) interest in the subject of effectiveness, how we got to where we are today, what our role has been, and what it will be in the future.

HCFA'S INTEREST IN EFFECTIVENESS

In late 1987, HCFA Administrator William Roper and several others of us at HCFA became aware that research had been going on for years in the area of effectiveness, but for some reason it had not yet gotten into the consciousness of health policy officials, at least in Washington, D.C. We began to review the available information and to talk to some of the people who have been and continue to be so actively involved—people like John Wennberg, Robert Brook, and David Eddy. We began to learn more and more about effectiveness, or, as we say, "what does and does not work in the practice of medicine." We became increasingly concerned about the lack of empirical data to support so much of what occurs in the practice of medicine. To paraphrase what a prominent person in this field likes to say, "Most procedures in medicine, whether they be surgical procedures, diagnostic tests, or whatever, are not subject to the same elementary scrutiny for safety and effectiveness that drugs must undergo before they are approved for public use."

Given the fact that we have responsibility for 33 million Medicare beneficiaries, an indirect responsibility for many millions of Medicaid recipients, and a responsibility for the quality of care, particularly for Medicare beneficiaries, we became increasingly interested in the area of effectiveness.

We also began to recognize the wealth of information that we have at our

disposal within HCFA—information on those 33 million Medicare beneficiaries and what happens to them as they go through the entire spectrum of the health care system. We have information on their treatment by the 500,000 or so physicians who provide services to Medicare beneficiaries, the 6,000 hospitals, the 15,000 nursing homes, and so on. The claims data that we have from our fiscal intermediaries and carriers contain a wealth of information.

In addition, through our Peer Review Organizations we have the capacity to generate more and more clinical information, which computers can marry with or connect to claims data. Together they provide an incredibly rich source of data for analysis, research, and evaluation.

Unfortunately, there is a certain risk in looking at effectiveness as a panacea, whether it be for improving quality of care or as a way of ensuring that all services are appropriate and that we are not wasting our health care dollars. We have already seen evidence that effectiveness can be used as an excuse for avoiding reform or other systemic changes. Nonetheless, above is a thumbnail sketch of the reasons that HCFA became so interested in the subject of effectiveness.

THE EFFECTIVENESS INITIATIVE

The first public HCFA effectiveness activity was a meeting convened by Dr. Roper in June 1988. He brought together many of the major players in the health policy arena to talk about effectiveness and the ability of HCFA and other researchers to use the data that we have available to learn more about what works in the practice of medicine. I believe the June 1988 meeting was a historic development because there was a consensus that effectiveness was an idea whose time had come and that the uses that HCFA and other researchers were making of claims and clinical data were the way to proceed.

Another critical event, which followed the June 1988 meeting, was an article published in the *New England Journal of Medicine* by Dr. Roper, Dr. Krakauer, and others outlining the HCFA approach to effectiveness and a very interesting companion editorial written by Arnold Relman. Dr. Relman heralded effectiveness as the third revolution in health policy in this country.

HCFA also collaborated with the Institute of Medicine (IOM) to develop a series of meetings on effectiveness. The first meeting was held in October 1988 to look at the broad areas of medicine that we should consider in our effectiveness work. It was, in effect, an agenda-setting meeting. Obviously we could not take on the entire spectrum of medicine. The meeting was very helpful in recommending priorities, and it resulted in a series of three follow-up workshops. The workshops addressed in more detail the areas of acute myocardial infarction, breast cancer, and hip fracture. All of this

culminated in the proceedings published in this volume, which give an idea of where we go from here.

I have mentioned William Roper several times, and I believe he deserves recognition for what he has accomplished. Dr. Roper would be the first person to acknowledge that he did not invent effectiveness and that long before we even heard of the word or the concept several people had spent many productive years of work in this area. However, I do think it is fair to say that Dr. Roper "discovered" effectiveness from the standpoint of the health policy agenda. Along with the IOM and the *New England Journal of Medicine*, he has helped to popularize the concept of effectiveness. Largely as a result of his efforts, the Bush administration and Congress have put effectiveness high on their list of priorities; as a result, we are clearly going to have increased funding from the federal government for effectiveness activities.

There are several other important activities within HCFA related to effectiveness. We will shortly be publishing our third annual hospital mortality release, showing for the nation's hospitals that participate in Medicare the actual and expected mortality rates in general and the rates for a number of specific conditions.

We are also working on our second annual release of nursing home information, which shows certain performance indicators for the 15,000 nursing homes across the country that participate in Medicare and Medicaid. While not directly part of effectiveness work, these efforts demonstrate the power of putting good information and data into the hands of both providers and consumers of health care.

FUTURE ACTIVITIES

I see the first meeting in June 1988, which produced an amazing degree of consensus within the health policy community, and this meeting today as bookends in terms of HCFA's leadership in the Effectiveness Initiative. This has been the critical first leg of the effectiveness race, and I believe we have completed it successfully. Now it is time for HCFA to pass the baton to the Public Health Service, which will assume leadership in the area of effectiveness within the Department of Health and Human Services.

I think that we can look forward to outstanding results. James Mason, Assistant Secretary for Health, has been a distinguished public health official for many years, having, among other things, headed the Centers for Disease Control in Atlanta. Dr. Mason has the full support of Louis Sullivan, himself a distinguished physician and academician who has established effectiveness as one of a small number of priorities for his tenure as Secretary of Health and Human Services.

Of course, we continue to have the good offices of Dr. Roper, now

serving in the White House as Deputy Assistant to the President. Finally, as evidenced by his remarks in the swearing-in ceremony for Dr. Sullivan, President Bush is interested in the subject of effectiveness.

Certainly HCFA will be working in close cooperation with the Public Health Service, continuing in many of the activities that we have started. For example, we are operating and expanding our health care information resource center, in which we will increasingly make available to qualified researchers more and better data, both from Medicare and from other sources.

We will continue to work with the American Hospital Association, the American Medical Association, the Joint Commission on Accreditation of Healthcare Organizations, and others to complete the uniform clinical data set. This will provide, on a regular, systematic basis, a wealth of clinical information to us and to other researchers.

A final critical task is dissemination of information. We can all do wonderful work on outcomes, have all kinds of great information, but if it is not put into the hands of real practitioners and real patients, we will not really have accomplished very much.

Looking to the future, I would hope that effectiveness will ultimately supersede traditional quality assurance and peer review activities, perhaps as suggested by Paul Ellwood in his outcomes management approach. I hope that effectiveness will produce a quantum leap forward in quality of care, not focusing just on the few bad actors, but rather improving all practice of medicine so that all services to all people can be as effective as possible.

Practitioners and consumers alike look forward to the fruits of your activities and the activities of your colleagues. What we are looking forward to is not by any means cookbook medicine, but rather an informed and empirically based practice of medicine that can ensure the best possible outcomes for all Americans.

ACKNOWLEDGEMENTS

I would like to thank the Institute of Medicine, particularly its president, Samuel Thief, for the leadership they have provided in the effectiveness area and the work they have done with the HCFA. I would also like to thank the IOM staff who have been so heavily involved, Kathleen Lohr, Richard Rettig, Karl Yordy, and others. Finally, I am grateful to Kenneth Shine for the leadership he has exerted in chairing these various events. He has a remarkable ability to keep the trains running on time and still allow for free and open discussion on the part of all of the participants.

5

The Effectiveness Initiative: Retrospective and Prospects

William L. Roper

I have a great affection for the Effectiveness Initiative, now an initiative not just of the Health Care Financing Administration (HCFA) or even just of the Department of Health and Human Services (DHHS) but of the entire government. It is a far-reaching endeavor that I think will bear fruit long into the future, and I am delighted to have had some small part in beginning it.

One of the things that I am most proud of concerning my tenure at HCFA is the attention that the agency gave and is continuing to give to the whole area of quality and quality measurement, and effectiveness and outcomes research. I think this work has boosted the image of the agency not only before the outside world, but also in our own eyes.

HCFA is no longer seen as an agency that does bad things to people. It is one that is in the process of producing very worthwhile information and good things for the American people.

THE GOVERNMENT'S ROLE IN EFFECTIVENESS RESEARCH

Let me comment on what it seems to me the government is doing in this Effectiveness Initiative. First, government has the unique capacity to call the nation's attention to an area of interest, to set something high on the nation's list of priorities. In the business of health care and health services research, I think that is what the government has done over the last couple of years. A matter that was formerly of passionate interest to only a few pioneering health services researchers is now an item of national importance.

The government is also in the business of setting priorities as to how this research is going to be done. That is what Michael Fitzmaurice, at the National Center for Health Services Research, and HCFA are doing in partnership with the IOM and other organizations around the country. We are

saying that these are the areas of greatest interest; this is where research should be done first. Yes, we would like the best information that could be developed on how to practice medicine, what works in medical practice across the board now, but we cannot do that, and so we have got to set priorities. That is what this whole exercise with the IOM is about.

A second thing that the government is doing is committing unique resources to this research enterprise. My former colleagues at HCFA are the custodians of immense amounts of data, information that is of very great value to health services researchers. Henry Krakauer and others are shaping that information in ways that will be of real utility.

Of course, we aspire to create data sets that will be even more useful in the effectiveness research that will be done in the future. The government has resources that go well beyond those of anyone in the private sector. What has now happened is the government has said we are going to make this information widely available and spur research in that fashion.

The third thing that the government is doing is funding research in medical practice and clinical effectiveness. The Congress is working its way through appropriations for DHHS for next year. Thus far it has not seen fit to fund fully what the President asked for in his budget for effectiveness research. We hope that Congress will become convinced of the need to go yet higher in this area. We need to have a loyal cadre of people across the country pushing for the notion of health services research.

I have for the last several years been convinced by my own rhetoric of the essential virtue of this endeavor. But I think it is going to take some arm-twisting as well. Anyone who has an interest in this had better get to work.

A fourth thing that the government is doing is developing partnerships with a wide array of organizations and individuals to carry out effectiveness research. There are partnerships with foundations that have an interest in funding research themselves; with other payers, such as insurance companies, corporations, and Blue Cross plans; with organizations such as IOM and others; but especially with practicing doctors.

It is absolutely essential that doctors across the country take hold of this idea and push it forward. It cannot just be an ivory tower matter, and it surely cannot be just a gleam in a bureaucrat's eye, however dedicated and smart that bureaucrat may be. It has got to be something that the average doctor in America sees as useful to him or her. That is why I think the American Medical Association's embracing this concept is such an important element in the whole Effectiveness Initiative.

PROSPECTS

The effectiveness and outcomes research enterprise has been given a high priority, not just by those of us who used to be in DHHS but by Louis

Hays, at HCFA; by James Mason, Assistant Secretary for Health; by Secretary of Health and Human Services Louis Sullivan in a number of respects; and, indeed, by President Bush. The President at Dr. Sullivan's swearing-in in March named research into cost-effective medical practice as a priority. I think this sets the tone for what the Bush Administration is going to be pushing for, from the top down.

Second, I think Congress will increase support for research into medical practice. The leadership of the Congress—Willis Gradison among the House Republicans, Henry Waxman and others of the House Democrats, Senate Majority Leader George Mitchell, David Durenberger, a leading Republican—are all pushing hard to further this enterprise. Individual members, even those not on the relevant health committees or appropriations committees, are being convinced that we must develop a much better knowledge base for medical practice and health care financing than we now have. It is not sufficient to say that we do not like the fact that we are spending 12 percent of our economy on health care. We have got to say, "How can we spend that money better?" The answer, it seems to me, is to develop a research base for guiding medical practice in the future.

Finally, as I alluded to earlier, the future of medical effectiveness and outcomes research does not depend on leaders in government, whether they are in the executive branch or the legislative branch. These people are distracted with other things and have transient tenures. The long-term future of this depends on broad support across the country in academic research centers, among practicing physicians, and the American public generally. Support is growing, but it needs to grow much more widely, much more quickly. That is why conferences like this one are so very important.

6

The Social Perspective

Uwe E. Reinhardt

THE PRESENT SCENARIO

Expenditures on health care outgalloped the Gross National Product (GNP) by about 3 percentage points per year, on average, during the 1980s. At that rate, it will take 82 years for 100 percent of the GNP to be eaten up by health care.

In response to these prognostications, people tend to say, in effect, "What is the big deal? We have got a long time to figure this problem of health care expenditures out—82 years in fact." It is true: it would take only 82 years. Even if we did end up spending 100 percent of our GNP on health care, what would be wrong with that? When people ask, "What would it be like?" I say, "Very simple—king-size beds from coast to coast, two Americans in each, giving each other health care, and the Japanese feeding us intravenously, as they do now."

COSTS OF HEALTH CARE

This scenario sounds comical, but not everyone is laughing at it. One person who is not laughing is an employee benefits manager of a typical American corporation. This person is 30 years old, works for General Motors, and just made a payroll entry crediting cash for $800 million in health care for people who do not work for General Motors. One can imagine him asking, "Where do I put the debit? If these people aren't working, it can't be payroll. It has got to be something else." That $800 million is in fact what General Motors pays per year for retired General Motors workers and their families, and indeed there is no clear place to record the debit.

In 1980, employers paid $60 billion in health care premiums for working and nonworking employees; now they pay $135 billion. This does not include what corporations pay toward Part A Medicare. In 1987, U.S. Steel spent $125 million of its $219 million net income on retirees—57 percent. Anything over 20 percent will catch a chief executive officer's attention: that 57 percent is being noticed by the CEOs of big corporations.

The Health Care Financing Administration (HCFA) forecasts expenditures of $1.5 trillion in the year 2000. We can afford that amount, but is it really worth spending? That question is what gets us into this Effectiveness Initiative.

Much outcomes research has shown that there seems to have been no indication for many of the procedures that have been done. In fact, patients would actually have been better off medically if they had not been done.

The issue, then, is really one of appropriateness. On some cost-quality curve, is point B, which identifies a point of diminishing marginal returns to care, appropriate? Is it more appropriate than point A, where the curve is still rising? Physicians say, "If you go past B it is not appropriate, but up to B is always appropriate." Economists would not agree, for the very reason that the National Academy of Sciences gave for not putting seatbelts in school buses: it would cost $40 million a year to save the life of one 10-year-old. Therefore, an economist would say, if that is true for youngsters, it must be true for the aged, too, and we ought to stay at point A. Appropriateness means we stop short of the maximum attainable quality. Citizens have said so with votes on roads, on seatbelts, and on many other things, and they will learn to say so for health care. That is, a raging debate will soon be upon us regarding the rationing of health services: Going from B to A means withholding beneficial services, and we will ration them. But since we are willing to ration safety on school buses, we should be willing to ration services in health care, too.

VARIATIONS IN MEDICAL COSTS AND PRACTICE

Why did this come about? Well, in 1972, John Wennberg found that Part B Medicare spending by counties in Vermont varied enormously, and he could not explain it.

Those geographic variations in practice persist to the present. For example, hysterectomy rates vary inexplicably across counties. In 1982, age-adjusted Part A hospital expenditures in Iowa City for hysterectomy were $734 per patient; in Des Moines, they were $1,300. How can the health care in Des Moines be twice as expensive as the health care in Iowa City, particularly when residents of each city insist that health care there is the best in the world? The answer will be "practice style." Why are there so many more operations in Boston than in New Haven? Why are there far more coronary bypasses in New Haven than in Boston? Physicians answer,

"In Boston, they have different theories about what works than in New Haven, and after all, New Haven is some 200 miles south of Boston." Which explains a lot.

Practice patterns vary along other dimensions, too. The Health Care Coalition of Florida has some interesting data, by hospital, on the cesarean section rate for commercially insured women and for Medicaid patients. In many hospitals, 42 percent of commercially insured women have cesarean sections, whereas women on Medicaid have none. What theory is compatible with these data? Is it a medical theory—are poor women thought to be more robust than richer women? Could it be an economic theory, because Medicaid reimbursement is likely to be one-third of commercial reimbursement? Or could it be a legal theory—do poor women sue for malpratice and poor obstetric outcomes less often than rich women? No one knows.

REASONS FOR THE EFFECTIVENESS INITIATIVE

Canadians spend a lot less than Americans as a percentage of GNP, and that raises the question, What do Americans get in health care that Canadians do not get? We know what Canada does not get; it does not have 30 million uninsured—or any uninsured—but what do they miss that we get? That, again, leads to the question of what we actually buy for all this money.

Why has this country, alone in the world, undertaken an Effectiveness Initiative? One naive theory is that American medicine decided, "We had better search what we are about. We want to be the best physicians in the world, and we will research this and do good for mankind." One could hold that null hypothesis, and one could wish it were so, but in fact I do not think it is so.

I read a lovely little essay by John Ball, Executive Vice President of the American College of Physicians. He lists the mistakes organized medicine has made, and he reminded me of something that happened early in the 1980s. Medicine argued for and got the elimination of the Health Care Technology Council that [Secretary Joseph] Califano had put in with the idea that it should do outcomes and effectiveness research. The minute President Reagan was elected, the Council's budget was zeroed out, and it never met again. Up to that time, that was the only federal agency that evaluated the appropriateness of medical technology from a medical perspective and made recommendations about payment. Partly as a consequence of its disappearance, organized medicine has had little voice when decisions about payments are made, and a real opportunity to affect standards of practice has been lost.

The alternative hypothesis is this: Concern over cost is without question one of the major drivers in this field, but there is also concern over quality.

Many researchers, led by John Wennberg, Robert Brook, David Eddy, Barbara McNeil, Kathleen Lohr, and others who worked in the 1970s and 1980s, were simply interested in quality. They realized something was wrong and studied it, with support from the National Center for Health Services Research and HCFA, to the credit of those federal agencies. These efforts were certainly a harbinger of the Effectiveness Initiative.

HOW TO PROCEED IN EFFECTIVENESS AND OUTCOMES RESEARCH

What is to be done? I am not expert in this particular type of research at all, but I do have an idea. The traditional model was that the physician held a theory: I do X, and Y happens. The "ideal" was that if every physician is allowed to have his or her own theory and is just left alone, health care will lumber toward the optimum. No one believes that notion any longer.

The next level up was to take a bunch of the smartest people, put them in a room together, and let them come out with a consensus about good practice. However, if we had done that 20 years ago, the smartest people would have said that gastric freezing is a nifty idea. We now know that gastric freezing should not be done. Period. One hundred years ago the smartest people would have said, "Under these conditions, you bleed the patient." Thus, the pure consensus approach is not adequate either. It might be better than this current free-for-all, but it is not enough, in my view.

Ultimately what we need is an empirically tested hypothesis that links medical intervention with observable outcome. We then must ask: How is a good outcome defined? In whose mind? Some surgical procedures might enable the patient to play golf or kick a soccer ball, but render him impotent. The patient will have an opinion about this. Tell a German that he can play soccer but will be impotent, and he will say, "Give me a Mercedes, I am all right." Tell a Frenchman that and all hell will break loose.

I exaggerate a little to make a point: Patients' perspectives must be included in any definition of a good outcome. We will need data, as Paul Ellwood and others say, that track other data and allow us empirically to examine medical practice issues without randomized trials. It should be possible statistically to test a hypothesis without randomized clinical trials. We do it in economics all the time, and we do it in other spheres.

This is the direction in which I see outcomes research going. We need empirical tests, and they will take a lot of money. There has to be a sizable investment in data bases, on the order of $50 million just for the first phase. If we in this country end up spending $100 to $200 million a year on outcomes and effectiveness research, we would not be wasting money, I assure you. It is a trivial percentage of the national expenditures on health care, and it is one of the finest investments we could make.

7

The Clinical Perspective

Paul F. Griner

The IOM's core committee, which generated the report on the Effectiveness Initiative, suggested a number of objectives that would be achieved by effectiveness research. First, the knowledge gained would help clinicians in their day-to-day management of patients. Second, it would improve the peer review process. Third, it would aid policymakers in the allocation of Medicare resources. I believe this knowledge will also affect patient participation in decision making and the organization and delivery of health care.

The extent to which the knowledge gained from this initiative will be applied in the real world will be determined by a number of very complex factors. Some will affect the provider, some the patient, and some the system within which they both operate. My purpose in this chapter is to suggest some of the issues that I believe require attention if the knowledge gained from the Effectiveness Initiative is to be used to its fullest advantage.

ISSUES IN THE USEFULNESS OF EFFECTIVENESS RESEARCH

There are five issues in particular that I would like to focus on. First, the knowledge must have attributes that are important to the provider; second, it must be readily accessible; third, it must facilitate patient involvement; fourth, financial incentives must be in line with the directions suggested by the new knowledge; and finally, as I see it, the problem of unbalanced regulation within the health care industry must be corrected.

MEETING PHYSICIANS' NEEDS

The first requirement bears on the credibility and usefulness of the information to the provider. The need is obvious for accuracy, relevance, and measures

of outcome that go well beyond morbidity and mortality, measures that are meaningful and that do not suffer the constraints of data collected principally for purposes of payment.

Not enough has been said, however, about the need for knowledge that is not limited by age and about which practicing physicians feel a sense of ownership. We must have clinical data that span the age distribution of the diseases of interest. Otherwise, we limit our understanding of the natural history of the illness, we miss opportunities to intervene early, and we fail to recognize age-dependent differences in treatment options. All of these deficiencies reduce the usefulness of the knowledge to the practicing physician. This point was made previously by the committee, but it bears repeating.

Another reason for the generation of disease-specific data bases that span time and are not age-limited is that these data will require years to amass and to evaluate. We presume, at least we hope, that by that time universal access to health care will have been achieved. Whether such access is financed centrally or is in large part pluralistic, as it is today, insurance benefits cannot be determined equitably if clinical knowledge is limited by age considerations.

The need for ownership of the knowledge by practicing physicians is my second point on this first issue. In my opinion, it is a critical element. Knowledge that is generated and evaluated solely by payers or by health services researchers, or both, will be suspect. Medical organizations need to be involved, they need to be empowered, and they need to be ready to promote the knowledge among their members.

I note that the committee report reflects on how little is known about what influences the behavior of health care providers. I suggest that there is a fair amount of anecdotal information to support the premise that, if competent providers are given relevant and accurate data, data that they have had a hand in developing, their behavior will be influenced accordingly.

Still on the subject of physician ownership and empowerment, we need to recognize that the primary care physician must be the principal recipient of knowledge acquired through the Effectiveness Initiative, whether as a general internist, a family physician, or a general pediatrician. These are the groups that currently feel most disenfranchised as the result of intrusive regulation and draconian reimbursement policies. We need to consider the practice of the generalist: how it is organized, how generalists are reimbursed, and what elements must be addressed if their practices are to embrace the knowledge gained from the Effectiveness Initiative.

Accessibility of Results

The second theme is the need for ready access to data that will aid in clinical decision making. Everyone recognizes the difficulty that physicians

have in keeping up with the extraordinarily rapid advances in medical knowledge and the technology to apply that knowledge. It seems clear that the same can be expected with regard to the findings of the Effectiveness Initiative unless the knowledge can be provided in real time and in a usable fashion.

It is one thing to impart a few salient points in a peer-reviewed journal; it is quite another to provide a comprehensive data base from which to extract information bearing on the many variables that need to be considered for diagnostic or management decisions regarding an individual patient. We are all aware of prepackaged information currently available for use in a personal computer. Such information is likely to be most effective for relatively straightforward tasks, such as choosing the most cost-effective antibiotic for a particular infection. The more difficult task is the use of information about the host of patient variables that need to be considered in evaluating treatment for options such as balloon angioplasty, bypass surgery, or medical management for the 67-year-old diabetic who has angina, emphysema, and hypertension.

For the everyday considerations, primary care physicians will benefit immeasurably from the availability of a comprehensive data base that can be accessed quickly. We have a few prototypes of this kind in this country, such as Duke University's cardiovascular data base. Referring again to physician ownership of new data, a very important step will be to secure the input of practicing physicians into decisions bearing on the nature of the data to be collected, how they are analyzed, and how they might be made most usable.

Aiding Patient Invol Vement

The third point has to do with patient participation in decision making. I am going to be brief here because Albert Mulley goes into this subject later in this volume (1).

Most patients continue to defer to their physicians in selecting the proper treatment option. There would undoubtedly be more patient participation in decision making if we had better knowledge concerning the outcomes of various treatment options. This knowledge would need to be packaged in such a way that the patient could fully understand the issues, explore the benefits and risks of each option, and choose among them according to his or her unique values, values that the physician must not assume.

It is quite exciting to see people such as John Wennberg, Albert Mulley, Michael Barry, and others beginning to take advantage of the technology that currently exists by preparing an educational program on treatment options for patients with benign prostatic hypertrophy. Some of my colleagues in urology in Rochester are now in the process of evaluating the efficacy of this approach. The findings will be of great interest to all of us.

Relation to Payment Mechanisms

Now to the fourth point, one of particular concern to me. It has to do with the potential conflict between the findings of effectiveness research and how providers are paid for their services. Under Medicare, hospitals and physicians are paid by the number of units of service they render. Because the price per unit of service is controlled, whether through annual increments below the rate of inflation or whatever, the system responds by attempting to increase the number of units provided to ensure financial stability. Among the results are unnecessary hospitalizations, most of which are not picked up through utilization review; unhealthy competition; unnecessary duplication of technology and other health resources; and, perhaps most important, a slowing down of change in the way services should be organized and delivered to take advantage of out-of-hospital alternatives to care and to enhance continuity of care.

Twenty-five years ago, the medical chief resident at Yale-New Haven Hospital, Eli Schimmel, wrote an article published in the *Annals of Internal Medicine* under the title, "The Hazards of Hospitalization." I have always carried that article with me, physically and in my mind. It is just as relevant today as it was then. No patient should be in the hospital unless it is required. Hospitalization poses a risk.

I have the unusual challenge, as well as opportunity, of wearing two hats at the same time—the hat of a professional who understands what we should be doing in patient care and the hat of a hospital administrator who has a fiduciary responsibility to ensure the financial vitality of our hospital. These present a conflict of interest. For example, for every open-heart surgery case above a given volume, the hospital averages a profit of $20,000. But it also costs the hospital $45,000 for the treatment of infants weighing less than 2 pounds in the neonatal intensive care unit.

We had 43 such babies last year. Forty of them left the hospital alive. That $45,000 average cost is reimbursed at a much lower figure and has to be underwritten one way or another. The surplus from the individual open-heart surgery cases helps to do that.

Our volume-driven system obviously has adjusted well, because much of the hospital care that is not necessary can still be shown to be appropriate— or at least not inappropriate. The work of Robert Brook and his colleagues has shown us that. The strength of the Effectiveness Initiative is that, for the diseases that are to be studied, it should be possible to find out what is both appropriate and necessary; that will be particularly helpful, given our interest in increasing patient involvement in decision making.

Such findings will almost certainly indicate that fewer rather than more procedures and hospitalizations are in order, at least for patients who are currently receiving care. Unless the reimbursement system is changed from

one that is driven by volume to one that provides incentives for more discriminating and coordinated use of health resources, the findings from effectiveness research are going to be accepted grudgingly and implemented slowly. We will continue to see patients hospitalized unnecessarily for cardiac catheterization and many other procedures, or discharged early after their hip fracture without adequate provision for rehabilitation services.

Application of the fruits of the Effectiveness Initiative demands reform of the current price-based reimbursement system, reforms that avoid the incentives to do too much. We need to be sure, however, to avoid a response that is too far in the other direction, one that is occasionally seen with global budgeting systems, where too little care can become the risk.

Unbalanced Regulation

My fifth point has to do with regulation. We have, in my opinion, a problem of unbalanced regulation in the health services industry. We have extremely tight regulation of hospitals and of physicians for the hospital component of their practices. We have loose regulation in the out-of-hospital marketplace. The proliferation of freestanding diagnostic and treatment centers is an excellent example of unnecessary duplication of facilities, where opportunities for unneeded services are greatly increased. Future health policy should pay attention to the issue of balanced regulation if we are going to achieve the objectives of the effectiveness initiative.

IMPACT OF THE EFFECTIVENESS INITIATIVE

Let me close with a word or two about where I think the Effectiveness Initiative will have its greatest impact. For two of the three diseases at the top of HCFA's priority list, there is a disturbing underuse of services. Four of five women at the ages where screening is known to be effective for early detection of breast cancer do not undergo such screening. The majority of poor or near-poor persons have low rates of utilization of diagnostic and therapeutic procedures for their underlying coronary artery disease. I believe the findings of the Effectiveness Initiative will result in an even more striking picture of missed opportunities among these populations.

The initiative should help change for the better the general approach to medical practice. After all, the Effectiveness Initiative can address only a limited number of illnesses, perhaps 20, perhaps 50. While better knowledge regarding treatment of these conditions will obviously improve quality and limit health care inflation, it will still account for only a very small fraction of total health care.

Built into the Effectiveness Initiative are approaches that eventually should change the very fabric of medical practice. Some of them I have already

referred to. They include a much stronger role for patients in decisions relating to their health care and greater physician support, perhaps even enthusiasm, for a systematic study of outcomes of care once the value of such heretofore unavailable information is recognized.

The Effectiveness Initiative will also facilitate the incorporation of functional assessment and quality-of-life measures into the day-to-day practice of medicine, measures that are so greatly lacking now. I believe that, in the long run, these results of the Effectiveness Initiative are going to be among its greatest contributions, contributions that go far beyond the knowledge gained through specific attention to given illnesses.

I conclude by simply repeating the caution that I began with: the ultimate impact of the products of this initiative will be determined by the extent to which they address important requirements of the provider, are readily accessible, promote patient involvement, are accompanied by a reimbursement system that provides incentives, not constraints, for their application, and are facilitated by more balanced regulation.

Reference

1. Mulley, A.G. Applying Effectiveness and Outcomes Research to Clinical Practice. Pp. 179-189 in *Effectiveness and Outcomes in Health Care*. Heithoff, K.A. and Lohr, K.N. eds. Washington, D.C.: National Academy Press, 1990.

8

The Legislative Perspective

John D. Rockefeller, IV

We on the Pepper Commission [U.S. Bipartisan Committee on Comprehensive Health Care] are charged with developing a public consensus on long-term care. Catastrophic illness costs $5 or $6 billion a year, a mere pinch. Long-term care costs $40 to $60 billion. And for the uninsured, one must simply grab some figure in the tens of billions. The bill for our nation's unmet health care needs is just extraordinary. The legislation on catastrophic illness coverage has been a subject for many speeches. It is an extraordinary thing, isn't it, that a program which is so directed, so progressive, and so precisely right—a rare thing—could be rejected by precisely those people who have no business rejecting it, on behalf of all those people who benefit from it. But that is what has happened.

UNMET HEALTH CARE NEEDS

I was recently in Chicago, where I attended a hearing of the National Commission on Children. Much more important than that, perhaps, was a visit of commission members into ghetto areas, into housing projects, to see how it is there. First we started at Cook County hospital and saw premature, low-birthweight babies who weighed a pound and a half. Some got to two pounds, some to two and a half. When they got to four pounds, they looked like they were really healthy, and one rejoiced; but we did not see many of them weighing four pounds. There were endless numbers of low-birthweight babies. Should they emerge—at a cost of some $50,000 to $100,000 per child—from the intensive care unit, most of them will have permanent developmental disabilities.

We went into the housing projects to see what is happening there—if it is possible to get in, if the gangs are not already there, which they are in most

of the housing projects. And there we saw how it is that teenagers who discover they are pregnant at the age of 14 or 15 do not go looking for something called prenatal care, especially since many of them do not even know what it is. If they do know, and if they have applied for Medicaid (or should be on Medicaid, which only reaches 42 percent of poor people anyway) and Medicaid keeps sending the forms back, they will have already had their babies by the time they get signed up. And so on and on and on.

We are in a crisis. The dimensions of it—not just the financial dimensions, but the human dimensions as well—are awesome.

I am a relative newcomer to health care. I asked somebody the other day, whom I very much respect and who has been in this business for a long time in Washington (not in Congress), how many people he thinks there are in Congress who understand health care. He said, "Six." I was not one of them, unfortunately. But give me a little time, and I will be, because I am intensely determined about it. Since I am not one of those six, and since I am a relative newcomer, it is probably not very good for me to presume to make observations. Nonetheless, I will do so.

It strikes me that we are at a dangerous crossroads. We have literally hundreds of billions of dollars' worth of health care needs that are as yet unmet, some as yet unthought-of. I completely agree with Uwe Reinhardt about the concept that a civilized nation simply will not tolerate having 37 million citizens who are uninsured, not to mention those who are underinsured.

How do we rectify this? America has apparently fastened onto the concept of no new taxes, after having reduced taxes and having removed $150 billion from the revenue base of this economy every year since 1981. On further reflection, however, some Americans have decided that lowering taxes was not a good idea and that we should think about raising them. So the two parties argue about whether we should lower them again and how—capital gains or IRA? It is all sheer madness, if one cares about health care.

Health care costs are rising 14 to 15 percent annually. By 2003, Medicare costs will be larger than Social Security costs. Defense spending, interest on the national debt, and Social Security and Medicare costs together account for 85 percent of the entire federal budget.

Those who would cut costs, including Medicare payments, face the wrath of the providers. Those who would add coverage, and thus the payment for it, face the wrath of the taxpayer and, we now discover in the case of catastrophic care, of the beneficiary.

We risk gridlock because everyone with a vested interest in our current system—providers, employers, patients, insurers, and taxpayers—has something to lose from the changes that must be made. On the other hand, I think that all of these vested interests have a lot more to lose if the changes are not made.

There is evidence of stress in the system. The budget reconciliation

process was tied in a knot over catastrophic illness coverage. The coal strike in West Virginia and the "baby Bell" strikes this summer grew out of arguments over who would pay health care costs. One-pound babies like those I saw in the neonatal intensive care unit are still being born. The list goes on and on. We have to find a consensus somehow, and we have to do it very quickly.

We need, as never before, those persons who are most knowledgeable about our health care system to help us reach that consensus. We need people who know what is at stake and who stands to lose in the process of trying to forge this consensus. That is where scientists fit in, at least in my judgment.

GROWING AWARENESS OF EFFECTIVENESS RESEARCH

This conference is incredibly timely. The answer to the question "What works in health care?" probably holds the key, or at least part of the key, to the many conflicts that we will wade into.

Two years ago, nobody on Capitol Hill was talking about medical outcomes and effectiveness research. Nobody. Nobody was really talking about the trade deficit until 1984, at which point we were already in the tank so deeply —but that's the way we are as a country. The crisis has to overwhelm us before we recognize it—and then sometimes we can get out of it. Now outcomes and effectiveness research, if not quite the talk of the town, are the talk of a good deal of it.

A recent survey of over a dozen studies on the appropriateness aspect of medical care included these findings: "Research finds high incidence of unwarranted pacemaker implantation." "Inappropriate coronary artery bypass surgery is frequent." "Rate of inappropriate hospital use is high." "[There is] evidence of anti-psychotic drug misuse in nursing homes."

The potential benefits of outcomes and effectiveness research have jolted many of us into what I hope is the realization that evaluation of health services may not mean rationing in a pejorative sense. Some of us even hope that the dual efforts of controlling health care costs and improving quality are complementary.

In the Senate, George Mitchell, who is not an inconsequential figure as Majority Leader, introduced a companion bill to the legislation authored by Willis Gradison on the question of federal financing of a national research effort on medical outcomes and effectiveness. Now Senator Mitchell has requested, as have many physicians' groups, that his bill be included in a much larger physician payment reform bill that David Durenburger and I are working on.

The provision is integral to the success of our package. Our package does not have integrity without that research in it. Physicians and policymakers

understand that the process of rationalizing payments for services must proceed apace, along with the process to understanding the value of the services.

THE CHALLENGE TO SCIENTISTS

Today, the medical community and the political community are ready for research. We hope it will improve the quality of care; we hope it will save lives; and we hope it will make cost control more rational and more consistent with good medicine. I encourage your scientific efforts enthusiastically, pleadingly, but with two important caveats.

First, do not take political support of this concept for granted. To wit: last week, the research community came within a hair's breadth of losing its congressional backing. Senator Mitchell believed that a minimum of $35 million was necessary to get this national outcomes and effectiveness research underway and he made a very strong pitch as Majority Leader to the Appropriations Committee. But that committee did not vote out the hoped-for funds. Only the personal appeals and efforts of Senator Mitchell won a last-second restoration of funds. Without him, it would not have happened. So understand, please, when I say that you cannot take support for granted, because nobody understands it.

All of us will need to work diligently to protect those funds, because the bill has yet to go to the House-Senate Conference Committee. Be aware of that conference. Summon whatever influence you have and exercise it on that House-Senate Conference to make sure that outcomes research money stays.

That leads me to my second caveat, which is that, as scientists and as experts, your advocacy is imperative. Last spring, Frank Press, President of the National Academy of Sciences, called upon his colleagues to unite behind specific scientific efforts. Do not throw them out Gatling-gun style; they have to be given priorities—you must set priorities. The work that you are about here clearly must be one of those priorities. A nation that spends over half a trillion dollars for health care can afford to spend—in fact, certainly cannot afford not to spend—$35 million to begin learning what works.

Not only must your case be made to the public and to the Hill on the need for this research, you must be persistent in your advice on the nature of the research. The politics of where these studies should begin and what they should seek will certainly overwhelm you unless you give us your best advice on how to proceed.

Because of the budget crunch in Medicare, Congress will urge you in one direction; that is, toward the effectiveness of the "big ticket" services. Providers of those same services will urge you in a different direction. You

need to give Congress your objective advice on the best way to proceed. We cannot do it ourselves.

David Durenburger is a Republican and I am a Democrat; we do everything on the Pepper Commission together. That is our policy. I am chairman of the commission, but he knows much more than I do—and I tell him that frequently. We and our staffs do nothing without talking to each other. We are trying to make our actions not only bipartisan, but bicameral. Most people in Congress do not operate that way. We are determined to, and you can help us.

CONCLUSION

Two years ago, health was not center stage; it was all a question of what are we going to do with ASATs and Star Wars. But interestingly enough, defense—and all of the raging passions it inspires—has receded in the last year and a half, and other things have come to the fore. That has happened, I suppose, because of the Gorbachev window, and I pray that that window will stay open for a while.

People now are really onto the cost of health care, are really scared about the annual increase in the cost of that health care, and really do understand that there are very few people in Congress who understand health care. We need you scientists; more importantly, we know that we need you and what you are doing. You are called upon not only for the right answers, but also for politically compelling evidence that those answers are right, evidence that will help us forge a consensus and bring about needed actions. It is a tall order, but I know that you will do it.

PART III

THE IOM CLINICAL CONDITION WORKSHOPS

The IOM Condition Workshops: Introduction

Kenneth I. Shine

The individuals who participated in the IOM clinical condition workshops on breast cancer, hip fracture, and acute myocardial infarction believe that the Health Care Financing Administration (HCFA) has taken a very positive step in seeking to make its vast administrative data bases available for effectiveness research. They recognize, however, that HCFA has taken only a first step in what will be a very long journey. Many participants—and all of them on this panel—come from academic medicine and from a tradition of rigorous science and analysis. Not surprisingly, they caution prudence in the use of such data because of a realistic awareness of pitfalls to be avoided and obstacles to be surmounted.

The following three papers reflect this prudent caution. In breast cancer, for example, treatment often involves in-hospital surgery and subsequent radiation therapy or chemotherapy administered in an outpatient setting. However, existing Medicare data are primarily hospital-based, and HCFA does not yet have good ambulatory data. Evaluating mammography for its effectiveness in breast cancer screening and diagnostic uses, to take another example, will require long-term follow-up data, not simply data on acute care encounters between the patient and a provider. On the other hand, Medicare data can help immeasurably in focusing attention on the similarities and differences between Medicare-age women and younger women.

Valerie D. Jackson has a special research interest in breast imaging. She brings that expertise to bear on diagnosing and treating breast cancer in the elderly.

Hip fracture was approached as a relatively straightforward clinical problem. A single bone is involved, diagnosis is clear and consistent across different practitioners, and surgery is the recommended intervention, followed by rehabilitation. Yet as the committee delved into the issues, we realized how

complicated it was to assess the effectiveness of prevention, of different surgical and medical interventions, and of rehabilitation programs. Prevention, for instance, requires that effective efforts begin long before an individual reaches the age of Medicare eligibility. Research along these lines, therefore, must link Medicare data bases to Medicaid and private insurance data bases for a younger population, linkages that span federal and state as well as public and private boundaries. In addition, different sites of care are required for treating hip fracture, and unexplained geographic differences exist in rates of fracture. David G. Murray, an orthopedic surgeon, examines these and other effectiveness issues related to treating hip fracture.

One of the major issues confronted in the clinical workshops was how to make the best use of administrative data bases for effectiveness and outcomes research. Barbara J. McNeil, a radiologist and investigator with experience in using large data bases, addresses the opportunities and the limits of using claims data in acute myocardial infarction and other conditions.

9

Breast Cancer

Valerie P. Jackson

Breast cancer, the second leading cause of cancer death in American women, is a major health problem for women in the Medicare age group because its incidence increases with advancing age. Currently, the American Cancer Society estimates that 1 in 10 American women will be affected by this devastating and highly emotional disease during her lifetime. Several studies have shown that screening mammography can detect breast cancer at a more favorable stage, resulting in improved prognosis for screened women found to have breast cancer (1-10).

THE PROBLEM OF POOR COMPLIANCE

The American Cancer Society and the American College of Radiology have recently been joined by most of the medical groups in this country in recommending the following guidelines for screening mammography:

- Baseline mammogram by age 40
- Mammogram every one to two years between ages 40 and 49
- Yearly mammograms after age 50

The overall mortality statistics for breast cancer in the United States have changed little in several decades, however, largely because of poor compliance with screening mammography guidelines. Thus, screening mammography has proven efficacy, but its effectiveness is diminished because it is underutilized. Although we can detect tumors as small as 3 to 5 millimeters with mammography (Figure 1), we continue to see too many large, clinically obvious carcinomas that carry a poor prognosis (Figure 2). Noncompliance is a problem for all groups of women, but it is particularly prevalent in our elderly and indigent populations.

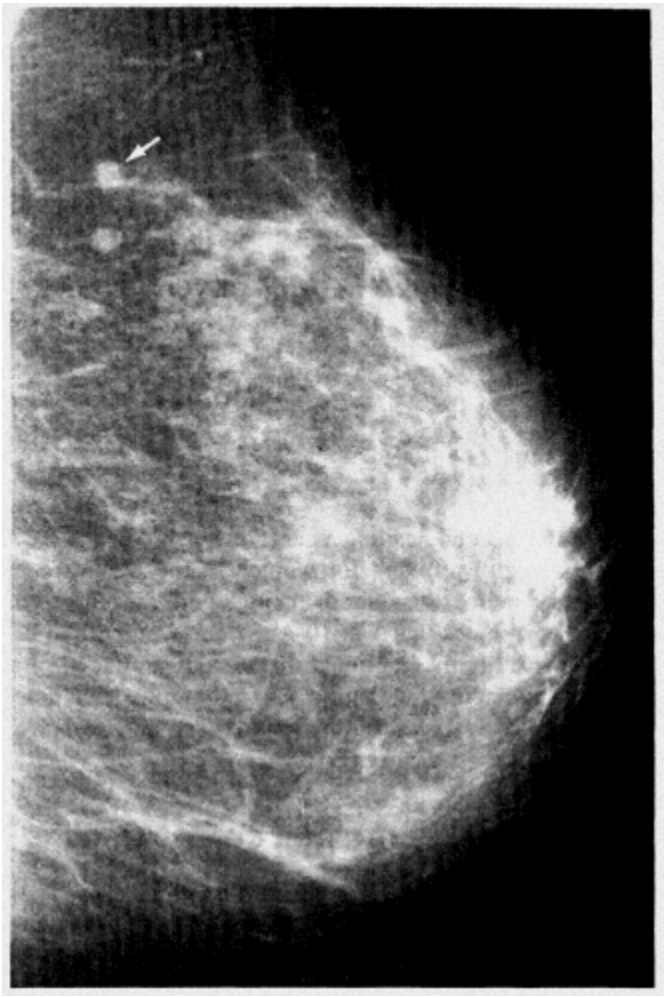

Figure 1
Oblique left mammogram of an asymptomatic 60-year-old woman. There is a 4 millimeter slightly irregular mass in the upper portion of the breast (arrow), which was found to be a small invasive ductal carcinoma at surgery. She has had no evidence of spread to the axillary lymph nodes or elsewhere in her body in one year of follow-up, and her prognosis is excellent.

As shown in Table 1, there are a number of potential reasons for poor compliance. Cost is a major factor, particularly for elderly women on fixed incomes. Many mammographers are working to decrease the cost of screening mammography to approximately $50. If Medicare paid for screening (as opposed to just diagnostic) mammograms, cost might no longer be a deterrent and compliance would improve.

Figure 2
Oblique left mammogram of a 60-year-old woman with bloody nipple discharge and a large, hard, palpable mass behind the nipple. The mammogram demonstrates a 6 centimeter mass, found to be invasive ductal carcinoma with positive axillary lymph nodes at surgery. She died 6 months after mastectomy.

Vigorous compression is necessary in order to minimize radiation dose and maximize image quality. Many women worry about pain from this compression, and pain has been cited as a possible deterrent to mammography. However, compression is adequately tolerated by the majority of women (11) and is unlikely to be a significant factor in compliance.

Many women have difficulty finding the time to get a mammogram. There are probably many psychological reasons for this, but we need to

make mammography facilities more accessible for patients and minimize the time necessary for the examination.

TABLE 1 Noncompliance with Screening Mammography Guidelines

Potential Factor	Potential Solution
1. Cost of mammograms	Medicare payment for screening Lower cost of mammograms
2. Pain from compression	Education about advantages of compression Compassionate technologist
3. Time	Improved access to screening mammography facilities Less time necessary for examination
4. Radiation exposure	State-of-the-art equipment, appropriately used, monitored, and maintained Patient and physician education
5. Interpretation errors	Consistently high-quality mammograms Greater mammographer experience and education Improved diagnostic criteria Adjunctive use of ultrasound in selected cases
6. Fear of finding cancer	Patient education

Exposure to radiation was a serious consideration in previous years (12); however, mammography equipment and film systems have improved markedly, and the radiation dose from properly performed mammography is so small that radiation exposure is no longer a problem (13). It has been estimated that, for women over age 50, the risk of having yearly mammograms is one-tenth the risk of early death caused by failure to diagnose breast cancer by screening mammography (14).

Both women and their physicians are worried about interpretation errors (that is, false negative and false positive studies). Unfortunately, there is overlap in the appearances of benign and malignant processes, necessitating biopsy or mammographic follow-up for differentiation. In spite of the extensive experience with mammography in this country, there are some cancers that are missed, either by negligence or because they are not mammographically visible, even with a good quality film. Although education and careful attention to technique and quality control will minimize these unfortunate occurrences, it is unlikely that we can completely eliminate false negative mammograms in the foreseeable future.

Breast cancer is a very emotional disease, and many women are afraid of having breast cancer discovered. This fear may paralyze a woman to the point where she will not undergo screening or diagnostic evaluations. Increased

patient education regarding the benefits of early detection and treatment should be targeted at specific geographic or socioeconomic groups, or both.

CONTROVERSIES

Screening

A number of controversies surround breast cancer screening. First, who should be screened and at what intervals? The American Cancer Society guidelines represent our "best guess" at appropriate intervals for women over age 35. However, these are likely to be modified as our knowledge of the biology of breast cancer increases. Currently, the majority of controversy surrounds screening for women age 40 to 49 (15,16). Although the efficacy of screening mammography for women over age 50 is well established, we do not know at what age screening should stop. Obviously, this will depend on the patient's physiological status and whether she would benefit by having a small, potentially curable, cancer detected, in light of her other disease processes.

Is it effective to screen all women over the age of 35 or only those at high risk? Unfortunately, most women with breast cancer do not have identifiable risk factors, other than the fact that they are women who are getting older. Thus, targeting specific 'thigh-risk" groups for screening is of limited value.

Why is screening mammography underutilized, and how can compliance be improved? These are major issues in determining the effectiveness of breast cancer screening in this country. Current investigations are studying the reasons women do not go for mammograms and the reasons their physicians do not order them. In the future, we must define and test interventions that will improve compliance and improve our overall breast cancer mortality statistics.

Treatment

A number of controversies surround treatment of breast cancer as well. Debates rage over appropriate surgical approaches to various types of breast cancer. In the past, the standard surgical treatment was a mastectomy (generally, a modified radical mastectomy), which obviously left the woman without a breast. In recent years, many surgeons have offered some women a less mutilating approach: segmental resection with an axillary node dissection, usually followed by radiation therapy. Large randomized controlled trials have shown that these two treatments result in equal prognoses for most women (17,18). Unfortunately, many women, particularly elderly women, may not be given any choice of surgical therapy (19).

Chemotherapy and hormonal therapy are relatively new interventions for women with breast cancer and have improved the prognosis for selected patients. Until recently, chemotherapy was generally reserved for women with positive axillary lymph nodes at the time of surgery or for very young women with a generally poor prognosis. However, recent studies (20-24) have led the National Cancer Institute (NCI) to recommend that all breast cancer patients, even those with node-negative tumors, have chemotherapy or hormonal therapy, depending upon their age and a number of other tumor factors. This has provoked considerable controversy among oncologists and may be a deterrent to compliance with screening mammography. In the past, we were able to tell women that if they had early cancer detected by screening, they probably would not require chemotherapy, with its unpleasant side effects. If the NCI recommendations are followed, women may feel there is no advantage to early detection. We need to find new tests to identify subgroups of women with node-negative breast cancer who are most likely to benefit from chemotherapy.

THE ROLE OF EFFECTIVENESS STUDIES IN BREAST CANCER

There are many long-term considerations for women with breast cancer. For example, what are the appropriate methods and intervals for follow-up of women with cancer? What are the psychological needs of these patients, particularly of elderly women? What are their reconstructive surgery options? Most important, what therapies improve quality of life as well as mortality statistics? Surprisingly, these areas have received relatively little attention in the past. Effectiveness studies may provide answers to these questions.

Initial breast cancer effectiveness studies should involve mammography. For optimal studies, we need much more data than are currently available from the Medicare data bases. Because Medicare currently reimburses only for diagnostic mammograms (meaning that the patient has some sort of a problem, such as a palpable lump), we do not have data on screening mammograms done on asymptomatic women. If Medicare pays for screening mammography, it will be a golden opportunity to study utilization and effectiveness of breast cancer screening; however, we must accurately determine and record the reason for the mammography (that is, screening vs. diagnostic examination). We need to track the interpretations and outcomes for all women who have mammograms. Some mammograms are going to be interpreted as "negative" for cancer; others are going to be called "positive." Of those women with negative mammograms, we must find out how many subsequently go on to have a breast cancer that was missed on the mammogram (false negative mammogram). Tracking this information may require years. Of women who have a positive mammogram (where a lesion

is called suspicious), we must find out how many women actually have cancer (true positive rate), how many have a benign lesion (false positive rate), and how many women do not undergo further evaluation. Occasionally a mammogram is interpreted as abnormal, but the patient and her physician are never notified or they choose not to do anything about it.

It is also crucial that we identify the temporal relationship between mammography and biopsy. For example, if a woman has a mammogram in January 1989 and has a biopsy positive for cancer in December 1989, the study and the surgery may have very little relationship. We must know if the mammogram prompted the biopsy and the length of time between the mammogram and the surgery. Ideally, we should record inpatient and outpatient diagnosis, treatment, and follow-up data on all of these women. It is also very important that we have standardized terminology and recording of data on tumor stage, cell type, and hormone receptor status. All of these factors must be analyzed in order to obtain accurate data about the effectiveness of breast cancer screening in the United States.

References

1. Shapiro, S. Evidence on Screening for Breast Cancer from a Randomized Trial. *Cancer* 39:2772-2782, 1977.
2. Shapiro, S., Venet, W., Strax, P., et al. Ten-to Fourteen-Year Effect of Screening on Breast Cancer Mortality. *Journal of the National Cancer Institute* 69:349-355, 1982.
3. Shapiro, S., Venet, W., Strax, P., et al. *Selection, Follow-up, and Analysis in the Health Insurance Plan Study: A Randomized Trial with Breast Cancer Screening*. NCI Monograph 67:65-74, 1985.
4. Baker, L.H. Breast Cancer Detection Demonstration Project: Five-Year Summary Report. *Cancer* 32:194-225, 1982.
5. Seidman, H., Gelb, S.K., Silverberg, E., et al. Survival Experience in the Breast Cancer Detection Demonstration Project. *Cancer* 37:258-290, 1987.
6. Tabar, L., Fagerberg, C.J., Gad, A., et al. Reduction in Mortality from Breast Cancer After Mass Screening with Mammography: Randomised Trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. *Lancet* 1:829-832, 1985.
7. Verbeek, A.L., Hendriks, J.H., and Holland, R. Reduction of Breast Cancer Mortality Through Mass Screening with Modem Mammography: First Results of the Nijmegen Project, 1975-1981. *Lancet* 1:1222-1224, 1984.
8. Collette, H.J.A., Day, N.E., and Rombach, J.J. Evaluation of Screening for Breast Cancer in a Non-Randomised Study (the DOM Project) by Means of a Case-Control Study. *Lancet* 1:1224-1226, 1984.
9. Palli, D., DelTurco, M.R., Buiatti, E., et al. A Case-Control Study-of the Efficacy of a Non-Randomized Breast Cancer Screening Program in Florence (Italy). *International Journal of Cancer* 38:501-504, 1986.
10. Andersson, I., Aspegren, K., Janzon, I., et al. Mammographic Screening and

Mortality from Breast Cancer: The Malmo Mammographic Screening Trial. *British Medical Journal* 297:943-948, 1988.

11. Jackson, V.P., Lex, A.M., and Smith, D.J. Patient Discomfort During Screen-Film Mammography. *Radiology* 168:421-423, 1988.
12. Bailar, J.C. Mammography: A Contrary View. *Annals of Internal Medicine* 84:77-84, 1976.
13. Feig, S.A. Radiation Risk from Mammography. Is it Clinically Significant? *American Journal of Roentgenology* 143:469-475, 1984.
14. Ritenour, E.R. and Hendee, W.R. Screening Mammography. A Risk Versus Risk Decision. *Investigative Radiology* 24:17-19, 1989.
15. Eddy, D.M., Hasselblad, V., McGivney, W., et al. The Value of Mammography Screening in Women Under Age 50 Years. *Journal of the American Medical Association* 259:1512-1519, 1988.
16. Moskowitz, M. Breast Cancer Screening: All's Well That Ends Well, or Much Ado About Nothing? *American Journal of Roentgenology* 151:659-665, 1988.
17. Fisher, B., Bauer, M., Margolese, R., et al. Five-Year Results of a Randomized Clinical Trial Comparing Total Mastectomy and Segmental Mastectomy With or Without Radiation in the Treatment of Breast Cancer. *New England Journal of Medicine* 312:665-673, 1985.
18. Veronesi, U., Saccozzi, R., del Vecchio, M., et al. Comparing Radical Mastectomy with Quadrantectomy, Axillary Dissection, and Radiotherapy in Patients with Small Cancers of the Breast. *New England Journal of Medicine* 305:6-11, 1981.
19. Greenfield, S., Bianco, D.M., Elashoff, R.M., et al. Patterns of Care Related to Age of Breast Cancer Patients. *Journal of the American Medical Association* 257:2766-2770, 1987.
20. The Ludwig Breast Cancer Study Group. Prolonged Disease-Free Survival After One Course of Perioperative Adjuvant Chemotherapy for Node-Negative Breast Cancer. *New England Journal of Medicine* 320:491-496, 1989.
21. Mansour, E.G., Gray, R., Shatila, A.H., et al. Efficacy of Adjuvant Chemotherapy in High-Risk Node-Negative Breast Cancer. *New England Journal of Medicine* 320:485-490, 1989.
22. Fisher, B., Costantino, J., Redmond, C., et al. A Randomized Clinical Trial Evaluating Tamoxifen in the Treatment of Patients with Node-Negative Breast Cancer who have Estrogen-Receptor-Positive Tumors. *New England Journal of Medicine* 320:479-484, 1989.
23. Fisher, B., Redmond, C., Dimitrov, N.V., et al. A Randomized Clinical Trial Evaluating Sequential Methotrexate and Fluorouracil in the Treatment of Patients with Node-Negative Breast Cancer who have Estrogen-Receptor-Negative Tumors. *New England Journal of Medicine* 320:473-478, 1989.
24. Early Breast Cancer Trialists' Collaborative Group. Effects of Adjuvant Tamoxifen and of Cytotoxic Therapy on Mortality in Early Breast Cancer: An Overview of 61 Randomized Trials Among 28,896 Women. *New England Journal of Medicine* 319:1681-1692, 1988.

10

Hip Fracture

David G. Murray

Although a fracture of the hip is not in and of itself potentially fatal, the mortality associated with the occurrence of this injury in the elderly is significant, and the associated morbidity and negative effect on quality of life are important. Moreover, the incidence, which increases rapidly in the Medicare population, places a major demand on health resources, social institutions, and the budget for health care. Any changes that could be brought about to decrease the incidence of hip fracture, facilitate improved treatment, reduce hospitalization, and increase the number of individuals restored to their prefracture lifestyle would have impressive benefits for society.

PREVENTION

Prevention of fracture of the hip in the elderly involves an increased understanding of etiological factors. Osteoporosis, which to some-extent is a natural accompaniment of aging, is an obvious predisposing condition. The extent to which the normal decrease in bone density that occurs during aging plays a role in the predisposition is poorly understood and requires further study. Pathological osteoporosis (itself poorly understood) is an obvious predisposing condition. The various factors affecting this condition, such as alcoholism, smoking, steroids, sedatives, anticoagulants, and diet, need further study. Mechanisms for modifying osteoporosis through diet, activity, or drug therapy are currently being investigated.

The vast majority of fractured hips are associated with falls. It has never been clear whether the individual falls because the hip fractures or the hip fractures as a result of the fall. Probably both play a role. Falls in the elderly are influenced by external and internal factors. The external environment,

which includes obstacles to ambulation such as furniture, slippery floors, and carpets, can obviously be modified once the relationship to falls is clearly understood. The internal factors such as Parkinsonism, malnutrition, Alzheimer's disease, balance problems, and visual impairment may be more difficult to modify. On the other hand, once such internal factors are clearly identified as being associated with an increased incidence of falls and fracture, some modifications of the external environment may be able to compensate for them.

Data also suggest that there is a geographic variation in the incidence of fracture of the hip. Whether this is due to dietary differences, differences in demographics, or some other factor remains to be explained and deserves further investigation.

TREATMENT

The diagnosis of fracture of the hip is straightforward. The history of a fall with associated disability in an elderly person is suggestive. X-ray examination confirms the diagnosis and characterizes the fracture as either a fracture of the femoral neck or an intertrochanteric fracture (one involving the upper end of the femur just below the femoral head). The location of the fracture influences the treatment and the prognosis. Fractures of the fern oral neck may impair the blood supply to the bone of the femoral head and therefore compromise the results of treatment that retains the femoral head. Intertrochanteric fractures may be complex, and the damage to the bone may preclude replacement with a prosthetic device.

Since the 1930s, surgery has been the preferred method of treatment for fractures of the hip. Fixing the fracture in some manner has been shown not only to diminish the length of hospitalization but also to lower significantly the mortality rate and improve the chances of the patient's returning to the prefracture lifestyle. At this point, nonsurgical treatment is reserved for those patients who cannot undergo surgery for medical reasons.

Initially, fractures of the hip were treated surgically by internally fixing the fracture with a nail, plate, screws, or some other means of holding the bone ends together. Because the bone ends frequently failed to unite, a prosthesis was introduced to replace the femoral head. Subsequently, total hip replacement was used to treat certain fractures of the hip.

At this point no ideal treatment for hip fracture has been established. The method used varies with the preference of the individual surgeon. To some extent economics enters the picture as well. Simply fixing the fracture with a nail or plate carries the previously mentioned risk of nonunion or loss of position of the fractured fragments. Replacement of the femoral head with a prosthesis is sometimes associated with persistent pain in the hip and gradual erosion of the bone of the pelvis by the metallic femoral head.

Total hip replacement is a somewhat more complex procedure and is more expensive, both in terms of operating time and equipment and device costs.

There is a definite need for outcome studies to clarify the relative advantages and disadvantages of each treatment method. Such outcome studies would include length of hospitalization, in-hospital complications, rate of reoperation, and the overall recovery of the patient.

IN-HOSPITAL CARE

In addition to the surgical procedure itself, a number of other factors are associated with the initial hospitalization of a patient with a fractured hip. These factors need to be reviewed in terms of their impact on the outcome or effectiveness of treatment. Preoperative evaluation by consultants, including internists, geriatricians, family physicians, cardiologists, urologists, and so on, may have a beneficial effect on the mortality or morbidity associated with the surgical procedure. Following surgery, the involvement of a rehabilitation team has been shown in other countries to have an effect on the length of hospitalization. Hospitalization in the United States is significantly shorter than in other countries, but similar studies should be done to clarify the impact of associated special services on the outcome of the patient's hospital treatment.

REHABILITATION

A multitude of factors are associated with the ultimate rehabilitation of the patient. Currently it is known that the mortality associated with a fractured hip is elevated over that of a matched population group during the first 6 to 12 months after fracture. In addition, the percentage of individuals who are converted from independent to dependent lifestyles is sizable. This has been well documented in the literature, along with other factors that may play a role in this conversion. Obviously, the number of persons who become dependent upon the institutions of society affects the overall costs associated with the problem of hip fracture. Mechanisms need to be developed to reduce the number of such individuals. Further data are needed to characterize this group and to show modification of outcome by intervention. This will require improved data collection, including collection of information after hospitalization, and an effective method for assessing function.

CONCLUSION

The Medicare data bank already provides a mechanism for accumulating information concerning the effectiveness of various types of treatment for hip fracture. By extrapolation, information may be derived concerning epi

demiology, predisposition, and prevention. Many factors that may play a role in predisposition occur far in advance of the age of 65, however. Clarification of some of these factors depends upon expanding the data collection to younger people. If worthwhile data on long-term functional outcome are to be gathered, the data set must be augmented. Ways of doing this have been identified and appear feasible.

If the occurrence of hip fracture is reduced significantly and treatment and rehabilitation of persons with fractures are improved, the quality of life of a large number of elderly persons will be improved. The commensurate savings in health care dollars will more than justify the cost of the effectiveness studies.

11

Claims Data and Effectiveness: Acute Myocardial Infarction and Other Examples

Barbara J. McNeil

The question of effectiveness of medical treatment is an extremely important one and one that will benefit from close collaboration between physicians and social scientists. In this chapter, however, I confine my discussion to a limited aspect of that collaboration—that is, to the analysis of claims data, particularly analysis of Medicare claims data. My discussion is based on the claims data as they exist today. It is important to note, however, that since these data have begun to be used for prospective payment, their accuracy has improved considerably. I think we can expect improvements of similar magnitude once these data are used to a greater extent for research on effectiveness and outcomes, particularly as they relate to medical technology.

The original definition of medical technology from the Office of Technology Assessment (OTA) considers two types of technologies. The first is any medical device, drug, or surgical procedure used in the care of patients. The second is any organizational or support system within which medical care is delivered. It is unlikely that claims data in their current form will be usable in the latter, so I will restrict my comments to the first type of technology.

STRENGTHS OF CLAIMS DATA

Large claims data bases have a number of strengths. To illustrate these I draw upon the experience of many other researchers and on my own experience as a researcher and as a commissioner with the Prospective Payment Assessment Commission (ProPAC). The following list illustrates the most notable strengths. It applies principally to Medicare Part A data (primarily hospitalization data) because that is where most of our experience has been thus far. Part B (ambulatory) data, particularly when linked to Part A data, expand these

strengths still further. Such linkage is costly, however, and initial efforts are just being completed. Strengths of claims data include the following:

1. They can be used to provide usage rates.
2. They can be used to indicate variations in use of technology by geography, hospital type (e.g., teaching, nonteaching; urban, rural), age, sex, and so on. This is the area in which John Wennberg has worked so successfully over the years.
3. They can be linked to mortality data in order to define mortality rates as a function of the above items and as a function of key diagnostic and procedure codes. This is the basis of the Health Care Financing Administration's (HCFA) initiative in providing mortality rate data to hospitals.
4. When linked with the Medicare Cost Report, claims data can be used to estimate the costs of hospitalization. Comparative data can also be obtained across types of institutions. If patients' records were linked over time and Part B data were linked with Part A data, they could be used to provide information on the costs of an episode of care. (This assumes that it is possible to define an episode accurately.)
5. They can provide information on home health services.

Although this list of strengths is long, for our initial activities in the Effectiveness Initiative, we will largely be talking about items 1, 2, and 3.

GENERAL LIMITATIONS

There are four serious limitations to these Medicare claims data. First, there is very limited information on comorbidity and disease severity. Thus, it is difficult, if not impossible, to define an "inception cohort"—that is, a homogeneous group of patients whose identity is clearly and reproducibly defined at a particular time and who are then followed into the future. Second, there is limited information on socioeconomic status, and much recent literature has shown that socioeconomic status correlates well with usage of certain health services and medical technologies.

Third, data on outcome are sparse. Currently, they allow us to measure mortality rates and readmission rates; however, it is not always possible to determine whether a readmission is related to the prior admission, is a consequence of suboptimal care, or is an unrelated event. Because much of medical care is designed to reduce morbidity rather than mortality, omission of data on postdischarge functioning of the patient and on alleviation of the symptoms that generated the hospitalization limits the usefulness of current outcomes data to research. Moreover, as we think about incorporating outcomes data, we should think about obtaining data at times after discharge that reflect the expected results of the hospitalization. For example, outcomes

after a cholecystectomy should probably be obtained at 3 months, but outcomes after hip replacement surgery should probably wait for 6 to 12 months. Fourth, many codes used to describe diagnoses and procedures are nonspecific, as discussed below.

Recent work by Lisa Iezzoni and her colleagues on coding of acute myocardial infarction illustrate some of these limitations (1). This study reports that more than one-quarter of the patients assigned an acute myocardial infarction code from the International Classification of Diseases (ICD-9-CM) at the time of discharge did not have the condition or receive active treatment for the condition during hospitalization. Miscoding resulted most often when patients were admitted with a "rule-out infarction" diagnosis. Misspecification (that is, the physician failed to note explicitly the absence of acute myocardial infarction) or failure of the medical abstracter to note subsequent explicitly documented exclusion of the infarction resulted in the largest number of coding errors. Admission of patients for cardiac catheterization with coronary angiography within 8 weeks of acute myocardial infarction (thus technically permitting the acute myocardial infarction code) was cited as another major reason for misclassification.

The difficulties raised by the coding guidelines for the ICD-9-CM and the diagnosis-related group (DRG) codes are further compounded when a secondary diagnosis of acute myocardial infarction is used to assign the infarction DRG to cases where another cardiac condition is the principal diagnoses. The study supports the conclusion that previous hospital discharge data on acute myocardial infarction lack sufficient validity in themselves to define an inception cohort for effectiveness and outcomes research. As coding rules change over the next year, however, to minimize some of the above-mentioned problems, identification of an inception cohort from the discharge codes will become more accurate.

Inaccuracy of diagnostic codes is not unique to acute myocardial infarction. In the next section, I amplify on the four general limitations of claims data in the context of assessment of three technologies: diagnostic devices, drugs, and clinical trials.

Limitations for Diagnostic Devices

This is probably the area in which claims data are likely to be least useful, in the absence of significant changes. The first limitation derives from the fact that, for inpatients, Medicare claims files code for only three procedures. Ill patients usually have significantly more than three diagnostic procedures, and hence the list of coded diagnostic procedures is frequently incomplete and biased. It is biased because sicker patients will not have room on the claim for the diagnostic code, whereas healthier patients will.

An example of this phenomenon occurred when ProPAC tried to track the use of magnetic resonance imaging (MRI) among Medicare beneficiaries. There were far fewer MRIs reported than we estimated had been done. In addition, there were far fewer done on sicker patients in the DRGs most likely to make use of MRI. This is analogous to the phenomenon described by Stephen Jencks regarding concurrent diagnoses among ill patients (2).

The second problem in the evaluation of the effectiveness of diagnostic devices relates to time-lags between the use of new technologies and the development of codes for them. Development of codes can take years, thus preventing us from identifying the use of new devices. Although this is a limitation primarily of inpatient records, it can occur in outpatient records as well. Examples of coding omissions that hinder evaluation include MRI, electrophysiological studies, and positron-emission tomography (PET). Third, claims data do not provide any information on the type of equipment used. For imaging technologies this is critical: major differences in effectiveness can result from use of older generations of equipment. Fourth, it is seldom possible to differentiate between tests done for diagnosis and those done for screening. This is obviously important in the case of mammography. Finally, there is no correlation of diagnostic test results with information from an independent source (for example, pathology).

It is important to emphasize that, to the extent that we have information from inpatient care (ICD-9-CM codes) and ambulatory sources (ICD-9-CM or Current Procedural Terminology [CPT] codes) some of the above problems can be alleviated. In any case, the limitations described above regarding inpatient data have prompted the National Cancer Institute to conduct a major *prospective* study of the effectiveness of diagnostic imaging procedures in patients with one of five types of cancer. Nine institutions are currently collaborating in this study, and six more are expected to be added next year.

Limitations for Drugs

The problems of claims data for drugs are similar to those for diagnostic devices. Codes for new drugs may lag their availability by many years. The classic example of this relates to thrombolytic therapy. Most physicians, policymakers, and researchers identified this as an extremely important area for study two years ago; however, there were no codes for thrombolytic therapy. There are still no codes for the therapy per se—it can be identified (and then not always) only when done in connection with an angioplasty.

Drugs are very complicated to evaluate because of multiple doses and multiple forms, and it is going to be tricky to get information on outpatient drug use. The repeal of the Medicare Catastrophic Coverage Act, with its drug coverage, will make information on Medicare beneficiaries more difficult to obtain. However, a number of researchers have been extraordinarily

successful in using claims data from selected states (for example, New Jersey) (3).

Limitations for Therapies

There has been a tremendous amount of discussion about the use of claims data for therapies, and much of it has been very negative. I think we should recognize, however, that a number of useful things can be accomplished with claims data for therapy. For one, we may not need to resort to randomized trials for all interventions.

Some limitations remain, however. The first one is that the coding for therapy is not always current. For example, two years ago ProPAC was interested in studying cochlear implants as a new therapy for patients with deafness. At the time there was no way of identifying these patients from hospitalization claims data alone. The second problem with the coding for therapies is that the code may not be specific enough. This is particularly troublesome for ICD-9-CM codes used on inpatient records. CPT codes are considerably more specific in reporting procedures although they have little or no diagnostic information. Thus, if bills for physician services or outpatient services are linked with hospitalization records, specificity is improved.

Failing that linkage, there are problems in four areas:

1. The ICD-9-CM codes do not reflect *refinements* in a procedure (for example, a cementless instead of a cement hip prosthesis).
2. The codes frequently do not indicate whether a procedure was a *repeat* one (for example, a first or a second coronary artery bypass graft). Linking patient records over many years (for example, 10 years) would solve this problem if the payer were the same during the entire period.
3. The codes are sometimes incomplete. A one-year study of total parenteral nutrition (TPN) conducted by ProPAC illustrates this. At that time it was believed that DRGs 296 and 182 (nutritional disorders and miscellaneous digestive disorders) would contain many patients having TPN. A review indicated that approximately 1,200 patients that year (less that 1 percent of all patients in those DRGs) were identified from the claims records. Independent estimates suggested a number more like 100,000 to 200,000 patients. In this case, as with MRI, sicker patients had enough other procedures done to them that TPN never reached the claims records.
4. Claims data seldom allow identification of an inception cohort. This was mentioned under general limitations, but I repeat it here because of its particular importance for evaluation of therapies. Elliot Fisher and John Wennberg emphasize this in their discussion of the claims analyses of transurethral prostatectomies (4). In general, it will be easier to define an inception cohort for an acute event, such as acute myocardial infarction, than for a chronic one.

CONCLUSION

Finally, where are we? I think we are at a point in the careers of a number of health services researchers that is really quite rosy. We have a data base that is constantly being improved and will continue to be improved as a result of our research interests. Over the short term, I believe that these data will be used primarily for generating hypothesis. Our resultant analyses and studies will have an obvious impact on our ability to measure the effectiveness of medical practice. Over a longer term, it is likely that some of our results will be used to identify access problems. Who is not getting what? For what reason? To accomplish both short-and long-term objectives, we must work closely with policymakers on activities related to improving the data base and the training of individuals capable of using it.

References

1. Iezzoni, L.I., Burnside, S., Sickles, L., et al. Coding of Acute Myocardial Infarction: Clinical and Policy Implications. *Annals of Internal Medicine* 109:745-751, 1988.
2. Jencks, S.L. Issues in the Use of Large Data Bases for Effectiveness Research. Pp. 94-104 in *Effectiveness and Outcomes in Health Care*. Heithoff, K.A. and Lohr, K.N, eds. Washington, DC: National Academy Press, 1990.
3. Avorn, L., Dreyer, P., Connelly, K., et al. Use of Psychoactive Drugs and Quality of Care in Rest Homes. *New England Journal of Medicine* 320:227-232, 1989.
4. Fisher, E.S. and Wennberg, J.E. Administrative Data in Effectiveness Studies: The Prostatectomy Assessment. Pp. 80-93 in *Effectiveness and Outcomes in Health Care*. Heithoff, K.A. and Lohr, K.N., eds. Washington, D.C.: National Academy Press, 1990.

PART IV

METHODOLOGICAL ISSUES AND WORK IN PROGRESS

Use of Large Data Bases: Introduction

Emmett B. Keeler, Session Moderator

Although it is not entirely clear what is meant by large data bases, we know that to administer its programs, the Health Care Financing Administration (HCFA) collects enormous amounts of data that contain information on the location and use of medical services, both inpatient and outpatient, and information on everyone covered by and mortality associated with Medicare and Medicaid. To keep the costs of administration down, HCFA does not collect all the clinical detail that researchers might want. However, the data are fairly universal in scope, and there are lots of possibilities for using them as a resource: linking them to outside data, putting together different HCFA files (such as hospital records with outpatient records), and so forth. Used creatively, they are an invaluable resource for anybody interested in studying what is actually occurring in the United States.

Janet B. Mitchell is president of the Center for Health Economics Research in Needham, Massachusetts. She and her institute are both well known for their studies of payment mechanisms and their effects on physicians. Dr. Mitchell gives a general methodological overview of the things that can be done with administrative data sets.

Elliot S. Fisher is a physician at Dartmouth Medical School and was involved in the large data set analysis of the Wennberg study, which is the prototype for effectiveness research. (John Wennberg is director of the Center for Evaluative Clinical Sciences.) Dr. Fisher and Dr. Wennberg highlight the problems and achievements of the original study and describe the use of administrative data in the ongoing assessment of treatments for benign prostatic hyperplasia.

Stephen F. Jencks is a physician and chief scientist at the Office of Research in HCFA. Dr. Jencks has extensive experience in sponsoring, critiquing, and performing a number of studies looking at postadmission mortality. He discusses the uses and limitations of claims data for outcomes research.

12

The Role of Large Data Bases in Effectiveness Research

Janet B. Mitchell

The first question in any consideration of the use of large data bases in effectiveness research is: what is a "large data base"? Usually, it refers to administrative records, or insurance claims data, regarding patients receiving various treatments. The nice thing about using claims for research purposes is that someone else actually collects the data, namely, providers filling out the claims forms. By the time the researcher receives the claims, the data are already computerized in a consistent format.

SIZE OF LARGE DATA BASES

One of the major difficulties in working with these data bases is that they are indeed large—enormous or gargantuan might be more appropriate descriptors! It is not uncommon to work with millions of claims on hundreds of reels of tape. I am sure many of you have conducted clinical research involving hundreds of patients, and you may be wondering why I or anyone else would want to get involved with millions of records in the first place. The reason, of course, is that these records do not represent individual patients, but rather pieces of information describing the medical services received by each patient. These pieces of information need to be put together in order to obtain a picture of an episode of care. During a single inpatient episode, for example, a patient might incur anywhere from a dozen to a hundred bills. For longer periods of care, the number of records would be considerably larger, especially for sicker patients.

Why so many claims? In Medicare, for instance, inpatient hospital and skilled nursing facility stays are billed using a single claim, but physician and other Part B services are billed individually. Thus, there will be a claim for every discrete service: for every surgical procedure, for every

visit, for every X-ray, for every laboratory test. The detailed nature of these claims data bases is one of their greatest strengths; the creative researcher can use them in an almost infinite variety of ways.

USES OF LARGE DATA BASES

Probably the most common use of claims data for effectiveness research is to follow patients with a specific diagnosis or patients receiving a specific therapy. Diagnoses are available on institutional claims; procedures are documented on all physician bills. For example: What happens to patients receiving percutaneous transluminal angioplasty? What services do those patients receive afterwards and in what kinds of settings? Some services will suggest that complications have arisen, say, if the procedure is followed closely by repeat angioplasty or bypass surgery.

Outcomes, such as readmission and mortality rates, can also be studied. Besides studying individual patients or episodes of care, claims data can also be used to evaluate effectiveness at the level of individual providers, such as hospitals. Thus they provide an opportunity to examine questions such as whether mortality rates for a given procedure depend in part on a hospital's surgical volume, for example.

MEDICARE DATA BASES

Medicare claims files are particularly valuable, for several reasons. First, every beneficiary has a unique identification number based on his or her Social Security number. Because this number is attached to every Part A and Part B claim, it is easy to construct episodes of care for individual patients. Sometimes, however, these numbers are slightly different on the Part A and the Part B claims. Fortunately, there are fairly straightforward algorithms that can be used to equate them.

Second, the Health Care Financing Administration (HCFA) maintains claims data on samples of patients for research purposes. These samples are selected, based on their identification numbers, and remain in the data base until the patient dies. This enables researchers to follow the same patients over a period of years. In addition, HCFA maintains eligibility files that include information on dates of death. Because of the need to prevent Social Security checks from being mailed to deceased beneficiaries, these deaths are verified and the dates are believed to be reasonably valid.

Historically, researchers have primarily used Part A hospital records to study effectiveness issues. Only relatively recently have they discovered the value of Part B claims, either in their own right or as supplements to Part A data. One major limitation of hospital claims for effectiveness research is the absence of detailed information on what was actually done to the

patient in the hospital. Part A claims do include information on surgical procedures, but this information is generally limited to procedures that affect assignment to diagnosis-related groups (DRGs); thus, many diagnostic surgeries are missing. The only data available on ancillary diagnostic tests, furthermore, are simply charges per revenue center, that is, charges for radiology with no indication of how many X-rays were performed or which ones. There is also no information on physician visits and consultations.

Except for some services performed by residents, however, every physician service will show up as a Part B bill. These bills provide the researcher with an in-depth look at the mix of services provided during the hospital stay. Because each physician bill includes the date of service, we can also look at the timing of various tests. This can be useful in trying to infer the clinical decision-making process that took place during the hospitalization.

The Part B detail can also be used to define the universe of patients receiving a specific therapy of interest. Not all patients undergoing coronary bypass surgery will be identified through DRGs 106 and 107, for example; a surprising number will show up in other DRGs, such as those involving valve replacements. This is important, as geographic variation has been found in the frequency with which bypass operations are combined with other open-heart surgery. Thus, how a study sample is selected could have profound effects on the research findings.

Anesthesiologists and assistant surgeons frequently report a different procedure than that billed by the surgeon. Usually, they are reporting an operation in the same general anatomic area, but not always. My rule has always been to assume that the primary surgeon is right and use what this surgeon reports to define the sample.

Using claims data to examine outcomes associated with ambulatory episodes of care is more problematic because of the absence of diagnostic information on the Part B claims. Thus it is not possible to determine the reason for a given office visit or to trace referral patterns accurately. Beginning this year, however, physicians are being required to assign diagnoses a code number from the International Classification of Diseases (ICD-9-CM) and to include those numbers on their claims, so it is possible that such analyses will be feasible in the future.

It is possible to identify specific illnesses indirectly, using the procedure codes on the Part B claims. Services provided under Medicare Part B are billed using the Common Procedural Terminology (CPT-4) or, in the case of nonphysician services, a system developed by HCFA known as HCPCS (HCFA Common Procedure Coding System). There are over 10,000 codes available for billing purposes. This wealth of codes is the despair of many policymakers, who feel it helps fuel the inflation in physician spending. However, it is a boon to researchers.

Unlike the ICD-9-CM procedure codes, which are often vague concerning

the precise nature of the surgical procedure or diagnostic test, CPT-4 records that information in excruciating detail. We can tell, for example, not just that a patient received a total hip replacement, but whether it was an original replacement, whether it was a conversion of previous hip surgery to a total hip replacement, or whether it was a revision of an earlier replacement. In the latter instance, we also know whether the revision involved the acetabular part of the hip, the femoral component, or both. Some examples of identifying outpatient treatments through the procedure codes would include hemodialysis for end-stage renal disease patients and chemotherapy for cancer patients.

A particular interest of many researchers is how the utilization of services varies around the country. Unfortunately, only the institutional claims include information on exactly where the service was provided. The only geographic identifiers on Part B claims are the carrier (which generally corresponds to a state) and the reasonable charge locality. The reasonable charge locality is a fairly arbitrary geographic entity used by the carriers to determine allowed charges. It provides a finer breakdown than the state, but it is still fairly crude. In fact, for 16 states, only a single statewide locality is used.

The Part B claims also lack any information on where the patient lives. This means that population-based measures of utilization and outcomes can be easily created only for hospital services. The researcher who wants to study the utilization of ambulatory services must obtain information on the patient's residence from HCFA's eligibility files and merge it.

Let me mention here an additional consideration when analyzing Part B claims data. Although Medicare is a national program, each carrier has considerable flexibility in how it actually processes and pays claims. These idiosyncracies can lead the unwary researcher astray.

Permanent pacemaker insertion is a good example of the potential problems that can be encountered. A number of physicians use the team approach to pacemaker insertion; a surgeon makes the pocket to hold the device and a cardiologist inserts the electrodes. Carriers have attempted to recognize the team approach and reimburse it in a number of different ways. In some states, each physician submits a bill for pacemaker insertion without any indication that another physician was involved. The carrier knows which physicians practice in this way and pays each physician less than if he Or she had performed the procedure independently. The researcher Cannot tell this from the claims data, however, and it will appear as if twice the number of pacemakers were inserted in that area.

One carrier has dealt with the team approach by having one physician bill for the insertion, while the other physician bills for pacemaker repair. If a researcher did not know this ahead of time, it would appear that there were a lot of pacemaker failures in that particular state.

So far, I have been talking about Part B physician and Part A hospital

claims, but Medicare claims are also available for other types of services, such as skilled nursing facility and home health care. These claims can be particularly valuable for examining rehabilitative treatment; one example might be to look at the care received following hip fracture.

MEDICAID DATA BASES

To date, most research has focused on Medicare patients, for two reasons. The Medicare program is consuming an increasingly large share of the federal budget, and the claims data have been readily available (more or less) from HCFA. Because of problems in data acquisition, the services received by Medicaid patients have historically received less attention. HCFA is working on some new data bases that will eventually provide Medicaid claims in a consistent format for all states. I believe data from about a half-dozen states are available at the present time.

There are several advantages in using Medicaid claims to analyze effectiveness, either in conjunction with or in place of Medicare claims. For one, the Medicaid-eligible population encompasses a much wider age range, thus permitting study of pregnancy and pediatric illnesses. In addition, there are other important conditions whose incidence is simply not sufficient to study in the Medicare population. Substance abuse is one example; another is AIDS. Although the permanently and totally disabled are also eligible for Medicare coverage, most AIDS patients simply do not survive long enough to qualify for benefits. A large number do become eligible for Medicaid, often early in the disease process, and Medicaid claims can be used to help track the effectiveness of various treatment regimens.

Another advantage of Medicaid claims is that the Medicaid program covers a wider range of benefits than does Medicare, especially in the areas of long-term care and prescription drugs. A major disadvantage of Medicare claims has been that, although the program serves the elderly, it covers only a small part of long-term care—only 150 days of nursing home care per year, and that care must be in a skilled nursing facility. This means that studies of patients with chronic conditions requiring ongoing custodial care (for example, Alzheimer's disease, stroke, or spinal cord injury) will be able to paint only a partial picture of health care use. Because state Medicaid programs do cover these services, however, Medicaid claims can be used to fill some important gaps.

Similarly, because Medicaid pays for most prescription drugs, these claims can be used to evaluate alternative treatments or to identify a sample of patients undergoing a given treatment regimen: for example, all AIDS patients receiving AZT. Data on prescription drugs can be used in many ways. An obvious one is to compare the effectiveness of drug therapy to surgical intervention. Another is to look at adverse or unintended consequences of

specific medications. One researcher, for example, examined the incidence of hip fracture in patients receiving psychotropic drugs.

One of the main disadvantages of Medicaid claims is that Medicaid recipients are not representative of the population at large. This is in contrast to Medicare recipients: a sample of Medicare patients with myocardial infarction is virtually synonymous with a sample of elderly persons with myocardial infarction. Another disadvantage is that, unlike Medicare beneficiaries, Medicaid patients are not always continuously eligible for care. This is particularly true of recipients of Aid to Families with Dependent Children, who may be eligible for only some months in a year.

The Medicare Catastrophic Coverage Act passed by Congress last year would have given Medicare many of Medicaid's data advantages, and thus research advantages, by expanding coverage. Both the skilled nursing facility benefit and the home health care benefit were extended, for example, providing more data on these components of postacute care. Screening mammography was a brand-new benefit. Most important, the legislation expanded Medicare coverage to outpatient prescription drugs. Repeal of the Act in late 1989 deprived researchers of the opportunity to broaden the questions that could be addressed using Medicare claims data and thus expand effectiveness and outcomes research.

13

Administrative Data in Effectiveness Studies: The Prostatectomy Assessment

Elliott S. Fisher* and John E. Wennberg

Comprehensive, population-based administrative health care data bases provide an increasingly accessible and important source of data for studies of the effectiveness of health care (1). To illustrate their potential uses, their strengths, and their limitations, we describe the role that administrative data have played in the ongoing assessment of treatments for benign prostatic hyperplasia, one of the more common conditions affecting elderly men.

OVERVIEW OF THE PROSTATECTOMY ASSESSMENT

Analyses of administrative health care data bases have long documented marked variations in population-based rates of prostatectomy (2,3). To understand the causes of these variations, a multidisciplinary team composed of practicing urologists from Maine and researchers from academic medical centers in the United States, Canada, and Europe was assembled. The assessment team, funded under the Patient Outcome Assessment Research Program of the National Center for Health Services Research, undertook a comprehensive program of evaluation, the early findings of which are described in a series of recent publications (4-8). These findings are briefly summarized in Table 1 to provide a context for the description of the analyses based on administrative data.

The first goal of the assessment process was to identify possible explanations for the observed variations in utilization rates. This entailed both a

* The paper was presented by Dr. Fisher, but it represents the ongoing research of many investigators in the Prostatectomy Patient Outcomes Research Team of which Dr. Wennberg is the principal investigator. The work is now part of the Patient Outcome Research Team program of the Agency for Health Care Policy and Research.

review of the scientific literature and discussions with practicing urologists in Maine. Two conflicting theories concerning the indications for prostatectomy were identified. Many physicians believed that prostatectomy should be performed early in the course of symptomatic prostatism on the theory that if the operation is delayed, the patient will be at higher risk when the surgery becomes unavoidable. Because overall life expectancy would be reduced by delay, those who held to this *preventive theory* believed that watchful waiting was not a reasonable option. In contrast, urologists who believed the *quality of life* theory argued that prostatectomy is not inevitable. For patients without evidence of actual or impending renal dysfunction, the primary indication for the procedure should be improvements in functional status and quality of life. According to this theory, watchful waiting is a reasonable option.

TABLE 1 Aims, Methods, and Data Sources for Assessment of Treatments for Benign Prostatic Hyperplasia

Aim	Method and Data Source
Describe patterns of use of treatments and characterize the theories of efficacy advanced by their proponents	Geographic variation studies using insurance claims and other large data bases Structured literature review and focus groups with practicing physicians
Identify, define, and develop (where necessary) measures for the full spectrum of relevant outcomes	Literature review and semi-structured interviews with patients, physicians Identification or development of valid and reliable outcome and case-mix measures
Establish the best estimates for probabilities of the relevant outcomes of alternative treatments	Claims-based cohort studies; linkage of claims and other data bases Prospective cohort studies (Maine Interview Study)
Assess the efficacy of alternative treatment theories	Decision analysis, meta-analysis Observational studies, randomized trials where appropriate
Integrate results, identify questions for further research	Publication of results and impart findings to practicing physicians Development of interactive video for Shared Medical Decision-making Procedure

To evaluate these competing theories, the assessment team identified all relevant outcomes through discussions with patients and physicians. A

review of the medical literature demonstrated serious gaps in existing knowledge about these outcomes. Claims-based analyses made possible reliable measures of the likelihood of mortality in the postoperative period and of reoperation (4). The probabilities for other outcomes—such as incontinence, impotence, and postoperative symptom relief and improvement in functional status—required the development of new measurement instruments and the implementation of a prospective interview study of patients undergoing prostatectomy in Maine (6).

The findings of the literature review, the claims-based analyses, and the interview study provided sufficient data to assess the efficacy of watchful waiting versus transurethral prostatectomy (TURP) through decision analysis (5). The decision analysis demonstrated that for most patients the decision to undergo prostatectomy results in a slight decrease in life expectancy. These findings confirmed the opinion of those physicians who believed that the operation was justified primarily for its value in reducing symptoms. However, the assessment also demonstrated (a) that improvements in symptoms were only available to those willing to accept the risks of the surgery, and (b) that patients with identical symptoms differed greatly in their attitudes toward those symptoms and, presumably, toward the risks of surgery.

The assessment thus revealed that variations in utilization rates induced by practice style were primarily a function of differences in providers' attitudes toward the preventive theory and of difficulty in integrating patients' preferences into the decision to undergo prostatectomy. To help address these difficulties, the assessment team developed a computer-assisted, interactive video presentation that provides a comprehensive description of the risks and benefits of the alternatives and is tailored to the individual patient viewing the presentation. This Shared Medical Decision-making Procedure (SMDP) has been implemented in several participating centers, with both surgical and watchful waiting patients being followed up to provide further refinements in the probability estimates for outcomes.

The assessment steps described above required the application of multiple research methodologies. In the remainder of this chapter, we describe the role that administrative data bases played both in the assessment of TURP versus watchful waiting and in addressing a specific question that emerged from early analyses.

OVERVIEW OF METHODS

Because we are describing the results of a series of studies conducted over many years, it is impractical to present in detail the methods used in each of the analyses. The general approach followed, which was similar in all analyses, will be reviewed briefly. The reader is referred to the primary publications for additional detail (3,4,8,9).

All of the studies relied on administrative or health insurance data bases as primary sources of data. The Health Care Financing Administration (HCFA) maintains comprehensive files on inpatient, outpatient, and skilled nursing home care for virtually the entire U.S. population over the age of 65 (10,11). Similar files have long been maintained by the Manitoba Health Services Commission, in the Oxfordshire Region of England, and in Denmark (8).

Three features of these files are essential to the analysis. First, the eligible population can be precisely defined, the date of death can be ascertained independent of health care utilization, and patients can be located for long-term follow-up studies. Second, administrative procedures in each system ensure that virtually all hospital utilization is documented. Third, unique personal identifiers allow utilization files to be linked to each other, to the population files, and to other sources of data.

The methods used to define cases for inclusion in the study population and to define relevant variables were similar in all claims-based analyses reported here. They thus represent a generalizable approach to the use of administrative data bases for cohort studies.

Case Identification and Variables

All patients were initially identified on the basis of computerized hospital discharge abstracts or physician claims documenting a prostatectomy during the various study periods encompassed by the assessments. Where both physician and hospital claims were available (HCFA and Manitoba), potential cases were identified, consistency checks carried out, validity of claims determined, and appropriate exclusions applied. For each case, the first prostatectomy during the study period was defined as the index operation. Based on the claims data, three classes of variables were defined.

Outcomes

The population file was searched to determine whether and when patients might have died. Reoperation was defined based on the presence of subsequent claims for prostatectomy. Other possible complications were defined based on combinations of diagnoses and procedures coded on inpatient hospital records and on physician claims for both inpatient and outpatient services.

Patient Covariables

Diagnoses recorded on the index hospitalization claim and on physician and hospital claims preceding the index prostatectomy were used to measure comorbidity.

Treatment Variables

The specific codes recorded on hospital and physician claims were used to define the type of prostatectomy received by the patient (open versus transurethral).

USES OF ADMINISTRATIVE DATA IN PROSTATECTOMY ASSESSMENT

Variations in Utilization Rates

First, and perhaps most important, studies of small-area variations in prostatectomy rates provided the initial stimulus for the research project and were critical to engaging the interest of practicing urologists in the assessment. Early studies documented age-adjusted population-based utilization rates for prostatectomy that varied by a factor of four across small areas of New England (3). Other studies documented variations across large geographic regions (12) and between and within countries with different health care financing and organizational structures (13). Discussed extensively elsewhere (14,15), small-area analyses have highlighted the clinical uncertainty surrounding many decisions in medicine and underlined the need for comprehensive assessments of the risks, benefits, and alternatives to specific treatments.

Population-Based Estimates of Adverse Outcome Rates

As mentioned above, urologists in Maine disagreed in their understanding of the risks and benefits of prostatectomy. Some of this disagreement could be attributed to gaps and flaws in the existing medical literature. Physicians usually rely on reports from clinical trials and case series to estimate the risks of adverse outcomes following specific surgical or medical interventions. Unfortunately, these sources suffer from several limitations. One problem with case series is reporting bias: only where the results are better than previously reported is there a strong incentive to publish. Consequently, published rates of adverse outcomes may underestimate the risk in most clinical settings. Clinical trials usually report findings on highly selected populations and therefore may be difficult to generalize. Moreover, sample sizes are usually limited, and follow-up and choice of outcomes for study vary among case series. Consequently, confidence intervals (CIs) are likely to be wide, rare events may not be documented at all, and results are difficult to pool. Claims data can overcome these limitations.

in the mid-1970s, a review of the literature on prostatectomy stated that mortality rates following TURP were under 1 percent and that patients rarely required reoperation (16). Wennberg, Roos, and colleagues, using claims data from Maine and Manitoba for 1974 through 1976, found that over 3 percent of patients died within 90 days of surgery and that the overall rate of reoperation following TURP was 20.2 percent at eight years (4). While these findings demonstrate the difficulty of relying on small, highly selected samples to estimate the likelihood of various outcomes, the data are by now quite old. What are the current risks of prostatectomy?

We have used Medicare data for New England to examine mortality and morbidity following prostatectomy in the 1980s (Tables 2 and 3). Because of the large sample sizes, mortality rates for prostatectomy can now be precisely estimated: 30-day mortality ranges from 0.3 percent for patients between the ages of 65 and 69 to 2.6 percent for patients age 80 and over. Studies of morbidity are more difficult when relying on claims data alone, because few of the diagnostic and procedure codes used on hospital discharge abstracts or physician claims specify that a given complication or procedure is the direct consequence of a prior prostatectomy. Consequently, using methods similar to those described by Roos et al. (17), we asked physicians to group codes into those that were possibly complications of the procedure (that is, outcomes occurring with increased frequency following any operation) and those that were probably complications (because they are more directly related to prostatectomy). More than 10 percent of patients had a probable complication, while 16 percent had a possible complication. In all, almost one-quarter of patients had significant adverse outcomes in the 90 days following prostatectomy.

TABLE 2 Mortality Rates Following Prostatectomy Among Medicare Enrollees Who Were New England Resident Patients Without Indication of Prostate or Bladder Cancer, 1984-1986

Age Group	Cases (No.)	Patient Dead Within 30 Days of Surgery (Percent)	Patient Dead Within 90 days of Surgery (Percent)
65-69	6,428	0.3	1.2
70-74	6,946	0.8	2.3
75-79	5,740	1.2	3.2
80 and over	5,652	2.6	6.8
Total	24,766	1.2	3.3

Note: Based on Medicare Part A and Part B claims and Medicare Enrollment (HISKEW) files.

TABLE 3 Morbidity Rates Within 90 Days of Transurethral Prostatectomy Among Patients Without Indication of Prostate or Bladder Cancer Who Were New England Resident Medicare Enrollees, 1984-1986

Possible Complications	Percent	Probable Complications	Percent
Myocardial infarction	1.0	Bladder infection	2.0
Pulmonary embolus	0.3	Kidney infection	0.1
Respiratory infection	3.0	Prostate infection	0.2
Wound infection	0.3	Other urinary infection	0.3
Congestive heart failure	1.5	One or more urinary infection	2.3
Phlebitis	0.2	Stricture treatment	3.7
Deep venous thrombosis	0.3	Retention treatment	1.8
Arterial embolus	0.4	Other invasive testing	4.5
Bleeding	7.5	Second prostatectomy	0.6
Miscellaneous	3.3	One or more invasive procedures	8.7
One or more of above	16.0	One or more of above	10.3

One or more possible or probable complications, 23.4 percent
Note: Based on Medicare Part A and Part B claims files.

Although these data demonstrate that a variety of adverse events may be detected through the claims data, several limitations must be acknowledged. First, the completeness and accuracy of the coding in claims data bases has been questioned (18,19). However, if the accuracy of the data could be confirmed and if administrative safeguards were enacted to ensure their complete and accurate documentation, then claims-based measures could be used to monitor the outcomes of care for patients undergoing prostatectomy. Second, the scope of the data is limited. Many outcomes critical to the prostatectomy assessment, such as disease-specific functional status and quality of life, could not be ascertained from the claims data. The next section provides examples of how these specific limitations of the claims data can be overcome.

Comparisons of Transurethral and Open Prostatectomy

The initial claims-based analyses of prostatectomy outcomes in Maine and Manitoba also compared the long-term results of TURP with those of open prostatectomy. Both operations have the same purpose—to relieve urinary obstruction. The open procedure is usually performed through an incision in the abdominal wall, whereas the transurethral procedure is performed through the urethra. Because of its less invasive nature, TURP was believed by urologists to be both safer and more effective than the open

operation. Although a randomized clinical trial has never been conducted, TURP has gradually replaced open prostatectomy to the point where, in the 1980s, only about 5 percent of prostate operations in our data base were open.

The claims data provided an opportunity to compare the long-term outcomes of the two procedures. Controlling for both patient and hospital characteristics, our study showed that patients undergoing TURP were twice as likely to require reoperation within eight years and appeared to face a significantly elevated long-term risk of death, compared to patients receiving the open procedure (4). These findings raised potentially important questions about both the safety and the efficacy of TURP compared with open prostatectomy.

To evaluate further the association between the type of operation received by patients and their long-term outcomes, several additional studies were conducted. The first study sought to determine whether the increased risk associated with TURP would be found across different time periods and in different countries. Retrospective cohorts were assembled; these cohorts consisted of all patients aged 55 through 85 (except those with bladder or prostate cancer) who underwent prostatectomy between 1977 and 1985 in Denmark, between 1972 and 1985 in Manitoba, and between 1963 and 1977 in the Oxfordshire region of England (8). The risk of reoperation was consistently higher among patients who received a TURP, ranging from a relative risk of 2.7 at eight years in Denmark to 6.7 at eight years in Oxford. Also, the risk of death following TURP was consistently higher at five and eight years, the relative risk of TURP to open being 1.2 to 1.3 at eight years.

There remained the possibility that physicians were selecting only relatively healthy patients for the open procedure and that increased severity of illness among TURP patients might explain the excess mortality observed. Data from a teaching hospital in Manitoba were reviewed to investigate this possibility. All patients who underwent prostatectomy at the hospital between July 1974 and December 1983 were identified through the claims data. Those with bladder or prostate cancer were excluded. All claims records before and after prostatectomy were identified and used to define patient covariables, including age, the presence of cancer diagnoses, prior hospitalizations with high-risk diagnoses, and nursing home residence. A clinical data base collected by anesthesiologists for a study of all surgical patients at this hospital was identified, and key clinical variables were extracted and linked to the prostatectomy records. The linked variables included the American Society of Anesthesiologists' risk score and medication use.

Among all cases the adjusted relative risk of death within five years was 1.45 (95 percent CI, 1.15, 1.84) (see Table 4). Similarly, after excluding all cases with evidence of significant comorbidity, the relative risk remained

elevated at 1.60 (95 percent CI, 0.93, 2.77), although the confidence limits increased because of the smaller sample size.

TABLE 4 Relative Risks of Death for Patients Receiving Transurethral (TURP) and Open Prostatectomy, Operated On at Manitoba University Hospital, 1974-1983, by Selected Demographic and Clinical Characteristics

Characteristics	All Patients (N = 1650)	Healthiest Patients[a] (N = 557)
TURP vs. Open prostatectomy	1.45 (1.15, 1.84)[b]	1.60 (0.93, 2.77)
Age Groups		
85+ vs. 55-69	3.75 (2.75, 5.09)	5.92 (2.44, 14.40)
80-84 vs. 55-69	2.77 (2.07, 3.72)	5.22 (2.57, 10.60)
75-79 vs. 55-69	2.35 (1.78, 3.10)	3.54 (1.85, 6.79)
70-74 vs. 55-69	1.48 (1.12, 1.96	1.60 (0.79, 3.24)
Cancer diagnosis prior to surgery	3.93 (2.92, 5.28)	NA[c]
Hospitalized with high-risk diagnoses prior to surgery		
Within 6 months	1.46 (1.14, 1.87)	NA
Within 7-12 months	1.54 (1.13, 2.10)	NA
Nursing home resident	1.17 (0.76, 1.80)	NA
ASA Score 3+	1.91 (1.57, 2.36)	NA
On digitalis	1.40 (1.10, 1.78)	NA
High-risk diagnosis	1.42 (1.15, 1.76)	NA
Prostatic hyperplasia only diagnosis	0.54 (0.38, 0.77)	.0.41 (0.22, 0.74)

Note: Cox regression results based on linked claims and anesthesia data bases.
[a] Healthiest defined as not resident in nursing home, had no current or previous diagnosis of cardiovascular disease, had no diagnosis of cancer, took no medications preoperatively, had no other high-risk diagnosis, and had a physical status score of 1 or 2 (healthy or mild disease).
[b] 95 percent confidence intervals in parentheses.
[c] Not applicable.
SOURCE: Roos et al. (8).

There remained a concern that the elevated risk might reflect subtle characteristics of patients known to their physicians and recorded in the medical record but not in either the claims data or the anesthesiologists' study. To address this concern, the medical records of a sample of TURP and open patients were abstracted to obtain a broad range of clinical data from patients'

histories, physical examinations, and laboratory findings at the time of surgery.

The medical record data were used to determine an index of comorbidity and a measure of functional health, both of which have been previously demonstrated to predict long-term survival (20,21). Two Cox regression models were developed. In one, we used the indices of comorbidity and functional health to control for differences in illness levels. In the other we allowed all variables significantly associated with long-term survival into the model. Using these models, the relative risk of death within five years of operation was elevated for patients undergoing TURP compared to open prostatectomy, and it was similar in magnitude to the relative risk obtained from the claims data alone (Table 5).

These analyses suggest several conclusions. First, they confirm our initial observation of increased mortality and reoperation rates among TURP patients in the original small sample from Maine and Manitoba. Second, measures of case mix that were obtained retrospectively did not explain the findings. However, it is important to note that patients may appear similar based upon retrospective review of their charts, but that the measures obtained retrospectively may not identify significant prognostic differences. For example, physicians may record characteristics of patients differently, based on their own assumptions about the relative safety of TURP compared to open prostatectomy. Nevertheless, because of the large numbers of patients undergoing TURP and the potential public health importance of the observed increased mortality following TURP, the evidence we found should not be

TABLE 5 Relative Risk of Death for Patients Receiving Transurethral (TURP) versus Open Prostatectomy, Operated On at Manitoba University Hospital, 1974-1983

Variable	Adjusted Relative Risk (95% confidence interval)
TURP vs. Open	1.59 (1.06, 2.37)
Age 70-74 vs. under 70	1.69 (1.05, 2.64)
Age over 75 vs. under 70	2.23 (1.38, 3.58)
Comorbidity index ≥ 2	2.52 (1.74, 4.08)
Decreased functional status[a]	2.66 (1.74, 4.08)

Note: Cox regression results based on linked claims and chart review data, N = 485.
[a] Decreased functional status defined as a Karnofsky score ≤ 70.
SOURCE: Malenka et al. (9).

ignored. We are pleased that the American Urological Association has joined with our assessment team to undertake the prospective clinical trials needed to resolve the issue.

IMPROVING THE USEFULNESS OF ADMINISTRATIVE DATA BASES

Administrative data have played an important role in stimulating the current interest in studying the effectiveness of medical care and offer an important resource for assessments of current treatment patterns. To make use of their full potential, we should build on their strengths and make the investment necessary to overcome their limitations.

Strengths

As recognition of the importance of further evaluation of medical practice has grown, so has advocacy of the Medicare claims files and similar data bases as sources of data for technology assessment. The assessment of prostatectomy exploited four major strengths that Medicare data offer for outcomes research. First, the enrollment file provides not only the population counts required for epidemiological studies, but also a means to efficiently ascertain death, eligibility status, and change of residence for long-term follow-up studies. Second, universal coverage offers the opportunity to study populations that are free from selection bias and are of sufficient size to document rare outcomes. Virtually all health care utilization by the covered population is identified in these files.

Third, individual identification numbers allow records to be linked across time and providers. Such linkage is essential to longitudinal studies of health care outcomes and utilization. Finally, the individual identification numbers provide a mechanism to link Medicare data to other sources of data. Potential sources of supplemental data include those reported here, existing clinical data bases, and medical records. It is also feasible to obtain names and addresses so that individuals could be surveyed to ascertain outcomes not recorded in either the claims themselves or the medical records, such as functional status and quality of life.

Limitations

As with any source of data, limitations in Medicare data must be acknowledged and, when possible, overcome. Treatments and diagnoses in the claims files are recorded in nonresearch settings using International Classification of Diseases (ICD-9-CM) codes (hospitals) and Common Procedural

Terminology (CPT-4) codes (physicians). The precision of the codes themselves and the accuracy with which they are recorded limit the kind of studies that may be successfully undertaken. Major surgical procedures have been found to be accurately coded, and the precision of these codes allows reasonable cohorts to be defined. In contrast, fine distinctions among different subgroups of patients with medical conditions are poorly documented within the existing coding conventions; it would be difficult, for example, to define a cohort of patients with unstable angina. Similarly, the records do not document either the timing of the onset of medical conditions within a hospitalization or the affected side (left vs. right) for procedures or conditions that may affect either side of the body.

Codes for many new technologies and treatments are rarely introduced in a timely fashion. Specific codes for coronary angioplasty were introduced several years after the widespread adoption of the technique in practice. Finally, the scope of data recorded is limited, and utilization rather than the incidence of a medical event is recorded. Some patients with adverse outcomes may not bother or be able to afford to see their physicians. Certain events (mortality, reoperation) can be accurately measured, but other variables (clinical risk factors, functional status, quality of life) cannot be ascertained directly from the claims data.

Suggestions

These limitations suggest several steps we could take to enhance the value of administrative data bases for health care research and outcomes assessment. First, we should improve the completeness and accuracy of the coding used in claims data bases. Establishing codes for new technologies as soon as they become eligible for reimbursement would markedly enhance assessment efforts. Documentation and publication of the accuracy of coding in administrative data bases by the agency responsible for collecting the data would enhance the utility of the data bases to all users.

Second, because the scope of the data is limited, additional data will be required for many analyses. We should be cautious in our strategies for supplementing data, however. There is a tension between the desire to collect all possibly relevant data on each patient and the needs of a given assessment. For example, the specific variables required to study angioplasty or prostatectomy are not likely to be included in even the most comprehensive data set. Consequently, we should determine efficient, flexible means of supplementing the data base. These might include not only facilitating access to medical records to supplement claims data, but also developing strategies for routine posttreatment interviews to determine functional status and quality of life.

References

1. Roper, W.L., Winkenwerder, W., Hackbarth, G.M., et al. Effectiveness in Health Care: An Initiative to Evaluate and Improve Medical Practice. *New England Journal of Medicine* 319:1197-1202, 1988.
2. Wennberg, J.E. and Gittelsohn, A.M. Small-Area Variations in Health Care Delivery. *Science* 183:1102-1108, 1973.
3. Wennberg, J.E. and Gittelsohn, A.M. Variations in Medical Care Among Small Areas. *Scientific American* 246:120-134, 1982.
4. Wennberg, J.E., Roos, N.P., Sola, L., et al. Use of Claims Data Systems to Evaluate Health Care Outcomes: Mortality and Reoperation Following Prostatectomy. *Journal of the American Medical Association* 257:933-936, 1987.
5. Barry, M.J., Mulley, A.G., Fowler, F.J., et al. Watchful Waiting vs. Immediate Transurethral Resection for Symptomatic Prostatism: The Importance of Patients' Preferences. *Journal of the American Medical Association* 259:3010-3017, 1988.
6. Fowler, F.J., Wennberg, J.E., Timothy, R.P., et al. Symptom Status and Quality of Life Following Prostatectomy. *Journal of the American Medical Association* 259:3018-3022, 1988.
7. Wennberg, J.E., Mulley, A.G., Hanley, D., et al. An Evaluation of Prostatectomy for Benign Urinary Tract Obstruction: Geographic Variations and the Assessment of Medical Care Outcomes. *Journal of the American Medical Association* 259:3027-3030, 1988.
8. Roos, N.P., Wennberg, J.E., Malenka, D.J., et al. Mortality and Reoperation After Open and Transurethral Resection of the Prostate for Benign Prostatic Hyperplasia. *New England Journal of Medicine* 320:1120-1124, 1989.
9. Malenka, D.J., Roos, N.P., Fisher, E.S., et al. Further Study of the Increased Mortality Following Transurethral Prostatectomy. *Urology*. In press.
10. Lave, J., Dobson, A., and Walton, C. The Potential Use of Health Care Financing Administration Data Sets for Health Care Services Research. *Health Care Financing Review* 5:93-98, 1983.
11. Hatten, J. Medicare's Common Denominator: The Covered Population. *Health Care Financing Review* 2:53-63, 1980.
12. Chassin, M.R., Brook, R.H., Park, R.E., et al. Variations in the Use of Medical and Surgical Practices by the Medicare Population. *New England Journal of Medicine* 314:285-290, 1986.
13. McPherson, K., Wennberg, J.E., Hovind, O.B., et al. Small-Area Variation in the Use of Common Surgical Procedures: An International Comparison of New England, England and Norway. *New England Journal of Medicine* 307:1310-1314, 1982.
14. Wennberg, J.E. Dealing with Medical Practice Variations: A Proposal for Action. *Health Affairs* 3:6-33, 1984.
15. Paul-Shaheen, P., Clark, J.D., and Williams, D. Small Area Analysis, A Review and Analysis of the North American Literature. *Journal of Health Politics, Policy and Law* 12:741-809, 1987.
16. Grayhack, J.T. and Sadlowski, R.W. Results of Surgical Treatment of Benign Prostatic Hyperplasia. Pp. 125-134 In *Benign Prostatic Hyperplasia.* Grayhack,

J.T., Wilson, J.D. and Scherbenske, M.J., eds. DHEW Publication No. (NIH) 76-1113. Washington, DC: Government Printing Office, Washington, D.C., 1976.

17. Roos, L.L., Cageorge, S.M., Austen, E., et al. Using Computers to Identify Complications After Surgery. *American Journal of Public Health* 75:1288-1295, 1985.
18. Greenfield, S., Aronow, H.U., Elashoff, R.M., et al. Flaws in Mortality Data: The Hazards of Ignoring Comorbid Disease. *Journal of the American Medical Association* 260:2253-2255, 1988.
19. Jencks, S.F., Williams, D.K., and Kay, T.L. Assessing Hospital-Associated Deaths from Discharge Data: The Role of Length of Stay and Comorbidities. *Journal of the American Medical Association* 260:2240-2246, 1988.
20. Charlson, M.E., Pompei, P., Ales, K.L., et al. A New Method for Classifying Prognostic Comorbidity in Longitudinal Studies: Development and Validation. *Journal of Chronic Diseases* 40:373-383, 1987.
21. Stanley, K.E. Prognostic Factors for Survival in Patients with Inoperable Lung Cancer. *Journal of the National Cancer Institute* 65:25-33, 1980.

14

Issues in the Use of Large Data Bases for Effectiveness Research

Stephen F. Jencks

We have an enormous opportunity to move forward with outcomes analysis, particularly outcomes analysis based on claims and other large data sets. At the same time, I think there is a real risk of promising more than these approaches can deliver and compromising the future of this research.

DEFINING LARGE DATA SETS

I begin by explaining what a large data set is because I think the term has been too narrowly construed at times. Certainly, size is a feature of a large data set, but two other characteristics may be more important.

Population Base

First, a large data set usually contains, in some sense, data for a population or a random sample of a population. This can mean payer administrative data, such as claims data from the Medicare program (and thus all the data about services available from that source). It can also mean:

- All hospitalizations occurring in a state. A number of states have all-payer data bases, and some of these are developing considerable clinical richness.
- All persons with a given disease in certain geographic areas. An example is the SEER (Surveillance, Epidemiology, and End Results) data bases maintained by the National Cancer Institute.
- All persons born or dying in a state. State vital record systems are rich with data, and the National Mortality Registry provides an index for those state death records.
- All persons in a random sample. Examples are the National Medical

Expenditure Survey, the National Long-Term Care Survey, and the Uniform Clinical Data Set of the Health Care Financing Administration (HCFA) (which will include a random sample of Medicare discharges).

- Various complex populations, such as Medicaid data sets, where people wander in and out of eligibility in complicated ways but nevertheless comprise a population.

Original Purpose

Large data bases are typically collected for some purpose other than that for which researchers wish to use them, and they often lack, therefore, some features or data researchers want. They are typically strong on size, on longitudinal detail, and on linkability to other data sets, but they are particularly likely to be thin on clinical detail and on functional status. Large data sets, then, tend to be unbiased pictures of patients and practice in the real world, but they rarely have just what researchers want and they are not randomized for treatment. In large data bases, many descriptors of health events are fairly good. There are errors in assigning codes to health events, but overall the data are highly usable. The researchers can find surgeries, hospitalizations, office visits, and many kinds of health events, such as myocardial infarctions.

Outcomes data tend to be pretty good for certain outcomes and not so good for others. It is usually possible to get some information about outcomes other than death, costs, and resource utilization from billing data. These include:

- Morbid events, such as rehospitalization, extended stay, and complications that are indicated by diagnoses and procedures;
- Kinds of service utilization that indicate health status;
- Information on nursing home tenure (Medicare data, for example, include not only bills for care in skilled nursing facilities, but also physician bills, which indicate, by the location of service code, that the patient was in a nursing home); and
- Causes of death (these data from death registries can be hard to obtain and difficult to link, but investigators at HCFA's Office of Research and elsewhere have succeeded in doing so).

Risk adjustors tend to be weak. Although previous diagnoses and use of services are available, and multiple concurrent diagnoses are available for hospital care, physiological risk adjustors are rarely available. On the other hand, there are data sets emerging, such as that being created in Pennsylvania, with substantial physiological data for inpatients, and there will be HCFA's Uniform Clinical Data Set (1).

Functional status data are almost unobtainable in large data sets. There

has been intensive discussion about whether one can define a functional status instrument that should be collected on every patient, but this issue has not been resolved.

USES OF LARGE DATA BASES

What does one use a large data set for? The Office of Research at HCFA has a triple agenda in the area of effectiveness, namely, to look at the closely linked issues of the comparative effectiveness of providers, of procedures, and of payment systems. Large data sets have a variety of applications in these areas.

Sampling Frames

Large data bases are valuable as sampling frames for more intensive studies. The Office of Research and Demonstrations used the Medicare hospital discharge file in this way when it developed the Medicare mortality predictor system. We chose discharges from the discharge file and then went back and pulled those records and got supplementary information in order to develop risk adjustment tools.

Rates and Outcomes Surveillance

Elliot Fisher and John Wennberg have described an example of how informative this kind of surveillance can be (2). The Office of Research is using Medicare data to study diagnosed complications, rehospitalizations for apparently related conditions, and mortality for eight major surgical procedures. These analyses will be broken down by race, locality, and age. We are not analyzing these data by hospital, both because there are scientific problems involved and because the response of the professional community might well be so hostile as to interfere with effective use of the data. We will, however, be consulting with a number of groups about how to make the data more useful.

Variations in Outcomes

There are three issues in variation of outcomes:

1. the amount of variation in outcomes across providers doing the same procedure;
2. variations in outcomes among different procedures for "similar" patients, such as those described by Fisher and Wennberg for transurethral versus open prostatectomy; and
3. variations in the effectiveness of different providers (for example, comparison of rates of various outcomes).

As we move along this spectrum, methodological problems multiply and our ability to be confident in the conclusions we can draw from large data sets becomes progressively weaker.

Linking Large Data Sets

An important feature of large data sets is that they can often be linked so as to increase information about a patient or an event. Data sets with Social Security numbers on them can be linked to one another, and many data sets have or easily could have Social Security numbers. Linkages can broaden many kinds of research. The following are examples.

- The Office of Research is linking Medicare files to the SEER registry in an effort to increase information about what happens to patients who are diagnosed with cancer (the registries contain information on stage and treatments). This is a powerful way to enrich a smaller data base: although the SEER data base is not small by most definitions, it is small compared to the Medicare data base.
- Katherine Kahn, Robert Brook, Emmett Keeler, and others at The RAND Corporation have been studying the impact of the Medicare Prospective Payment System on quality of care. In that study, the Medicare claims data base has been linked to individual cases selected at random from hospitals in order to provide information on rehospitalization and mortality; this information would otherwise be very expensive, perhaps even unobtainable.

In summary, the range of uses for large data sets is extraordinarily broad. We should be careful not to limit our thinking to analyses of Medicare hospital claims data, which are only a very thin slice of the pie.

LIMITATIONS OF LARGE DATA BASES

Large data sets have obvious and less obvious limitations.

Data Quality

Many of the quality issues in large data sets will be familiar to any investigator who has used such secondary data. One feature of secondary data sets, however, requires special emphasis: unused data tend to be useless. Unless the people who create a large data set make use of an item in a way that provides feedback to those who collect it, the risk is very high that the item will contain so much error as to be unusable. We have found this to be true for items ranging from Social Security number to discharge destination. Thus, careful coordination between creators of data sets and investigators can be critical.

Timeliness

Because they are collected for other purposes, large data sets tend not to be available in a timely fashion. This is a special problem in using them for assessing individual providers because it is hard to get hospitals and physicians interested in data that are basically archival.

Access

There are three kinds of problems in getting at the data: administrative access, processing, and understanding.

- There is a perfectly straightforward way of getting the income data that Barbara McNeil discusses (3) by linking Internal Revenue Service (IRS) data from tax returns using the Social Security numbers. The problem with this elegant solution is that, by law, IRS cannot release the data. To come a little closer to the possible, one can link to employment data in the Social Security Administration files, again using Social Security numbers. That is technically feasible and has been done, but because of privacy rules it can only be done by the people at Social Security, which means they must invest staff effort. That requirement really restricts what researchers can do with that large data set.
- The National Mortality Registry records the fact that a death certificate exists for an individual, but you have to deal with each state's vital records officer to obtain information from the death certificates. That process costs blood, sweat, tears, and money. The problem could be solved by legislation or some other means; such a solution would promote important research.
- The greatest access problems, however, are not getting copies of a data set or getting computer time, despite the costs of spinning 20 to 200 reels of tape. Access includes learning how to use these data well once one has them. The big access problem is understanding the intricacies, flaws, quirks, and limitations of these data. It is knowing that a frequency code for a procedure is generally good but that it is unreliable in Illinois in one year because the carrier counted all of the rejected claims when figuring out how many cases were done. There is a lot of detail that is terribly nit-picky, but ignorance can lead to the wrong conclusions. That kind of mastery is hard to acquire.

METHODOLOGICAL ISSUES

The real controversies in using large data sets are methodological. They focus on whether large data sets can be used to assess the effectiveness of a procedure or a provider or the relative effectiveness of procedures or providers.

The fundamental theorem in using risk-adjusted data to examine effectiveness is that one can infer the relative effectiveness of two treatments from the risk-adjusted difference in outcomes. This requires not only that one know the outcome, but also that one be able to adjust for the risk. Our limited capability for risk adjustment, relative ignorance about how providers select procedures, and ignorance about the interactions between providers and procedures create very serious difficulties when we try to employ this fundamental theorem in the real world.

Risk Adjustment

Our best risk adjustment instruments account for less than 30 percent of variation in mortality among individuals with the same condition, which leaves 70 percent or more to be explained by other factors. We can attribute this 70 percent to "luck" or define it more formally as some combination of:

- things we do not measure about patients,
- things we do not measure about the care we give,
- things we do not know about the care we give, and
- the various mistakes we make in providing care that we do not quantify very well and rarely record.

Accounting for 30 percent of the variance might be sufficient to allow us to apply the fundamental theorem if we were confident that the remaining sources of variation were not different among patients getting different treatments or treated by different providers. But we often do not know how good the adjustors are in terms of the kinds of variation we might see among the treatment groups.

Our risk adjustments are probably not much better than clinical judgment. Expert systems can do a bit better than experts, but not much better. We as clinicians cannot say very accurately which patient will live or die or which patient will be bed-ridden a year after hip surgery. Our instruments are probably weakest for outcomes other than death.

Risk adjustment tools are extremely interesting. I spend a lot of my time working on them, developing them, and assessing them, but I think that they are still not fully developed medical technologies. Indeed, considering the very limited evidence we have for risk adjustment systems as tools for identifying ineffective procedures or ineffective providers, I doubt if the Food and Drug Administration would let them be marketed if they were drugs or medical devices. This analogy is appropriate because these systems are being used in settings where they may have a major impact on the health care system. They may be good, but we do not have sufficient evidence yet, and we ought to be generally cautious.

Treatment Selection

To make inferences from these observational data sets, we must understand something about how treatments are selected, particularly about the unmeasured risk factors that physicians may consider when they select treatments. Such factors would confound an assessment of outcomes.

Provider-Procedure Interaction

The effectiveness of a procedure is inextricably linked to the effectiveness of the provider who performs it. Both may be influenced by the payment system under which the procedure is performed.

Let me give an example of how that interaction might be important. Suppose we had done the recent trial of antiarrhythmic drugs in myocardial infarction using claims-based data; suppose those data were infinitely supplemented so that we had *perfect* risk adjustment. We would, I think, have found that most uses of these drugs occur in more sophisticated and advanced settings, such as teaching hospitals. If those sophisticated and advanced settings generally have better outcomes for their patients, yet patients on antiarrhythmic drugs experienced worse results in those settings, that worse outcome would have been confounded by the general pattern of better results in those settings. Although multivariate techniques may control for this effect, the problem requires further study.

This is not a selection phenomenon resulting from unmeasured variables used by the physician in choosing a treatment for a patient. This selection phenomenon involves interaction between the competence of the people who perform a procedure and the effectiveness of the procedure.

Statistical Issues

Without becoming highly technical, I wish to note two statistical issues that are important in using large data sets. One relates to evaluating providers, the other to evaluating procedures.

Multiple Hypotheses

If one uses large data sets to examine outcomes for individual providers, the sheer number of providers and tests can create problems of interpretation. The Medicare Hospital Mortality Information release, for example, examines about 20 categories in more than 5,000 hospitals, a total of about 100,000 outcomes. Although HCFA's Health Standards and Quality Bureau has taken a number of steps to deal with evaluating so many results, such as

publishing three years' data and using sophisticated statistical techniques, the best way to take advantage of these data remains unclear.

Processes in Control

If one wants to assess a procedure, especially if one wants to compare two procedures, it is necessary to have a process that is in statistical control. This means that the variation in outcomes is highly predictable and is distributed in a statistically predictable fashion. Available evidence suggests that, in routine practice, procedures are not in such control and that, for example, outcomes are different for different providers.

RANDOMIZED CONTROLLED TRIALS VERSUS ANALYSIS OF LARGE DATA SETS

Few people think that randomized controlled trials (RCTs) alone or analyses of large data sets alone are sufficient to meet research needs.

What we really need to know is how large data sets can complement RCTs. It is this complementary role, in which large data sets are used to extend and replace RCTs, that we must pursue by analyzing large data sets and comparing the results to those from RCTs. Inferring the relative effectiveness of procedures from large data sets alone is risky at our present level of understanding.

It is important to realize that, although the results of many studies with large data sets will not be definitive, the data that clinicians are working with at the moment are not definitive either. Analysis of large data sets can add probabilistic data to RCTs, thus bringing clinicians closer, in a Bayesian mode, to smart clinical choices. From that point of view, what can be done with large data sets is exceedingly important.

Data almost never speak with great clarity. There is almost always a substantial confidence interval around the results, a lack of certainty as to whether the investigators really did the study exactly correctly. There are conflicting data from other studies. There is a constant problem in evaluating immediate clinical evidence, whether the decision rules to be applied to that evidence come from RCTs or large data sets. Large data bases introduce more problems in evaluating data, but these problems arise in evaluating data that would not otherwise be available to clinicians at all.

PROSPECTS

What can we clearly use large data sets for now, and how might we be expand those uses in the future?

First, these sets are clearly very good as sampling frames.

Second, if one either supplements the large data sets or uses large data sets to supplement other data sets, one can obtain very powerful information about risk, physiology, and disease process.

Third, one can learn from relatively crude outcome rates. I think that John Wennberg, Elliot Fisher, and Noralou and Leslie Roos have really done a signal service here. The outcomes we are going to be publishing for a variety of surgical procedures will follow the direction they have set. Their argument is that the rates of not-so-good events or bad outcomes are important because those rates are much higher than the literature suggests and much higher than physicians and patients believe. They argue that better understanding of the real risks of procedures would lead to more conservative and better practice, and their argument seems very reasonable to me. Therefore, these large data sets have immediate practical importance.

Fourth, large data sets are useful for looking at certain adverse events that we cannot study in other ways. Consider Wayne Ray's study, which showed an association of hip fracture with the use of various psychotropic drugs in the elderly. Given the strong suspicion that a lot of that use is inappropriate, we could not ethically mount an RCT to examine this relationship. So, we have to look to large data sets for such evidence. There are many other kinds of data about practices and procedures that can only be obtained from large data sets.

AN AGENDA

Let me try briefly to set out an agenda.

First, we need to validate the input, that is, the diagnostic and other data that are in these large data sets. We have some information about validity, but it is lying around in funny places, and we need to bring it together. Investigators need access to the results of administrative examinations of the diagnoses recorded in the Medicare data, but further validation is also needed: we need to know how well procedures and complications are recorded. Although we may reasonably infer that a patient admitted with a hip infection after a total hip replacement has developed that infection as a result of surgery, it is much more speculative to link other subsequent events or nonevents to procedures.

Second, we need to increase access to large data sets. This includes creating data centers, changing rules in some cases, and making the data sets easier to understand and use.

Third, we need to do a lot of linking of various kinds of data sets. The HCFA Office of Research has been experimenting with SEER, has worked with the Social Security Administration, and has done a bit with mortality registries. We have to think more carefully about this. If there is some

linkage that would really improve health research and that linkage requires a change in the law, let us find some way of preserving the confidentiality of the data and go to Congress to ask for a change in the law.

Fourth is the creation of new public data sets. I am very ambivalent about proposing this because I fear that unmeetable promises are being made to promote some state data bases. Nevertheless, I think the data sets that are being created in Pennsylvania, Colorado, and Iowa are extremely interesting sources of information. Investigators should be thinking now and talking to state people now about how to use them. I will give you an example of the importance of this thinking and talking. Pennsylvania is collecting the entire MedisGroups data set of more than 200 items, but current plans are, I understand, to provide access only to a summary score. Earlier communication might have made it easier to change the situation so researchers would have access to that entire data set. The Uniform Clinical Data Set is an even more interesting and flexible source of data.

Fifth is the creation of public reference data sets that have been carefully validated. If researchers are going to try to determine the functional status of people after surgical procedures, there is a lot to be said, for example, for selecting certain centers, whether randomly assigned or recruited, from which these data will be collected and in which special efforts will be made to guarantee data quality.

Sixth, we need to learn much more about risk adjustment, and I do not mean just better instruments. For example, there is some evidence that one can do fairly accurate risk adjustment for routine elective surgery from diagnosis and previous treatment data. We need to know how true that is and when it is sufficient. We also need to know when the variations across providers will be adequately measured by the risk adjustment tools we have and when there are major variations that those tools cannot get to.

Seventh, we need to look at *when* risk adjustment can help us to identify the relative effectiveness of providers or procedures. Two examples follow.

- HCFA is presently designing a study to determine when risk-adjusted mortality can be used to screen for cases where peer review will find problems with care. There are many other areas in which the validity of large data bases must be determined before research on them can be validated. Specifically, we need to know when risk adjustment can replace randomization in evaluating either a procedure or the relative effectiveness of two procedures.
- A possible validation study would be to expand a clinical trial by asking the physician to record, before opening the assignment envelope, the treatment he or she would have selected had there not been a randomization process. With such a design one can ask, "How well would risk adjustment have been able to correct for the selection bias that physician judgment would have introduced in a retrospective, risk-adjusted study?" There are probably a lot of other useful approaches, and the Institute of Medicine

might make helpful suggestions in this area. What is really needed is empirical evidence, not people saying, "This isn't randomized, therefore it isn't truth," and not people saying, "We have controlled for the relative risk, so it is true."

Finally, we need to develop some consensus on how large data sets can be used. This problem extends beyond how studies using these data should be carried out and exactly what should be done in the studies. For example, we need some consensus on how the HCFA mortality data can be used. Major health organizations are beginning to work toward such a consensus, and developing that consensus may be an important step in working toward consensus on how to use data from other large data bases.

Cutting across all these issues is the challenge of doing as much as we can without promising more than we can deliver. I hope the issues discussed in this chapter will move us forward in the narrow but important path between the risks of promising too much and attempting too little.

References

1. Krakauer, H. The Uniform Clinical Data Set. Pp. 120-133 in *Effectiveness and Outcomes in Health Care*. Heithoff, K.A. and Lohr, K.N., eds. Washington, D.C.: National Academy Press, 1990.
2. Fisher, E.S. and Wennberg, J.E. Administrative Data in Effectiveness Studies: The Prostatectomy Assessment. Pp. 80-93 in *Effectiveness and Outcomes in Health Care*. Heithoff, K.A. and Lohr, K.N., eds. Washington, D.C.: National Academy Press, 1990.
3. McNeil, B.J. Claims Data and Effectiveness: Acute Myocardial Infarction and Other Examples. Pp. 65-70 in *Effectiveness and Outcomes in Health Care*. Heithoff, K.A. and Lohr, K.N., eds. Washington, D.C.: National Academy Press, 1990.

Collection of Primary Data: Introduction

Harold C. Sox, Session Moderator

In the context of the Effectiveness Initiative, primary data are those obtained from sources other than administrative claims data sets. Thus, the purpose of primary data collection is to supplement the information that is obtained from the administrative data sets. There are several reasons, which were discussed earlier in this volume, why claims-based data are often not adequate for medical research.

- To attribute an improved outcome to an intervention, patients who had the intervention should have been identical prior to the intervention to patients who did not have the intervention. There are multivariate statistical methods for adjusting for baseline differences between the intervention group and those who did not have the intervention. Administrative data sets typically do not have sufficiently detailed clinical information for this purpose. This information can sometimes be obtained by reviewing the patients' hospital records.
- Studying an intervention in a subset of patients may reveal effects that are not observed in the entire population. To create useful subsets of patients, one must have clinical information that is often not available in administrative data sets.
- The range of outcomes that can be measured with administrative data sets is limited. Administrative data sets have information about whether a patient is alive or dead, as well as whether the patient was rehospitalized or required an intervention. Information about disease status, functional status, or the patient's preferences must usually be obtained by other means, such as reviewing the patient's hospital record or interviewing the patient.

John E. Ware is a senior scientist at the Institute for Improvement of Health and Medical Care at the New England Medical Center. His chapter

focuses on gathering data directly from patients and emphasizes practical issues in primary data collection, as well as issues of precision, reliability, and validity.

Henry Krakauer is Director of the Office of Program Assessment and Information in the Health Standards and Quality Bureau (HSQB) of the Health Care Financing Administration (HCFA). In 1987, HSQB began a complex project to develop a data set for use by Medicare Peer Review Organizations (PROs) and the wider research community. The data set was intended to contain far more detailed clinical data than were available heretofore in the HCFA data files. Dr. Krakauer discusses the part of this project known as the Uniform Clinical Data Set.

15

Measuring Patient Function and Well-Being: Some Lessons from the Medical Outcomes Study

John E. Ware, Jr.

Among the important developments in the health care field during the past decade is the recognition that the patient's point of view, in monitoring the quality of medical care outcomes, is central. Indeed, the goal of medical care today for most patients is the achievement of a more "effective" life (1) and the preservation of function and well-being (2,3,4,5). The patient is the best source on the achievement of these goals. However, information about patients' experiences of disease and treatment is not routinely collected in clinical research or medical practice. This information is not part of the medical record and consequently is not typically available for analysis in the current health care data base.

We are entering a new era in which information from patients about functional status, well-being, and other important health care concepts will be added to the health care data base. Included are data bases used to compare costs and benefits of various financial and organizational aspects of health care services, by organizational managers who try to provide the best value for health care dollars, by clinical investigators who evaluate new treatments and technologies, and by practicing physicians and other providers who try to achieve the best possible outcomes for their patients.

The primary source of this information will be from standardized patient surveys that have served research well over the past decade. The most efficient way to monitor functional status and well-being for most adults is via scoring of carefully constructed sets of survey questions. Advances in assessment and measurement, particularly in terms of surveys of patient perspectives, have facilitated this kind of data collection (see, for example, 6 and 7), although their use on a large scale has not been practical.

It is clear that the field of health care needs more cost-effective ways to obtain new data about patient outcomes. The methods must be practical and

they must satisfy the most crucial psychometric standards. The trade-off between practical considerations and psychometric standards has led to a rethinking of measurement strategy. Better measurement is measurement that has information one absolutely has to have, and no more. I am going to emphasize practical issues as much as precision and reliability and validity, and I plan to do so without using any numbers whatsoever. Numbers provide some form of authority, but they also can be restrictive.

CONCEPTS AND DATA SOURCES

Health care providers collect data about functioning for virtually every body organ, but none of these measures tells about the function of the entire individual—which is certainly affected by disease and treatment (see Figure 1). Further, these measures of biologic phenomena cannot be used to characterize human phenomena. There simply are not good algorithms for combining diverse biologic information to predict functioning, and such algorithms are doomed to leave too much about quality of life unexplained. The most comprehensive models I have seen to date might explain 10-25 percent or so of the reliable variance in, for example, physical functioning. Thus, biologic indicators are not adequate proxies for measures of functional status or well-being or to changes in these variables over time.

Biologic indicators must be supplemented if we are to use outcome data to achieve the goal of providing the best value for the health care dollar. Specifically, we must consider how individuals experience disease as well as treatment. The current data base also has information about death. However, to quote Jack Elinson, formerly of the National Center for Health

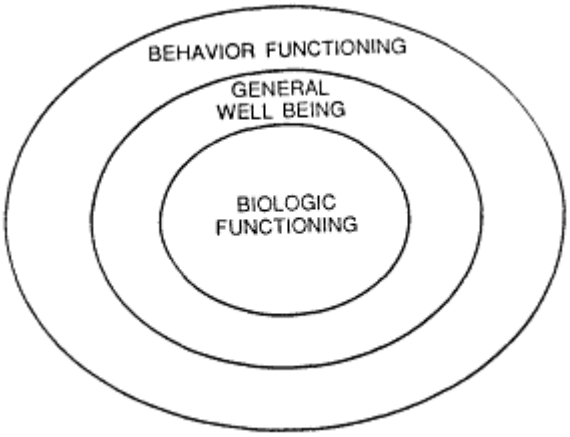

Figure 1
Health Status Concepts

Statistics, there isn't very much information about the health of a population from mortality data in a developed country. Consider heart surgery: for patients with heart disease the mortality rates are approximately 5 percent or less. Thus, for nearly all patients, that particular indicator provides no information about variations in outcomes.

Essentially, we need a new data base. In addition to some other things, the new data base should add two types of information: patients' experience of *health care*, and the patients' experience of *health outcomes*. Our task is to find ways to incorporate this information into the total health care equation. In this regard, I believe that the effectiveness initiative (8) involves more than going from efficacy to effectiveness. It really involves going from one relatively limited set of variables that has been used in judging efficacy to a completely different set of variables not traditionally used to evaluate alternate treatments and technologies. We have not routinely assessed the effect of treatment on quality of life, or functioning, or well-being from the patients' point of view.

DISEASE-SPECIFIC VERSUS GENERIC MEASURES

Before proceeding, let me define what I mean by generic measures. They measure concepts that are relevant to everyone. They are not specific to any age, disease, or treatment group. Generic measures focus on such basic human values as emotional well-being and the ability to function in everyday life.

Should we use disease-specific or generic measures? The overwhelming answer should be to use both and to use them together. We should not reject one data base in favor of the other. We went through a period in the mid-1960s during which the validity of a patient rating a generic health concept was questioned when it did not agree with what was in the record or with what the provider said. The logic of validity has since been turned around. We are now entering an era in which the same findings are accepted as evidence for the necessity of including patient assessments as part of the evaluation process. The record and provider judgments are not valid proxies for patient ratings of functioning, well-being, or other aspects of the quality of life.

We should not always expect assessments of different health components or clinical versus generic measures to agree, and often they do not. One example comes from a study of the effects of antihypertensive therapy on quality of life (9). Therapies shown to be equally efficacious in terms of medical efficacy (i.e., blood pressure control) had significantly different quality of life profiles. In other words, it is possible to work with a patient in therapy to achieve a better quality of life outcome without compromising biologic function. There is also evidence accumulating that shows that

differences in biologic function often have quality of life implications. These two concepts are distinct; they are affected by different processes and they interact with each other. To understand patient health outcomes, different health components need to be measured and interpreted separately and in combination (the latter when trade-offs are involved).

The greatest progress is going to occur, not by substituting one measurement or assessment strategy for another, but by mastering them in concert. We should not underestimate the power of a data base that includes clinical measurements familiar to medical providers, measurements that they believe in because they have clinical validity, in parallel with other measures, such as measures of generic health concepts, not typically linked with such measures in clinical practice or research. This is the most powerful strategy for analyzing and understanding outcomes and for diffusing recent advances in methods for assessing patient outcomes.

A MINIMUM SET OF GENERIC HEALTH CONCEPTS

I would like to take this opportunity to recommend what a minimum set of generic health concepts might look like. At the risk of oversimplifying the past 40 years of health assessment research, I think most health measures can be classified into one of three major categories: functional status, well-being, and general health perceptions. I have defined these categories elsewhere and have illustrated them with sample questionnaire items from widely used measures (10). Functional status, which includes disability assessment, refers to behavioral dysfunctions due to health problems. It is the concrete, observable, tangible, and objective category of health measures. Measures in this category use a standard external to the individual, such as usual role activity, walking at a certain rate, or customary self-care behaviors. This is the functional status axis in a multidimensional conceptualization of health. It is the concept that has been preferred and best understood until now. There are a number of well-developed measures of functioning available.

Interestingly, almost completely orthogonal to the functional status axis is the well-being axis, which includes psychological distress, psychological well-being, and life satisfaction. In most populations we observe all levels of each of these axes at all levels of the other. In fact, in most populations, the correlation between them is only 0.20 or less (assuming confounding of measures across axes has been removed). The implication is that we cannot know how people feel by observing what they are doing. Consider two people sitting on a fencepost, for example. One may be experiencing a lot of pain and may have difficulty just sitting there. The other person may sit in ecstasy. In order to know, we have to ask them. It used to be thought that well-being could not be measured reliably. We have learned that quite reliable scores for this continuum can be obtained and that they add a

completely different perspective to that gained from functional status assessment.

Finally, there is a third axis, which cuts across the other two and brings still another perspective beyond both of the other two axes; that is the category of general health perceptions. It includes measures that are personal evaluations of health, based on whatever health means to the respondent. This category of measures brings each person's own health values to the equation. He might be a mental-health-oriented person or a physical-health-oriented person. Health perceptions represent the third axis or category that I would recommend for inclusion as a minimum standard for generic health measures.

THE MEDICAL OUTCOMES STUDY

A hallmark of the Medical Outcomes Study (MOS) is its reliance on a broad array of outcome measures, including parallel assessments of disease-specific clinical endpoints traditionally measured by clinicians (biologic functioning in Figure 1) as well as generic measures of functional status, well-being, and satisfaction with health care as reported by patients. This more encompassing assessment of outcome increases the likelihood of detecting the consequences to patients of policies that modify the structure of the health care system or the process of care. Measuring disease-specific end points, as well as a common set of generic health outcomes for various conditions, will also contribute a new data base that will allow physicians to inform patients about the trade-offs involved in different treatment.

I am focusing here on health outcomes. Patients should also be involved in assessing the quality of the medical care process (11). I am going to give you a brief summary of some of our experiences to date in the MOS (12). One of our intentions in the MOS was to test the feasibility of implementing the same primary data collection system in very diverse systems of care for purposes of monitoring the results of that care over time. By design, we included very different health care settings and very different patient populations.

We sampled different health care settings to vary the structure and process, and we are measuring variations in the outcomes of care. Structural features of care include, for example, whether the provider is an HMO, an insurance plan, a subspecialist, or a more generally trained physician. These traditionally stable attributes of the health care system are now among the many tools of cost containment. People are experimenting with such structures in efforts to reduce medical expenditures. In the MOS we are looking at how structural differences affect the process of care in the two major categories, *technical process* and *interpersonal process* (12).

Sponsors of the MOS are undoubtedly interested in whether the expenditure

for the study, which is approaching $12 million, can also be justified in terms of addressing whether different ways of organizing and financing health care affect patient outcomes. What is the best way to organize, finance, and deliver care? If you have hypertension, for example, does it matter whether you are treated by a cardiologist or a family practitioner? Does it matter whether depressive disorders are detected and treated (13)?

Again, one of our primary interests has been to work towards advancing the state-of-the-art in outcome assessment methods. One lesson we have learned is that it is feasible to create the new data base defined above and to add to it routinely on a rather large scale in very diverse health care settings. MOS analyses in progress will be quite informative about the more cost-effective ways of creating such a database and which variables are most important for what kinds of analyses. Before commenting on some of the MOS lessons to date, let me discuss briefly some study design features. Additional details are given elsewhere (12, and in some references cited there).

The study was done in three sites. At each site we sampled physicians and patients from three different kinds of organizations: traditional prepaid group practice form of health maintenance organizations, multispecialty groups, and solo practices. From the latter two we sampled both fee-for-service and prepaid patients treated by the same physicians. The result is five "systems of care" that differ in organization and financing. We sampled 523 physicians trained in family practice, general internal medicine, endocrinology, cardiology, and psychiatry. Other mental health providers were also sampled (13). These are the specialties that treat the MOS tracer conditions: hypertension, diabetes, heart disease, and depressive disorders. The conditions were chosen primarily because they are prevalent, costly, treatable with variations in practice style, and have an impact on the outcomes of interest.

We looked at adult patients who were seen during a nine-day period in these physicians' offices. We gathered screening data from both the patient and the doctor at that time. We then took approximately a 10 percent random subsample of those patients who had one or more of our chronic tracer conditions. As of October 1988 we had followed these patients for over two years. We hope to continue to follow them. We look at the care they receive and we monitor transitions in their clinical status as well as functional status and well-being, the latter at six-month intervals.

FEASIBILITY AND COST

Much of the expense of the MOS was attributable to the cost of identifying and recruiting providers and patients, designing instruments and data collection methods, and making sure that they would work. We spent nearly two

thirds of the study's total funds before the panel started. Measuring and analyzing patient outcomes over time has not been the major expense.

Clearly what one concludes about what things cost depends upon what is charged to a given cost category. If primary data collection is considered a marginal cost in a framework for monitoring patient outcomes, the marginal cost would be relatively small. If other costs are charged to this category, including for example, defining what diabetes is, determining whether somebody has it, determining how to sample doctors and patients, dealing with differences in patient case mix, the cost per patient followed is much higher. Again, once we had identified both patients and providers and had measured differences in patient case mix, the cost of following patients over time was *relatively* small. With state-of-the-art short forms and processing methods, the cost to process a patient health assessment in a doctor's office is less than the least expensive lab test.

There has been at least one other lesson about feasibility. We oversampled Medicare patients because of the policy relevance of that group. We hypothesized that they would be treated differently in different practice settings as a result of oversampling Medicare. The median age in our longitudinal sample was about 60. All of them had one or more chronic conditions. This population is very sick relative to, for example, the population we followed in the Health Insurance Experiment where we also used self-administered questionnaires as a primary data collection tool in comparing outcomes across different systems of care (14).

How well did these methods work in the MOS? When we conducted our two-year follow-up survey two years after enrollment, we again used a self-administered survey, a booklet with about 250 questionnaire items. Our response rate was over 80 percent for those who self-administered the full-length questionnaire. We used telephone interviews and the MOS short form for those who did not and raised the overall response rate to over 90 percent of those who were contacted and still alive. Whereas dollars can be saved early on by using self-administration, you must be willing to spend some of these savings for follow-up (e.g., by telephone) for people who do not complete a self-administered form. Nearly 70 percent of the panel has completed all surveys during the course of the study. The typical completed questionnaire has 1 percent or fewer of the items missing. Thus, it is possible to get high completion rates for long questionnaires even in a relatively elderly and relatively sick population. This experience has made me more enthusiastic than I had been after the Health Insurance Experiment (and I was enthusiastic then) about the feasibility of standardized surveys and self-administration as a primary data collection strategy for monitoring patient outcomes.

MOS providers were roughly evenly divided between group and solo practice settings. We found that it is more efficient to monitor outcomes in

group practices. Solo practitioners do not have equivalent support personnel. Thus our sampling rates were higher in groups (12). Our completion rates were roughly the same across practice settings once people had enrolled in the study. Thus, with centralized data collection and standardized forms and methods, the completeness and quality of the database need not vary by practice setting. This leads to the notion of centralized health assessment laboratories. With support from the John A. Hartford Foundation of New York, we are now developing and testing this concept.

Again, the MOS is a methodological study, and with support from the Henry J. Kaiser Family Foundation, my colleagues and I are now comparing the briefest forms of measurement, e.g., the best single-item measure, with longer but still short multi-item scales, and with full-length research versions of these scales. Our question is how well do shorter measures work relative to much longer measures used in research? Not surprisingly, preliminary findings indicate that longer measures do better. However, the question should be: "Do shorter measures do well enough?" The answer is very important because the most psychometrically elegant instrument is useless if it is impractical to use. Thus, we should be very interested in how briefer measures do in these comparisons.

STANDARDS FOR EVALUATING MEASURES

On what basis should measures be compared and how do we construct them in the first place? A number of things are wrong with what we traditionally do in psychometrics. Take reliability as an example. Reliability is important, but we learned quickly that although reliability is a prerequisite, other attributes of a score (scale) are equally or more important. In comparing scales, most important are tests that most closely approximate the intended use of the measure. Unfortunately, traditional reliability and validity coefficients have little or no relationship to actual applications of measures.

Some attributes of measures typically ignored include things like how many different scores are possible. An enumeration system that puts people into one of four levels or categories with a reliability of 0.90 is not as valuable as is one that puts people into 10 categories with the same reliability. This particular attribute of measurement may not prove critical in a cross-sectional analysis comparing things as different as Volkswagens and trucks. The latter is analogous to comparing disease groups that differ a lot. Most measures do well in that kind of comparison. When we start measuring change in health over time, however, this issue becomes crucial. How much change can occur *within* a given health category before the person changes to the next category? The number of levels of measurement is a very important attribute or measure. This attribute is almost never discussed in books on health assessment.

Another important and related issue is simply how many people get the lowest or the highest possible score for a given measure. If 90 percent of the people in a long-term care facility have the worst possible score before a cost-containment strategy is implemented, you do not have precision for showing a worsening in their condition as a result. If 80 percent of the people earn the highest possible score on a physical health index (as they did in the Health Insurance Experiment), and we randomly assign them to free care, we do not have much chance of determining, using that particular measure, whether there is any benefit of free care. Fortunately, this was not the only measure used in that experiment (15). I suggest that, when measures are published, we routinely report how many people get the lowest and the highest possible scores when measures are published as well as the number of scores possible, in addition to reliability coefficients.

The same logic should apply to results regarding validity. It is extremely important that the kind of analysis used to judge the validity of a measure approximate as closely as possible the intended use of the measure in medical practice, a clinical trial, or a policy study. Much published evidence bears little or no relationship to most intended applications. One example of what I mean is the issue of whether a given questionnaire is sensitive to the extent and nature of differences in functional status and well-being across groups of patients with different chronic conditions. My colleagues and I recently reported an example of such comparisons using the 20-item MOS short-form survey (16,17). Figure 2 presents examples of profiles for patients with four different chronic conditions at a point in time when the study began.

Each profile is expressed as standard score deviations from the averages for well patients (represented by the horizontal dotted line). The first three data points (columns) for each disease are defined by functional status scales, the last three by well-being scales. We have connected the points across scales for each disease to help identify a particular disease profile. These are scored so that the lower the profile on the scales the worse the profile.

Figure 2 generally confirms clinical wisdom about the impact of these diseases. Not surprisingly, patients with hypertension (the top profile in Figure 2) function no differently than well patients. The only significant decrement was their score for health perceptions. Patients with hypertension tend to believe their health is worse. Arthritis has the most pain. Physical, role, and social functioning is poor for survivors of myocardial infarction (MI) and tends to be as bad as, if not worse than, any of the nine chronic conditions we have studied to date.

One lesson from this is that, on average, the patient point of view is valid. Further, even very brief measures can be used to measure differences in health across groups of patients. The questionnaire used to estimate scores in Figure 2 was administered to about 12,000 patients while waiting in a doctor's office, in about three and a half minutes each.

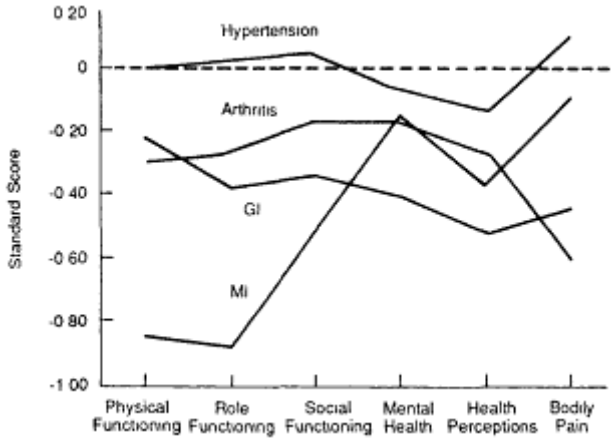

Figure 2
Health Profiles for Patients with Four Conditions. Dotted line indicates patients with no chronic conditions; GI, gastrointestinal disorder; MI, myocardial infarction.

Relative to available full-length research instruments, including our own, this short-form has one-fifth to one-tenth the number of questionnaire items. Yet it produced a pattern of results that make sense from a clinical point of view. One surprising finding (not shown in Figure 2) is the very low profile of scores for patients with depression, including those with a psychiatric diagnosis and those suffering with symptoms of depression. They scored very low on these scales relative to other chronic diseases, suggesting that the burden of depression may have been underestimated to date (13). [1]

Of course, other kinds of tests are necessary before conclusions are drawn about candidate measures. How well does a questionnaire distinguish differences in functional status and well-being across groups differing in severity within a diagnostic category? Research in progress within the MOS is encouraging in this regard. For example, in preliminary analyses of MOS data, average functional status scores for diabetics differing in severity (e.g., with or without renal failure) show ordinal consistency in relation to clinical severity. Thus, these measures may be sensitive to differences in severity within a diagnostic group.

This kind of analysis does not prove that if treatment moved people from severity level five to level three, we would see a corresponding change in functioning. This example is based on a cross-sectional analysis. We are currently linking measures of actual change in disease severity over time with measures of change in functional status and well-being. Again, preliminary results are encouraging.

[1] For another description of the MOS and results pertaining to depression, see Chapter 19 of this volume (20).

USE OF HEALTH SURVEY DATA

The potential of generic functional status and well-being scales, even short forms, to be used successfully across a wide range of purposes is illustrated by the vitality scale used in the MOS, a 4-item scale that takes about one minute to complete. It measures a continuum of energy versus fatigue. Its history, which is documented in part elsewhere (17), includes successful use in a population health survey about 15 years ago (18), the Health Insurance Experiment (17), a more recent clinical trial comparing antihypertensive therapies (9), and the MOS (19). Its track record defies the notion that completely different measures are needed for different applications. This one-minute vitality scale, for example, has been used successfully in describing the young and the old in the U.S. population, the sick and the well, and in measuring outcomes across homogeneous groups of patients receiving different treatments in a randomized trial.

With my recent move to the New England Medical Center, I have had occasion to contrast my own background and training with the needs of the next era of health assessment. I was trained in measurement theory and methods and for the prior 15 years I had been in a full-time research setting where we evaluated measures in traditional psychometric terms and used them for purposes of research. Now I focus much more on the information needs of health care delivery organizations and look at measurement and the use of data from a different perspective. Two years in a research institute in a health care delivery organization have convinced me that a new data base with information about patient outcomes is not the solution to the problem, it is just the next step. The challenge of implementing outcomes management or an effectiveness initiative is not a problem of measurement, and it is certainly not a problem of assessing outcomes from the patients' view. The methodological problems include determining (1) a coherent sampling strategy; (2) techniques for case-mix measurement and statistical control; (3) a meaningful schedule of assessments for different diagnostic groups; (4) analytic strategies for displaying results in a meaningful way; and (5) recognizing that conclusions are sensitive to these and other choices. Finally, the real challenge is the creation of a decision-making process capable of using dam about patient outcomes. Indeed, the collection of outcomes data from patients is one of the simplest steps ahead.

ACKNOWLEDGEMENTS

The MOS has been sponsored by grants from the Henry J. Kaiser Family Foundation, the Robert Wood Johnson Foundation, the John A. Hartford Foundation, the Pew Charitable Trusts, the Agency for Health Care Policy and Research, the National Institute on Aging, and the National Institute of

Mental Health, and by The RAND Corporation and the New England Medical Center from their own research funds.

The author gratefully acknowledges Albert P. Williams, at The RAND Corporation and especially the MOS staff and consultants, including Sharon Arnold, Sandra H. Berry, M. Audrey Burnam, Maureen Carney, Allyson Ross Davis, Sheldon Greenfield, Ron D. Hays, Elizabeth McGlynn, Eugene C. Nelson, Lynn Ordway, Judith Perlman, Edward B. Perrin, William Rogers, Cathy Sherbourne, Anita L. Stewart, Alvin R. Tarlov, Kenneth B. Wells, and Michael Zubkoff, and the secretarial and administrative support of Kathy Clark.

References

1. McDermott, W. Absence of Indicators of the Influence of Physicians on a Society's Health. *American Journal of Medicine* 70:833-843, 1981.
2. Cluff, L.E. Chronic Disease, Function and the Quality of Care. *Journal of Chronic Diseases* 34:299-304, 1981.
3. Tarlov, A.R. Shattuck Lecture. The Increasing Supply of Physicians, the Changing Structure of the Health-Services System, and the Future Practice of Medicine. *New England Journal of Medicine* 308:1235-1241, 1983.
4. Schroeder, S.A. Outcome Assessment 70 Years Later: Are We Ready? *New England Journal of Medicine* 216:160-162, 1987.
5. Ellwood, P.M. Shattuck Lecture. Outcomes Management: A Technology of Patient Experience. *New England Journal of Medicine* 318:1549-1556,1988.
6. Lohr, K.N. and Ware, J.E. Advances in Health Assessment, Special Issue. *Journal of Chronic Diseases* 40:Supplement 1:1S-193S, 1987.
7. Lohr, K.N. Advances in Health Status Assessment: Overview of the Conference. *Medical Care* 27 (3) Supplement:S1-S11, 1989.
8. Roper, W.L., Winkenwerder, W., Hackbarth, G.M., et al. Effectiveness in Health Care: An Initiative to Evaluate and Improve Medical Practice. *New England Journal of Medicine* 319:1197-1202, 1988.
9. Croog, S.H., Levine, S.M., Testa, M.L., et al. The Effects of Antihypertensive Therapy on Quality of Life. *New England Journal of Medicine* 314:1657-1664, 1986.
10. Ware, J.E. Standards for Validating Health Measures: Definition and Context. *Journal of Chronic Diseases* 40:473-480, 1987.
11. Davies, A.R. and Ware, J.E. Involving Consumers in Quality of Care Assessment: Do They Provide Valid Information? *Health Affairs* 7:33-48, 1988.
12. Tarlov, A.R., Ware, J.E., Greenfield, S., et al. The Medical Outcomes Study: An Application of Methods for Monitoring the Results of Medical Care. *Journal of the American Medical Association* 262:925-930, 1989.
13. Wells, K.B., Stewart, A., Hays, R.D., et al. The Functioning and Well-Being of Depressed Patients: Results from the Medical Outcomes Study. *Journal of the American Medical Association* 262:914-919, 1989.
14. Ware, J.E., Brook, R.H., Rogers, W.H., et al. Comparison of Health Outcomes at a Health Maintenance Organization with Those of Fee-for-Service Care. *Lancet* i(8488):1017-1022, 1986.

15. Brook, R.H., Ware, J.E., Rogers, W.R., et al. Does Free Care Improve Adults' Health? Results from a Randomized Controlled Trial. *New England Journal of Medicine* 309:1426-1434, 1983.
16. Stewart, A.L., Greenfield, S., Hays, R.D., et al. Functional Status and Well-Being of Patients with Chronic Conditions: Results from the Medical Outcomes Study. *Journal of the American Medical Association* 262:907-913, 1989.
17. Ware, J.E., Johnston, S.A., Brook, R.H., et al. *Conceptualization and Measurement of Health for Adults in the Health Insurance Study: Volume III, Mental Health*. R-1987/3-HEW. Santa Monica, Calif.: RAND Corp. 1979.
18. Dupuy, H.J. *The Psychological Section of the Current Health and Nutrition Examination Survey*. U.S. Dept. of Health, Education, & Welfare Publication. No. (HRA). 74-1214. Washington, D.C.: Government Printing Office, 1974.
19. Stewart, A.L. and Ware, J.E., eds. *Measuring Functional Status and Well-Being: The Medical Outcomes Study Approach*. Forthcoming.
20. Burnam, M. A. Studying Outcomes for Patients with Depression: Initial Findings from the Medical Outcomes Study. Pp 159-168 in *Effectiveness and Outcomes in Health Care*. Heithoff, K.A. and Lohr, K.N., eds. Washington, D.C.: National Academy Press, 1990.

16

The Uniform Clinical Data Set

Henry Krakauer

To properly assess the Uniform Clinical Data Set, it must be clearly understood that it was designed for a very specific purpose, namely, to meet the operational needs of the Health Care Financing Administration (HCFA) in assuring, through the Peer Review Organizations (PROs), the quality of the care that Medicare beneficiaries receive. It will, therefore, satisfy a limited array of needs for clinical data, but the extent to which it does will have to be determined empirically, that is, through experience with its use.

PROs exert considerable influence on the practice of medicine through the financial and disciplinary actions at their disposal. They are authorized to:

- deny reimbursement for inappropriate admissions,
- deny reimbursement for substandard care,
- initiate sanctions by the Inspector General, and
- correct aberrant patterns of medical care.

The difficulties inherent in these activities may be illustrated by the process of denial of reimbursement for care judged to have been substandard. Table 1 presents a theoretical but reasonable algorithm leading to such a denial. First, it is necessary to demonstrate that the patient suffered harm, that is, an adverse outcome. Beyond that, it is necessary to demonstrate that the adverse outcome was avoidable; that is, it should not have been predictable with a reasonably high level of probability from the condition of the patient at admission. Finally, it is necessary to establish that a breach of protocol occurred, in other words that there was negligence or incompetence, in order to establish culpability.

TABLE 1 Problems in Denying Reimbursement for Substandard Care

SPECIFIC
Denials for substandard care
Harm—death, disability, physiological impairment, increased intensity or duration of care
SCREENING
Avoidable—physiological condition at admission
Objective standards
CULPABILITY
Breach of protocol—negligence, incompetence

THE PRO CASE REVIEW PROCESS

The current process of case review by the PROs begins with a screening of the medical record to identify instances in which the care is suspicious enough to merit further expert attention. A case identified in the screening is then referred to a physician advisor for review. If the system of peer review is not to be perceived as arbitrary and capricious, it is necessary to (a) develop and uniformly apply well-defined screening criteria that identify where there is a reasonable probability of a deficiency and (b) develop and uniformly apply objective standards that would permit the physician reviewer to ascertain with a high level of probability that a deficiency in care did occur.

The best guide in these matters is actual experience. Experience permits one to judge that an act of omission or commission results in harm reasonably often and that it was therefore reasonably likely to have done so in the case in question. Given the realities of medical practice, that experience should be passed through the filter of consensus to make it acceptable. Once this has been accomplished, the devising of consistent screening criteria and objective standards and their uniform application become straightforward.

A strategy for the efficient accumulation and evaluation of experience in the Medicare environment is displayed in Table 2. It consists of a sequential process that begins with assessment of the health of the Medicare beneficiaries and of the time trends and geographic variations therein—two problem-finding tools—and proceeds with the assessment of the effectiveness of interventions, be they medical or administrative, as the problem-solving step. The final step is feedback of the results of the evaluations, coupled with disciplinary activities and financial incentives to ensure their proper and timely use. This approach makes extensive use of observational techniques to evaluate the natural history of conditions as they are currently being

treated and to begin identifying which of the available and competing courses of treatment appear most beneficial. Because the approach is sequential and begins with an assessment of the universe of patients at risk for a condition, every successive step will, while decreasing the number of patients and increasing the detail of the pertinent data (down to the randomized clinical trial), allow the researcher to generalize the findings to the patient universe. In addition, each earlier, broader step informs planning for the subsequent, more specific step in the analytic sequence.

TABLE 2 Strategy for Improvement of the Effectiveness of Medical Interventions for Medicare Beneficiaries

Health Care Financing Administration
1. Monitoring time trends
a. population-based
b. medical interventions (feasible with available billing and census data)
2. Analyzing geographic variations
a. population-based
b. medical interventions (feasible with available billing and census data)
3. Assessing effectiveness of interventions (longitudinal, to develop objective standards)
a. monitoring, as in step 1 above (retrospective, based on billing data)
b. use of data from medical records (retrospective, natural history) Uniform Clinical Data Set (also for case finding for PRO review [medicolegal])
National Center for Health Services Research, HCFA
c. clinical demonstrations (prospective, natural history, especially for emerging technologies)
National Institutes of Health, NCHSR, HCFA
d. randomized, controlled clinical trials
Health Resources and Services Administration, NIH, HCFA
4. Feedback (educational, disciplinary, financial)

THE UNIFORM CLINICAL DATA SET AND PRO NEEDS

The Uniform Clinical Data Set occupies a specific niche in this process. It is a tool for extracting data from medical records to permit effective risk adjustment in assessing the treatment of a given patient.

The composition of the Uniform Clinical Data Set is dictated, as was indicated above, by two requirements. It must enable the PROs to screen cases efficiently and uniformly in order to identify those in which the effectiveness of the care delivered was problematic, and it must enable reviewing physicians to develop objective tools by which to judge the cases. Thus, it

must satisfy an operational need, screening, and simultaneously must support an epidemiological activity, the development of objective standards through evaluation of the effectiveness of medical interventions. The contents of the data set were defined by a task force of clinicians and research personnel who had these two objectives clearly in mind.

Using the Uniform Clinical Data Set

The content of the Uniform Clinical Data Set is illustrated by Tables 3 through 7. These tables are excerpted computer screens, or menus, that prompt the entry of data from the medical record. Table 3 displays the main menu, which lists the major classes of information that are to be extracted from the hospitalization record. The order of the screens follows roughly the order in which the data are likely to be encountered in the medical record. The hardware and software are, however, so quick that the order of abstraction is dictated by the order of appearance of data in the medical record rather than the order presented on the list.

The richness of the data collected is best illustrated with specific examples,

TABLE 3 Excerpt from Uniform Clinical Data Set Computer Screen: Peer Review Screens (Main Menu)

A.	Sociodemographic data
B.	Admission status
C.	Admission medication history
D.	History permanent anatomic changes
E.	History and physical
G.	Laboratory: chemistry/blood gases
H.	Laboratory: hematology/urinalysis
I.	Laboratory: microbiology
J.	Laboratory: cytology/histology
K.	Admission diagnostic tests
L.	Endoscopy
M.	Operative episodes
N.	Treatment interventions
O.	Recovery phase
P.	Discharge status
Q.	Discharge planning
R.	HDI master
	Type letter corresponding to a menu item: __
	F3=Set HDI_ID F10=Leave

such as the screens which address the patient's history and physical examination. These are accessed by striking the letter that labels the "history and physical" menu entry. The subscreens (Table 4, the first of three "history and physical" screens) provide a menu from which more specific subscreens (which refer to organ systems) may be selected and indicate whether data pertaining to the organ system identified have been entered. When data have been entered, the flag "F" (false) next to that item changes to "T" (true), as shown for several entries in Table 4.

At the next (organ) level of data, using the cardiovascular examination as an example (Table 5), the data recorded include specific abnormalities, findings within normal limits ("normal"), and "other findings". Because more data may be recorded than can fit on one screen, an additional screen may be entered by striking "+". The abstractor toggles the F to a T for any finding by striking the appropriate letter, resulting in the recording of that datum (for example, for jugular venous distention in Table 5).

The results of diagnostic tests are also recorded for specific periods during the hospitalization. For the laboratory tests (chemistry, hematology, enzymes, urinalysis, microbiology, and cytology), the worst result obtained within the first 24 hours of hospitalization is recorded as the "initial" value; in the case of some enzymes, such as the cardiac enzymes, a window of 48 hours is specified. If no admission data are on the record, preadmission results obtained within a week before admission are accepted.

TABLE 4 Excerpt from the Uniform Clinical Data Set Computer Screen: Peer Reviews Screens—History and Physical A

A.	Chronic neurologic disease	F
B.	History of neurologic surgery	F
C.	Current neuroloic exam findings	F
D.	Chronic cardiac disease	T
E.	Chronic vascular disease	T
F.	History of cardiovascular surgery	F
G.	Current cardiovascular exam findings	T
		F
I.	Chronic pulmonary disease	T
J.	History of pulmonary surgery	F
K.	Current pulmonary exam findings	T
L.	Chronic psychiatric disease	F
M.	Current psychiatric exam findings	F
N.	History of cancer	F
+/- GO TO OTHER HISTORY AND PHYSICAL INFORMATION (+ = SCREEN B AND - = SCREEN C)		
	Blank=Criteria F10=Leave	Type letter corresponding to item __

TABLE 5 Excerpt from the Uniform Clinical Data Set Computer Screen: Cardiovascular Examination Findings

Enter item letter on 1st line to change or enter T or F beside item		
ENTER ITEM LETTER: CHANGE ITEM __ ENTER = ITEM BY ITEM, F10 = LEAVE PRESS "+" FOR OTHER CV EXAM FINDINGS		
Item	Description	Value
A	Normal	F
B	Shock	F
C	Pulmonary edema	F
D	Peripheral edema	F
E	Jugular venous distension	T
F	Tachycardia	F
G	Bradycardia	F
H	Murmur	F
I	Arrhythmia	F
J	Cardiomegaly	F
K	Gallop rhythm	F
L	Peripheral pallor	F
M	Bruit	F
N	Thrill	F
O	Friction rub	F
P	Pulse Deficit—peripheral	F

Current CV exam Findings

Enter item letter on 1st line to change or enter T or F beside item		
ENTER ITEM LETTER: CHANGE ITEM __ ENTER = ITEM BY ITEM, F10 = LEAVE PRESS "+" FOR OTHER CV EXAM FINDINGS		
Item	Description	Value
A	Ischemic ulcers	F
B	Stasis ulcers	F
C	Venous/varicose ulcer	F
D	Gangrene	F
E	Dependent rubor	F
F	Delayed capillary fill	F
G	Chest pain (steady)	F
H	Other findings	F

In addition to the laboratory findings that apply to the period immediately surrounding the admission, the results of the last test prior to discharge are recorded (although this may have occurred considerably before discharge) and, for selected tests, the worst value between the initial and final value (the "interim" value) are recorded. The date a test was drawn is also recorded, as well as whether, given the calibration of the equipment on which the test was performed, the result fell outside the hospital's normal limits. Although the worst interim result is not usable for epidemiological analyses, it is used in screening cases by means of the HCFA Generic Quality Screens and so must be collected.

Table 6 illustrates the kinds of information being collected about treatment, such as nonsurgical procedures and drugs administered during the hospitalization. The medications data sought include route of administration (self-administrable, thus not requiring specific skills, or invasive, requiring specific skills for administration), the start date (the date when the drug was first given), and the end date (the last date of administration). The entry of drugs is guided by a dictionary that contains about 6,000 trade and generic names. This ensures that the drugs recorded are recognizable and associates

TABLE 6 Excerpt from the Uniform Clinical Data Set Computer Screen: Peer Reviews Screens—Treatment Interventions

Nonsurgical procedures	
A Blood products	F
B. Inhalation therapy	T
C. Professional services	T
Medication therapy in hospital	
D. Prescribed medications	T
E. Adverse reaction to medications	F
F. Delivery systems for medications	T

Prescribed medication

Use Arrows, Home, End, Pgup, Pgdn and Enter to Choose. F10 to Cancel

Drug name	Route	Start	End
Heparin	2	05/09/88	05/12/88
Theophylline, anhydrous	1	05/09/88	05/15/88
Bactrim	1	05/12/88	05/15/88
Pepcid	1	05/13/88	05/15/88
Carafate	1	05/14/88	05/16/88
Codeine	1	07/08/88	07/10/88

Blank=Criteria F10=Leave Type letter for item desired __

the names with therapeutic categories used by the expert system that assesses the appropriateness of admission and the quality of care.

Application of Findings

This expert system represents one of the two applications of the abstracted data. Immediately following abstraction, a sequence of about 3,000 logical rules is applied to the data. These rules embody the criteria currently used by the PROs to (a) verify that the patient's illness was sufficiently severe to justify the admission and that services that require the patient to be hospitalized were in fact rendered (the "admission necessity" algorithms) and (b) ascertain whether breaches of protocol pertaining to inpatient management and the discharge of the patient may have occurred (the "generic quality" and the "discharge" algorithms). The product, placed on the screen of the computer in about two and a half minutes, is a case summary that contains the results of the evaluation (Table 7) and an ordered listing of all the findings abstracted (not shown).

TABLE 7 Excerpt from the Uniform Clinical Data Set: Flag Settings and Reasonsa

Algorithm		Flags
AD00	Admission necessary	PA
	CASE FAILS ADMISSION NECESSITY SCREENS. REFER TO PA.	
ES00	Elective admission	PR
	SP Elective admission flag.	
ES04	Cardiac revascularization	PR
	2B INDICATIONS FOR INPATIENT ELECTIVE SURGERY NOT PRESENT. PA FLAG	
ES04		PA
DP01	Ischemic heart disease / chest pain	PR
	8D APPROPRIATE HOSPITALIZATION AND SERVICES. OK FLAG	
DP01		OK
OP09	Central nervous system	MO
	A CASE REQUIRES MONITORING FOR SEVERITY OF ILLNESS	
DS01	Discharge status/disposition	MO
	N17 Surgical patient with final hemoglobin missing or result less than admission result with difference >= 3 and < 4 grams per deciliter, discharge pulse > 110	
DS01	Discharge status/disposition	MO
	N29 No creatinines	

[a] Example for actual hospitalization

The case shown represents an actual hospitalization in which a coronary angioplasty was performed, ostensibly for a malfunction of a prior coronary artery bypass graft. The "flag" report states that the admission was elective, for cardiac revascularization, but that sufficient indications for the procedure were not present. In fact, the results of cardiac catheterization, specified in the case summary, included a left ventricular ejection fraction of 56 percent and a 50 percent stenosis of the right coronary. To justify revascularization, the algorithms require at least 70 percent stenosis. Consequently, the case is to be referred to a physician advisor (PA) for further review of the necessity for the admission.

The result of the surgical procedure algorithm (those labeled ES) results, in turn, in the summary recommendation (AD00) on admission necessity because elective surgery was performed. Had the admission not been elective, the results of the "disease-specific" (DP) and the more generic "organ-specific" algorithms would also have been considered, and an "OK" in any of the admission necessity algorithms would have resulted in no referral for further review. In this instance, there were enough findings and services to justify the hospitalization (but not the angioplasty) for ischemic heart disease.

There were some signs of disease of the central nervous system (OP09), but not enough to either justify or deny the admission, resulting in a recommendation that the case be accumulated in a data base for monitoring (MO) for patterns, but only if the monitoring flag appeared in the absence of an "OK" flag. In this case, it is disregarded.

No generic quality screens (DS) were failed, but two problems were identified by the discharge screen. Neither of these was severe enough to merit referral to a physician advisor, but they were serious enough to merit tracking (MO) to ascertain whether they are recurring problems at the hospital that cared for the patient.

The results of the case-finding algorithms suggest the level of clinical judgment they incorporate: rather rudimentary, but sufficient to give rise to controversy. This is a problem that, in the current environment of medical uncertainty, will dog the application of any expert system to the evaluation of medical practices.

The other application of the data acquisition system, the development of an epidemiological data base, has as its ultimate purpose the reduction of that uncertainty so that case-finding rules and judgments rendered by PRO physician advisors might be more more objective and substantial. In a more general sense, when patterns of care and patterns of outcomes are compared among providers of medical care, adjustment for risks contributed by patients and therefore not attributable to care, must be made.

The first example of the epidemiological application of abstracted clinical data is, in fact, an adjustment of mortality rates of hospitalized patients, grouped by hospital (the hospital is the provider). Table 8 and Figure 1

present results obtained with data abstracted from medical records using MedisGroups, a commercial system.

Table 8 compares measures of goodness-of-fit of models that employ data available from the HCFA claims files (core model), the core model plus the MedisGroups admission severity grade (ASG) (a measure that makes use of the abstracted data but selects findings and weights them according to clinical judgment), and a model that consists of the core variables plus specific clinical findings as abstracted. The outcome is the probability of death of individual patients. The fit improves progressively as the comprehensiveness of the model increases. Variables that identify hospitals and whose regression coefficients estimate the contribution of the hospital to the probability of patient death (patient risk factors being equal) contribute little at any level, but progressively less as patient risk factors are included in greater detail.

The two measures of goodness-of-fit—the proportion of concordant pairs or area under the ROC (receiver operating characteristic) curve, and the rank order correlation coefficient of observed and predicted probabilities of death—are directly related. The area under the ROC curve ranges from 0.5, if the model is totally ineffective, to 1.0, if it is perfectly predictive. The value of 0.9 achieved by inclusion of the specific clinical findings is substantial. It indicates that in about 90 percent of pairs of patients, one of whom died and one of whom did not, the patient who died had the higher predicted probability of death. I am not expert in these matters, but I am

TABLE 8 Evaluation of Goodness-of-Fit of Regression Models of the Probability of Death of Individual Patients

Variables in Model	Proportion of Concordant Pairs[a]	Rank Correlation of Observed and Predicted Deaths
Demographic only	0.640	0.279
Demographic and hospital	0.689	0.378
Core	0.838	0.675
Core and hospital	0.852	0.704
Core and MedisGroups ASG[b]	0.883	0.767
Core, MedisGroups ASG, and hospital	0.890	0.781
Core and clinical findings	0.896	0.792
Core, clinical findings, and hospital	0.902	0.804

[a] Concordant pairs: in pairs consisting of one patient who did and one who did not die, those pairs in which the patient who died had the higher predicted probability of dying.
[b] ASG — admission severity grade.

rather impressed by the power of the clinical findings to improve the goodness-of-fit of the model.

A more precise indication of what improvement is achieved in assessing the hospital's contribution to the probability of patient death when clinical data are added to the model is suggested by Figure 1. The figure plots estimates of that contribution obtained with claims data alone (core model) and with claims and clinical data (full model). Further discussion of this matter is best left for another occasion, but the potential uses of detailed clinical data in risk adjustment should be clear.

A more compelling application of detailed clinical data is to the estimation of the influence of patient risk factors and of treatments on outcomes, illustrated in Table 9. It presents a very useful example because of its

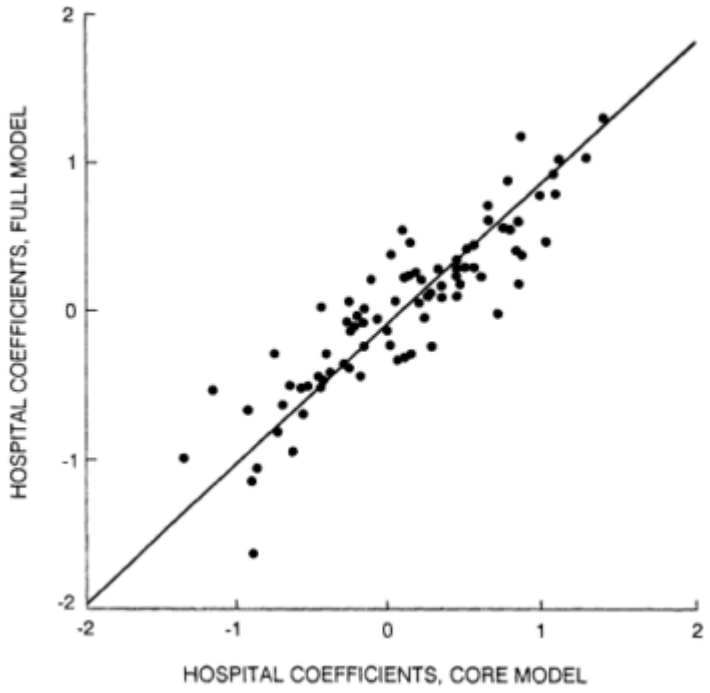

Figure 1
Comparison of Hospital Regression Coefficients from Core and Full Models

TABLE 9 Risk Factors for Death Up to Two Years After Admission for Acute Myocardial Infarction

Risk Factor	Relative Risk of Dying[a]
Age, 80 vs. 65 years	1.25
Leukocytosis, 20,000 vs. 7,000	1.15
Hypokalemia, 3.2 vs. 4.3	0.87
Alkalosis, pH 7.49 vs. 7.41	1.14
Prior admission within 30 days, yes vs. no	1.34
Myocardial ischemia (EKG), yes vs. no	0.83
Myocardial infarction, age undetermined, yes vs. no	0.84
Blood glucose, 300 vs. 90	1.19
Atrioventricular dissociation, yes vs. no	2.64
Congestive heart failure (X-ray), yes vs. no	1.54
Blood urea nitrogen, 60 vs. 15	1.24
Arterial oxygen pressure, 60 vs. 90 mm Hg	1.22
Disoriented, x2 or x3, yes vs. no	1.59
Coma/stupor, yes vs. no	2.46
Heart murmur, yes vs. no	1.48
Systolic blood pressure, 60 vs. 120	1.71
Tachypnea, 32 vs. 12 per minute	1.27
History of diabetes, yes vs. no	1.29
History of stroke or transient ischemic attack, yes vs. no	1.37
History of congestive heart failure, yes vs. no	1.22
History of myocardial infarction	1.16
Comorbidities (by ICD-9-CM codes) cancer, yes vs. no	1.88
Chronic renal disease, yes vs. no	1.49
Treatment Covariates	
Streptokinase (IV or IC), yes vs. no	0.87($P > 0.5$)
Coronary angioplasty, yes vs. no	0.47($P < 0.001$)
Coronary bypass surgery, yes vs. no	0.48($P < 0.001$)
Parenteral drugs (first 48 hours)	
Beta blocker, yes vs. no	0.66($P > 0.1$)
Calcium channel blocker, yes vs. no	1.12($P > 0.5$)
Digitalis, yes vs. no	0.92($P > 0.4$)
Intravenous nitroglycerine, yes vs. no	0.76($P < 0.003$)
Loop diuretic, yes vs. no	0.72($P > 0.1$)
Pressor agent, yes vs. no	1.48($P < 0.001$)
Short-acting nitroglycerine, yes vs. no	1.68($P > 0.2$)

[a] Based on the Cox proportional hazards model and stepwise regression. Follow-up is 12 to 24 months. All patient-specific risk factors are statistically significant at $P < 0.05$.

complexity and the difficulty of its interpretation. The results shown were obtained by the application of the Cox proportional hazards model.

The upper portion is straightforward, consisting of estimates of changes in the risk (relative risks) of death due to acute myocardial infarction (AMI) associated with the specified risk factors, all other risk factors being held constant. Only highly statistically significant ($P < 0.05$) predictors of death are listed in this portion. The risk factors include demographic characteristics, results of the admission physical examination, laboratory tests and other diagnostic tests carried out in the first 48 hours of hospitalization (or prior to surgery), and historical data.

The lower portion, which addresses treatments, is intriguing. The data suggest that coronary angioplasty and bypass are (or were in 1985, the year the treatments were administered) highly effective tools for the treatment of patients with AMI, controlling for the patient risks specified in the upper portion of the table. In fact, they reduced the probability of death by about half. This effect persists if patients who died on the day of admission are excluded, because they may not have lived long enough to become candidates

TABLE 10 Risk Factors for Rehospitalization Following Acute Myocardial Infarction

Risk Factor	Relative Risk of Rehospitalization[a]
Leukocytosis, 20,000 vs. 7,000	1.18
Hypocalcemia, 7 vs. 9.5	0.66
Hypokalemia, 3.2 vs. 4.3	0.85
Prior admission within 30 days, yes vs. no	1.54
Blood urea nitrogen, 60 vs. 15	1.23
Arterial oxygen pressure below 75 mm Hg	1.16
Edema, >2+	1.34
Systolic blood pressure, 60 vs. 120	0.78
Tachypnea, 32 vs. 12 per minute	1.19
Left ventricular ejection fraction, 35 vs. 62%	1.33
History of diabetes, yes vs. no	1.15
History of congestive heart failure, yes vs. no	1.23
History of chronic obstructive pulmonary disease, yes vs. no	1.30
History of immunosuppressive therapy, yes vs. no	1.26
Currently on anticoagulants, yes vs. no	1.41
Comorbidities (by ICD-9-CM code) cancer, yes vs. no	1.61

[a] Based on the Cox proportional hazards model and stepwise regression. Follow-up is 12 to 24 months. All patient-specific risk factors are statistically significant at $P < 0.05$.

for revascularization, and if the "center effect" is controlled for, because the superior outcome associated with revascularization may reflect the fact that patients admitted to hospitals that perform coronary revascularization may have received better care overall. The adverse effect of the use of pressor agents, controlling for hypotension, is also intriguing.

The analyses presented in Table 9 are observational and must, therefore, be approached with some caution. Nevertheless, the power of detailed clinical data in describing the natural history of conditions as they are currently being treated is well illustrated.

Of course, mortality is not the only outcome that may be addressed by the combination of detailed clinical and claims data. Table 10 illustrates, in a fashion analogous to Table 9, analyses addressing rehospitalization rates. In all, to characterize adequately the effectiveness of medical interventions, and therefore their impacts on the health of patients, at least from a public health perspective, it is necessary to measure mortality, morbidity, disability, and expenditures for health services, in order to try to track the effectiveness of interventions.

CONCLUSIONS

HCFA's objectives in assessing the effectiveness of interventions were as follows:

- to assess the overall merits of competing procedures,
- to provide information to assist clinicians in the management of patients,
- to provide information to assist in peer review of care,
- to guide in the formulation of policy on the allocation of resources.

The initial intent was to provide PROs with more effective tools for the review and evaluation of patient care. Clearly, the information generated in this process has broader applications, the most important being to assist clinicians in the treatment of patients by providing assessments of the relative merits of treatment strategies overall and for patients with specific risk factors. A further useful by-product is guidance for the allocation of resources, at whatever level such decisions are made, by providing measures of the impacts of those decisions on the health of patients.

Development and Use of Outcome Measures: Introduction

G. Richard Smith, Session Moderator

The primary thrust of the Effectiveness Initiative is to determine what works in the practice of medicine. One can determine if something works only by knowing what happens to the patient. Therefore, a major emphasis of the committee's work has been on the outcomes of patient care.

Previously, we could determine outcomes only in terms of whether a patient was alive or dead or by using some kind of medical test. We had very limited knowledge about what effect various medical interventions had on patients. Over the past 15 years, a new technology has been developed to enable us to understand some of the effects of medical care. This technology is called health status assessment, functional status assessment, or even at times quality of life. As a result of this new technology, we are now able to quantify a number of aspects of the state of a patient's health. For example, we can determine the effect on a patient's physical health of developing asthma or the effect on a patient's mental health of being told that she has breast cancer.

When these tools are applied in a systematic and thoughtful fashion, we can tell much about the effect of our medical care system—not only on our people as a whole, but upon our individual patients.

Donald L. Patrick is currently professor of health services and director of the Social and Behavioral Sciences Program at the University of Washington School of Public Health. Dr. Patrick discusses selection of outcomes; use of generic and disease-specific measures; progress toward short, reliable, valid, and responsive measures; and interpreting observed changes in measures and what these changes mean.

Paul D. Cleary is an associate professor in the Department of Health Care Policy at the Harvard Medical School. His chapter describes current

and recent research efforts in the development and use of outcome measures. The presentation highlights the use of patient self-reports.

Audrey Burnam is a senior behavioral scientist at The RAND Corporation. Her chapter concentrates on the depression part of the Medical Outcomes Study and summarizes the group's approach to studying depression outcomes. Initial findings from the study are presented.

17

Assessing Health-Related Quality of Life Outcomes

Donald L. Patrick

The hope and promise of the Effectiveness Initiative to curtail escalating health care costs remind me of the time an airline pilot made an urgent announcement during a transatlantic flight. His voice suddenly came over the public address system, and he said, "Ladies and gentlemen, I have two pieces of news for you. One of them is good, and one of them is not so good. First I'll tell you the bad news. The bad news is that we are lost. We don't have any idea of where we are. But, as I told you, there is good news, too. The good news is that we have a 200-mile-an-hour tail wind. In other words, we don't know where we're going, but we're getting there awfully fast." I think this story nicely describes our current situation. Rapidly rising costs are hastening our attempts to use whatever means possible to find solutions. The Effectiveness Initiative is one means of turning bad news into good news.

I wish to address four methodological issues involved in the assessment of health status outcomes: (a) selection of relevant outcomes; (b) use of generic and disease-specific measures; (c) progress toward short, reliable, valid, *and* responsive measures; and (d) methods for interpreting observed changes in measures.

Before reviewing these issues, however, I would like to identify two challenges to our reliance on outcomes assessment for controlling health care costs. First, I would remind us that health services are only one determinant of health status (1). When we talk about effectiveness of health care in terms of health status outcomes, we cannot forget that socioeconomic, political, and cultural systems have diverse and powerful influences on outcomes (2). Effectiveness of medical care is our focus, but the larger sociocultural context influences both provider and patient reports of outcomes. Effectiveness is often in the eye of the beholder; patient expectations range from efficacious

treatments that "cure" to "hugs from the doctor." Patient expectations, in fact, may well exceed our ability to provide the services that produce expected outcomes.

TABLE 1 Combinations of Cost and Effectiveness Outcomes

		Quality of Life Outcomes	
		-	+
Cost of Treatment	+	Worse quality of life Higher cost	Better quality of life Higher cost
	-	Worse quality of life Lower cost	Better quality of life Lower cost

A second cautionary note is that the increase in health technologies makes cost containment extremely difficult. Our decisions concerning cost and effectiveness outcomes can be described in Table 1.

There are four different combinations of cost and effectiveness outcomes. Ideally, new technologies such as pharmaceuticals will produce better quality of life outcomes at reduced cost (lower right quadrant). Medical treatment for back pain and new drugs for benign prostatic hypertrophy might produce similar results. Rationing, capping reimbursement, and other methods of cost containment may produce outcomes in the lower left quadrant, that is, worse quality of life and lower costs. The upper left quadrant (higher cost, worse quality of life) is obviously to be avoided, although life-extending treatments might well fall into this category. Most technological innovations probably fall in the upper right quadrant, that is, better quality of life outcomes at higher cost. Technological advance clearly challenges our initiatives to maintain or lower health care costs.

QUANTITY AND QUALITY OF LIFE

Quantity and quality of life are distinct but related concepts used to evaluate the present and future state of a person or group of people (3). Taken together, quantity and quality should represent a complete picture of the person or group. Quantity of life is assessed in terms of length of survival. For example, survival time is the number of days a patient lives after undergoing heart transplantation or while receiving drugs such as azidothymidine (AZT, a treatment for AIDS). Although such drugs may prolong the life of patients, they may produce concurrent toxic effects. It is

easy to assess quantity of life accurately in retrospective studies, but prognosis or duration of survival can only be estimated, often with considerable uncertainty. Furthermore, the value attached to a day of life differs from one person to another (4). For some persons, death is preferable to the lowest states of functioning, such as coma or profound pain or depression.

Quality of life has been assessed in a variety of different ways, including beauty in the landscape, close family life, environmental purity, and a spiritual understanding of existence. Health-related quality of life is more limited; it can be defined as the value assigned to the duration of life as modified by the *social opportunities, perceptions, functional states*, and *impairments* that are influenced by disease, injuries, treatments, or policy (5). This definition covers five broad concepts on a continuum of health-related quality of life, which is anchored at the top by an optimal value of 1.0 and at the bottom by a minimal value of 0. Specific dimensions of opportunity, perception, functional status, impairment, and survival fall along this continuum. See Table 2 for a more comprehensive description of the concepts and dimensions of health-related quality of life.

Dimensions of the five concepts may be negatively or positively valued in relation to one another. The value assigned to the particular state of individuals or groups defines health-related quality of life. The time spent in that state or the probability of moving from one state to another (that is, prognosis) defines quantity of life. Thus, a complete representation of health-related quality of life involves specification of relevant states or combinations of dimensions, the values or preferences assigned to these states, and the duration or probability of duration in different states. This definition of health-related quality of life is similar to the health-state utilities approach developed over the last two decades (6).

SELECTION OF OUTCOMES

The question arises whether there is a core set of outcomes that must be included in a quality of life assessment on theoretical, empirical, or judgmental criteria. Sol Levine has provided theoretical guidance in this area by focusing attention on two very important aspects of quality of life (7). The first is performance of the physical, psychological, and social functions and activities that people do in their everyday lives. The second is the satisfaction derived from performing these usual activities. Functional status and satisfaction with health are the core domains of health-related quality of life"

There are many measures currently available to assess these health-related quality of life dimensions. When we select these measures, we are attempting, in advance, to identify the potential effects of treatments as well as the potential side effects or unanticipated consequences of treatments. Table 3 contains a taxonomy of health-related quality of life measures.

TABLE 2 Concepts and Domains of Health-Related Quality of Life

Concept and Domain	Definition/Indicator
Opportunity	
Social or cultural handicap	Disadvantage because of health
Individual resilience	Capacity for health; ability to withstand stress; reserve
Health perceptions	
Satisfaction with health function	Physical, psychological, social
General health perceptions	Self-rating of health; health concern, worry
Functional status	
Social	
Limitations in usual roles	Acute or chronic limitations in social roles of student, worker, parent, household member
Integration	Participation in the community
Contact	Interaction with others
Intimacy	Perceived feelings of closeness; sexual
Psychological	
Affective	Psychological attitudes and behaviors, including distress and general wellbeing or happiness
Cognitive	Alertness; disorientation; problems in reasoning
Physical	
Activity restrictions	Acute or chronic limitation in physical activity, mobility, self-care, sleep, communication
Fitness	Performance of activity with vigor and without excessive fatigue
Impairment	
Subjective complaints	Reports of physical and psychological symptoms, sensations, pain, health problems, or feelings not directly observable
Signs	Physical examination: observable evidence of defect or abnormality
Self-reported disease	Patient listing of medical conditions or impairments
Psychological measures	Laboratory data, records, and their clinical interpretation
Tissue alterations	Pathological evidence
Diagnoses	Clinical judgments after "all the evidence"
Death and duration of life	Mortality; survival; longevity

SOURCE: Patrick and Erickson (5)..

TABLE 3 A Taxonomy of Health-Related Quality of Life Measures

Approach	Strength	Weakness
Scores for analysis		
Single index number Useful for cost-effectiveness	Represents net impact	Effects on different outcomes not possible May not be responsive
Profile of interrelated scores	Single instrument Effects on different outcomes possible	May not be responsive Length often problem
Battery of independent scores Wide range of outcomes	Can select relevant outcomes Multiple comparisons possible	Cannot relate different outcomes to common scale Need to identify major outcome
Objective of application		
Generic: across conditions and populations	Broadly applicable Summarize range of concepts May detect unanticipated	May not be responsive enough May not have focus of patient interest Length often problem Effects may be difficult to interpret
Specific: disease, population, function, or condition	More acceptable to respondents May detect unanticipated	Comparisons across conditions and populations not possible
Weighting System		
Utility: preference weights from patients, providers, or community	Interval scale Patient view incorporated	Difficulty obtaining weights May not differ from statistical weights that are easier to obtain
Statistical: items weighted equally or from frequency of response	Self-weighting samples More familiar techniques Appears easier to use	May be influenced by prevalence

SOURCE: Guyatt et al. (8).

Measures can be classified according to the scores they produce for analysis, their objective in application, and the weighting system used in scoring (8).

Sources For Analysis

Indexes

Measures such as the Quality of Well-being Scale (9), the Health Utilities Classification (6), and the Disability/Distress Scale (10) combine duration of life with specific dimensions of impairment, as well as physical, psychological, and social function. These measures yield a single index value, quality-adjusted life years (QALYs), that can be used to compare the cost per quality-adjusted life year gained from different health interventions. For example, the cost per QALY gained in 1986 U.S. dollars for coronary artery bypass surgery for left main coronary artery disease is $4,796, compared with $36,316 for neonatal intensive care for infants weighing 500 to 900 grams (6). The effects of a particular treatment on a single index, such as QALYs, however, remain hidden. There is considerable controversy over whether such an index can represent health in a sufficient manner to detect changes and, indeed, interpret where those changes have taken place. Nevertheless, QALYs, or years of healthy life, are gaining acceptance, as exemplified by their inclusion in the *Year 2000 Objectives for the Nation* (11).

Profiles

Other measures provide a profile of scores for different components of health-related quality of life. The Sickness Impact Profile (SIP) assesses sickness-related dysfunction in 12 different categories, producing a score for each category (12). Various categories may be aggregated into a physical dimension score, a psychosocial dimension score, and an overall score with independent categories of work, eating, sleep and rest, home management, and recreation and pastimes. Similarly, Part I of the Nottingham Health Profile (NHP) contains 38 items that cover six domains of experience, yielding individual scores for each; Part II contains perceived problems in seven areas of daily life (13). Unlike the SIP, however, the NHP does not yield an overall index score. The 59-item McMaster Health Questionnaire yields separate indexes for physical, emotional, and social function (14). Measures developed originally at The RAND Corporation—the 108-item Health Insurance Study battery and the 20 to 40-item Medical Outcomes Study short-form generic measures—cover a wide spectrum of health concepts for use in general populations (15). All these generic measures have been tested extensively with different patient populations.

Batteries

It is important to make a distinction between health profiles and batteries. Batteries are collections of health status measures with independent scores for each outcome. Specific measures of different health outcome domains are selected to make up an assessment battery. The recent evaluation of antihypertensive medications, sponsored by the Squibb Pharmaceutical Company, is an example of the battery approach (16). These investigators selected, among others, the latest "best available" specific measures of general wellbeing, physical symptoms, and sexual dysfunction. Improvement in quality of life was assessed for three different anti-hypertensive agents on each of these independent measures. Luckily, all measures chosen for this study showed some improvement for the new drug under consideration. Often, our results are more mixed.

The battery approach is an appealing assessment strategy because of the wide range and type of outcomes that can be assessed. A major outcome needs to be identified as the primary endpoint, however, to avoid conflicting findings and multiple comparisons of outcomes on different measurement scales.

GENERIC AND SPECIFIC MEASURES

Generic measures of health status are those that purport to be broadly applicable across types and severities of disease, across different medical treatments or health interventions, and across demographic and cultural subgroups. Visual analogue measures are designed to summarize a spectrum of the concepts of health or quality of life that apply to many different impairments, illnesses, patients, and populations.

Disease-specific measures are those designed to assess specific diagnostic groups or patient populations, often with the goal of measuring responsiveness or "clinically important" changes. These are changes that clinicians and patients think are discernible and important, have been detected with an intervention of known efficacy, or are related to well-established physiological measures (such as grip strength for arthritis patients or spirometry for those with chronic obstructive lung disease) (17). The term "disease-specific," here, refers to different adult patient populations with specific conditions or diagnoses.

Not all specific measures are disease-related. They may be specific to conditions (for example, back pain or dyspnea), functions (for example, sexual or emotional function), or populations (for example, older adults or developmentally disabled children). Specific measures of single concepts or conditions are the most numerous of all within the health status field. These single-concept measures range from the assessment of specific symptoms

such as nausea and vomiting to more global concepts of life satisfaction. Mental health measures of depression, anxiety, and other emotional states, for example, are frequently used in clinical research for assessing individual concepts of psychological status. Numerical estimates of subjective pain, such as visual analogue scales, have gained a wide following, partly because of their high correlation with verbal rating scales and their simplicity (18). Visual analogue scales are also gaining popularity in the measurement of symptoms and functional status.

Disease-specific measures, such as the Karnofsky Performance Status Scale for cancer (19), the American Rheumatism Association (ARA) functional classification for arthritis (20), and the New York Heart Association functional classification (21), have been used extensively over several decades. These measures were developed to meet the need for rapid classification of patients, and their sensitivity to small but clinically important change is limited. The ARA classification, for example, may detect large changes, such as those following hip replacement, but not smaller changes following drug therapy judged successful by other criteria. The popularity of disease-specific measures arises primarily from the need of clinical trials and practitioners to use scales that are most responsive to clinical changes that occur over time. Both discriminating improved from unimproved patients and accurately quantifying minimally important changes are particularly important measurement objectives for clinical research and clinical practice.

Generic and disease-specific assessments alike are useful for clinical research, clinical practice, and policy analysis. Selection of different measures depends on the objectives of measurement and the environment of the application. No single general-purpose measure is likely to meet all the needs of investigators and specific populations. Patients with different medical conditions have different concerns or place different emphasis on more generic concepts of health. Rather than develop disease-specific measures that incorporate generic concepts of health-related quality of life, the preferred strategy is to use standardized, generic instruments with disease-specific supplements.

Generic measures permit the comparison of different populations and different programs, a most important objective for policy analysis and decision making. Use of generic measures is necessary for comparing benefits of different health interventions and allocating resources. Cumulative knowledge of health and quality of life outcomes using generic measures will establish the relative burden of different diseases and the relative merit of different interventions.

By contrast, instruments specific to different diseases, conditions, and populations are critical for identifying important concerns of patients with particular conditions and for measuring small, clinically important changes from specific treatments. Experience with disease-specific measures to date

indicates their usefulness in discriminating among different conditions and in assessing changes. Rapid development of such instruments is to be expected for conditions and populations where few specific measures now exist, notably certain cardiovascular conditions, diabetes, gastrointestinal disorders, and sexually transmitted diseases including AIDS.

SHORT, RESPONSIVE MEASURES

The development of short-form generic measures for use in clinical practice is a welcome advance (22,23). Undoubtedly these measures will be tested against more comprehensive and detailed generic instruments to identify information that may be lost using brief assessments. The need to measure an increasing array of physiological, physical, psychological, social, and general outcomes within a single investigation cannot be met unless measures are short, efficient to administer, and highly acceptable to investigators and respondents. Both generic and disease-specific measures will be assessed against these practical constraints.

Self-administered, comprehensive measures that are sensitive to variations in health care organization and medical practice are also needed. Short, generic health status measures (1 to 60 items) have been developed from longer versions based on minimal psychometric criteria for internal consistency reliability and content and construct validity (22,24). The Short-Form Health Survey, derived from measures used in the Medical Outcomes Study, is currently being tested for its responsiveness, or ability to detect minimal changes of importance to interested parties. In the next decade, short-form generic measures need to be tested rigorously for their content validity, responsiveness, convergent and discriminant validity, and generalizability.

"Short" and "comprehensive" can be conflicting goals for some applications and populations. The full domain of health-related quality of life outcomes of interest to patients, providers, and payers simply cannot be represented in short measures. Some concepts, for example, cognitive function, sleep and rest behaviors, recreation, and satisfaction with health, are seldom represented in short, generic measures. These omissions may not seriously compromise the usefulness of short-form measures in relatively well populations, but outcomes assessment in specific populations such as older persons, mentally ill persons, and institutionalized persons may require long-form assessments.

Responsiveness, how well short-form measures detect subtle changes in behavioral and subjective health status, also requires testing and comparison with clinical measures. We can be encouraged by the data from the Medical Outcomes Study indicating that generic measures are very strong in detecting stable scores among a clinically stable group. Greater emphasis needs

to be placed on the assessment of responsiveness in comparing generic and disease-specific outcomes.

Responsiveness is an important consideration in all serial applications of health status measures. Including items sensitive to change is critical to such assessments. Responsiveness of health status measures has been assessed using the relative efficiency statistic (a ratio of paired t statistics) (25), correlation of scale changes with other measures (26), receiver-operating characteristic curves (26), and a responsiveness statistic (ratio of minimal clinically important differences to variability in stable subjects) (17).

Disease-specific measures with items selected to assess particular concerns or worded to attribute change to the condition of interest, for example, back pain in the modified SIP, may be particularly sensitive to within-subject changes and thus more responsive than generic measures, which contain items unrelated to change.

CHANGES IN MEASURES

Analyzing and interpreting changes in health status measures are problems in all longitudinal studies. These may be observational case studies, cohort studies, clinical trials, or health services evaluations. Changes in physiological measures such as blood pressure or cholesterol level may be interpreted in terms of prognostic implications and well-established or agreed cutoff points. Changes in generic health measures are more difficult to interpret, although even small changes in portions of such measures may be quite useful (for example, changes in physical mobility or self-care are meaningful in disabled populations). Changes in scores on the most general measures, such as health perceptions or global physical and psychosocial dimension scores, can be even more difficult to interpret.

The net changes observed may reflect a large number of different transitions or combinations of transitions within the population. Single-score or aggregated measures can make it difficult to identify which items or components are responsible for the change. Net changes must also be distinguished from random or systematic changes (learning effects, rumination) that may occur independently of an intervention. Although changes in these scores may reflect sensitive effects, the relative magnitude of the change may be difficult to assess. For example, is a 5-point difference more meaningful than a 3-point difference?

Changes in disease-specific measures may be easier to interpret because they are more specific or more closely associated with changes in clinical measures of disease activity such as blood pressure or joint inflammation. Clinician or patient assessments of improvement, which are common measures of change or effects, may be more closely associated with changes in disease-specific measures than with those in generic health status measures (27).

A THEORETICAL MODEL FOR HEALTH STATUS AND QUALITY OF LIFE

A major challenge facing developers and users of health and quality of life measures is to establish a testable theory of the expected relationships among the different concepts and domains of health-related quality of life. The problem is not confined only to the relationship between physiological measures and behaviors or perceptions, for example, blood pressure and functional status. Measures of various dimensions, such as symptoms, psychological function, and satisfaction with health have been shown to be only loosely associated or entirely dissociated within the same sample (2). Figure 1 depicts hypothesized relationships among different health-related quality of life concepts in a simple linear progression. The concepts are bounded by environmental determinants that influence disease and its consequences and by prognoses for improvement, maintenance, or decline in health-related quality of life.

The simple causal model suggested in Figure 1 does not represent the complexity or strength of the expected relationships among health-related quality of life dimensions. For example, persons can have an asymptomatic disease that affects prognosis without affecting functional status, perceptions, or opportunity. A person with hypertension or hypercholesterolemia may not have restrictions in activity but may be disadvantaged by fear of a stroke or death. Similarly, not all persons with impaired physiological capacity experience psychological dysfunction. Persons with rheumatoid arthritis and congestive heart failure may have high satisfaction with their health and positive well-being.

Figure 1 also indicates that the causal relationships among concepts can be reversed; for example, functional limitations and perceived health can be viewed as influencing impairment or physiological measures of chronic disease (29). Reversing the causal chain permits testing of the variable course of chronic disease, whereby impairments may become permanent and lead to changes in behavior and perceptions that, in turn, influence symptoms or level of impairment. The notion of an interplay between the psyche and the body is as old as medicine itself. Models of psychophysiological processes such as disruption in the regulation of blood volume and control of blood pressure by the kidneys can be invoked to explain sociobehavioral influences on disease processes. This evidence may not be sufficient to convince the most skeptical biomedical researcher, but the hypothesis has moved well beyond mere speculation.

At present, researchers tend to approach the relationship among end points inductively, by collecting data and examining the correlation among measures. Little hypothetical or deductive reasoning is involved in either the selection of measures or analysis of results. Head-to-head comparisons of different

dimensions will be important for determining the association between specific disease states or disorders and their behavioral, perceptual, and social consequences. Increasing our understanding of these relationships will help us realize the potential of health-related quality of life measures for identifying the intervention strategies that address the most important concerns of patients, their families, clinicians, and society in general.

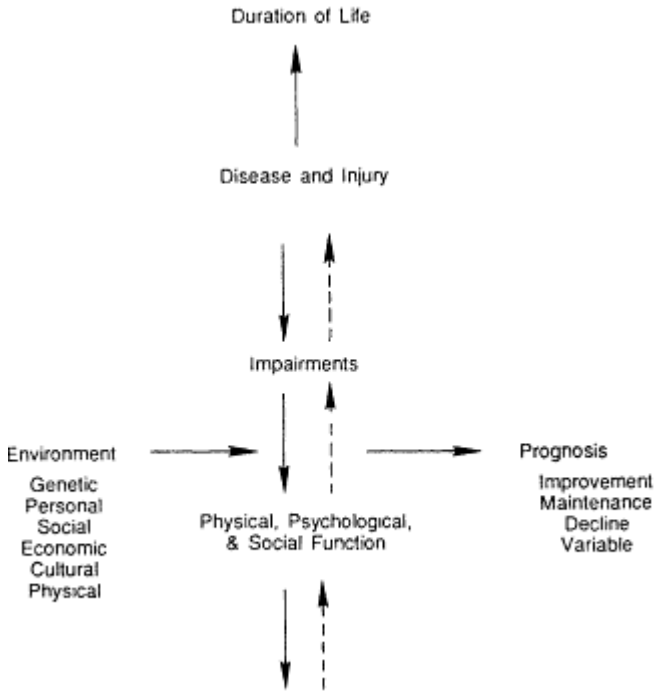

Figure 1
Relationships Among Health-Related Quality of Life Concepts
Source: Adapted from Patrick and Bergner (28).

CONCLUSIONS

The use of health-related quality of life measures, especially those based on function, is likely to increase during the next decade. This increase,

however, is most likely to occur in clinical research and clinical practice (30). Unless the necessary political will, resources, data, and policy researchers coexist, there will be relatively little advance in the use of health status measures for decision making and policymaking.

Policy research tends to rely on available national data, and currently these data provide limited information about health status. The Effectiveness Initiative will be successful only if it motivates data collection and methods that incorporate a broad spectrum of health outcomes (such as death, impairment, functional status, and perceptions) into a single assessment. Health and quality of life outcomes are what count, and these outcomes cannot be determined without appropriate and inclusive measures of health-related quality of life.

I hope that motivation and resources will be found to help resolve methodological issues in the measurement of health status and quality of life. I also hope that government agencies, employers, and private providers will begin to collect health-related quality of life data on the constituents and populations they serve. Even if these data are imperfect or primitive, the effects of improving accessibility and quality of health care can only be assessed adequately in terms of the health-related quality of life of the nation.

References

1. Levine, S., Elinson, J., and Feldman, J. Does Medical Care Do Any Good? Pp. 394-406 in *Handbook of Health, Health Care, and the Health Professions*. Mechanic, D., ed. New York: Free Press. 1983.
2. Patrick, D.L., Stein, J., Porta, M., et al. Poverty, Health Services, and Health Status in Rural America. *Milbank Quarterly* 66(1):105-136, 1988.
3. Patrick, D. and Elinson, J. Sociomedical Approaches to Disease and Treatment Outcomes in Cardiovascular Care. *Quality of Life* 1:53-65, 1984.
4. Patrick, D., Bush, J., and Chen, M. Toward an Operational Definition of Health. *Journal of Health and Social Behavior* 14:6-23, 1973.
5. Patrick, D.L. and Erickson, P. What Constitutes Quality of Life? Concepts and Dimensions . *Clinical Nutrition* 7:53-63, 1988.
6. Torrance, G.W. Measurement of Health State Utilities for Economic Appraisal. *Health Economics* 5:1-30, 1986.
7. Levine, S. The Changing Terrains in Medical Sociology: Emergent Concern with Quality of Life. *Journal of Health and Social Behavior* 28:1-6, 1987.
8. Guyatt, G., Feeny, D., and Patrick, D. Issues in Quality-of-Life Measurement in Clinical Trials. *Controlled Clinical Trials*. In press.
9. Bush, J.W. Relative Preference Versus Relative Frequencies in Health-related Quality of Life Evaluation. Pp. 118-139 in *Assessment of Quality of Life in Clinical Trials of Cardiovascular Therapies* . Wenger, N.K., Mattson, M.E., Furberg, C.D., et al., eds. New York: LeJacq Publishing, Inc., 1984.

10. Rosser, R.M. A Health Index and Output Measure. Pp. 133-160 in *Quality of Life: Assessment and Application*. Walker, S.R. and Rosser, R.M., eds. Lancaster, England: MTP Press, 1988.
11. U.S. Department of Health and Human Services, Public Health Service. *Promoting Health/ Preventing Disease: Year 2000 Objectives for the Nation*. Washington, D.C.: Government Printing Office. 1989.
12. Bergner, M.B., Bobbitt, R.A., Carter, W.B., et al. The SIP: Development and Final Revision of a Health Status Measure. *Medical Care* 19:787-805, 1981.
13. McEwen, J. The Nottingham Health Profile. Pp. 95-112 in *Quality of Life: Assessment and Application*. Walker, S.R. and Rosser, R.M. eds., Lancaster, England: MTP Press, 1988.
14. Chambers, L.W. The McMaster Health Index Questionnaire: An Update. Pp. 113-131 in *Quality of Life: Assessment and Application*. Walker, S.R. and Rosser, R.M., eds. Lancaster, England: MTP Press, 1988.
15. Stewart, A.L., Greenfield, S., Hays, R.D., et al. Functional Status and Well Being of Patients with Chronic Conditions. *Journal of the American Medical Association* 262:907-913, 1989.
16. Croog, S.H., Levine, S., Testa, M.A., et al. The Effects of Antihypertensive Therapy on the Quality of Life. *New England Journal of Medicine* 314:1657-1664, 1986.
17. Guyart, G., Walter, S., and Norman G. Measuring Change Over Time: Assessing the Usefulness of Evaluative Instruments. *Journal of Chronic Diseases* 40:171-178, 1987.
18. Scott, P.J. and Huskisson, E.C. Graphic Representation of Pain. *Pain* 2:175-184, 1976.
19. Karnofsky, D.A., Abelmann, W.H., Craver, L.F., et al. The Use of Nitrogen Mustards in the Palliative Treatment of Cancer. *Cancer* 1:634-656, 1948.
20. Steinbrocker, O., Traeger, C.H., and Batterman, R.C. Therapeutic Criteria in Rheumatoid Arthritis. *Journal of the American Medical Association* 140:659-662, 1949.
21. Criteria Committee of the New York Heart Association, Inc. *Disease of the Heart and Blood Vessels: Nomenclature and Criteria for Diagnosis* , 6th edition. Boston: Little, Brown, 1964.
22. Stewart, A.L., Hays, R.D., and Ware, J.E. The MOS Short-Form General Health Survey: Reliability and Validity in a Patient Population. *Medical Care* 26:724-735, 1988.
23. Nelson, E.C., Wasson, J.H., and Kirk J.W. Assessment of Function in Routine Clinical Practice: Description of the COOP Chart Method and Preliminary Findings. *Journal of Chronic Diseases* 40 Supplement 1:55S-63S, 1987.
24. Nelson, E.C. and Berwick, D.M. The Measurement of Health Status in Clinical Practice. *Medical Care* 27(3) Supplement:S77-S90, 1989.
25. Liang, M.H., Larson, M.G., Cullen, K.E., et al. Comparative Measurement Efficiency and Sensitivity of Five Health Status Instruments for Arthritis Research. *Arthritis and Rheumatism* 28:542-547, 1985.
26. Deyo, R.A. and Centor, R.M. Assessing the Responsiveness of Functional Scales to Clinical Change: An Analogy to Diagnostic Test Performance. *Journal of Chronic Diseases* 11:897-906, 1986.
27. MacKenzie, C.R., Charlson, M.E., DiGioia, D., et al. Can the Sickness Impact

Profile Measure Change? An Example of Scale Assessment. *Journal of Chronic Diseases* 39:429-433, 1986.

28. Patrick, D.L. and Bergner, M. Measurement of Health Status in the 1990s. *Annual Review of Public Health* 11:165-183, 1990.
29. Patrick, D.L. Commentary: Patient Reports of Health Status as Predictors of Physiologic Health in Chronic Disease. *Journal of Chronic Diseases* 40:37S-40S, 1987.
30. Bergner, M. Quality of Life, Health Status, and Clinical Research. *Medical Care* 27:S148-S156, 1989.

18

Using Patient Reports of Outcomes to Assess Effectiveness of Medical Care

Paul D. Cleary

In this chapter, I provide a brief overview of some of our current and recent work on the development and evaluation of outcomes measures based on patient reports. First, I make some general comments about the way we think about quality and effectiveness and discuss the range of outcomes that we think are important to consider in these types of studies. I also discuss a study recently conducted in six hospitals in California and Boston, including some results from the use of both generic and disease-specific measures in that study. I illustrate several of my points using data from patients with total hip replacements and conclude with some observations that bear on the strengths and weaknesses of these measures for assessing effectiveness and outcomes in health care.

LINKING PROCESS AND OUTCOME

Donabedian has described three approaches to the assessment of the quality of medical care: observation of structure, of process, and of outcomes (1). Most programs for evaluating quality focus primarily on process, and we tend to use terms such as "consensus," "norms," "standards," "criteria," and "appropriateness" when we describe the ways in which we evaluate process. These are all part of our lexicon and are an integral part of the way we think about quality.

Donabedian pointed out that we typically evaluate quality on the basis of observations of process, but he also asserted that judgments of quality tend to rest on what is known about the relationships between process and outcome. I would emphasize that point and argue that evaluations of process and outcomes are inextricably linked. Unless we know what the outcome of a

particular process is, we cannot determine whether it represents quality care.

The purpose of measuring outcomes is to help establish the relationships between process and outcomes. Little is learned by studying variations in outcomes by themselves, and little is gained by developing better measurement tools in isolation. These types of research activities are undertaken to help develop a practical, valid model of the linkages between process and outcomes so that we can improve quality of care.

Early Studies of Outcome

Outcomes are so widely considered to be the ultimate indicator of quality that it is surprising how infrequently we analyze them carefully. In the 1830s, a physician named Pierre-Charles-Alexandre Louis started a group in Paris that discussed the use of statistics to examine patterns of medical care. In 1838, a physician from that group named George Norris returned to the United States and looked at 55 cases in which an amputation had been performed. Norris found that 21 of the patients who had had an amputation had died. This was an important finding, and it challenged many people's assumptions about the dangers associated with amputation.

In subsequent work, Norris compared how surgery outcomes at the Pennsylvania Hospital compared with those of hospitals in other cities and counties. It is interesting that although we think of the publication of mortality statistics as a very recent phenomenon, they have been published on a hospital-specific basis for 150 years! It is disappointing that we have not improved our methods for assessing the relationships among case mix, process of care, and outcomes, given the long history of work in this area and the importance of the issues.

The Need For Disease-Specific Measures

One of the central issues in outcomes assessment concerns the range of outcomes that should be assessed and whether it is better to use disease-specific or general measures or both. I think that we should definitely include both measures in outcome batteries. A number of researchers have argued that assessing disease-specific outcomes is not necessary, that if one measures generic outcomes, specific measures will not help explain additional variance. I disagree strongly with that position. If we want to be able to detect differences in outcomes that are related to the process of care, it is usually necessary to include measures of outcomes specific to the condition studied and, in some cases, specific to the process of care being examined.

The domain of general outcomes that we think are important to study includes general health perceptions, disability, activities of daily living, role performance, well-being, fatigue, cognitive functioning, and satisfaction with care. It is not always necessary to measure all of these outcomes. Depending on the application, one may want to measure only one or two dimensions. It is often useful to have measures of all of the areas mentioned above, but to develop a practical system that can be used for policy-related relevant research and for quality assessment and assurance programs, it may be necessary to be more selective with respect to the measures used.

SCOPE OF THE STUDY

The specific study I describe here is an investigation of variations in case mix, patterns of care, and outcomes at six hospitals. The investigators besides myself were Barbara McNeil, Sheldon Greenfield, Albert Mulley, Steven Pauker, Steven Schroeder, and Lewis Wexler. The hospitals were not-for-profit, university-affiliated teaching hospitals. Three of the hospitals are in California and three in Boston. The sample was about 3,000 patients receiving treatment for one of six medical or surgical conditions: acute myocardial infarction (AMI); rule-out AMIs; total hip replacements; cholecystectomies; coronary artery bypass graft surgery (CABG); and transurethral prostatectomy (TURP).

During a one-year enrollment period, all eligible patients were sent a letter after discharge explaining the purposes of the research and encouraging them to take part in the study. For each patient who agreed to participate, we obtained from the medical records data on disease severity, comorbid conditions, and the process of care during the index hospitalization. Information about sociodemographic characteristics, health-related quality of life before and after hospitalization, perceived improvement in health status, health care utilization, and satisfaction with care were collected using a self-administered questionnaire mailed after discharge (2). The timing of the follow-up questionnaire was determined by a panel of experts and varied, depending on the condition, from 3 to 12 months. Patients who did not return the original questionnaire within two weeks were sent a second questionnaire, which was followed up with a telephone call reminding them to return the questionnaire. Patients who still did not return a completed questionnaire were interviewed over the telephone, when possible.

A medical record reviewer abstracted information about sociodemographic characteristics, indicators of severity, comorbid conditions, surgical procedure, occurrence of in-hospital complications, and use of services in the hospital (for example, laboratory tests, days in the intensive-care unit, and so on). To record information on comorbid conditions, we used the approach developed

by Greenfield and colleagues (3,4). Another measure used was the physical status classification of the American Society of Anesthesiologists (5). Using the medical record, we coded disease-specific indicators of severity and generic, as well as disease-specific, complications. To synthesize the information on complications, a panel of experts selected a subset of complications that they considered to be "serious." For these complications, we created an index representing a count of the number of these complications experienced by the patient.

Patient Questionnaire

The outcome questionnaire asked about perceived general health, number of days disabled, use of health services, symptoms related to the hip replacement, current social activities, activities of daily living, well-being, satisfaction with medical care and health, whether patients thought the operation made them feel better, whether their health was better or worse than expected, whether they felt "back to normal," employment, and role functioning, as well as indicators of socioeconomic status such as education and income. The questionnaires also had questions about condition-specific outcomes. For example, the questionnaire sent to total hip replacement patients contained questions about the amount of pain experienced doing a range of activities, degree of limping, and use of walking supports. Finally, the questionnaire also asked about daily activities, limping, use of walking supports, wellbeing, employment, and role functioning in the month *preceding* surgery.

The measures of social activities, functioning, and well-being were adapted from the Functional Status Questionnaire (6). The pain scale was also derived from a measure used by Jette and colleagues. The questions about use of walking supports and limping were developed for this study. The questionnaires were designed to be easy to read and answer, and took approximately 30 minutes to complete. The psychometric properties of the generic components of this scale, when used with different groups of surgical patients, have been described elsewhere (2).

We collected billing data primarily to look at process. We extracted comparable computerized information on 101 resource elements or process of care variables at each of the hospitals. I discuss primarily data from the patient questionnaire, but I want to emphasize again that those data are part of a system of quality assessment.

Although many studies have used patient questionnaires to assess outcomes, this study was unusual in a couple of important ways. The first is that these patient questionnaires were administered a substantial period after the hospitalization. The second is that we asked about these variables before as well as after hospitalization. What we wanted to know about was not just outcomes or simply variations in outcomes, but how postdischarge

health status differed from preadmission status and how those changes related to differences in the process of care.

I would like to emphasize that it is possible, in some cases, to get very important information with a few simple questions. Although a single question about general health status may appear to have limited validity to some clinicians, empirical studies have shown that asking such a question can elicit information related to clinical measures of health status.

Other types of issues are also easy to assess. For example, one question on the health interview survey was, "During the past month, how many days did illness or injury keep you in bed all or most of the day?" National data on how people respond to this question are available, as are data from a variety of studies in hospitals, clinics, and communities. A simple question of this kind can be extremely informative.

When we talked to orthopedic surgeons and internists, they invariably said that an important outcome for hip replacement patients is pain. They operate for pain, they try to relieve pain, they try to get people back to functioning without pain—and so we included a pain scale on our questionnaire. It is important to note that the scales used in this study are in most cases closely related to existing scales: it is not the content of the scales, but rather the way they are applied, that is different.

Orthopedic surgeons also usually say that one of their primary goals is to enable people to walk without support again. To assess this outcome, we included simple questions: "What type of walking supports do you use now? What kind of limp do you have now?" Again, this is not a long battery, it is not complicated, it is very easy to answer, and it is very easy to administer. Surgeons often do not know what proportion of their patients are still limping a year after surgery or what proportion of their patients are using walking supports, so they often find the responses to such questions very informative.

To assess psychological well-being, we used five items from the Functional Status Questionnaire (6) that are the same as those used in the Medical Outcomes Study. They include questions such as "Have you been a very nervous person?" and "Have you felt calm and peaceful?" To measure role functioning, we ask a series of questions about how the patient is doing at work or at home.

In the six hospitals study, we asked about patient satisfaction using traditional questions such as "How satisfied were you with your hospital stay in general?" I am now conducting a different study, the Picker/Commonwealth Study of Patient-Centered Care, in collaboration with Tom Delbanco at Beth Israel Hospital in Boston, Tom Moloney at the Commonwealth Fund, and a number of other colleagues at Harvard and the Commonwealth Fund. We are collecting information from a national probability sample of about 6,000 patients nationally and 2,000 of their caregivers and asking them very

specific questions about the process of care that we think one would want to know about when evaluating quality and effectiveness.

We do not usually think of patient satisfaction as a measure of outcome, but I think after hospitalization we would like one of the outcomes to be an informed, involved, cooperative patient. Thus, in the Picker/Commonwealth Study, we ask a series of questions such as: "Were you involved in the decisions about your care as much as you wanted?" "Were the important side effects of the medicines that you were getting explained to you in a way you could understand?" and so on. We have a sample of 62 hospitals nationally, and I think we will be able to make some very interesting observations about the differences among hospitals. That study is almost completed and the results should be published this fall.

Findings

One of the first things we wanted to know about our outcomes study was whether it is feasible to distribute a questionnaire like ours to patients from multiple institutions. We found that it was a very practical way of collecting information. The questionnaire we used was 30 minutes long; we probably could make it much shorter. Patient acceptance was high. In most surveys there tends to be a reluctance to participate, but in this study there was a great deal of interest in the study. Rather than feeling burdened by the questionnaire, many patients reported that they were pleased that the hospital was checking on how they were doing. We got a response rate of approximately 80 percent. About 10 percent of patients said they do not want to participate in a research study, and about 90 percent of the remaining patients returned a usable questionnaire.

It is important that measures be reliable and valid. Our scales were very reliable, with coefficients ranging from 0.64 to 0.92. For the more established scales, the reliabilities were quite high. Data on the correlations among measures and the correlations with other health measures indicate that ours are valid measures of health status. One common concern is that these scales may reflect, to a great extent, differences in general psychological well-being: that is, if patients are depressed, they will say they are doing poorly; if they are feeling good, they will say they are doing well. We did not find that to be the case in our study, probably because we focus on questions that are as concrete as possible. Questions about specific activities, such as limping and the use of walking supports, are less likely to be confounded

Another important feature of our health status measures is their responsiveness to changes in health status. For most of the conditions we studied, there is a ceiling effect; that is, everyone is doing well and the observer cannot see any difference. For total hip replacement patients, however, there was a

dramatic improvement in functioning. One could see that the changes are similar to what a clinician would predict—after hospitalization, patients' basic activities are largely back to normal.

An important question is whether it is necessary to measure different dimensions separately or whether it is possible to use a combined measure. The data demonstrate why I think it is better to measure components separately. If one measures intermediate activities separately from basic activities, one can see a very different pattern emerge. Patients who have had a total hip replacement still show quite dramatic improvements, but there is a slightly different picture for patients who have had CABG surgery. These patients show a very strong and statistically significant improvement in functioning on the intermediate level that we would not have picked up with a basic activities scale.

The data on work performance also show a different pattern. With total hip replacement patients there is a dramatic improvement in performance. That contrasts with the perplexing but fairly consistent clinical finding that CABG patients do have impaired work performance and do not return to work as much as one would expect them to.

Among the AMI and rule-out AMI patients, postdischarge functioning is worse than predischarge. Again, a separate scale picks up an important phenomenon that I think would have been obscured in a combined measure. The data on psychological well-being indicate that most patients are doing pretty well, and everyone shows slight improvement.

It is difficult to describe the relative improvement across these scales. I have taken each scale and calculated an improvement score, which is basically how they are doing before hospitalization minus how they are doing later, divided by the standard deviation of the change. Using this statistic as a gauge of responsiveness, we find that the question about limping gives us the best sense of how people are doing. Use of walking supports is not quite as good. As we would expect, there are big improvements in intermediate and basic activities: among hip replacement patients the basic and intermediate scores show a .78 correction. This provides more evidence that in the future we might be able to shorten our questionnaire.

The mental health scores did not show much change, and I frequently hear an argument that one should not include social or psychological components that are not directly related to the condition being studied. I would make a plea for not discounting such measures so quickly. First of all, mental health is a very, very important component of case mix. Another reason is that it may be very important for interpreting good and bad outcomes. For example, we are now engaged in analyzing older and younger patients, and we have found that mental health status is related to perceived health status in both groups and that it may be critical in determining differences.

CONCLUSIONS

The first conclusion from these data is that we have adequate measures for most constructs. We have available a series of comprehensive batteries or instruments. They are brief and can be made briefer. They are acceptable to patients. They meet or exceed our normal standards for reliability. They are very valid, and they are responsive to changes in health status.

Outcomes assessment should be an integral part of quality assurance activities because it is not possible to assess fully the quality of processes of care without data on associated outcomes. There has been a substantial amount of research on how to assess case mix, and there are many systems for monitoring the process of care. A fair amount is now known about variations in certain outcomes, such as mortality, and I think the main factor that limits us at this point is a lack of understanding about the linkages among case mix, process of care, and outcomes.

I would argue that if we understood these linkages better, cost containment and regulation would again become an administrative inconvenience rather than a threat to the practice of medicine as we know it today, a frequently expressed concern. We have the tools; we have the creativity; and we have the will to address these issues. It is up to us to seize the day.

References

1. Donabedian, A. Explorations in Quality Assessment and Monitoring, Volume *I. The Definition of Quality and Approaches to Its Assessment* . Ann Arbor, MI: Health Administration Press, 1980.
2. Cleary, P.D., Greenfield, S., and McNeil, B.J. Assessing Quality of Life After Surgery. *Controlled Clinical Trials*, in press.
3. Greenfield, S., Blanco, D.M., Elashoff, R.M., et al. Patterns of Care Related to Age of Breast Cancer Patients. *Journal of the American Medical Association* 257:2766-2770, 1987.
4. Greenfield, S., Aronow, H.U., Elashoff, R.M., et al. Flaws in Mortality Data: The Hazards of Ignoring Comorbid Disease. *Journal of the American Medical Association* 260:2253-2255, 1988.
5. Owens, W.D., Felts, J.A., and Spitznagel, E.L., Jr. ASA Physical Status Classifications: A Study of Consistency of Ratings. *Anesthesiology* 49:239-243, 1978.
6. Jette, A.M., Davies, A.R., Cleary, P.D., et al. The Functional Status Questionnaire: Reliability and Validity When Used in Primary Care. *Journal of General Internal Medicine* 1:143-149, 1986.

19

Studying Outcomes for Patients with Depression: Initial Findings From the Medical Outcomes Study

M. Audrey Burnam

My purpose is to. describe work that my RAND colleagues and I have conducted to examine outcomes for patients with depression. I will summarize our approach and then some initial findings from the study.

THE MEDICAL OUTCOMES STUDY

Our work was done as part of the National Study of Medical Care Outcomes (the Medical Outcomes Study, or MOS). The MOS was designed to examine the impact of different health care systems on the processes and outcomes of care for patients with specific chronic conditions. Four conditions were selected to be the focus of the study: depression, coronary heart disease, diabetes, and hypertension.

Health Care Setting

Because we wanted to understand the outcomes of care as practiced in usual circumstances and did not want to disrupt naturally occurring relationships between patients and providers, this was an observational study. Clinicians and patients were selected on the basis of the health care systems that they had chosen. As a result, there were likely to be differences in patient characteristics—for example, severity of the target condition, stage of treatment, and complicating comorbidities—that could affect outcomes, independently of the quality of care received. To estimate the effect of the health care system on outcomes in this study, it was necessary to assess patient characteristics that might affect these outcomes. The plan, then, was to control for patient differences across health care settings by statisti

cally adjusting for these differences, a strategy sometimes referred to as case-mix adjustment.

The study was designed to compare care received in three types of health care systems: (a) single-specialty small group and solo practices representing the traditional, largely fee-for-service, private practice sector; (b) health maintenance organizations (HMOs), large health care organizations representing the major prepaid alternative to traditional private practice care; and (c) large multispecialty group practices, a rapidly growing alternative that includes significant prepaid as well as fee-for-service financing. The study was conducted in three cities—Boston, Chicago, and Los Angeles—with each system of care studied at each site.

Initial Samples

More than 500 providers were recruited. They were selected to represent specialty groups providing the majority of care to patients with the four target conditions. The medical providers included in the study were internists, family practitioners, cardiologists, endocrinologists, and diabetologists. Mental health specialty providers included psychiatrists and psychologists. The outpatient practices of these clinicians provided the patient sample. Patients visiting these practices over a short period (nine days on average) were screened in the initial, baseline phase of the study to determine whether they had one of the target conditions. Persons identified by the study as having one of the targeted chronic conditions were recruited into a two-year longitudinal panel to follow their outcomes. Over 22,000 patients were screened initially.

THE STUDY OF DEPRESSION

Depression was selected to be studied in the MOS because of its importance from a health policy perspective. Some background information will illustrate this.

First, it is clear from recent epidemiological studies that depression is a very common mental disorder. One in 20 persons has experienced it at some time, and one in 40 persons is currently experiencing it (1,2). I am not referring here to transient spells of depressed mood or demoralization, but to distinct, clinically defined syndromes that are characterized by multiple and persistent symptoms and that tend to occur as repeated episodes of illness lasting from a few months to years. Second, depression has serious consequences for the affected individual and his family and for society. About 15 percent of depressed individuals commit suicide within 10 years after onset of the illness (3,4). Depression can often be socially and occupationally debilitating (5,6). Depressed persons use considerable health

care resources (7) and may present with somatic symptoms or nonspecific complaints when seeing a provider in primary care settings (8). Unless the depression is recognized and treated, inappropriate use of services is likely to result (9).

Third, most depression can be successfully treated. Sufficient evidence has accumulated to support the efficacy of a variety of pharmacological and psychosocial therapies (10).

Finally, about two-thirds of persons with depression are not receiving treatment (11). Although most people with depression do visit medical care providers (12), the literature suggests that medical providers often fail to detect depression in their patients (13).

Taking all these points together, we can hypothesize that important differences exist across health care settings in the detection of depression and the subsequent quality of care provided to depressed patients. We may further hypothesize that such differences have important implications for patients and for society.

The MOS focused on two specific types of depressive disorders, major depression and dysthymia. The definitions of these were based on the diagnostic criteria of the American Psychiatric Association. Major depression is characterized by persistent depressive mood or loss of interest in nearly all usual activities. It is accompanied by such symptoms as disturbances in appetite, weight, and sleep; psychomotor agitation or retardation; decreased energy; feelings of worthlessness or guilt; difficulty concentrating or thinking; and thoughts of death or suicide or attempts at suicide. A cluster of such symptoms must be present nearly every day for a period of at least two weeks.

Dysthymia is also characterized by depressed mood or loss of interest in nearly all usual activities. However, dysthymia lasts longer than major depression (it must last at least two years to meet diagnostic criteria) and the symptoms are less severe. The two disorders commonly coexist. That is, a major depressive episode may be superimposed upon underlying dysthymia.

Identifying Patients With Depression

Because primary care providers, in particular, may underdetect depression in their patients, it was important to base our case identification method on direct assessment from the patient. To screen over 22,000 patients for the presence of depression, we used a two-stage case identification strategy. At the first stage, we administered a very brief (eight-item) screen for depression that patients completed themselves while waiting in their providers' offices (14). To patients who exceeded a specified score, we subsequently administered a structured diagnostic interview by telephone. The interview was designed to help us determine a specific diagnosis and to collect infor

mation on history and severity of depression for use in case-mix adjustment. About one-third of those who screened positive for depression at the first stage were determined to have met criteria for current major depression or dysthymia.

Assessing Outcomes

Once we had identified depressed patients, a sample was recruited to the longitudinal study. Both generic and depression-specific outcomes were assessed periodically in the longitudinal study. Generic outcomes were assessed initially, to provide a baseline, and once every six months thereafter. The generic outcomes consisted of brief, self-administered measures of functional status and well-being that have been developed and extensively tested at RAND (15). The functioning scales encompass physical, social, and role functioning. Items on the physical functioning scale ask about limitations due to health in activities such as sports, climbing stairs, walking, dressing, and bathing. Role functioning refers to the extent to which health interferes with work, housework, or schoolwork. Social functioning is the extent to which health interferes with social activities such as visiting friends or relatives. Well-being measures include general perceptions of current health (such as feeling well or ill) and the degree of body pain experienced. There is evidence that each of these measures reliably represents a single outcome dimension (16).

Depression-specific outcomes were assessed once every year by means of a structured telephone interview. This interview elicited information on number and duration of spells of depression during the past year, including whether each spell met criteria for major depression or dysthymia. In addition, the interview determined whether a complete recovery from depression had occurred during the past year, and if so, for how long. This information was used to construct a number of outcome indicators. Some indicators reflect the current level of depression at the time of follow-up: these include type of depression diagnosis (if any) and number of current symptoms. Other indicators represent the course of the disorder during the past year: whether a recovery occurred, and in the case of recovery, whether there was a relapse (onset of a subsequent depressive episode). Finally, we examined the number and persistence of depressive symptoms during the past year.

As I mentioned earlier, to compare patient outcomes across different health care settings using an observational design, one must identify baseline patient characteristics that may affect the course of depression. In the baseline phase of the MOS, we comprehensively assessed factors that are believed to be of some prognostic significance in depression. These included demographic and socioeconomic characteristics, medical comorbidity, the presence of other psychiatric disorders (particularly anxiety disorders,

psychotic symptoms, and substance use disorders), the type and severity of depression at baseline, and lifetime history of depressive symptoms and episodes. We also, of course, controlled for generic measures of functioning and well-being at baseline.

RESULTS OF THE DEPRESSION STUDY

Findings From the Baseline Data

I would like to summarize some results from our analyses of the baseline data. We have arrived at estimates of the prevalence of depression among patients in these health care systems (17). In practices of mental health specialists, about 25 percent of visiting patients on any given day currently had depression. The treatment of this disorder thus occupies much of mental health specialty practice. The prevalence of depressed patients in practices of general medical physicians was lower, as we would expect, but even so it was strikingly high—present in 5 percent of patients. This high rate of depression in medical practices was similar for each of the three health care systems and was similar across sites. It was similar in practices of family practitioners, internists, and medical subspecialists. The rate of depression in medical outpatients is double the rate found in the general population.

We also learned that medical providers detected depression in only one-half of their currently depressed patients (18). The rate of detection was significantly lower for patients in prepaid care than for patients in fee-for-service care. These results—the high prevalence and low rates of detection of depression in medical practices—suggest that one important determinant of depression outcomes across health care settings may be the extent to which it is detected and any treatment provided.

Another set of baseline findings illustrated the importance of case-mix adjustment. Among patients with current depression, those visiting mental health providers had a more severe pattern of depressive symptoms than did those visiting medical providers (19). Depressed patients of medical providers, on the other hand, were more likely to have chronic medical conditions. The differences were not great—patients of both mental health and medical providers had, on average, severe depression, a pernicious history of past depression, and much medical comorbidity. For example, patients of mental health providers typically had 14 depression symptoms, compared to 12 symptoms among patients of medical providers. We know, however, that differences of this magnitude will have a substantial impact on the course of depression (20).

We also examined the levels of functioning and well-being experienced by patients with depression, compared to those experienced by patients with various chronic medical conditions (21). In this analysis, we estimated the

levels of functioning and well-being that were uniquely associated with depression and with each specific chronic condition (holding other factors, such as demographic characteristics and comorbidity, equal).

Figure 1 illustrates the results. The zero level on the vertical axis represents the average level of functioning and well-being of patients with no chronic medical or mental health conditions. Positive numbers along the vertical axis represent the extent to which patients with depression and chronic medical conditions have poorer functioning and well-being than those with no chronic conditions. For example, the physical functioning of patients with depression is 10.5 points poorer than that of patients with no chronic condition. The figure also shows results for some of the other chronic medical conditions that we examined—angina, advanced coronary artery disease, arthritis, diabetes, and hypertension.

The physical functioning of patients with depression is worse than that of patients with most other conditions (including diabetes, arthritis, and hypertension) but better than that of patients with advanced coronary artery disease or angina. Social functioning of patients with depression is worse than that of patients with any of the other chronic conditions we studied. Role functioning of patients with depression is about the same as for patients with angina. Depressed patients perceived their general health as poorer than did patients with most other conditions and about the same as patients with

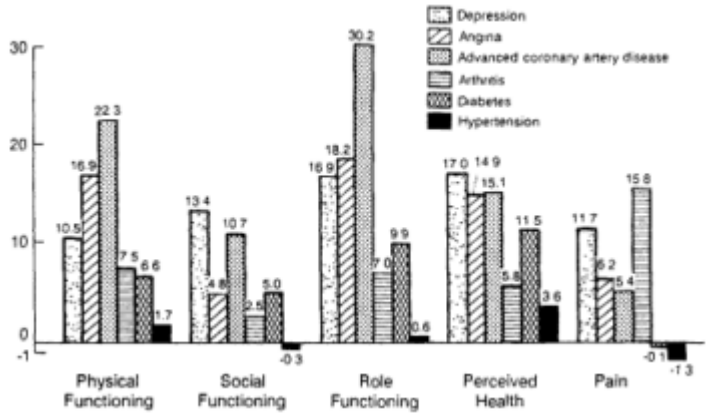

Figure 1
Levels of Functioning on Five Measures of Health Status Among Patients Enrolled in the Medical Outcomes Study.
Note: Higher scores imply poorer functioning.

heart conditions. Finally, patients with depression experience more pain than patients with most other medical conditions, except for arthritis. The overall pattern of results across these measures indicates that the functioning and well-being of depressed patients is similar to or worse than that of patients with other major, chronic medical conditions.

Besides measures of functioning and well-being, which can be affected by cognitive biases known to be associated with depression (such as pessimism), we also looked at a more "objective" measure of functioning—days spent in bed in the past month. What we found is that depression is associated with more days in bed than any other chronic medical condition except current advanced coronary artery disease.

Preliminary Findings From the Longitudinal Data

At this point, we are in a preliminary stage of analyzing the longitudinal data. We have begun to examine baseline predictors of depression-specific outcomes one year later, including the probability of recovery, and the severity and persistence of symptoms throughout the year (20). We have discovered that these measures of the clinical course of depression are quite sensitive to the severity of depression at baseline and also the severity of prior history of depression. Finally, we know that the presence of certain chronic medical conditions at baseline also affects the subsequent course of depression (22).

We have not yet compared depression-specific outcomes across health care settings, but we have learned two things that are important for undertaking these comparisons, which are the next step in our work. First, we have identified some depression-specific indicators that should be relatively sensitive outcomes for our comparisons across health care settings. Second, we have identified a number of baseline patient characteristics, particularly severity of depression, which need to be included as case-mix adjustment factors in comparisons of health care settings.

CONCLUSION

I will end with a couple of thoughts. First, it is dangerous for us to forget about mental health when we start to think about health effectiveness and outcomes. I was happy to see that, although depression is not on the short list of conditions for the HCFA initiative, it is on the long list.[1] Depressive disorder is highly prevalent in medical care settings, and there is

[1] Editors' Note: The reference is to the list of clinical conditions recommended by an Institute of Medicine committee for high priority attention in the Effectiveness Initiative. See Institute of Medicine. *Effectiveness Initiative: Setting Priorities for Clinical Conditions* . Washington, D.C.: National Academy Press, 1989.

much that we can learn from an examination of the effectiveness of care for depression. If we ignore depression, its impact on general outcomes such as functioning and well-being are nonetheless going to emerge in our studies of other health conditions.

A second issue is whether, from a measurement perspective, we are ready to begin studying outcomes as a part of health care effectiveness studies. With respect to generic measures of functioning and well-being, I agree with John Ware that we are ready to begin using generic measures in large-scale efforts.[2] There exist brief, patient-administered generic measures that have established reliability, that are responsive to changes in patient state, and that are responsive to differences across conditions. I think these measures are ready to be used. I also think that the field is ready, at least for certain conditions, to assess disease-specific outcomes.

We may not, however, be quite able to determine the factors responsible for differences in outcomes across care settings when using observational study designs. Although we want to be able to attribute outcomes to quality of care, outcomes can be a function of patient case-mix differences. To make inferences to quality of care we will have to make sure that we have controlled well for case-mix. So far, brief case-mix measures are not available, and there are difficulties in developing measures of case-mix differences. We have to isolate, from all of the possible confounding patient selection factors, those that are relevant for the specific outcomes of interest.

One way to approach this problem is to continue to do observational studies in which we have comprehensively assessed case-mix, so that we can begin to learn which case-mix factors are important. I think we can also begin to distinguish effects of case selection and effects of quality of care by looking very closely at the process of care in any study of patient health outcomes. We can have greater confidence in attributing differences in outcomes to differences in health care delivery systems once we understand how the process of care varies across systems.

References

1. Robins, L.N., Helzer, J.E., Weissman, M.M., et al. Lifetime Prevalence of Specific Psychiatric Disorders in Three Sites. *Archives of General Psychiatry* 41:949-958, 1984.
2. Regier, D.A., Boyd, J.H., Burke, J.D., et al. One-Month Prevalence of Mental Disorders in the United States. *Archives of General Psychiatry* 45:977-986, 1988.
3. Guze, S.B. and Robins, E. Suicide and Primary Affective Disorders. *British Journal of Psychiatry* 117:437-438, 1970.
4. Coryell, W., Noyes, R., and Clancy, J. Excess Mortality in Panic Disorder—

[2] For more discussion of this point and for further elaboration of the MOS, see Chapters 15-17 (23, 24, 25) in this volume.

A Comparison with Primary Unipolar Depression. *Archives of General Psychiatry* 39:701-703, 1982.

5. Stoudemire, A., Frank, R., Hedemark, N., et al. The Economic Burden of Depression. *General Hospital Psychiatry* 8:387-394, 1986.
6. Weissman, M.W. and Paykel, E.S. *The Depressed Woman: A Study of Social Relationships*. Chicago: University of Chicago Press, 1974.
7. Houpt, J.L., Orleans, C.S., George, L.K., et al. The Role of Psychiatric and Behavioral Factors in the Practice of Medicine. *American Journal of Psychiatry* 173:37-47, 1980.
8. Klerman, G.L. Other Specific Affective Disorders. Pp. 1305-1309 in Kaplan, H.I., Freedman, A.M., and Sadock, B.J., eds. *Comprehensive Textbook of Psychiatry III*, vol. 2. Baltimore: Williams & Wilkins, 1980.
9. Katon, W. Depression: Somatic Symptoms and Medical Disorders in Primary Care. *Comprehensive Psychiatry* 23:274-287, 1982.
10. Paykel, E.S., ed. *Handbook of Affective Disorders*. New York: Guilford Press, 1982.
11. Shapiro, S., Skinner, E.A., Kessler, L.G., et al. Utilization of Health and Mental Health Services. *Archives of General Psychiatry* 41:971-982, 1984.
12. Regier, D.A., Goldberg, I.D., and Taube, C.A. The De Facto US Mental Health Services System. *Archives of General Psychiatry* 35:685-693, 1978.
13. Kessler, L.G., Amick, B.C., and Tompson, J. Factors Influencing the Diagnosis of Mental Disorders Among Primary Care Patients. *Medical Care* 23:50-62, 1985.
14. Burnam, M.A., Wells, K.B., Leake, B., et al. Development of a Brief Screening Instrument for Detecting Depressive Disorders. *Medical Care* 26:775-789, 1988.
15. Stewart, A.L., Greenfield, S., Hays, R.D., et al. Functional Status and Well-Being of Patients with Chronic Conditions . *Journal of the American Medical Association* 262:907-913, 1989.
16. Stewart, A.L., Hays, R.D., and Ware, J.E. The MOS Short-Form General Health Survey: Reliability and Validity in a Patient Population. *Medical Care* 26:724-735, 1988.
17. Burnam, M.A., Wells, K.A., Rogers, W., et al. *The Prevalence of Depression in General Medical and Mental Health Outpatient Practices in Three Health Care Systems*. Santa Monica, CA: RAND Corporation, in preparation.
18. Wells, K.B., Hays, R.D., Burnam, M.A., et al. Detection of Depressive Disorder for Patients Receiving Prepaid or Fee-for-Service Care: Results from the Medical Outcomes Study. *Journal of the American Medical Association*, 26:3298-3302, 1989.
19. Burnam, M.A., Wells, K.B., Rogers, W., et al. *Severity of Depression in Prepaid and Fee-for-Service Practices of Mental Health Specialists and General Medical Providers*. Santa Monica, CA: RAND Corporation, in preparation.
20. Wells, K.B., Burnam, M.A., and Rogers, W. *One-Year Course of Depression for Adult Outpatients: Implications for Psychiatric Nosology* . Santa Monica, CA: RAND Corporation, in preparation.
21. Wells, K.B., Stewart, A., Hays, R.D., et al. The Functioning and Well-Being of Depressed Patients. *Journal of the American Medical Association* 262:914-919, 1989.
22. Wells, K.B., Rogers, W., Burnam, M.A., et al. *Are There Differences in the*

Medical Comorbidity of Depressed Patients by Type of Payment for Services and Type of Treating Clinician? Santa Monica, CA: The RAND Corporation, in preparation.

23. Ware, J.E., Jr. Measuring Patient Function and Well-being: Some Lessons from the Medical Outcomes Study. Pp. 107-119 in *Effectiveness and Outcomes in Health Care*. Heithoff, K.A. and Lohr, K.N., eds. Washington, D.C.: National Academy Press, 1990.
24. Patrick, D.L. Methodologic Issues in Assessing Health-Related Quality of Life Outcomes. Pp. 136-151 in *Effectiveness and Outcomes in Health Care*. Heithoff, K.A. and Lohr, K.N., eds. Washington, D.C.: National Academy Press, 1990.
25. Cleary, P.D. Using Patient Reports of Outcomes to Assess Effectiveness of Medical Care. Pp. 151-158 in *Effectiveness and Outcomes in Health Care*. Heithoff, K.A. and Lohr, K.N., eds. Washington, D.C.: National Academy Press, 1990.

Application to Clinical Practice: Introduction

J. Sanford Schwartz, Session Moderator

The ultimate objective of effectiveness research is to improve the health of our patients and the public. To accomplish this goal, we need to do several things: (1) we must define what we mean by effectiveness; (2) we must be able to measure effectiveness in a valid and reliable way (that is, in a way that is clinically meaningful); (3) we must be able to interpret the results in a way that will be useful to those delivering and receiving health care services; and (4) we must present the information to providers and patients in such a way that its adoption and application are facilitated.

The next four writers discuss how the results of effectiveness research can be best implemented to change provider and patient behavior, thereby improving the health of the public. They address such questions as: How does one change behavior among physicians and patients? What information is needed to address the concerns of providers and patients? Once this information is obtained, how can it be presented to patients and providers in in a way that will get them to change their practices?

Harold C. Sox is chairman of the Department of Medicine at Dartmouth Medical School. He examines the question of what to do, given valid and important effectiveness data, to modify the practice patterns of practicing physicians.

Albert G. Mulley is an associate professor of medicine and health care policy and chief of the Section of General Internal Medicine at Massachusetts General Hospital and Harvard University School of Medicine in Boston. Dr. Mulley addresses medical decision making from the perspective of patient preferences and outcomes. His chapter focuses on how to combine this information in a way that actually changes physician and patient practices.

Stephen C. Schoenbaum is deputy medical director of the Harvard Community Health Plan (HCHP). He discusses a clinical program evaluation

and management system at HCHP that attempts to measure and manage variations in clinical practice.

Eugene C. Nelson is director of quality-of-care research at the Hospital Corporation of America. In his discussion of outcome measures to improve care delivered by physicians in hospitals, he focuses on what works to improve the practice of medicine and addresses the question of outcomes measurement from a system perspective.

20

Effectiveness Research and Changing Physician Practice Patterns

Harold C. Sox

Our goal has been to learn the circumstances in which technologies are effective. Once that goal has been accomplished, however, we must meet a second goal: that of altering physicians' behavior so that they implement research findings appropriately and consistently.

My purpose is to discuss two assertions. First, it is difficult to be sure that changes in doctors' practice habits are due to published recommendations. Second, some of the resources of the Effectiveness Initiative should be earmarked for studying the factors that influence physicians to adopt new ways of practicing medicine and for testing interventions designed to promote change.

EFFECTS OF RESEARCH ON MEDICAL PRACTICE

The relationship between a specified research result and changes in medical practice is very complex. Diffusion, which is a term for the adoption of new medical technology, also applies to altered ways of using technology, such as might result from an effectiveness study. The determinants of diffusion include the following (1, pp. 178-181):

- Prevailing medical theory: A change in medical practice is more likely to be adopted if it builds on existing theory and medical logic.
- Ease of learning a new practice style: How much effort is involved in changing habits that have been ingrained and polished through years of practice?
- The importance of the clinical problem: Is the problem one that is likely to lead to death or disability for one's patients? If so, the physician is more likely to make the effort.
- Advocacy by a professional leader: There is evidence that opinion

leaders in the medical community can influence their colleagues to adopt new practices.

- Characteristics of the adopting physician: Have physicians' training prepared them to grasp new concepts quickly and see the implications for their patients? Do they have the ability to change from one style to another?
- The practice setting: Does the physician belong to a group practice in which there is a lot of peer pressure to change? Are there financial pressures to change, either to do more procedures in fee-for-service practice or to do fewer in a prepaid practice? Are the new technologies available in the practice setting?
- The physician's control over decision making: Do physicians have direct control over the decisions to acquire new technology or to make it easier or more difficult to obtain access to the technology?
- The results of formally evaluating the technology: This component is the one with which the Effectiveness Initiative is most concerned. The evidence that formal evaluation affects medical practice will be discussed later.
- The effectiveness of the channels of communication of evaluation findings: If physicians are not aware of the results of a formal evaluation of a technology, its influence will be much diminished or delayed in taking effect. Both the professional and the popular media are important in disseminating information about technology, and the influence of the popular media on patients' expectations of their physicians is a topic that is particularly neglected.

Clinical Trials

The recognized standard of evidence for clinical effectiveness is the clinical trial with randomized controls. Do clinical trials influence medical practice? Fineberg examined several studies that attempted to trace the influence of a clinical trial (1, pp. 185-195). To evaluate these studies, Fineberg first established standards of evidence that change in practice style was attributable to research results. First, what is the baseline pattern of using the technology? Is there a trend among practicing physicians that is due to factors unrelated to the research results? Second, is there evidence, perhaps obtained through surveys, that physicians are aware of the research results? Third, do the research results imply that a change in practice style should occur? Fourth, is there a temporal relationship between the assessment appearing in the medical literature and the subsequent changes in medical practice?

Fineberg applied these research standards to 28 studies, of which only ten were suitable for analysis. The others failed because the study results did not have clear implications for practice, because there were no data on practice style both before and after the assessment was published, or because

there were no quantitative data on the frequency of using the technology that was being studied.

Fineberg's study showed that only two of these ten studies contained strong evidence that the published technology assessment affected practice style. These two studies were among four in which there was evidence of a marked change in practice style. In two of these studies, the randomized trial preceded the change in practice, which is fairly strong evidence that the randomized trial had something to do with the change in practice; in the other two studies, the randomized trial did not precede the change in practice. In five studies, there were small changes that were consistent with trends in practice style, and in one there was no shift in practice style at all.

Fineberg's study shows that one of the more powerful forms of medical knowledge, the results of a randomized clinical trial, had little measurable effect on practice style.

Consensus Development Conferences

The second example shows that a program aimed at effecting change, the National Institutes of Health (NIH) Consensus Development Conferences, had little measurable effect on practice (2). The goal of these conferences is professional and consumer consensus about the best way to use a technology. The RAND Corporation studied the effect of four of these conferences on clinical practice in hospitals. Table 1 shows the four conditions studied and a selection of the recommendations of the NIH Consensus Development Conference on these topics.

The RAND investigators used the recommendations of these conferences as the standard of care against which to compare what they observed in patient records in a randomly selected sample of hospitals. They measured

TABLE 1 RAND Study of NIH Consensus Conferences

Condition	Selected Recommendations
Breast cancer	Standard is total mastectomy with axillary dissection in Stage 1 or 2
Breast cancer	An estrogen receptor assay should be performed on each primary tumor
Cesarean section	A trial of labor in low-risk women with a previous C-section
Unstable angina	Unstable angina patients should have a coronary angiogram on their first hospital admission for the condition

SOURCE: Adapted from Kosecoff et al. (2).

compliance with these recommendations in three time periods: for one year starting two years before the conference; for one year starting one year before the conference; and for one year starting nine months after the conference.

Table 2 shows the percent of cases in which there was compliance with the recommendations of the consensus conference. There was no trend toward increased compliance with the recommendation to perform total mastectomy with axillary dissection in Stage I and Stage II breast cancer. The RAND investigators studied compliance with a recommendation to test for estrogen receptors in breast cancer. There was a strong trend among practicing physicians toward increased compliance during all three periods of observation. In addition, there was a significant increase in compliance following the consensus conference, as compared with the entire period prior to the conference. The consensus conference appeared to have made a difference.

A trend could be observed throughout the three periods toward compliance with a recommendation that low-risk pregnant women with a previous Cesarean section be allowed a trial of labor. However, the consensus conference had no measurable effect on this decision. There was no trend toward increased use of angiography in patients with unstable angina and no evidence that the consensus conference had any effect.

The RAND investigators made three additional observations. First, compliance was less than 50 percent during the year following the conference for 6 of the 11 criteria. Therefore, compliance with these criteria was low. Second, there was an overall trend toward increased compliance throughout the three time periods. Thus, physicians were generally aware of the changing standards of practice, regardless of how much attention they paid to the consensus conference recommendations. Finally, when the RAND investi

TABLE 2 RAND Study of Compliance with Recommendations of NIH Consensus Conferences

	Compliance (Percent)		
Case	Period 1	Period 2	Period 3
Mastectomy and axillary dissection	74	79	84
Estrogen receptors	54	78	86[a,b]
Trial of labor	6	11	29[a]
Angiography	14	29	24

[a] $p < 0.05$ over entire period only.
[b] $p < 0.05$ for before-after.
SOURCE: Kosecoff et al. (2).

gators examined compliance for each indication and each condition (not just the sample illustrated in Table 1), the rate of change in compliance actually slowed when time period 2, before the conference, was compared to time period 3, after the conference. All in all, the NIH consensus conferences had a limited immediate effect on practice style.

Doctors do change. They no longer do gastric freezing or Halstead radical mastectomies. They discharge patients with uncomplicated myocardial infarction in one week rather than three weeks, which was the practice 20 years ago. Each of these changes is consistent with the findings in a series of empirical studies. Medical practice is moving ahead, albeit slowly, in a direction that is consistent with research results. How do we reconcile this change with our difficulty in establishing a cause-and-effect relationship between specific studies or consensus recommendations and change in practice style? The resolution of this paradox will require better understanding of the factors that influence adoption of new practice styles.

WHAT SHOULD BE DONE

The present circumstances provide an opportunity that may not come our way again soon. There has never been greater motivation to understand how to change medical practice, and there is adequate support to begin the task. We need much more research on the determinants of change in physicians' practice style, and now is the time to begin.

Second, we should look to the professional societies to identify effective clinical policies. Their recommendations should be based on research results, with clear delineation of the logic leading from the research findings to the recommendations. If the professional societies are to play this central role, they should spend some time working together on a common approach to developing guidelines. At present, no two organizations use the same methods. When two organizations come to different conclusions about the same technology, fruitful discussion about how to interpret the data may be impeded by disagreement about the methods that were used in coming to a conclusion. A common methodology should promote respect for each other's efforts and lead to useful dialogue.

We need to intensify efforts to motivate change. Organizations that pay for health care will be doing their part to motivate physicians, but we need more vigorous programs to teach physicians how to deal responsibly with fiscal pressures. The most efficient way to accomplish this task may be to aim these programs at physicians who are recognized by their peers as the opinion leaders in their professional community.

We need to keep our patients informed of research results that will affect them. Well-informed patients can exert considerable influence at the time of a close-call decision, either encouraging or frustrating attempts to practice

a lean style of medicine. There have been well-documented successes in using the popular media to influence people to reduce their cardiovascular risk. The same approach might help people to acquire a more realistic understanding of the benefits and risks of patient care technologies.

Leaders in government, industry, and medicine must look upon the Effectiveness Initiative as a long-term investment. Finding the truth about what works in the practice of medicine will be an ongoing task, in which constantly improving research methods are aimed at evolving technologies. The challenges are as daunting as those involved in basic biomedical research. It will take a decade or more for the Effectiveness Initiative to achieve research results that physicians can use to change the way that they practice medicine. These research findings will not have their intended impact unless there is an intensive effort to understand the factors that influence physicians to change.

References

1. Fineberg, H.V. Effects of Clinical Evaluation on the Diffusion of Medical Technology. Pp. 176-210 in *Assessing Medical Technology* , Mosteller, F. ed. Washington D.C.: National Academy Press, 1985.
2. Kosecoff, J., Kanouse, D.E., Rogers, W.H., et al. Effects of the National Institutes of Health Consensus Development Conferences on Physician Practice. *Journal of the American Medical Association* 258:2708-13, 1987.

21

Applying Effectiveness and Outcomes Research to Clinical Practice

Albert G. Mulley

Wide variations in medical practices for seemingly similar patients have called into question the adequacy of the knowledge base that supports clinical decision making (1). Such variations have also fueled concerns about both the cost and the quality of medical care. The research community has responded with proposals for a new focus in clinical research on outcomes of patient care, and the National Center for Health Services Research has recently announced a new program to sponsor such research (2). To meet its responsibility for ensuring the quality of care provided to Medicare beneficiaries, the Health Care Financing Administration launched a program within the Department of Health and Human Services. This Effectiveness Initiative will systematically gather information to improve our understanding of the relative effectiveness of alternative therapeutic approaches to conditions that commonly afflict Medicare beneficiaries (3).

My purpose is to examine the methodological issues associated with the application of effectiveness and outcomes research to clinical practice. What are these issues? The answer depends on how one defines effectiveness and outcomes research and how one distinguishes them from the clinical research that has informed, however well or poorly, clinical practice until now.

WHAT IS DIFFERENT ABOUT EFFECTIVENESS AND OUTCOMES RESEARCH?

First, the Effectiveness Initiative and outcomes research are motivated by concerns about the quality and cost of medical care. New knowledge alone may not be enough to declare success. The current wave of enthusiasm and support will be sustained only if the initiative has a demonstrable impact on quality, cost, or both.

Second, the initiative insists on measurement of outcomes that are important to patients. Five-year survival is too coarse a measure. Physiological measures are fine, but they are often irrelevant. However, detailed measures of health states and of subjective responses to those states will be just as irrelevant unless they can be communicated to the persons responsible for clinical and policy decisions.

Third, effectiveness and outcomes research must recognize clinical practice as a source of information in the production of new knowledge rather than as a passive and not always attentive consumer of knowledge produced by a separate enterprise called clinical research. This distinction between more traditional research and the new initiative is evident in the use of claims data and other administrative data bases to capture more of the collective experience of clinical practice (3-5).

Each of these premises about how outcomes and effectiveness research are different from traditional research raises a different set of methodological issues. The first set of issues is related to the need for dissemination of information. How do we get new information about effectiveness and outcomes to the point where it can have a positive impact on quality or a restraining effect on inflation? It is no accident that each of the bills supporting outcomes research also has provisions for practice guidelines.

The second premise, that the new research is different because of a focus on patient-oriented outcomes, complicates matters. Results of this new research must include the subjective responses of patients that determine their quality of life, as well as the trade-offs between quality and quantity that are acceptable to them. Communication of such subjective value judgments involves a set of methodological issues that must be addressed if we are to preserve the responsiveness of health care to the wants and needs of individual patients.

The third premise raises pragmatic issues about the possible integration of research and practice, not to mention epistemological questions. Which elements of clinical investigation are essential for valid inferences to be drawn about effectiveness? When can collected experience in clinical practice be an acceptable, or even preferable, source of information? Methodologists will recognize this as a variant of the tension between the "internal validity" of a clinical study and the limits that requirements for internal validity place on the "external validity," or generalizability, of a study finding (6).

A CLINICAL EXAMPLE: OUTCOME PROBABILITIES IN DECISION MAKING

These methodological issues will be less abstract in the context of a clinical example. A 72-year-old man, married and sexually active, has increasing symptoms of benign prostatic hyperplasia. He gets up twice

each night to void, and during the day he voids frequently, with a sensation of urgency. The patient experiences a clinical process, and the result is a health outcome that we can define simply—or in great detail, if we include the physical, psychological, and social dimensions.

Any simple model that goes from patient, through process, to outcome is too deterministic. There is no single, discrete clinical process that is uniquely suited to a particular patient. The path followed through clinical practice is the result of a decision or series of decisions (in this case, whether to proceed with prostatectomy), with the outcomes contingent on each decision being more or less uncertain (see Figure 1). This may seem too obvious to belabor so, but the critical link between "outcomes research" and effectiveness is the ability to make valid comparisons between outcomes produced by the alternative pathways. This may be forgotten by the outcomes researcher assembling a cohort from claims data procedure codes or the guideline developer whose frame of reference is a particular procedure rather than a particular condition. The irreducible uncertainty, or stochastic element of medicine in any individual case, may be forgotten by the quality assurance reviewer who equates a bad outcome with bad care.

It is worth taking a closer look at the decision-making process (Figure 2). The patient faces, with the help of his or her physician, a choice between two alternative treatment strategies. The first alternative is a bit risky because the eventual outcome is uncertain. Although there is a chance that it will produce the most valued outcome (in this case, relief from symptoms), there is also a chance that it will produce the outcome that is least valued (operative death). An intermediate outcome is also possible (for example, impotence). The second alternative is less risky: the only possible outcomes are the most valued and an intermediate health state that happens to be the

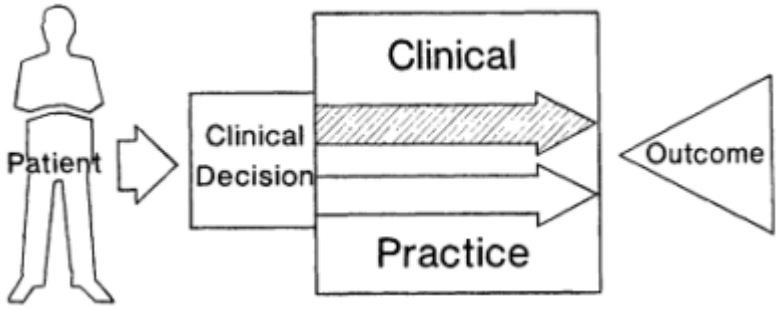

Figure 1
A health outcome can be viewed as the product of a clinical process that begins with a decision or series of decisions about the intervention(s) most likely to meet a particular patient's health care needs and wants.

patient's current health state. A more elaborate model of the prostatectomy decision has been developed, but this simpler model suffices to illustrate the process (7).

What is needed to make this choice? First, the patient and physician need to know how likely each of the outcomes is. These probabilities can be depicted as pie diagrams, as seen in Figure 3, where our hypothetical 72-year-old is referred to as Patient A. Alternative 1 has a 90 percent chance of producing the most valued outcome, a 1 percent chance of catastrophe, such as operative death, and a 9 percent chance of a bad but not fatal result, such as incontinence or impotence. Alternative 2, which looked so good without these numbers, now looks less promising: there is only a 10 percent

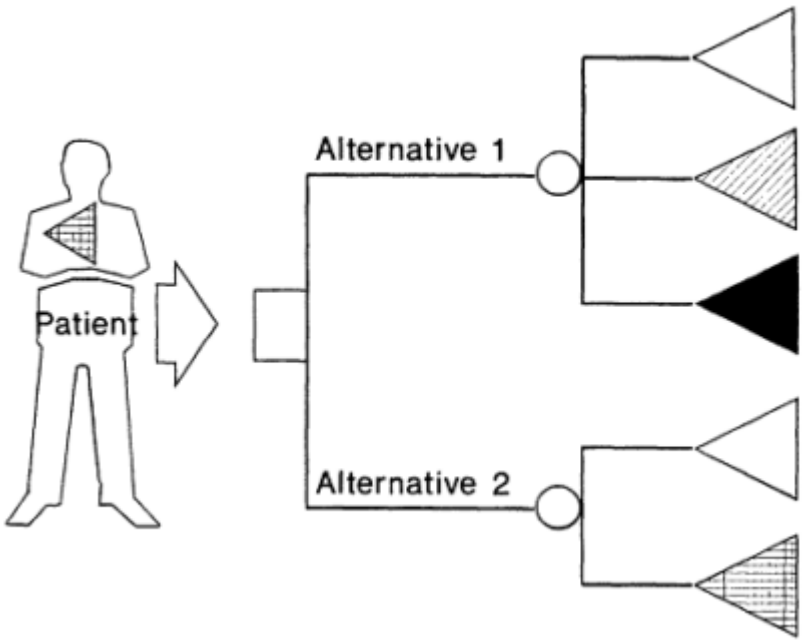

Figure 2
An Abstract Representation of a Simple Clinical Decision. The cross-hatched triangle superimposed on the patient represents the health state that has prompted medical care and the current decision. The square node represents a choice between Alternatives 1 and 2. Alternative 1 offers a chance, indicated by the round node, of dramatic improvement (represented by the white triangle) but with a risk of death (the black triangle) or a serious complication (the diagonally hatched triangle). Alternative 2 offers a chance of improvement with the only other outcome the baseline symptom state.

chance of improvement— the odds are 9 to 1 that the health state that was bad enough to bring the patient to the doctor will persist. It can be said that knowledge is power because it confers the capacity to predict. Accurate estimation of outcome probabilities, as represented in these simple pie diagrams, captures the essence of professional knowledge related to the practice of medicine.

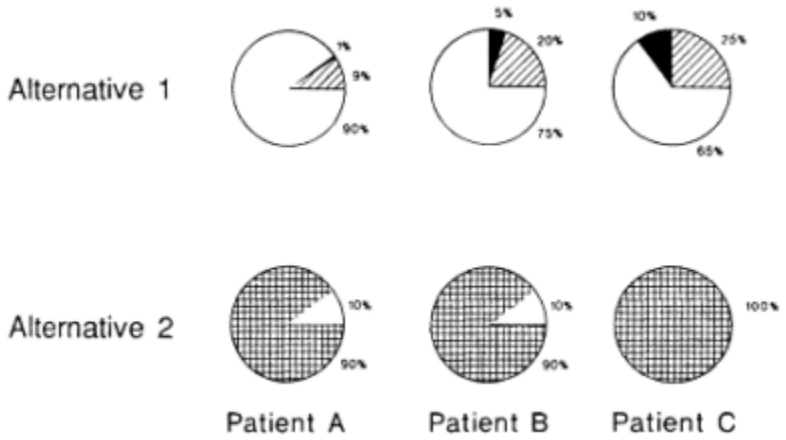

Figure 3
Outcome Probabilities. Probabilities of each of the outcomes in Figure 2 for both alternatives for three hypothetical patients (including the 72-year-old man cited in the text, here designated patient A).

SOURCES OF PROBABILITIES AND A ROLE FOR OUTCOMES RESEARCH

Where does this knowledge come from? The most obvious source of probabilities is the experience of previous patients. This constitutes the "clinical experience" of the provider that is so important to "clinical judgment." There are problems, however, with this source of information. First, there are problems with the way clinicians characterize individual patients. Second, clinical practice is not standardized. Interventions are not carefully defined and uniformly applied. Third, there is no routine mechanism to define outcomes with the appropriate level of detail or to aggregate and organize the information that could be derived from collective clinical experience. Without such systematic aggregation and analysis, the cognitive heuristics that we all use routinely may mislead the clinician's unaided, intuitive probability estimate (8).

Recognizing these problems, the profession relies heavily on published

clinical research when it is available. The randomized trial is the standard against which other clinical studies are measured. Information about patients entering the trial is systematically collected. The group is made homogeneous by applying exclusion and inclusion criteria. The alternative interventions are carefully defined and their elements carefully segregated. Outcomes, at least one or two of the more objective outcomes, are carefully catalogued. The scientific requirements of research designed to determine the effectiveness of one intervention relative to another, which is nothing more than the relative outcome probabilities, include: similarity of the initial states; the integrity of the interventions; and similarity of detection or measurement of outcomes.

Unfortunately, clinical research that meets these requirements is the exception rather than the rule. In the case of benign prostatic hypertrophy there are no randomized trials. Studies published in English describe outcome probabilities for very few men with symptoms who elected not to have surgery, and there are many methodological problems that call the accuracy of these few data into question (9).

Even when well-conducted randomized trials are available, problems arise in using the results to estimate outcome probabilities. Clinicians may forget about differences between the circumstances of the clinical trial and the circumstances of clinical practice. They may also forget about the patients excluded from the clinical trial. These exclusions are not trivial; they commonly represent more than 90 percent of the patients for whom the intervention would be used in practice.

The exclusions are also important because different patients face different outcome probabilities, even when the care rendered is identical. Figure 3 represents different outcome probabilities for three hypothetical patients. Clearly, a choice made by or for one of these patients should be based on probabilities derived from the experience of similar patients. Any inference about the effectiveness of a particular intervention must adjust for different mixes of patients with different outcome probabilities.

Outcomes research is an opportunity to integrate clinical research and clinical practice (Figure 4). Obviously, there will still be a place for rigorously controlled trials. What outcomes research gives up in terms of internal validity it more than makes up for in enhanced external validity and relevance to clinical practice. We need also to characterize patients, that is, determine disease severity, comorbidity, and other variables that affect prognosis. We need to characterize the processes of care. We must in addition describe and sort outcomes by the alternative care processes used and by patient type. Each of these tasks presents a challenging set of methodological issues that we must deal with if we are to realize the potential of outcomes research.

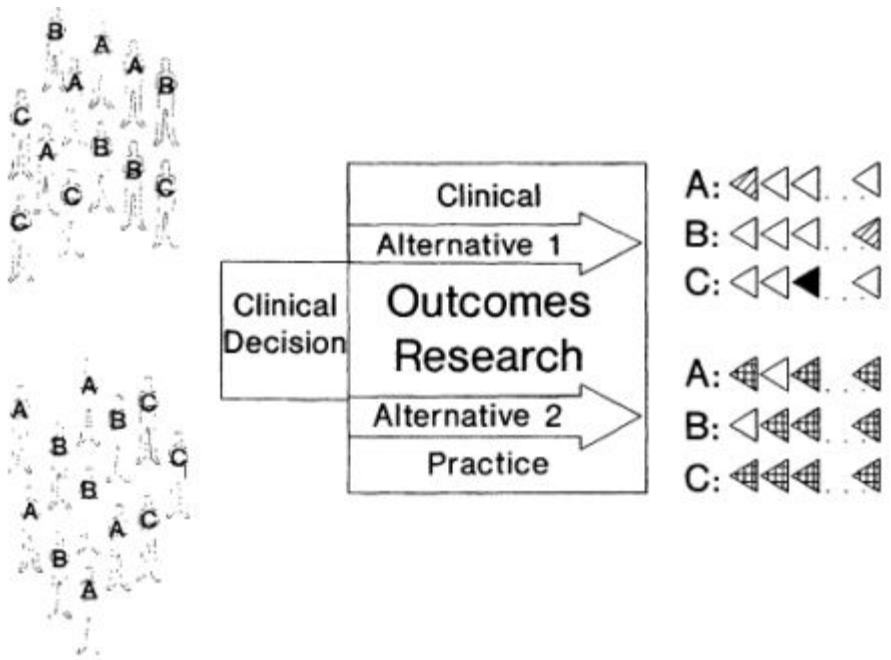

Figure 4
A Model for the Integration of Outcomes and Effectiveness Research with Clinical Practice. Patient characteristics, including disease severity and comorbidity, must be well understood; interventions must be described and their integrity maintained; and outcomes must be monitored and described in an unbiased manner so that outcome probabilities can be defined for different alternatives and different patient subgroups.

VALUE JUDGMENTS IN DECISION MAKING

Outcomes and effectiveness research has the potential to improve dramatically the clinician's ability to estimate clinically relevant outcome probabilities. Probabilities alone, however, are insufficient for informed decision making. Whether the pie diagrams in Figure 3 represent probabilities of outcomes for a health care decision or a simple game of roulette, information about the likelihood of the outcomes must be accompanied by information about their relative values in order to be helpful to a decision maker.

The top bar in Figure 5 represents a scale on which we can register the value judgments of the hypothetical patient with prostate disease. It is anchored by the least and most desirable outcomes. The markings on the scale indicate that he prefers his current state to one that would be imposed by a complication of Alternative 1 (for example, impotence). This patient might, therefore, opt for the less risky Alternative 2. The bottom two scales display different value judgments of different hypothetical patients; these patients are similar enough to face the same outcome probabilities, but with different preferences. For the second patient, the same health state diminishes life's quality more; for him, Alternative 1 may be preferable despite the risks. For the third patient, Alternative 1 would almost certainly be the best

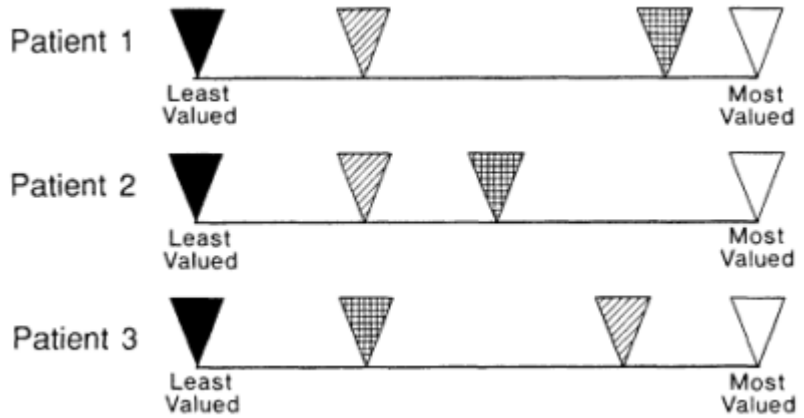

Figure 5
Value Judgments for Three Patients. Value judgments for three hypothetical patients, all of whom face the same outcome probabilities. Patient 1 values the baseline health state highly (the cross hatched triangle), relative to the state associated with a complication of alternative 1 (the diagonally hatched triangle [e.g., impotence]). Patient 3 prefers the latter to the former.

choice. The current health state is perceived as a serious hardship, and the state associated with a complication of Alternative 1 is not.

How confident can a patient be about these value judgments? He or she may be more confident in making a determination about the goodness or badness of a state that he or she has experienced than one that must be imagined. Such imaginings may be helped by hearing about the experiences of other patients. Physicians can provide such vicarious experience, but it severely tests their communication skills. Furthermore, there is no systematically collected body of experience on which to draw.

VALUE JUDGMENTS AND THE ROLE OF OUTCOMES RESEARCH

The assessment of values or preferences is extraordinarily difficult and raises a new set of methodological issues (10). As indicated in Figure 5, preferences for the same health states vary widely among patients. This has been demonstrated in a number of important studies that used hypothetical case scenarios (11,12) and in a large patient interview study of men undergoing surgery for prostate disease (13). Varying medical practice to reflect accurately these differences is both appropriate and desirable. These value judgments also change over time and are influenced by the context of the decision or the scaling task used. Even when preferences are measured accurately, there are difficulties in communicating them to other patients who might benefit from them. At this interface between outcomes research and clinical practice, the methodological issues relate more to the physician-patient relationship and its effect on care and outcomes than to the scientific basis of medicine as defined by the biomedical model.

New information about subjective responses to health states could also be of value to policymakers. It could help bridge the gap between the statistical person of cost-effectiveness analysis and the real patient when making coverage decisions or choosing those conditions for which restrictive boundary guidelines may be more or less appropriate.

CONCLUSIONS

The methodological issues in the application of effectiveness and outcomes research to clinical practice depend on the form that the new research takes in the coming years. Dissemination of results to decision makers will be of critical importance. Clinicians must be provided with information that will allow them to estimate accurately the outcome probabilities for different patients. Clinicians and their patients must be provided with information that will allow them to make informed value judgments about different potential health outcomes.

A more ambitious view of outcomes research would see it take full advantage of clinical practice as a source of information to inform future practice (Figure 6). Aggregate outcomes of individual decisions would inform professional knowledge regarding outcome probabilities and would inform value judgments made by professionals and patients. In this case, the list of issues expands to include accurate baseline description of patients, measures to ensure the integrity of the therapeutic interventions used, and the unbiased monitoring and measurement of outcomes, including patients' subjective responses. Closing this loop of practice and research will require unprecedented cooperation between clinicians and investigators, but both have much to gain in the form of a more robust and relevant knowledge base for the practice of medicine and the delivery of health services.

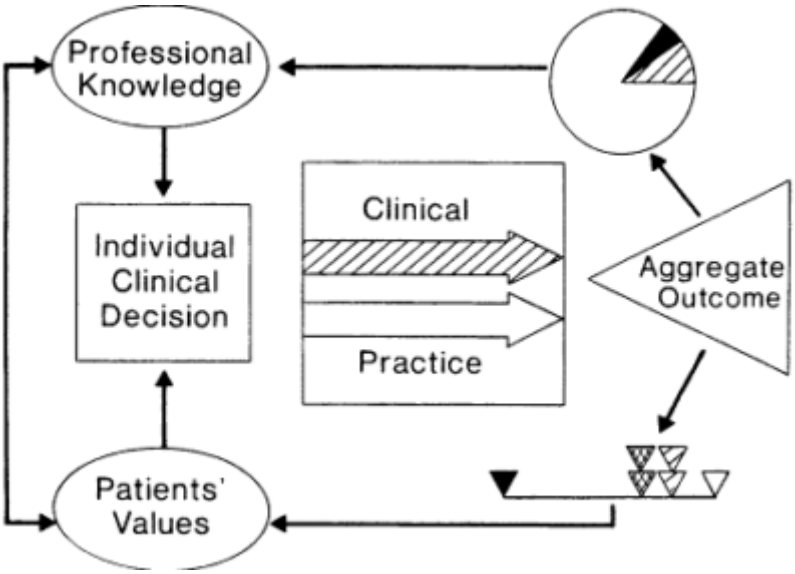

Figure 6
A Model of Clinical Practice and Outcomes Research Functioning as a Feedback Loop. Aggregate outcomes of many individual clinical decisions serve as an information base that informs professional knowledge with outcome probabilities and simultaneously informs patients' and professionals' value judgments with previous patients' subjective responses to those outcomes.

References

1. McPherson, K., Wennberg, J.E., Hovind, O.B., et al. Small-Area Variation in the Use of Common Surgical Procedures: An International Comparison of New England, England, and Norway. *New England Journal of Medicine* 307:1310-1314, 1982.
2. Patient Outcomes Assessment Research Program: Extramural Assessment Teams. *NCHSR Program Note*. Rockville, MD: U.S. Department of Health and Human Services. November, 1988.[1]
3. Roper, W.L., Winkenwerder, W.L., Hackbarth, G.M., et al. Effectiveness in Health Care: An Initiative to Evaluate and Improve Medical Practice. *New England Journal of Medicine* 319:1197-1202, 1988.
4. Wennberg, J.E., Roos, N.P., Sola, L., et al. Use of Claims Data Systems to Evaluate Health Care Outcomes. Mortality and Reoperation Following Prostatectomy. *Journal of the American Medical Association* 257:933-936, 1987.
5. Fisher, E.S. and Wennberg, J.E. Administrative Data in Effectiveness Studies: The Prostatectomy Assessment. Pp. 80-89 in *Effectiveness and Outcomes in Health Care*. Heithoff, K.A. and Lohr, K.N., eds. Washington, D.C.: National Academy Press, 1990.
6. Rothman, K. *Modern Epidemiology*. Boston: Little, Brown, 1986.
7. Barry, M.J., Mulley, A.G., Fowler, F.J., et al. Watchful Waiting vs. Immediate Transurethral Resection for Symptomatic Prostatism: The Importance of Patients' Preferences. *Journal of the American Medical Association* 259:3010-3017, 1988.
8. Tversky, A. and Kahneman, D. The Framing of Decisions and the Psychology of Choice. *Science* 211:453-458, 1981.
9. Barry, M.J. Medical Outcomes Research and Benign Prostatic Hyperplasia. In: *The Prostate*, in press.
10. Mulley, A.G. Assessing Patients' Utilities: Can the Ends Justify the Means? *Medical Care* 27(3) Supplement:S269-S281, 1989.
11. Sackett, D.L. and Torrance, G.W. The Utility of Different Health States as Perceived by the General Public. *Journal of Chronic Diseases* 31:697, 1978.
12. McNeil, B.J., Weichselbaum, R. and Pauker, S.G. Fallacy of the Five-Year Survival in Lung Cancer. *New England Journal of Medicine* 299:1397, 1978.
13. Fowler, F.J., Wennberg, J.E., Timothy, R.P., et al. Symptom Status and Quality of Life Following Prostatectomy. *Journal of the American Medical Association* 259:3018-3022, 1988.

[1] Editors' Note: Now the Patient Outcomes Research Teams of the Agency for Health Care Policy and Research.

22

An Attempt to Manage Variation in Obstetrical Practice

Stephen C. Schoenbaum

The work presented here is not research; rather, it is clinical program evaluation and management. Compared to the conduct of a controlled clinical trial, this approach is awkward. It is, however, a practical attempt to measure variation in clinical practice and to manage the apparent variation. The scenario below is a specific example from which it is possible to draw more general lessons about outcomes measurement and management.

THE HARVARD COMMUNITY HEALTH PLAN EXPERIENCE

Several years ago, Donald Berwick, Vice-President for Quality of Care Measurement at Harvard Community Health Plan (HCHP), decided that as part of his quality measurement activities it would be important to develop data bases on common clinical activities. The data bases would contain information on outcomes of interest. They would also contain some additional variables describing the population and processes of care. Dr. Berwick's hopes were that these data bases could be used to analyze outcomes in relation to process and that the data could be adjusted for population differences so that data from different sites within HCHP and external data might be compared.

The Choice of Obstetrical Care

Health maintenance organizations (HMOs) have large numbers of young members of childbearing age, and it was logical that the first HCHP data base should be one on obstetrical care. The Health Centers Division of HCHP is currently a 275,000-member staff model HMO with care delivered in 10 centers around the Boston area. In 1986-1987, the first year of the

obstetrical data base, there were eight health centers, seven of which used one of two Harvard teaching hospitals with large obstetrical services (Table 1). To obtain data for the period July 1, 1986 through June 30, 1987, HCHP staff reviewed and abstracted the hospital and ambulatory records of all HCHP members who had a delivery.

TABLE 1 HCHP Deliveries July 1, 1986 - June 30, 1987

	Health Center/ Hospital Used							
Deliveries	1/A	2/B	3/A	4/B	5/A	6/C	7/A	8/A
Number	785	627	489	510	497	180	376	118
Total per hospital								
A 2265								
B 1137								
C 180								

Table 1 shows the number of deliveries among members of each health center and the hospital (A, B, or C) used by each center. Only center 6 used a nonteaching hospital distant from Boston; because it had a relatively small number of deliveries, I will not consider it further. The aggregate number of HCHP deliveries in hospitals A and B is substantial, even though in both instances, HCHP deliveries were less than one-third of the total deliveries in each hospital.

Variations in Practice Between Two Hospitals

Hospitals A and B have fully developed academic departments of obstetrics and gynecology headed by full professors at Harvard Medical School. They are believed to have comparable outcomes in terms of neonatal mortality for comparable populations, although hospital A serves as a regional perinatal center, takes care of more referred high-risk obstetrical patients, and has a much larger and more sophisticated neonatal intensive-care unit (ICU). Neither HCHP, the hospitals themselves, nor the state's department of vital statistics can provide appropriate, comparable neonatal mortality figures.

Only in rare instances of extremely high-risk mothers does HCHP distribute patients to hospital A because it can provide more intensive care. Almost always, the hospital of delivery is determined by the health center in which the member receives her prenatal care. In 1986-1987, each HCHP health center had its own department of obstetrics and gynecology; that is, each center had its own chief and staff, and each arranged its own coverage schedule. To the extent that cross-center coordination of schedules and

combined educational sessions occurred at all, they occurred along the lines of hospital use.

TABLE 2 HCHP Deliveries July 1, 1986 - June 30, 1987, by Type of Delivery and Health Center and Hospital Used

	Deliveries (%) per Health Center/Hospital Used						
Type of Delivery	1/A	2/B	3/A	4/B	5/A	7/A	8/A
Spontaneous vaginal	57	76	59	78	57	55	59
Cesarean section	27	20	22	19	26	28	21
Forceps	15	1	18	1	15	16	18
Vaginal birth after cesarean section	1	3	1	2	2	1	2

TABLE 3 HCHP Deliveries July 1, 1986 - June 30, 1987, by Type of Delivery

	Deliveries (%) per Hospital	
Type of Delivery	A	B
Spontaneous vaginal	57	77
Cesarean section	26	19
Forceps	16	1
Vaginal birth after cesarean section	1	3

Table 2 shows one of the initial analyses from the obstetrical data base. Type of delivery was the first object of attention, and it is the focus of subsequent attempts at intervention. Substantial variation is noted in the percentage of women in the various centers who had a spontaneous (nonoperative) vaginal delivery. This is due to higher rates of cesarean section and forceps deliveries in some centers than others.

The lowest rates of operative deliveries are for health centers using hospital B (see Table 3). The occurrence of forceps deliveries is much lower in the hospital with the lower cesarean section rate; this is an unexpected finding, since one might predict that in order to avoid a cesarean section for problems such as cephalopelvic disproportion, the physician would have to extract the baby with forceps. In addition, vaginal birth after cesarean section was somewhat more commonly performed in hospital B. The difference in rates between the two hospitals is even more dramatic if one considers only those women who are primiparas—that is, those having their first

delivery. In hospital A, 29 percent of primiparas had a cesarean section vs. 22 percent in hospital B. In hospital A, an additional 24 percent of primiparas had a forceps delivery vs. only 2 percent in hospital B. All of these findings suggested to us that practice style might differ significantly between the two hospitals. There are no "right" rates of operative delivery, but we believe that the differences should not be so great.

Possible Causes of the Variations

A preliminary version of these data was shown to the center-based chiefs of obstetrics. They reasoned that there *are* differences between health centers and that, since they and their staffs were all *equally* competent, the differences in rates of operative deliveries must be due to differences in membership. Table 4 highlights these differences and shows the distribution by health center of obstetrical patients who are very young, relatively older, nonwhite, or not married.

Although substantial center-to-center differences exist, no single characteristic correlates with a high cesarean section or forceps delivery rate. A multivariate analysis with a large number of potential confounding variables was unable to demonstrate any important contributor to the observed variation in type of delivery by health center other than the hospital at which the delivery occurred.

Another consideration is whether the variation in delivery rates was related to characteristics of the offspring. When the data were adjusted for low birthweight (which is the only adverse neonatal characteristic that occurs

TABLE 4 Patient Characteristics and Types of Delivery of HCHP Members, by Health Center and Hospital Used

	Members (%) per Health Center/Hospital Used						
Measure	1/A	2/B	3/A	4/B	5/A	7/A	8/A
Patient characteristic							
Less than 18 years old	3.0	0.7	1.1	1.9	0.5	3.6	—
Over 35 years old	11.2	13.5	18.4	6.1	4.7	5.8	6.5
Nonwhite	48.6	24.2	2.6	7.1	9.5	51.9	29.0
Not married	24.0	9.4	3.6	8.5	4.8	37.2	23.3
Type of delivery							
Cesarean section	27	20	22	19	26	28	21
Forceps	15	1	18	1	15	16	18

with enough frequency in this population to permit adequate analysis), the variation in type of delivery by hospital persisted.

Cesarean section is associated with significant maternal morbidity, including infections, increased length of stay, and higher hospitalization costs. In recent years, despite continued increases in cesarean section rates nationwide, it has not been possible to show a continued concomitant improvement in neonatal outcomes. Accordingly, although little information on neonatal outcomes in HCHP patients in hospitals A and B existed, it was reasonable to consider operative deliveries a relatively independent outcome of obstetrical care.

There can be several determinants of outcome, including host (that is, the patient), environment in which care is delivered, and the process of care itself, the latter being within local control. We hypothesized that variation in operative delivery rates ought to have some relationship to the process of care, and we hypothesized that the following components of process of care might affect type of delivery:

- Prenatal education,
- Obstetrical care (prehospital, in-hospital),
- Nursing care,
- Anesthesia care, and
- Other

In a series of interviews we tried to determine from our chiefs of obstetrics and from others in hospitals A and B what the differences in the process of care might be for patients in these institutions. They identified several areas:

- Location of labor and delivery suite,
- Ratio of nurse to patient in labor,
- Obstetrical policies (labor curves, forceps),
- Epidural anesthesia in labor, and
- Relationship of HCHP obstetricians to hospital obstetrics department.

We did find differences in these areas. The labor and delivery suite in hospital A is below ground, windowless, relatively noisy, and unattractive compared to that in hospital B. There was a nurse-to-patient ratio of 1:2 or greater for the labor suite in hospital A, compared to a 1:1 ratio at hospital B. In both hospitals the chairman of the department of obstetrics gave strong guidance to clinical policy but in hospital B a graphics tool was used to follow the progress of labor (with a strict definition of failure to progress) and the use of forceps was frowned upon and required specific justification. Consequently, the staff of hospital B did not get much instruction or experience in the use of forceps and may have been less comfortable than the average staff in using them. In contrast, a type of forceps instrument was developed

at hospital A many years ago, and there was no injunction against using them.

Anesthesia practices also differed. In hospital A, the practice was to use epidural analgesia for most patients in labor (about 70 percent) and not to reduce use of it prior to delivery. In hospital B, only about half the patients received epidural analgesia in labor, and the practice was to keep use of it light, especially as labor progressed.

The obstetrical staff also differed. The HCHP obstetricians in hospital B had, for the most part, trained there and were considered "insiders," whereas the HCHP obstetricians in hospital A tended to have trained elsewhere and were more likely to be considered "outsiders." HCHP obstetricians in hospital A had poorer morale than HCHP obstetricians in hospital B. The chairman of obstetrics in hospital A had expressed concern about the coordination of care for HCHP patients by HCHP obstetricians. Finally, residents interacted closely and directly with HCHP obstetricians for virtually all HCHP patients in hospital B, but only for the highest risk or most complicated patients in hospital A.

One thing that did not differ between the two hospitals was patient satisfaction (except for the rating of the ambiance of the labor and delivery suite in hospital A).

The Attempt to Change Process of Care

To attempt to alter the process of care for HCHP patients in hospital A, HCHP felt that it needed a multifaceted programmatic intervention. Such interventions do not necessarily take the form of those that might be incorporated into controlled clinical trials. The first intervention was to appoint a single Plan-wide chief of obstetrics and to arrange for him to have an office in hospital A. The person who was appointed happened to be HCHP's most senior and experienced center-based chief. He had been trained at hospital B, but he had also practiced in hospital A a decade earlier and was respected by the chairman in hospital A. He had moved back to hospital B and had developed the departments of obstetrics in the two health centers using hospital B.

Over the next year, the workload of the new chief was very heavy. It included an enormous effort to recruit new obstetricians for a growing HMO and to improve HCHP's central infertility services in response to marked increases in demand. (In that year, Massachusetts mandated infertility benefits of all health insurers, including *in vitro* fertilization services.) The new central chief also began to work on changing the process of care in hospital A.

Several things occurred almost simultaneously. The chief, in order to recruit successfully, convinced HCHP to increase obstetrical salaries. This, and his successful personal interactions with subchiefs and staff, seemed to

improve morale among existing obstetricians. Three older obstetricians using hospital A ceased obstetrical practice and devoted themselves to gynecology. Thus, a high percentage of the persons doing obstetrics for HCHP at hospital A became direct recruits of the chief and could be thought of as "his people." He also recruited a senior perinatologist from the academic staff of hospital B to work for HCHP, based primarily at hospital A. This had at least two effects: it led the HCHP obstetricians at hospital A to realize that their performance was being monitored more closely (which might have an effect on issues such as continuity of care), and it provided them with an experienced and friendly consultant who could support them in a tough decision to wait it out with a patient rather than moving quickly to a cesarean.

The chief also began to work directly with the anesthesia staff of hospital A. In April 1988 an agreement was reached, in the form of a memorandum from the clinical chief of anesthesia to the entire obstetrical staff (HCHP and others), that obstetricians could have a say in the degree of analgesia provided to their patients in labor. The chief also worked to make nurses in hospital A aware that HCHP obstetricians might want longer and more forceful pushing by their patients in labor than had been the usual practice in hospital A in the past.

Preliminary Results

Table 5 shows *preliminary* results. The burdens of obtaining data for the obstetrical data base were sufficiently great that no additional data were collected for deliveries between July 1987 and September 1988. Data collection resumed in the fall of 1988, but only for a 50 percent sample of deliveries; the data shown in Table 5, therefore, are for only a small number of patients. They may not prove stable, and we will not have additional data, on January

TABLE 5 HCHP Deliveries in 1986-1987 and in October-December 1988, by Type of Delivery

	Deliveries (%) per Hospital			
	A		B	
Type of Delivery	1986-1987 (N = 2265)	1988 (N = 319)	1986-1987 (N = 1137)	1988 (N = 143)
Spontaneous vaginal	57	68	77	76
Cesarean section	26	21	19	19
Forceps	16	9	1	1
Vaginal birth after cesarean section	1	2	3	4

through March 1989, until approximately November 1989, because of a continued backlog of work. Nevertheless, as we look at the available data, we see an encouraging trend toward fewer operative deliveries in hospital A; anecdotal evidence suggests that the trend is continuing.

TABLE 6 HCHP Deliveries For Primiparas in 1986-1987 and in October-December 1988, by Type of Delivery

	Deliveries (%) per Hospital			
	A		B	
Type of Delivery	1986-1987 (N = 1213)	1988 (N = 168)	1986-1987 (N = 586)	1988 (N = 67)
Spontaneous vaginal	47	59	76	77
Cesarean section	29	25	22	22
Forceps	24	16	2	1

TABLE 7 Type of Delivery Among Candidates for Repeat Cesarean Section at HCHP

	Deliveries (%) per Year	
Type of Delivery	1986-1987 (N = 268)	1988 (N = 37)
Cesarean section	87	63
Vaginal birth	13	37

Table 6 shows a decrease in the rate of operative deliveries in primiparas, and Table 7 shows that the practice of vaginal birth after cesarean section is also increasing at HCHP. There has been a significant contribution from hospital A, and this, too, will lead to decreased cesarean section rates.

LESSONS

As I stated at the outset, this is not research, but rather a description of work in progress to manage variations in practice by altering process of care. What lessons might we derive from it?

First, the data collection process is difficult and expensive. Even though HCHP has automated medical records for all but two centers, and thus easy access to prenatal records, the information presented here does not come from routinely collected data.

It is hard to collect and analyze the data for enough potential confounders

to satisfy doubters that demonstrated variation is not attributable just to some unanalyzed confounding variable.

Process interventions in the real world are complex. They tend to be different from situation to situation, institution to institution. They often do not come directly from controlled trials, and they may take the form of appointing a new chief, or firing an old one, or changing a reimbursement scheme, or threatening to move one's business to another vendor or hospital. The complex intervention leads to multiple changes rather than to the single, sometimes unrealistically simple, changes of controlled trials.

Another lesson is that the unit of data collection and organization for process improvement is relatively small by statistical or epidemiological standards. This makes it hard to sort random from significant variation—it is not easy to convince caregivers or managers that there is a problem worth working on. It also makes it harder to analyze interventions—results will not necessarily be statistically convincing, as they are in a controlled trial, and the results will be confounded by time-trend differences in care that would have occurred anyway.

Despite these problems, it may be important to act in the face of apparent variation, as we did, even without the most solid data. Such action needs to be accompanied by a commitment to watch, nonjudgmentally, what happens over time. Some observers will undoubtedly feel we acted too fast; others may be uncomfortable in concluding at this time that we are making a difference.

Another instructive point from this example is that we took a very different tack in assessing variation in cesarean section rates than we might have taken if we had followed the appropriateness approach (as The RAND Corporation did in its work on variation in surgical practices). Table 8 shows the reasons for cesarean sections in hospitals A and B, as extracted from the records. Although the distributions differ somewhat, over 50 percent of the procedures are attributed either to cephalopelvic disproportion or failure to progress, 20 percent to breech presentation, and 15 percent to fetal distress. I believe that there is enough softness in the definitions of cephalopelvic disproportion, failure to progress, and fetal distress that an independent group of experts assessing *only* the records of patients having a cesarean section in the two hospitals would have concluded that a similar percentage of the cesarean sections in the hospitals was "appropriate." The departments of obstetrics in these hospitals have regular reviews of their own cesarean sections and rarely conclude that one is inappropriate.

Accordingly, had we followed the appropriateness approach to assessing variation in surgical practices, we might have concluded that the overall difference in cesarean rates between these two hospitals was most likely due to some occult underlying difference in the populations rather than to a difference in the process of care. Although it is important to eliminate truly

inappropriate procedures, some, perhaps many, procedures *that in retrospect are judged appropriate* may actually be unnecessary.

TABLE 8 Reason For Cesarean Section (% of Total Per Hospital)

	Hospital	
Reason	A	B
Breech	17	23
Failure to progress	36	23
Cephalopelvic disproportion	19	29
Fetal distress	16	14
Multiple pregnancy	1	0
Placenta previa	1	2
Abruptio placentae	2	1
Maternal indications	2	1
Herpes	3	3
Other	3	4
TOTAL	100	100

A CAUTION

I would like to end on a word of caution for those who think that outcomes measurement and management are "the way to go." It is clearly important to assess what we are doing in medical care and to try to determine how, in real life and real time, we can do it better. These efforts are, however, going to be slow, difficult, and costly. There is a whole science of program evaluation that needs to be developed and learned by providers of health care. Experienced epidemiologists will need to be recruited to these efforts. Journals will need to begin to report program evaluations so that we can learn from them.

Leaders in health care will have to learn what is realistic and what is not. There are important priorities to be set and the Effectiveness Initiative is a step in this direction. Experts have to put their heads together to consider what information might be obtained from routinely collectable data and what information might be collected at a low marginal cost.

Most important, health care experts and providers will have to learn how information, once obtained, can be used to generate process improvements. The trick is in getting from Health Care Financing Administration mortality data or HCHP cesarean rates to *some* intervention. I believe that regulatory approaches are not very conducive to ongoing, creative, process improvement,

although regulation or accreditation may play an important role in getting the process going. In general, process improvements require skilled assessment and skilled management. Until and unless the efforts I have just mentioned occur, "outcomes" is just a buzzword. It will wear thin and disappear from our vocabulary. That would be very unfortunate, for we will have lost a major opportunity to examine, evaluate, and improve the way we give medical care.

ACKNOWLEDGMENTS

I wish to thank the following persons on the HCHP obstetrical data base staff who developed and maintained this data base and who kindly made available the data in this chapter: Kay Larholtz, statistical specialist; Debra Cookson, project coordinator; Diana Parks Forbes, obstetrical database consultant; and Donald M. Berwick, Vice-President for Quality of Care Measurement.

23

Using Outcome Measures to Improve Care Delivered by Physicians and Hospitals

Eugene C. Nelson

The question "What works in the practice of medicine?" is very important. It is largely methodological and focuses on measurement. Yet an even more critical question is this: What works to improve the practice of medicine? It is one thing to use measurement to find out what works, but it is quite another thing to know what to do to improve that work.

"If you always do what you always did, you will always get what you always got." This simple saying, spoken by a factory worker to W. Edwards Deming, the father of continuous improvement, makes that point (1). Improvement in outcomes requires change upstream in the process. Measurement is part of a process of change—it can help the process get started in the right direction and monitor the effect of efforts, but measurement alone will not create improvements.

If effectiveness is to be increased, process improvement thinking must be included while constructing outcomes measurement systems. The challenge is not to create outcomes measurement systems, but to construct outcomes measurement/improvement (MI) systems for use by clinicians, hospitals, and other health care organizations. In this chapter, I will cover four points briefly. First, I describe two MI systems for medical practices. Second, I introduce two MI systems for hospitals. Third, I highlight the hallmarks of these systems, and fourth, I offer guidelines for using outcomes measures to make improvements.

Before moving to point number one, I wish to illustrate the concept of an outcomes measurement/improvement system. Figure 1 illustrates an MI system for individual patients. The cycle begins with a patient visiting the physician or entering the hospital. The patient's baseline health outcomes are measured (disease-specific measures, general health status indicators, and patient expectations for care) and assessed by the clinician; the patient's

regimen is planned; care is implemented and follow-up is instituted; and outcomes measures are periodically gathered. The cycle continues making adjustments as the patient's status changes.

Figure 1
A Measurement/Improvement System for Individual Patients

OUTCOMES MEASUREMENT/IMPROVEMENT SYSTEMS FOR MEDICAL PRACTICES

The Coop Charts

The first two systems I describe might be thought of as early attempts to develop MI systems for doctors' offices. One of these is the Dartmouth COOP Chart system. The following are the vital facts about the system:

What?	Illustrated posters of health status
How?	Patient rates health
	Patient scores self
	Resource guide for clinician to prompt a regimen
Benefits?	Better communication
	Discovery of important problems
	Ease of use

The COOP Charts (Figure 2) are similar to the Snellen charts that physicians have used in their offices for decades to test vision quickly. In fact, the Snellen charts were the inspiration for the COOP Charts (2). The idea was to construct simple charts that could be used to measure some 10 key dimensions of overall health rapidly—physical function, mental health, social function, pain, quality of life, and so on. Recently, we have begun to link COOP measurements with a functionally oriented resource guide. The

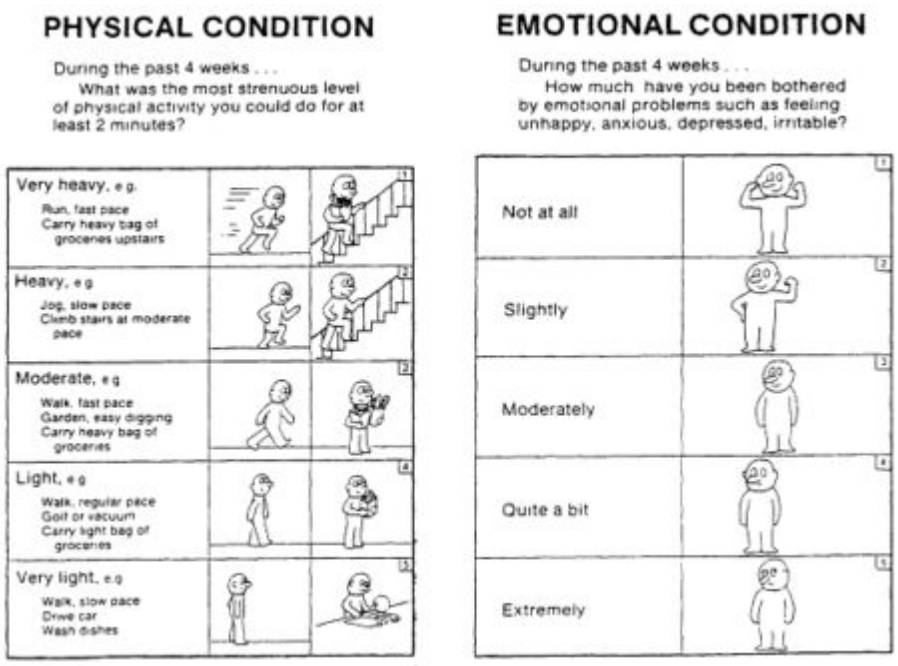

Figure 2
Illustrations of Two COOP Charts
Source: Copyright Trustees of Dartmouth College/COOP Project 1988.

objective is to link measurement of the patient's functioning with suggestions for improvements that the physician can use to plan the regimen, thereby building improvement into the process of patient care delivery.

Studies of the COOP Charts show that they are very easy to use in busy medical practices, that they are reliable and valid, and that use of them has several benefits (3,4). Both patients and physicians believe that the charts improve communication and frequently lead to the discovery of important problems that would otherwise be missed. In a study conducted in about a dozen medical practices, physicians said that when the COOP Charts are used for case-finding, new, important information is produced for approximately 25 percent of patients; physicians also said that this leads to new treatment in two of five of these patients, providing a better fit between the patient's problems and the physician's plan of treatment. In addition to case-finding, COOP Charts can be used to monitor the overall functioning of patients with serious chronic diseases. Research suggests that the charts are able to show what impact discrete medical events, such as falls and adverse drug reactions, have on the patient's basic physical and mental function. Thus, use of the charts may help the doctor to understand better the effect of disease on the "whole" patient and thus to deliver more comprehensive care.

The COOP Chart system for measuring and improving health outcomes holds great promise. The Henry J. Kaiser Family Foundation is sponsoring a large randomized trial at the Harvard Community Health Plan to document the system's case-finding utility in clinical practice, and the charts are being field tested in 20 countries to determine their value in other parts of the world.

THE RUBENSTEIN FUNCTIONAL HEALTH STATUS APPROACH

The second MI system for use in medical practice was developed by Lisa Rubenstein and her colleagues at the University of California, Los Angeles. The following are vital facts about the system:

What?	Questionnaire on health status
How?	Patient rates health
	Computer scores and profiles patient
	Resource guide for clinician
Benefits?	Better mental health
	Better social function

The development of the Rubenstein functional health approach is an interesting and important story for anyone interested in improving outcomes. Several years ago, Dr. Rubenstein and colleagues at UCLA, BIAC (Beth Israel Ambulatory Care Center), RAND, and Harvard collaborated on a

randomized trial. Their goal was to show that functional assessment of elderly patients visiting the offices of general internists would improve outcomes of care (5). The measurement strategy was based largely on the short-form general health status tools developed by John Ware and his colleagues at RAND. Two rather large randomized trials were conducted, one in Los Angeles and the other in Boston (6,7). The results were disappointing. Measurement of functioning of the internists' patients did nothing to improve outcomes a year later.

In analyzing the reasons for these negative results, the investigators discovered that the MI cycle had been broken. Patients' baseline functioning had been measured and the results placed in the medical records; however, there was very little evidence that physicians had used this new information to add to their assessment or to plan treatment. As noted earlier, measurement of outcomes alone may produce no gains: "If you always do what you always did, you will always get what you always got."

Dr. Rubenstein conducted a second randomized trial using the same measurement tool, but this time adding a function-oriented resource guide to the system. The resource guide was designed to link the patient's problem with specific treatments that would be appropriate and effective. It provided site-specific "tips" on what the physician might do for an elderly patient with a physical disability such as poor balance or a mental health condition such as depression. The results from this second randomized trial, which included more than 76 physicians and 571 patients, were positive. Patients in the test group had significantly better mental health and social activity scores than patients in the control groups who received customary care after one year (8). This time, the entire measurement/improvement cycle had been completed, and patients' outcomes had improved.

MEASUREMENT/IMPROVEMENT SYSTEMS FOR HOSPITALS

The first MI system for hospitals that I will discuss is being used in my organization, Hospital Corporation of America (HCA), and other hospitals around the country.

HCA Patient Judgment System

Here are the vital facts on the Hospital Quality Trends (HQT) Patient Judgment System:

What?	Random sample of patients rating hospital quality and health status on questionnaire
How?	Patient rates quality and health status
	Computer scores and profiles hospitals
	Results show improvement opportunities

Benefits?	Trends in quality over time Benchmarks across hospitals Linked to process improvement method (FOCUS-PDCA)

The HQT Patient Judgment System was developed by a multidisciplinary design team that included practicing physicians, hospital administrators, nurses, and quality research leaders such as Paul Batalden, Donald Berwick, and John Ware from HCA, Harvard, and RAND, respectively. The system was tested in eight hospitals in 1987 and is now in use in approximately 100 hospitals. About 65 of these hospitals are owned by HCA; the others are a mix of large and small voluntary hospitals. An article describing the system was published in the June 1989 issue of *Quality Review Bulletin* and a monograph summarizing the development work is in press at *Medical Care* (9,10).

The aim of the system is to provide hospitals with valid, reliable, and *useful* trends in hospital quality, based on the voice of the patient. A random sample of discharged patients judges 10 dimensions of hospital quality (for example, admissions, nursing, physicians, information, daily care, and discharge) that are measured with a 68-item questionnaire. Patients also evaluate their health benefit from the stay and complete selected COOP Charts showing postdischarge functioning. Each hospital receives reports twice a year. The reports use graphic techniques to reveal longitudinal trends in quality. Hospitals use the reports to monitor trends and to identify (or focus on) high priority areas for improvement. These areas can then be addressed by Quality Improvement Teams using a structured improvement method, FOCUS-PDCA, that takes advantage of the scientific method in planning and managing process improvement (11).

An example of how this MI system is used can be found at West Paces Ferry Hospital in Atlanta, Georgia. The senior leadership team there identified those aspects of quality that were most important to patients—clinical outcome, nurse response time and caring, nurse skill, the admitting process, and the discharge process—and that were candidates for improvement. The leadership then "chartered" several Quality Improvement Teams (composed of members from different departments involved in the process). It challenged every department in the hospital to identify which of its processes influence these key areas of quality and to begin FOCUS-PDCA on one or more of them. West Paces will use the HQT Patient Judgments System to monitor the overall success of its quality improvement efforts.

In October 1989, 50 hospitals began using the HQT system for adult psychiatric patients. In addition to everything in the system described above, it includes the clinician's assessment of mental health at admission and discharge plus the patient's rating of his or her own physical function, mental health, and quality of life at admission, discharge, and one month after discharge.

SOUTH SHORE HOSPITAL GERIATRIC ASSESSMENT AND PLANNING PROGRAM

The final example of a measurement/improvement system is the Geriatric Assessment and Planning (GAP) program. A thumbnail sketch follows:

What?	Functionally oriented hospital record system for managing and following elderly patients
How?	Nurse rates functioning at admission, midstay and discharge Functional ratings are linked to treatment plan
Benefits?	Replaces nursing notes Basis for comprehensive discharge planning Better match between patient function and treatment plan Frail patient follow-up after discharge

The GAP system was developed by leaders at South Shore Hospital, Carolee DeVito and William Zubkoff, with the assistance of external consultants in functional assessment such as Paul Densen and Charlotte Hamill (12). The purpose of the GAP program is to provide a standard method of comprehensive patient assessment that will enable the hospital to improve the match of its services to the changing needs of elderly patients (13).

Starting in about 1983, South Shore began modifying the processes for admitting, nursing, and discharge planning to include full assessment of the patient's clinical and functional status at admission, midstay, and discharge. The GAP program involves all patients age 65 and older admitted to the hospital. Assessment includes standard data on such aspects of health as clinical parameters; Activities of Daily Living; Instrumental Activities of Daily Living; social, emotional, and cognitive function; and continuing care needs after discharge. The entire caregiving team—physicians, nurses, discharge planners, and home health professionals—builds and uses the assessment/ management form to update the patient's status and to match services to patient needs. Patients with continuing care needs who are discharged to their homes are checked to see if the ordered services are being delivered, if their needs have changed, and if they need to be "relinked" with services.

The GAP system is being extended and applied to new areas. For example, it serves as the backbone of a major demonstration program sponsored by the Centers for Disease Control to prevent falls leading to hip fractures in frail elderly patients.

HALLMARKS OF MEASUREMENT/IMPROVEMENT SYSTEMS

It is probably fair to say that none of the MI systems discussed above possess all of the desired features needed to be as good as it could possibly be. It is a fact that all systems can—and should—be improved continuously

(14). Nevertheless, these systems share certain characteristics that medical practices, hospitals, health maintenance organizations, and other providers could use to both measure and improve outcomes. Chief among them are the following:

- commitment by leaders in the provider organization to use measurement to foster improvement,
- valid and reliable measures of outcomes,
- systematic, repeated assessment of outcomes,
- easy to fit into day-to-day pattern of care delivery,
- ease of administration, scoring, and interpretation of measures,
- directly linked between outcomes measures and improvement efforts,
- direct benefit to individuals and groups of patients,
- high value placed on system's utility by patients and clinicians
- ability to pass information "up-line" and to aggregate it for multisite efficacy studies and appropriate comparisons, and
- ability to compare outcomes against those of other providers.

These features, when combined into a working system that is part and parcel of the caregiving routine, can be very powerful. Such a system creates a new way of processing and using measures to manage and improve outcomes. In the right environment—one that promotes cooperation on quality improvement—clinicians can work together to improve the system.

USE OF OUTCOMES MEASURES TO BENEFIT A PATIENT POPULATION

The use of outcomes measures in the aggregate to benefit an entire patient population, as opposed to benefiting an individual patient, produces special challenges. The improvement cycle for a patient population is illustrated in Figure 3. The cycle begins with a population of patients with a selected health problem or condition. Measurements of structure, process, and outcomes are taken and then the relationships among them are analyzed to attempt to determine the "best" upstream settings (elements of structure) and the "best" upstream actions (processes) that appear to yield the "best" downstream results (outcomes). A field trial of the "best" upstream conditions is conducted to determine if they will produce the desired results in multiple settings. Finally, if the results are positive, this new information is disseminated to providers.

Even a casual comparison of this cycle with the simpler one for individual patients (Figure 1) shows that it is far easier to make improvements for an individual patient than for a population of patients. It is still harder to construct an outcome MI program that can help improve an entire system of care composed of autonomous health care providers.

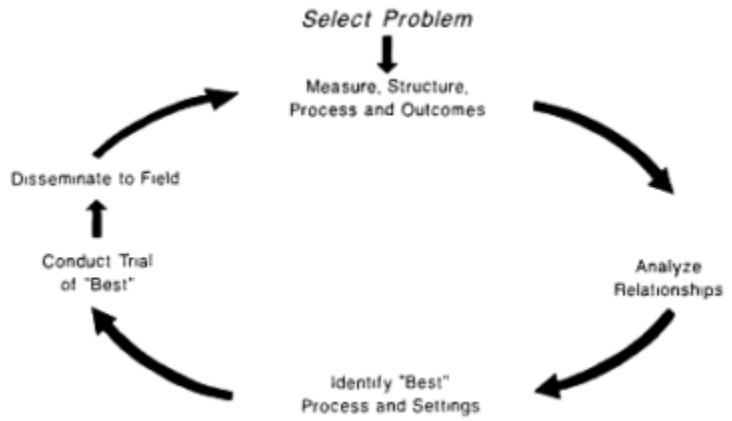

Figure 3
A Measurement/Improvement System for a Population of Patients

Recognizing that the challenge—that is, how best to use outcomes measures for improvement—is very great, one might be wise to look outside the health care industry for guidance. There one would find a new way of thinking about what quality is and how best to improve it that stresses continuous improvement of processes (15). One tool that is being used widely in quality improvement circles is an activity called "benchmarking." A recent book by R.C. Camp, an executive at Xerox, describes what benchmarking is and how to practice it (16). Camp defines benchmarking this way: "Benchmarking is the search for industry's best practices that lead to superior performance." The term "best practices" is equivalent to the term "best processes" and the term "superior performance" is analogous to "superior outcomes." Hence the purpose of benchmarking is to search upstream for the best processes that lead to superior outcomes. Note that the aim is not to find out who is best able to achieve superior ends. Rather, the goal is to spot superior outcomes as a way of flagging providers who employ outstanding processes that might be adapted for use in one's own organization.

Benchmarking, in my opinion, could be a powerful vehicle for improvement in health care if it is a voluntary, provider-based, "from-the-bottom-up" activity. Benchmarking could succeed if it is undertaken with zeal by physicians, hospitals, and other providers as a search for the conditions and processes that are most likely to produce the best outcomes.

Benchmarking, however, is unlikely to be helpful in health care if it is imposed from the top down. In fact, such a strategy for benchmarking might be counterproductive. Why? There are many reasons: (1) top-down benchmarking does not begin with the genuine need felt by most providers

to find a "better way"; (2) it is likely to produce fear, a desire to protect one's own position and to discredit the information and its source; (3) the focus will be on the ends—the outcomes—as opposed to the process and the means for achieving the end; and (4) top-down benchmarking is likely to foster blind competition among providers rather than useful cooperation.

With these thoughts about the potential power of benchmarking and some sense of the pitfalls if it is launched in the wrong way, I would like to offer a few guidelines on how to use outcomes measures for improvement.

- When measuring outcomes over time, one must measure related *upstream conditions* in order to understand the outcomes measures.
- It is essential to separate the *technical results* (often termed clinical end points or parameters) from the benefits desired or achieved by patients.
- Strive to understand all relevant *upstream conditions* (settings, processes, practices, and events) when interpreting outcomes measures.
- Identify the *key features* of the upstream conditions most likely to yield superior outcomes and conduct a trial to determine if the new way is more efficacious than the old.

CONCLUSION

Measuring outcomes is important. Improving outcomes is even more important. Outcomes can be improved by developing dual-purpose measurement/ improvement systems that are useful for individual patients, physicians, and other providers of care. These systems should link measurement of health outcomes directly with the care-giving process. They can best be assembled using a bottom-up, rather than a top-down, approach. This will be more likely to stimulate the curiosity of providers to make constructive clinical comparisons and thereby discover better ways for continuously improving patient care.

References

1. Deming, W.E. *Out of the Crisis.* Cambridge, MA: MIT Center for Advanced Engineering Study, 1988.
2. Nelson, E.C., Conger, B., Douglass, R., et al. Functional Health Status Levels of Primary Care Patients. *Journal of the American Medical Association* 249:3331-3338, 1983.
3. Nelson, E.C., Wasson, J.H., and Kirk, J.W. Assessment of Function in Routine Clinical Practice: Description of the COOP Chart Method and Preliminary Findings. *Journal of Chronic Diseases* 40(Supplement 1):55S-63S, 1987.
4. Nelson, E.C., Landgraf, J.M., Hays, R.D., et al. *The COOP Function Charts: A System to Assess Functional Health Status in Physicians' Offices* . Final report to

the Henry J. Kaiser Family Foundation. Hanover, NH: Dartmouth Medical School, 1987.
5. Jette, A., Davis, A., Cleary, P., et al. The Functional Status Questionnaire: Reliability and Validity When Used in Primary Care. *Journal of General Internal Medicine* 1:143-149, 1986.
6. Rubenstein, L.V., Calkins, D.R., Young, R.T., et al. Improving Patient Functional Status: Can Questionnaires Help? *Clinical Research* 34:835a, 1986.
7. Calkins, D.R., Rubenstein, L.V., Cleary, P.D., et al. The Functional Status Questionnaire: Initial Results of a Controlled Trial. *Clinical Research* 34:359a, 1986.
8. Rubenstein, L.V., McCoy, J.M., Cope, D.W., et al. Improving Patient Functional Status: A Randomized Trial of Computer-Generated Resource and Management Suggestions. Paper presented the annual meeting of the American Federation of Clinical Research, Washington, D.C., May 1989.
9. Nelson, E.C., Hays, R.D., Larson, C., et al. The Patient Judgment System: Reliability and Validity. *Quality Review Bulletin* 15:185-191, 1989.
10. Meterko, M., Nelson, E.C., and Ruben, H.R. Patient Judgments of Hospital Quality: Report of a Pilot Study. *Medical Care,* in press.
11. Batalden, P.B. and Buchanan, E.D. Industrial Models of Quality Improvement. Pp. 133-159 in *Providing Quality Care: The Challenge to Clinicians.* Goldfield, N. and Nash, D.B., eds. Philadelphia: American College of Physicians, 1989.
12. W.K. Kellogg Foundation. *Patient Assessment for Continuing Care: Executive Summary. Westchester Patient Assessment Program.* Battle Creek, MI: W.K. Kellogg Foundation, 1987.
13. DeVito, C.A. and Zubkoff, W. Discharging the Frail Elderly: One Hospital's Model Program. *Continuing Care* 42:26-31, 1989.
14. Berwick, D.M. Continuous Improvement as an Ideal in Health Care. *New England Journal of Medicine* 320:53-6, 1989.
15. Walton, M. *The Deming Management Method.* New York: Dodd, Mead & Company , 1986.
16. Camp, R.C. *Benchmarking: The Search for Industry's Best Practices that Lead to Superior Performance.* Milwaukee, WI: ASQC Quality Press and White Plains, NY: Quality Resources, 1989.

PART V

WHERE DO WE GO FROM HERE?

24

The Need for Reasonable Expectations

Henry J. Aaron

I should like to begin by posing a question. Then, I shall simply take a couple of points and beat the living daylights out of them.

Suppose one were given a multiple choice question, a very easy one with only two alternatives. The question reads as follows: "This conference and the work that preceded it have occurred because (a) key decision makers have become devoted to the improvement of knowledge about the linkage between medical interventions and medical outcomes OR (b) key decision makers have become persuaded that many medical interventions are useless and that effectiveness studies will document ineffectiveness and sharply lower medical expenditures."

The best answer to that question is "Both." But if one were forced, in the fashion of the Educational Testing Service, to choose the better answer, it would have to be b.

Most people involved in effectiveness and outcomes studies were drawn by scientific curiosity, unsullied by great concern about the cost issues. They want to see improved medical care and effective use of resources to promote improved health. In fact, many have been voices crying in the wilderness on this issue for years, if not decades. Others are relative newcomers, drawn into the field of effectiveness analysis by funding for it, which is newly abundant and may, if Senator Rockefeller gets his wish, become still more abundant in the future (1).

But my question was framed in terms of why the conference occurred and the work that preceded it occurred.

The problem of effectiveness in medical care has been around for a very long time. And despite the need for care in framing questions and in thinking about how they should be posed to patients and providers, the techniques involved in carrying out effectiveness research have, by and

large, also been around for a very long time. So I think one has to ask why the push for effectiveness research is coming only now.

The answer to that question, I think, is that the people who determine budgets in Congress, in the executive branch, and perhaps even, to some degree, in foundations think that the studies of effectiveness will save a lot of money and ameliorate or solve the vexing problem of rising medical costs, and that such studies will thereby render unnecessary most of the rather difficult choices that rising costs seem to pose for the general population.

THE LIKELIHOOD OF UNMET EXPECTATIONS

The theme of my remarks is that this expectation is almost certain to be frustrated and that the hope of avoiding the difficult questions is almost certain to be disappointed. If I am right, we face some very difficult problems involving what to do if the results of effectiveness studies, on balance, would boost rather than cut costs.

The first point I would stress is that a clearly defined, precise benefits curve such as the one Uwe Reinhardt lays out (2) is not really the right way to envision the problem. In fact, in the minds of individual practitioners that curve is a wide range of very fuzzy curves. Furthermore, those curves are not lines at all; rather they are shadowy expanses along which benefits rise as the intensity of care increases, until they reach some point beyond which they turn down. The point at which they turn down is a matter about which disagreement is widespread, deep, and passionate.

The aim of effectiveness research, of course, is to convert those shadowy blobs into something that looks more like a line. That process will lead, in some cases, to less care, in other cases to more care, and probably in a large number of cases to different care that may be roughly as costly as what we have now.

From the other chapters in this volume, I glean exactly the answer I expected to the question of whether implementation of the results of effectiveness studies would raise or lower costs: No one is really quite sure. "Some things will go up; some things will probably go down; we have to run the numbers to find out. And even then we may not be sure because the studies now under way include only a tiny part of the universe of possible studies."

The second reason I think expectations are bound to be disappointed is that, even if the direct result of effectiveness research is to save money on certain forms of care, the net saving will be reduced by the cost of the additional therapies that would prove necessary, either currently or at some time in the future. To illustrate the difficulty of deciding whether something reduces costs or not, consider the case of antibiotics. Did they by and large reduce or increase the cost of medical care? The initial response, of course,

is that they reduced costs. The correct answer, I think, is that they increased costs enormously by extending lives and enabling people to become ill from much more costly diseases at some time in the future.

A third reason that hopes for savings will be disappointed involves time. Effectiveness research will go on for decades. The results will accrue slowly. Even if, on balance, the results achieve the cost reductions that the most bullish supporters claim they will do, these results are going to come in over a period so long that I would suggest they are going to be almost undetectable against the background of other forces affecting medical care expenditures.

WHAT EFFECTIVENESS RESEARCH CAN DO

All of this leads me to conclude that effectiveness analysis will and should be expected to have no detectable effect on the rate at which health care spending changes in the United States. It promises something far more important than that, however: it promises improvements in the efficacy with which we use medical care resources. It promises an improvement in the quality of medical care.

I think the truth of the matter is that most of the people involved in the Institute of Medicine's effectiveness effort are involved for the right reasons. But the forces that led to the particular timing of this effort are predicated, at least in some degree, on expectations that are going to be disappointed in the future. If so, this disjunction between hope and reasonable expectation raises an acutely difficult problem for persons who believe, correctly, that effectiveness research is worth doing. If those persons tell funders what they want to hear, they are going to be lying and the funders will find out sooner or later. If those persons tell funders the truth, they risk cooling the enthusiasm that makes the research possible.

The latter course is the one I think people are going to have to accept. I must confess that I make this forecast hesitantly—after all, persons who advocate the former may have as much success as the advocates of competition have enjoyed, being able to live for years and years on unfulfilled promises of cost reductions.

References

1. Rockefeller, J. The Legislative Perspective. Pp. 44-48 in *Effectiveness and Outcomes in Health Care.* Heithoff, K.A. and Lohr, K.N., eds. Washington, D.C.: National Academy Press, 1990.
2. Reinhardt, U. The Social Perspective. Pp. 34-37 in *Effectiveness and Outcomes in Health Care.* Heithoff, K.A. and Lohr, K.N. eds. Washington, D.C.: National Academy Press, 1990.

25

Use of Effectiveness Research in Managed Care Plans

Howard L. Bailit

My approach to effectiveness research is from the perspective of the group health insurance industry, particularly that segment of the industry that operates managed care plans. I address four issues in this chapter:

1. The importance of effectiveness research to the group health insurance-managed care industry,
2. Current applications of effectiveness studies in managed care plans,
3. The contributions group health insurers can make to effectiveness research, and
4. Limitations of the effectiveness "strategy" in controlling health care costs.

IMPORTANCE OF OUTCOMES RESEARCH

It is important to emphasize that the group health insurance industry is under great pressure to control the rate of increase in health care costs. This issue is now the primary concern of employers because they cannot cope with 15- to 20-percent rate increases year after year. This pressure to control costs is causing major changes in the health insurance industry. First, a consolidation is taking place as insurers that are unable to provide employers with effective medical management services go out of business.

Second, insurers are concerned about the possibility of national health insurance or other government interventions to control costs that will adversely affect the industry. This is of special concern now because some of America's largest companies are advocating national health insurance. Traditionally, employers have been against greater government control of the delivery system, but some are becoming skeptical that the private sector can success

fully address the cost problem. As a result, insurers and others in the managed-care business have to demonstrate that they can control health care costs. Further, they must find a solution within the next five to seven years to prevent further government controls.

Within this general environment, insurers have two basic cost control strategies available to them. One is to increase patient cost-sharing in hopes of reducing utilization by fostering more prudent purchasing of health services. Until recently, patients' out-of-pocket costs were staying constant in real dollars. Employees were protected from the rapid increases in costs by having employers allocate a larger share of their total compensation to health benefits. Now, data from the Health Insurance Association of America suggest that cost-sharing is starting to increase, and this trend is expected to continue.

Greater cost-sharing alone is unlikely to solve the problem of rising costs. For one thing, this nation is about to enter a period of severe labor shortages, and companies may compete for skilled workers with richer benefit plans. Also, many Americans feel very strongly about their health benefits and will probably not tolerate major increases in out-of-pocket costs. This can be seen in the recent strikes at Pittston Coal and AT&T, where workers were not willing to accept reduced health benefits. In sum, then, some modest increases in cost-sharing will occur, but this option will probably not solve the problem of rising costs.

A second strategy open to the private sector is to establish a more competitive delivery system through the development of health maintenance organizations (HMOs) and other managed care plans. The basic idea is that by carefully selecting cost-effective providers, giving them an appropriate level of risk sharing, and carefully monitoring utilization, health care costs can be controlled. Risk-sharing is an important element in this strategy because it helps focus the attention of providers on the efficient use of resources. However, although risk sharing is a necessary part of a managed-care system, it is not sufficient.

Uwe Rheinhart presented a model where physicians' concern with maximizing their incomes was the driving factor in increasing health expenditures. He probably said this facetiously, because there is ample evidence that, even when physicians are at financial risk or are paid a salary, there is still a substantial amount of unnecessary and inefficient care delivered. A good illustration of this point was the case history presented by Steve Schoenbaum from the Harvard Community Health Plan (HCHP), a large staff-model HMO in Boston (1). He showed that HCHP obstetricians who used Hospital A had much higher rates of cesarean sections and forceps deliveries than those using Hospital B. He noted that this difference was not accounted for by variation in patient mix; more likely, it reflected differences in practice styles in the two obstetrics units. Thus, physicians' economic incentives are only one factor determining utilization patterns.

In the view of many persons in the managed-care industry, the key to the private sector's approach to controlling costs is utilization management, that is, programs that attempt to improve the effectiveness and efficiency of care delivered to individual patients. The utilization management strategy is based on the well-documented fact that a substantial amount of care is either unnecessary or only marginally beneficial. If inappropriate care is reduced, the quality of care will be improved and at the same time costs will come under greater control.

In part, responsible utilization management requires having explicit guidelines or protocols that define when a given procedure or test is necessary. In turn, the development of protocols depends on having data on the effectiveness of selected procedures in terms of health outcomes. As pointed out many times in this volume, there is a paucity of such data. Clearly then, effectiveness research is an important component of the health insurance-managed care industry's strategy to control health care costs.

CURRENT APPLICATIONS OF OUTCOMES RESEARCH

Clinical Protocols

AEtna has made a major investment in developing protocol-based utilization management programs, and several other insurers are moving in the same direction. AEtna now operates a precertification system that focuses on about 20 inpatient and 20 outpatient surgical procedures and diagnostic tests. The protocols were actually prepared by academic medical researchers and clinicians under contract to AEtna. We believe that the credibility and acceptance of protocols by employers, employees, and providers is enhanced by having academicians, who are focused on medical science rather than costs, prepare the protocols.

Even though AEtna did not develop the protocols internally, an obvious question is, Why are insurers rather than the medical profession taking the initiative in developing protocols? AEtna believes that organized medicine should, and eventually will, assume responsibility for developing national protocols, but for the time being AEtna is filling a gap.

AEtna's protocol-based utilization management programs have been running for about a year. They operate as a prior authorization system; that is, patients or providers call AEtna nurses, who use computer-based protocols to solicit specific information that is used to determine whether the proposed procedure is appropriate. If the procedure fails certification, the case is sent to an AEtna physician, who then discusses the details of the treatment with the attending physician. In this sense the protocols serve as screening tools to identify cases that do not meet current quality standards.

Technology Assessment

Another application of effectiveness research is technology assessment programs. In the past, insurers paid for procedures if they were in common use within the practicing community. Now, AEtna is taking a much tougher stand and has a large staff involved in trying to determine whether selected procedures are effective and should be covered benefits, regardless of local practice. This is done by reviewing the medical literature, consulting with nationally recognized clinical experts, and monitoring the positions of professional organizations (such as the American College of Physicians) that have active technology assessment programs.

This is just the beginning, and AEtna and other organizations committed to responsible cost management, including the government, are going to have to spend millions of dollars for technology assessment. Hundreds of procedures and tests now being used have never been carefully reviewed for effectiveness. Likewise, new procedures are being introduced into the delivery system with little, if any, scientific evaluation.

INSURERS' ROLE IN EFFECTIVENESS RESEARCH

Insurers can contribute to effectiveness research in several ways. First, they can assist the research community in obtaining congressional support for research funding.

Second, AEtna and other insurers can provide data on the population under age 65. In some respects, insurers' data are more extensive and detailed than data available from Medicare. In addition to the traditional data from paid claims, many insurers are now collecting from utilization management programs clinically detailed information that can be linked to paid claims. A good example is the extensive clinical data obtained in AEtna's protocol-based reviews of selected procedures.

Also, the quality of the data is getting much better. For example, AEtna captures International Classification of Disease (ICD-9-CM) codes for ambulatory visits and is working on ways to collect more detailed information on inpatient ancillary services. Further, some insurers' claims data systems have the capacity to include additional data elements. Thus, for example, a prospective study of several thousand patients could collect some information from hospital bills that is not usually captured on claims. On-site nurses could also collect concurrent data on selected patients. AEtna has nurses in many locations who use laptop computers to collect and transmit data on hospitalized patients.

Another enhancement of data that will be of interest to researchers is the ability some insurers have of creating episodes of care and linking claims

across settings (outpatient and inpatient) and services (drugs and outpatient ancillary services). Thus, some insurers can provide a fairly comprehensive clinical data set.

A third contribution insurers can make to effectiveness research is undertaking joint projects with university investigators. AEtna now employs several health services researchers and is actively seeking opportunities to join with established university groups in obtaining research grants from federal and private funding agencies.

The combination of AEtna's access to data, experience in insurance, internal research staff, and other resources with the expertise of university investigators offers a new model for applied health services research. This model should be attractive to funding agencies interested in supporting effectiveness research.

At some point, the information collected has to be used to effect positive changes in the delivery system. This is the fourth area in which insurers can contribute to the broader field of effectiveness research—that is, What are the best methods for changing the practice behaviors of providers? This is a very difficult problem, even with the necessary data on effectiveness. Steve Schoenbaum reported on the difficulty of trying to influence the behavior of several obstetricians employed by HCHP. Just imagine the problems faced by large insurers with HMOs and preferred provider organizations in 100 or more sites trying to modify the practice patterns of physicians.

The point is that effectiveness research needs to go beyond measurement and into applications. Because insurers operate many managed care plans in multiple locations, they offer an ideal natural laboratory for applications research.

LIMITATIONS OF THE EFFECTIVENESS "STRATEGY"

An underlying assumption of the effectiveness "strategy" is that with "hard" data on what medical treatments are cost-effective and with financial incentives and utilization management systems to influence provider practice behaviors, the rate of increase in health care costs can be substantially reduced.

From AEtna's experience with protocol-based review programs, HMOs, and other managed-care approaches, significant savings are possible. The still unanswered question is, "Are these one-time savings, or is the long-term rate of cost increases being reduced?" Only time will tell.

A related problem is the capacity to implement and operate effective managed-care programs in hundreds of different locations. Even if it can be demonstrated that one HMO can significantly reduce the long-term rate of increase in costs, it does not mean that this HMO can be replicated in every major medical market in the country.

Another concern with the effectiveness strategy is the liability issue. Just imagine the impact of two settlements of $40 million resulting from patients' being denied services, based on protocols, and later having adverse medical outcomes. It would have a profound effect on the whole managed-care industry and the use of protocols. So far, there have been few liability cases associated with managed-care programs, but the field is still relatively new, and we live in a very litigious society.

The final problem with the effectiveness strategy is having the time to make it work. Employers and legislators appear to want quick and easy solutions to complex problems. Certainly, all of us can sympathize with the desire to solve the cost problem within the next two years. Realistically, I believe that there are no easy answers, certainly no painless answers, and no answers that are likely to solve the problem within two years. These are my concerns. I am convinced that managed care can work and that effectiveness research will undoubtedly have a very positive long-term impact on improving health and making the delivery system more efficient. The health insurance industry is a strong supporter of this effort and is prepared to work with the research community to collect, analyze, and apply the results of effectiveness studies.

Reference

1. Schoenbaum, S.C. An Attempt to Manage Variation in Obstetrical Practice. Pp. 190-200 in *Effectiveness and Outcomes in Health Care.* Heithoff, K.A. and Lohr, K.N., eds. Washington, D.C.: National Academy Press, 1990.

26

Gaining Acceptance for Effectiveness and Outcomes Research

John D. Stobo

Effectiveness research has come a long way in terms of developing measurement scales that are reliable, somewhat easy to obtain, and pertinent. It is clear to me that further research in effectiveness and outcomes clearly needs to be done. It needs to be done, in my mind, for two reasons.

REASONS FOR PURSUING RESEARCH

First, as a profession, we physicians should be committed to providing the highest quality of care to our patients. Effectiveness and outcomes research will affect the question of quality and provide a rationale for deciding what the highest quality care is.

Second, it will provide a rationale for discussing the cost of health care. Like Henry Aaron (1), I am not convinced that outcomes research will substantially decrease the cost of health care. Nevertheless, it clearly will rationalize discussions of what is appropriate health care and what is not. The caution of Henry Aaron and others echoes the good advice that Holly Smith, my previous mentor and Chief of Medicine at the University of California at San Francisco, gave me: "Never promise more than you can give, and always give more than you can promise." Do not promise that outcomes research will significantly lower the cost of health care.

WHO MUST BE CONVINCED?

My concern is that effectiveness research be accepted by other groups who must be involved in it, of which there are three. First are the payers. I do not foresee any problem there. I think the payers of health care are thirsty for this information and will readily accept it.

The second group is the recipients of health care. Here, I think, is a major challenge: to provide the results of outcomes and effectiveness research to those individuals. A significant impact on cost can be achieved by educating recipients of care about utilization.

The last group that is crucial to outcomes research represents the biggest challenge. This is the providers of health care, particularly physicians. At my institution, Johns Hopkins, there is a lot of discussion about research into quality of care, effectiveness of care, outcomes of care, but it is done by a relatively small number of individuals. The majority of the faculty have not bought into outcomes research.

What will it take to get providers of care to accept this type of research? Again, I agree with others that there is going to be a pull and push phenomenon here. I think the pull will have to come from the persons who are already convinced of the value of outcomes research. Physicians have to buy into it. This is evident from the study of the Harvard Community Health Plan (2). Physicians have to be involved early on in these studies; they have to feel some ownership of them so that they are not always in a reacting mode.

It is important to train physicians in methodologies that are used in outcomes research. Most physicians, like myself, have been trained in areas related to biomedical research and are not conversant with methodologies that are important for carrying out and understanding other types of research. A major effort should be made to educate physicians about the methodology and interpretations of outcomes research.

WHO WILL PUSH FOR RESEARCH?

The push phenomenon is going to come from several areas, three in particular. One is the government: this prodding by the Health Care Financing Administration is important. There will be a push from employers. They, because of an interest in cost of care and also, I hope, because of an interest in quality of care for their employees, will be interested in effectiveness and outcomes research. Employers may push their employees in the direction of institutions that can document that they are as good as they say they are.

Finally, the push will come from hospitals, probably because they are being pressured by the government and by employers. Hospitals will pressure influential individuals to adopt practices that have been documented to provide the most effective care and the best outcomes.

Four years ago, effectiveness and outcomes research was an area that was completely foreign to me. It is one I have become interested in over the last two years and one I have become very excited about. It is going to be critical for American medicine in the future—and we are fortunate that there are such good people doing such good work in this area.

References

1. Aaron, H.J. The Need for Reasonable Expectations. Pp. 215-217 in *Effectiveness and Outcomes in Health Care.* Heithoff, K.A. and Lohr, K.N., eds. Washington, D.C.: National Academy Press.
2. Schoenbaum, S.C. An Attempt to Manage Variation in Obstetrical Practice. Pp. 190-200 in *Effectiveness and Outcomes in Health Care.* Heithoff, K.A. and Lohr, K.N., eds. Washington, D.C.: National Academy Press, 1990.

LIST OF AUTHORS

Henry J. Aaron, Ph.D.
Senior Fellow
The Brookings Institute
Washington, D.C.

Howard J. Bailit, D.M.D., Ph.D.
Vice President
Employees Benefits Division
AEtna Life Insurance Company
Hartford, Connecticut

M. Audrey Burnam, Ph.D.
Behavioral Scientist
The RAND Corporation
Santa Monica, California

Paul D. Cleary, Ph.D.
Associate Professor of Medical Sociology
Department of Health Care Policy
Harvard Medical School
Boston, Massachusetts

J. Jarrett Clinton, M.D.
Acting Administrator
Agency for Health Care Policy and Research
Department of Health and Human Services
Rockville, Maryland

Elliott S. Fisher, M.D., M.P.H.
Department of Community and Family Medicine
Dartmouth Medical School
Hanover, New Hampshire

Paul F. Griner, M.D.
Samuel E. Durand Professor of Medicine
Director, University of Rochester Medical Center
General Director, Strong Memorial Hospital
Rochester, New York

Louis B. Hays
Acting Administrator
Health Care Financing Administration
Washington, D.C.

Kim A. Heithoff
Research Assistant
Institute of Medicine
Washington, D.C.

Valerie P. Jackson, M.D.
Indiana University School of Medicine
Wishard Memorial Hospital
Indianapolis, Indiana

Stephen Jencks, M.D.
Chief Scientist
Office of Research and Demonstrations
Health Care Financing Administration
Baltimore, Maryland

Emmett B. Keeler, Ph.D.
Senior Mathematician
The RAND Corporation
Santa Monica, California

Henry Krakauer, M.D., Ph.D.
Office of Program Planning
Health Standards and Quality Bureau
Health Care Financing Administration
Baltimore, Maryland

Kathleen N. Lohr, Ph.D.
Deputy Director
Division of Health Care Services
Institute of Medicine
Washington, D.C.

Barbara J. McNeil, M.D.
Head, Department of Health Care Policy
Harvard Medical School
Boston, Massachusetts

Janet B. Mitchell, Ph.D.
President
Center for Health Economics Research
Needham, Massachusetts

Albert G. Mulley, Jr., M.D.
Chief
General Internal Medicine Unit
Massachusetts General Hospital
Boston, Massachusetts

David G. Murray, M.D.
Professor, Department of Orthopedic Surgery
SUNY Health Services Center at Syracuse
Syracuse, New York

Eugene C. Nelson, Sc.D.
Director, Quality of Care Research
Hospital Corporation of America
Nashville, Tennessee

Donald M. Patrick, Ph.D., M.S.P.H.
Professor
Department of Health Services
University of Washington
Seattle, Washington

Uwe E. Reinhardt, Ph.D.
James Madison Professor of Political Economy
Woodrow Wilson School of Public and International Affairs
Princeton University
Princeton, New Jersey

Richard A. Rettig, Ph.D.
Senior Staff Officer
Institute of Medicine
Washington, D.C.

John D. Rockefeller IV
Senator from West Virginia
Senate of the United States
Washington, D.C.

William L. Roper, M.D.
Director
Centers for Disease Control
Department of Health and Human Services
Atlanta, Georgia

Stephen C. Schoenbaum, M.D., M.P.H.
Deputy Medical Director
Harvard Community Health Plan
Brookline, Massachusetts

J. Sanford Schwartz, M.D.
Associate Professor
Section of General Internal Medicine
Hospital of the University of Pennsylvania
Philadelphia, Pennsylvania

Kenneth I. Shine, M.D.
Dean
UCLA School of Medicine
Los Angeles, California

G. Richard Smith, M.D.
Associate Professor of Psychiatry
Department of Psychiatry
University of Arkansas School for the Medical Sciences
Little Rock, Arkansas

Harold C. Sox, Jr., M.D.
Professor and Chairman
Department of Medicine
Dartmouth Medical School
Hanover, New Hampshire

John D. Stobo, M.D.
William Osier Professor of Medicine
Director and Physician-in-Chief
The Johns Hopkins University School of Medicine
The Johns Hopkins Hospital
Department of Medicine
Baltimore, Maryland

John E. Ware, Jr., Ph.D.
Senior Scientist
Institute for the Advancement of Health and Medical Care
New England Medical Center
Boston, Massachusetts

John E. Wennberg, M.D., M.P.H.
Professor of Epidemiology and Public Policy
Department of Community and Family Medicine
Dartmouth Medical School
Hanover, New Hampshire