Tools for Evaluating Health Technologies: Five Background Papers

September 1994

OTA-BP-H-142

Office of
Technology Assessment

Congress of
the United States

TOOLS
FOR
EVALUATING
HEALTH
TECHNOLOGIES

Five
Background
Papers

Recommended Citation: U.S. Congress, Office of Technology Assessment, *Tools for Evaluating Health Technologies: Five Background Papers, BP–I–I-142* (Washington, DC: U.S. Government Printing Office, February 1995).

For sale by the U.S. Government Printing Office
Superintendent of Documents, Mail Stop: SSOP, Washington, DC 20402-9328
ISBN 0-16 -045499-9

Foreword

Since the 1980s, the federal government has been actively pursuing a policy to encourage and sponsor medical effectiveness research: research that supports the evaluation of the effectiveness of existing medical technologies and practices. In September 1994, the Office of Technology Assessment issued its assessment of the successes and gaps in that endeavor to date with the report *ldentifiyng Health Technologies That Work: Searching for Evidence.*

This volume contains five background papers describing in greater detail some of the research techniques discussed in that report. Each of these papers was drafted by experts in the research technique. OTA gratefully acknowledges the contributions of the advisory panel to the overall study and the many other individuals who provided valuable information and reviewed preliminary drafts of these papers. As with all OTA documents, the final responsibility for the content of these papers rests with OTA.

ROGER C. HERDMAN
Director

Advisory Panel

William Fullerton, Chairman
Crystal River, FL

Robert Brook
Health Sciences Program
The RAND Corp.
Santa Monica, CA

Iain Chalmers
The Cochrane Centre
Oxford, England

Harold Cohen
Hal Cohen, Inc.
Baltimore, MD

David Eddy
Duke University
Durham, NC

Ruth Faden
Program in Law, Ethics,
and Health
Johns Hopkins University
Baltimore, MD

Diana Jest
Government Affairs
Group Health Association
Washington, DC

Robert Keller
Maine Medical Assessment
Foundation
Augusta, ME

Michael McCulley
Johnson & Johnson
New Brunswick, NJ

Jonathan Lomas
McMaster University
Hamilton, Ontario

Barbara McNeil
Department of Health Care Policy
Harvard Medical School
Boston, MA

Frederick Mosteller
School of Public Health
Harvard University
Boston, MA

Richard Peto
Clinical Trial Service Unit
Oxford, England

John Wennberg
Center for Evaluative Clinical
Sciences
Dartmouth Medical School
Hanover, NH

Note: OTA appreciates and is grateful for the valuable assistance and thoughtful critiques provided by the advisory panel members. The panel does not, however, necessarily approve, disapprove, or endorse this volume. OTA assumes full responsibility for these background papers and the accuracy of their contents.

Preject Staff

Clyde J. Behney
Assistant Director, OTA

Sean R. Tunis
Health Program Director

Louise Staley
Office Administrator

Carolyn Martin
Administrative Secretary

Chariotte Brown
Word Processing Specialist

Carolyn Swarm
PC Specialist

ELAINE J. POWER
Project Director

Arna M. Lane
Research Analyst

Cheryi Liechty
Research Assistant

Anne Meadows
Editor

Acknowledgments

OTA *wishes to express its appreciation for the following individuals and organizations that provided assistance with this volume. These individuals do not necessarily approve, disapprove, or endorse these background papers. OTA assumes full responsibility for the volume and the accuracy of its contents.*

USING PATIENTS' REPORTS TO EVALUATE MEDICAL OUTCOMES

Agency for Health Care Policy and Research
Rockville, MD

Pennifer Erickson
Centers for Disease Control and Prevention
Atlanta, GA

Marthe Gold
Office of the Assistant Secretary for Health
Washington, DC

John C. Hornberger
Department of Health Research and Policy
Stanford School of Medicine
Stanford, CA

Robert L. Kane
School of Public Health
University of Minnesota
Minneapolis, MN

Robert M. Kaplan
Department of Family and Preventive Medicine
Division of Health Care Sciences
University of California
San Francisco, CA

Kathleen N. Lohr
Institute of Medicine
Washington, DC

Susan Nayfield
National Cancer Institute
Bethesda, MD

Donald L. Patrick
School of Public Health and Community Medicine
University of Washington
Seattle, WA

George Torrance
Centre for Health Economics and Policy Analysis and McMaster University
Hamilton, Ontario

John E. Ware, Jr.
The Health Institute
New England Medical Center
Boston, MA

LARGE ADMINISTRATIVE DATABASE ANALYSIS

Agency for Health Care Policy and Research
Rockville, MD

Health Care Financing Administration
Baltimore, MD

Elliot S. Fisher
Dartmouth University Medical School
Hanover, NH

Floyd J. Fowler
Center for Survey Research
University of Massachusetts
Boston, MA

Mitch Greenlick
Center for Health Research
Kaiser Perrnanente
Portland, OR

William Harlan
National Institutes of Health
Bethesda, MD

Kathleen N. Lohr
Institute of Medicine
Washington, DC

Matthew Longnecker
UCLA School of Public Health
Los Angeles, CA

David B. Pryor
Duke University School of Medicine
Durham, NC

Patrick Romano
University of California at Davis Medical Center
Sacramento, CA

J. Stanford Schwartz
Institute of Economics
University of Pennsylvania
Philadelphia, PA

Earl Steinberg
Health Technology Associates
Washington, DC

Robert Temple
Center for Drug Evaluation and Research
Food and Drug Administration
Rockville, MD

LARGE AND SIMPLE RANDOMIZED TRIALS

Richard Doll
Department of Clinical Medicine
University of Oxford
Oxford, England

Floyd J. Fowler
Center for Survey Research
University of Massachusetts
Boston, MA

William T. Friedewald
Metropolitan Life Insurance Company
New York, NY

Lawrence Friedman
National Heart, Lung, and Blood Institute
Bethesda, MD

Michael A. Friedman
Division of Cancer Treatment
National Cancer Institute
Bethesda, MD

Peter Greenwald
Division of Cancer Prevention and Control
National Cancer Institute
Bethesda, MD

William Harlan
National of Institutes of Health
Bethesda, MD

Alessandro Liberati
Istituto Riccrche Farmacolgiche
"Mario Negri"
Mikmo, Italy

Matthew Longnecker
UCLA School of Public Health
Los Angeles, CA

Charles G. Moertel
Mayo Clinic
Rochester, MN

Robert Temple
Food and Drug Administration
Rockville, MD

Jeff Whittle
Section of General Internal
Medicine
Department of Veterans Affairs
Pittsburgh, PA

META-ANALYSIS

**Agency for Health Care Policy
and Research**
Rockville, MD

Jesse Berlin
School of Medicine
University of Pennsylvania
Philadelphia, PA

Graham Colditz
Harvard University
Boston, MA

Floyd J. Fowler
Center for Survey Research
University of Massachusetts
Boston, MA

William Harlan
National Institutes of Health
Bethesda, MD

Michelle Orza
U.S. General Accounting Office
Washington, DC

Donald Rubin
Harvard University
Boston, MA

Stephen B. Thacker
Centers for Disease Control
and Prevention
Atlanta, GA

CLINICAL-ECONOMIC TRIALS

John M. Eisenberg
Department of Medicine
Georgetown University
Medical Center
Washington, DC

Floyd J. Fowler
Center for Survey Research
University of Massachusetts
Boston, MA

Gilad Gordon
Synergen
Boulder, CO

Richard Green
Agency for Health Care Policy
and Research
Rockville, MD

William Harlan
National Institutes of Health
Bethesda, MD

Candy Littell
Health Care Technology Institute
Alexandria, VA

Matthew Longnecker
UCLA School of Public Health
Los Angeles, CA

Contents

Introduction 1

1 Using Patients' Reports To Evaluate Medical Outcomes 3
by Floyd J. Fowler

Summary 3
Current Applications 5
Measuring Health-Related Quality of Life 9
Reliability and Validity 14
Issues 17
Conclusion 21
References 21

2 Large Administrative Database Analysis 27
by Jeff Whittle

Summary 27
Descriptive Studies 31
Comparative Studies 37
Adjuncts to Other Research Methods 46
Other Issues 48
References 55

3 Large and Simple Randomized Trials 67
by Julie E. Buring, Michael A. Jonas, and Charles H. Hennekens

Summary 67
Principles of Large and Simple Trials 69
Other Applications of Large, Simple Trials 75
Impact of Large, Simple Trials on Clinical Practice 82
Limitations of Large, Simple Trials 83
Additional Issues 84
Conclusions 88
References 88

4 Meta-Analysis 93

by Matthew P. Longnecker

Summary 93
Conducting a Meta-Analysis 97
Reliability and Validity 101
Issues and Controversies 102
Future Applications 109
Conclusion 110
References 111
Appendix 4-A 116

5 Clinical-Economic Trials 125

by Neil R. Powe and Robert I. Griffiths

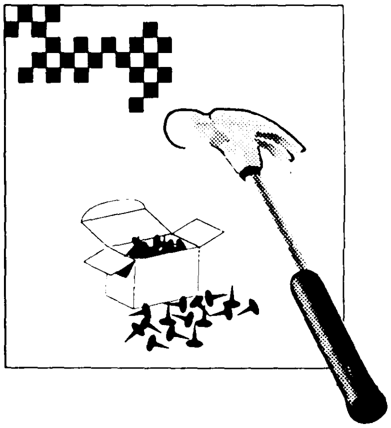
Summary 125
Methodological Considerations 131
Validity and Reliability 139
Future Prospects 142
References 145

Introduction

Scientific developments often occur alongside changes in the cultural and political environment, and this has certainly been true in recent years for the evaluation of health care technologies. New methods of research and evaluation have been developed, and new adaptations of existing methods are being applied. At the same time, the American health care system has undergone radical changes. The enormous expansion in managed care, the movement of many highly complex and sophisticated medical services into nonhospital settings, and the increasing willingness of physicians and patients to question the effectiveness of common procedures all bear witness to the tumultuous past decade in health care. And along with these changes has come an eager market for information on the value of existing medical technologies and the research methods that can supply this information.

Each of the five background papers contained in this volume describe new methods or new adaptations of existing methods to evaluate which health technologies work best. * These examples by no means describe the universe of changes in evaluative techniques. They do, however, demonstrate the great variety of areas in which methodological developments have been taking place.

The first of the five papers deals with one of the most basic questions in any health research endeavor: how to measure the outcomes associated with whatever is being studied. The devel-



[1] Summaries of these five papers, and a discussion of the methods in the broader policy context of the evaluation of medical technologies, appear in the Office of Technology Assessment report, "Identifying Health Technologies That Work: Searching for Evidence," OTA-H-608 (Washington, DC:U.S. Government Printing Office, September 1994).

opment of reliable techniques to measure health outcomes through *patients' reports* of how they feel represents not only anew research method but a subtle philosophical shift regarding which outcomes are important to measure.

The second and third of these papers deal with two methods of investigating the question: of two competing medical technologies, which is more effective? Both methods—the *analysis of large administrative databases,* and *large, simple randomized trials—have* been promoted as affordable, generalizable alternatives to the more costly, complex, and limited traditional randomized controlled trial. In fact, these two newly adapted techniques are not really substitutes for each other and fill somewhat different niches. They also differ in the amount of attention they have received by the U.S. research establishment. Large database analysis has gained prominence in the United States, as a method emphasized in the federal government's medical effectiveness initiative. In contrast, large, simple trials are a European development that has many potential applications but has so far seen relatively little use in this country.

Where previous studies of a technology 's effectiveness already exist, medical technology assessors must sift through the often obscure and sometimes contradictory literature on the topic. The fourth background paper in this volume describes the formal technique of *meta-analysis,* which structures a literature review by identifying relevant studies in a systematic, explicit fashion and combining the results quantitatively. Although many topics do not lend themselves to a quantitative meta-analysis, the systematic approach used to identify and evaluate studies is applicable to almost any review of the medical literature.

Along with the health system's new interest in documenting the value of existing medical technologies and practices has come a new, very pragmatic interest in techniques to determine the relative cost-effectiveness of competing technologies. The technique described in the fifth background *paper-clinical-economic trials—is an* increasingly popular method for analyzing a technology's cost-effectiveness early in its life cycle, at the same time that the technology's clinical effectiveness is being tested.

Few of the techniques described in this volume are fundamentally new. All are being applied with a new vigor and new twists, however, in the current drive to evaluate the worth of existing medical interventions. Understanding these techniques—their applications, their strengths, and their limitations—is a worthwhile endeavor for evaluators and policymakers alike.

Using Patients' Reports To Evaluate Medical Outcomes

Background Paper 1

SUMMARY

Most medical treatment is intended to improve patients' ability to function and their sense of well-being. Information about these outcomes can usually be supplied most accurately by the patients themselves.

Traditionally, medical conditions have been defined and treatments evaluated primarily through the results of diagnostic tests and clinical observation, but many studies of the outcomes of medical treatments now also routinely include protocols for asking patients questions about their health and well-being. Some outstanding examples of this phenomenon are:

- *the Medical Outcomes Study, which used a single set of measures to assess functioning and well-being of patients with any of six medical conditions; and*
- *the Patient Outcomes Research Teams and similar efforts, in which the researchers study a single medical condition and the outcomes of its treatment in considerable detail, including patient-reported health and well-being.*

The instruments used to measure these characteristics may be specifically tailored to the particular medical condition of interest. Or, they may be general measures of health-related quality of life that provide comprehensive views of the states of the patients' health at various points during the course of the treatment.

Health-related quality of life refers to those aspects of living that are affected by patients' medical conditions and to their finetioning and perceived well-being. Most survey instruments designed to measure health-related quality of life include questions related to four aspects: functional ability, perceived health, psychological well-being, and role limitations.

by

Floyd J. Fowler

University of

Massachusetts

Boston, MA

One way to analyze the effects of a particular treatment on patients' health is to describe the treatment results separately with respect to each of these aspects. The Sickness Impact Profile (SIP) and the RAND 36-item Health Survey (SF-36) are examples of instruments to measure health-related quality of life and at present are probably the major tools used to gather data from patients for this purpose in the United States. The SIP is perhaps the most comprehensive instrument for describing the effects of medical conditions on people, while the SF-36 attempts to strike a balance between comprehensiveness, validity, and parsimony.

Another approach to measuring health-related quality of life is to combine the ratings for all the various components of well-being into a single number that serves as a summary of the overall quality of life. The Quality of Well-Being Scale is probably the best-known example in the United States of the effort to produce quantitative summaries of people's health status.

Properly designed measures using patients' reports have proved as reliable and valid for describing the effects of medical conditions and treatments on patients as many other commonly accepted tests. Although the results of diagnostic tests and results based on patients' reports are difficult to compare directly, studies of diagnostic tests and of measurements taken in clinical settings almost always reveal considerable error across tests. The reliability and validity of measures based on patients' reports also vary, but the evidence is clear that measurement of medical conditions or health status, based on properly designed and evaluated questions, can be as reliable and valid as other measurements done in the clinical sciences.

At least three major conceptual and methodological challenges remain for researchers and users of patients' reports on medical outcomes:

How should prospective and retrospective designs be modified to ensure accurate measurements of the effects of treatment?

- *How should researchers collect information about the results that would have been expected had a particular treatment not been given?*
- *How should the effects of treatment be calibrated to facilitate comparisons across treatments and conditions?*

There is a clear need for better understanding of how best to conduct studies based on patients' reports so that they lead to valid conclusions, and how best to assess the significance of the results. Nonetheless, this tool is already a very useful one that produces considerable knowledge that neither patients nor researchers have had before. ■

Although the saving of lives may provide evidence of the value of treatments for serious ailments such as strokes and heart attacks, only a moderate amount of the medical care delivered in the United States is intended to prevent death. Most treatment is meant to improve patients' functioning or well-being.

Even the reason for performing most common hospital surgical procedures is not to save lives, at least in the short term. The conditions treated by back surgery, for instance, are virtually never life-threatening. Fewer than perhaps 10 percent of hysterectomies and a similar proportion of surgical procedures to treat benign prostate disease are performed on patients whose lives are at risk (48,55,78). Outpatient surgical procedures such as cataract surgery, the most common procedure covered by Medicare, fall into the same category. Even such a major procedure as coronary artery bypass graft surgery is performed as often to reduce angina symptoms as to save lives (2).

Compared with surgery, ambulatory care is sought even less often for life-threatening conditions. Most patients visit doctors for checkups, for acute but self-limiting conditions (such as respiratory infections), or for other nonfatal conditions (such as back pain and arthritis) (79).

Relieving symptoms is the goal of most common surgical procedures; ruling out more serious conditions and providing diagnoses that can lead to symptomatic relief are the goals of most ambulatory visits. Thus, to a large extent, the criteria justifying medical treatments rest on how the treatments affect the way patients feel or what patients can do. Ascertaining the value of the treatment, then, requires information about patients' perceptions of their well-being.

Patients' own reports of their well-being are vital to studies of medical care for at least two reasons. First, the studies often require information that only the patients can report well. When comparisons have been made, physicians have usually been found to be poor reporters of patients' symptoms or experiences in such diverse cases as enlarged prostates and toxic reactions to cancer treatment (1 1,62).

Second, some of the key information needed from patients—their perceptions, emotional responses, preferences, and values—is subjective. There is widespread agreement that no one can reliably report such things for another person (75). Studies comparing reports by individuals themselves with reports from proxies suggested that the less observable the characteristic, the less likely others can report it accurately (24,50,66,76). Thus, although good studies of the outcomes of medical treatment gather data from various sources, many also rely on accurate measurements from patients' reports.

Prior to the mid-1980s, few studies were designed to document the benefits of treatment from the patient point of view. Studies of medical outcomes tended to focus primarily on short-term risks, such as death and strokes, and on rehospitalization. If the broader benefit to patients was assessed at all, it was usually based on ratings by physicians.

Researchers conducting a meta-analysis of the literature published between 1964 and 1990 on surgery for lumbar spinal stenosis, for example, found 74 journal articles that purported to provide information about outcomes from laminectomy, but only 61 percent of the articles reported the prevalence of leg or back pain, the main reason for which the surgery is done (74). The analysis found no randomized trials comparing surgery and conservative treatment, and almost nothing was published on the response of spinal stenosis patients to conservative treatment. Most important, the researchers found that it was often impossible to tell for certain who rated the outcomes, but that it almost always appeared to be the surgeon, not the patient, who was describing the benefits of the surgery.

The poor quality of the data undoubtedly reflected the low priority placed on documenting the value of medical treatment. Medical treatment was presumed to be worthwhile if physicians, based on their training and clinical experience, thought it would be of value. Studies were concerned chiefly with whether complications arose and with how they could be minimized. A further limitation derived from the fact that studies using survival and short-term complications as measures of outcomes document only the risks, not the benefits, of treatment. To document the benefits of treatment, accurate information about patients' health and well-being must be collected from them in standardized ways.

CURRENT APPLICATIONS

Examples of the Use of Patients' Reports in Research

To fill the gaps in the medical literature, several recent studies have attempted to measure the effect of medical treatment by asking patients questions. Of particular note are the Medical Outcomes Study (69) and the work of the Patient Outcome Research Teams (PORTS, described below) and other related research studies that focus on the health outcomes associated with particular medical conditions.

Medical Outcomes Study

The-Medical Outcomes Study (69) is a good example of the current approach to studying the effects of treatments. In that study, patients who had any of six different conditions were recruited in physicians' offices throughout the United States. The patients filled out questionnaires about their health at the time they were first contacted; they then provided comparable data periodically so that changes could be measured. One of the important distinctive features of the Medical Outcomes Study was that a single set of measures (a predecessor to the SF-36, described below) was used in assessing overall functioning and well-being across all conditions (70). As a result, researchers could make three kinds of comparisons, each with its own value:

- how patients with the same health condition fared over time under different treatment protocols,
- how the lives of patients with different conditions were affected by those conditions, and
- how the benefits of treatments compared across conditions.

The Medical Outcomes Study was probably the first study that permitted all three of these types of analyses.

The PORTS: The Example of Lower Back Pain

The PORTS, interdisciplinary research teams funded by the Agency for Health Care Policy and Research, have been significant contributors to the development of measures of patient functioning and well-being. Each PORT studies the outcomes of medical care and treatment for one of 14 common medical conditions. The PORT study of lower back pain provides an example of how these research efforts are using a patient-oriented approach to evaluating medical outcomes.

Because the main goal in treating lower back problems is to reduce pain in the back and legs, patients are asked to describe the frequency and intensity with which they experience pain in these areas. One simple, straightforward analysis using these data is to determine the extent to which reports of back pain changed for better or worse over time with and without treatments (23). Back pain is significant because it can affect functioning, self-perceived health, and psychological well-being.

Deyo and his associates are conducting a study of back pain in cooperation with orthopedists and neurosurgeons in Maine. Patients who are being treated for back pain are asked by their physicians to participate. The physicians and patients together decide on the treatment to be used; the study does not affect the decision. Regardless of whether patients opt for surgery or nonsurgical treatment, the results are monitored.

Patients complete baseline questionnaires at the time of enrollment. The questionnaires cover the character and frequency of the back pain, the effect of the back pain, and the overall functioning and well-being of the patient. Physicians also complete forms describing the results of initial tests and the details of the treatment, but the patients' answers to the questions at 3, 6, and 12 months after enrollment are the main measures of the outcomes.

Other Outcomes Studies: *Indications for Hysterectomy*

In addition to the PORTS, other health researchers have been studying the outcomes associated with particular medical conditions and procedures using patients' reports.

Researchers recently completed a similar study of women who had conditions—such as excessive bleeding, abnormal pain, or large fibroids-that would make them candidates for hysterectomy (15,16). Whether they elected to be treated surgically or nonsurgically, the women completed questionnaires regarding their symptoms, including the frequency and intensity of their pain and bleeding. Both the conditions and the treatments have been reported to affect energy, sexual functioning, bowel functioning, frequency of urination, hot flashes, and anxiety level, so specific questions were included (either adapted from other survey instruments or newly designed) to monitor the patients' experiences in each of these areas. Other questions measured the women's general well-being, psychological well-being, perception of their health, and role limitations.

The protocol was very similar to that of the back study. Patients filled out questionnaires over the course of a year. Analyses evaluated the progress of the initial symptoms, the appearance of new problems, and the reported general functioning and well-being of patients who had been treated with or without surgery.

Six general characteristics of the Medical Outcomes Study and the studies of lower back pain and indications for hysterectomy mark important departures from most previous studies on the effects of medical treatments:

1. Combinations of data from patients and from medical records or physicians were used to describe the patients' initial condition and treatment.
2. Patients' reports were the primary measures of the effects of treatment.
3. Measures of patients' status were comprehensive, including changes in condition, various possible complications of treatment, and multiple measures of overall functioning and well-being.
4. Patients were followed for relatively long periods of time—a year in the lower back pain and hysterectomy studies, and several years in the Medical Outcomes Study.
5. Although patients were not assigned to different treatments as part of the protocol, the studies included patients treated in various ways, so that there was a context within which to evaluate the results of individual treatment approaches.
6. Numerous physicians who were in general community practice participated, thereby making results more likely to be representative than if the studies had been done only in university medical centers.

Role in Evaluating Effects of Medical Treatments

The role of patients' reports in evaluating the outcomes of medical care for a particular condition is to better understand the treatment effects on that condition, and to gain a broader understanding of the effects of care on patients' functioning and health-related quality of life overall.

Better understanding a treatment's effects on a medical condition has three components:

1. Assessing the characteristics of the condition. Traditionally, medical conditions have been defined through diagnostic tests and clinical observation. In some cases, however, patients' reports are needed in order to calibrate the severity of a condition. In other instances, patients' reports actually form the basis for defining the condition and its severity.
2. Measuring treatment complications. The value of treatment depends in part on whether the treatment has any negative consequences. Even when the treatment is aimed at saving lives or reducing strokes or heart attacks, the benefits must often be weighed against the risks of complications from the treatment.
3. Understanding how the condition affects patients' lives. Assessing the full value of a medical treatment requires understanding not only how a treatment affects a condition and what unwanted complications the treatment causes, but how much the condition affects patients' lives.

Assessing the Characteristics of the Condition

An example of a condition that is best measured using patients' reports is benign prostatic hyperplasia (BPH). Men's prostates tend to enlarge with age. As a consequence of this condition, some men experience a narrowing of the urethra, which obstructs urinary flow and produces such symptoms as frequent urination and difficulty in starting urination. Physicians can determine the size of the prostate gland through palpation and imaging; they can observe evidence of obstruction with cystoscopy; they can ascertain the rate at which urine flows and measure the extent to which the bladder completely empties after voiding. None of these physiological or clinical measures, however, correlates well with how patients experience symptoms or with the frequency of their symptoms (1,3,5,60).

From a medical point of view, there is no intrinsic reason to improve the rate at which urine flows, to reduce the obstruction that appears in a cystoscopy, or to make a prostate smaller. Although large post-voiding residual volumes of urine can lead to urinary-tract infections or to upper-urinary-tract pressure, which can cause deterioration of the bladder or affect renal function, such problems probably affect no more than 10 percent of men who undergo prostate surgery. Of the 350,000 men who have prostate surgery each year, about a quarter do so because of acute retention, whereas well over half do so to reduce their symptoms (55). For the latter group, the best indicators of the condition's severity are the patients' reports about their symptoms, and the goal of the treatment is to reduce the symptoms and their effect on the patients' quality of life (33).

The treatment of back pain is analogous. Image studies are commonly used to diagnose the cause of lower back pain. Among persons over the age of 40, the backs of as many as half appear on x-ray or other image studies to have serious problems, such as ruptured disks or stenosis, although the patients themselves experience no pain or disability (12,89). At the same time, image studies reveal no anomalies in other people who report experiencing pain in their lower backs and down their legs—pain that physicians are confident stem from stenosis or problem disks. Studies comparing symptomatic and asymptomatic patients consistently show they cannot be distinguished on the basis of images (10,61).

Patients' reports and the results of image studies are often complementary: the image study shows a ruptured disk or stenosis that corresponds well with the symptoms reported by a patient. When the two do not coincide, however, it is by no means clear that the clinical indicator should take precedence. To operate on a back when an image study indicated problems but the patient reported none would usually be inappropriate (40,68). The pain and dysfunction patients experience and report define whether the patients have back problems and are critical components of the indications for treatment; the relief of those symptoms and the restoration of functioning constitute the standard by which to evaluate whether medical care is effective or not. As is the case with BPH, the presence or severity of the condition is best defined by the patients' reports, not by clinical studies.

Patients' reports do not standalone in decisions about medical treatment. Although relieving symptoms is the focus of BPH treatment, the diagnosis of the reason for the symptoms and the likelihood that treatment will be effective depend on direct clinical evidence that the prostate is obstructing urination. If surgery is to be an effective treatment for back pain, a physiological problem that can be repaired by surgery must be identified. And some medical conditions are almost always defined by clinical examination and by test results. Patients' reports play little role in defining the presence of malignancies or hypertension, for example. Many common conditions, however, are best described by a combination of clinical observation, diagnostic tests, and patients' reports. Cataracts, arthritis, angina, and diseases of the uterus are particularly clear cases in which patients' reports play critical roles in defining the presence or severity of the conditions. Although the treatment for these conditions is physiological, the indications for treatment and the benefits of the treatment require assessing the status of the condition, in part, by asking the patients questions.

Measuring Treatment Complications

Comprehensive studies of treatments systematically estimate the frequency and severity of complications as well as their effects on the treated conditions. The presence or severity of many common complications cannot be characterized without patients' reports.

Accounting for the risks of complications is particularly important when the likelihood of a life-saving benefit of a treatment is relatively low. The treatment of mild hypertension, for example, is effective in preventing stroke; it reduces the probability that an otherwise healthy 50-year-old will have a stroke during the next five years from about 15 to about nine strokes per 1,000 men (19,56). The low overall probability of stroke means, however, that the great majority of men with mild hypertension would not have had strokes even without treatment. Because the medications used to lower blood pressure can reduce energy and sexual functioning and can produce depression, sleep disorders, anxiety, fainting, dizziness, and fatigue (21), a full evaluation of treatment for mild hypertension must include both the likelihood of stroke reduction and the rates at which patients report the various side effects.

Surgery is widely performed in cases of prostate cancer, but such surgery produces high rates of sexual impotence and significant incontinence (29). Furthermore, no studies have shown surgery to be more likely than less invasive procedures to save lives (88). Thus, the net value of the surgery cannot be assessed without taking these complications into account.

Ascertaining the Effects of a Condition on Patients' Lives

A condition or symptom that constitutes a major problem for one patient may be only a small problem for another (31 ,33). Differences in patients' roles and responsibilities account for some of this variation. A person whose job entails heavy physical labor, for example, may be affected by lower back pain to a greater extent than an office worker is. Even if the pain is the same, the office worker may be better able to avoid putting stress on his or her back and may be better able to perform despite the pain. In contrast, a physical laborer maybe unable to work at all if the back problem is severe.

The significance of health conditions also depends on the individuals' feelings or response styles, which may have nothing to do with roles. Women's responses to options regarding surgery for breast cancer demonstrate this concept. For the majority of patients, the probabilities of survival are the same whether they choose to have lumpectomy with radiation or to undergo mastectomy (28,83). The perceived cosmetic advantages of lumpectomy make that a clear choice for some women, whereas others choose more radical surgery (90) because they feel more secure with a more aggressive—though equally effective—treatment.

Thus, assessing the significance of a condition and the benefit of any treatment requires information about how much the condition matters to the patient. Because the answer to this question generally varies from one person to the next, the patients' own reports are crucial. To address this need, researchers have developed methods of measuring patients' health-related quality of life.

MEASURING HEALTH-RELATED QUALITY OF LIFE

I Concepts and Components

Studies of the outcomes of medical care often use condition-specific measures for describing the patients' medical conditions, the complications of the treatments, and the patients' perceptions of how the conditions and treatments have affected their lives. It is increasingly recognized, however, that medical outcomes cannot be fully determined without ascertaining the treatments' effects on the patients' quality of life. Thus, many studies now also include general measures of health-related quality of life to provide comprehensive views of the patients' health at various times during the course of the treatment.

In this context, *quality of life* refers to the aspects of living that are affected by patients' medical conditions and to their functioning and perceived well-being. As defined by Patrick and Erickson, "Health-related quality of life is the value assigned to duration of life as modified by the impairments, functional states, perceptions and social opportunities that are influenced by disease, injury, treatment, or policy" (64).

Experts do not entirely agree on exactly what constitutes health-related quality of life, but most surveys designed to measure it include questions related to four basic aspects of functioning and well-being (58,64,7 1):

1. Functional ability. Questions aimed at discovering functional ability ask what people can do. The most common questions inquire about such physical activities as walking across a room, climbing a flight of stairs, or walking around a block. Other questions may cover such things as the patients' abilities to read a newspaper, to watch television, to hear well enough to talk on the telephone, or to hold a pen. All such questions are independent of patients' role expectations, resources, or responsibilities.

2. Perceived health. The simplest question about self-perceived health asks people to rate how healthy they think they are. Such a question has been a staple of the National Health Interview Survey for many years, and is perhaps the most widely used measure of health status (52). Other commonly measured aspects of self-perceived health are the degrees to which patients worry about their health and to which they are satisfied with their health.
3. Psychological well-being. Measures of psychological well-being usually focus on the extent to which patients see themselves as distressed-where they would place themselves on an emotional continuum with depression at one end and happiness at the other end, or with anxiety at one end and calmness at the other (9,85). Although they disagree about what specific questions should be asked, most researchers accept psychological well-being as fundamental to the issue of quality of life.
4. Role functioning. How health conditions affect people's lives depends on their roles—what is expected of them, what kind of work they perform, what resources they possess, and what they must do on a day-to-day basis (e.g., 82). Questions about role functioning are self-adjusting. A condition that would seriously limit a young professional athlete might not limit a retired person at all. A condition's effect on mobility might be severe for a person who must ride buses, moderate for a person with a car, and minimal for a person with a chauffeur. Common questions about role functioning in measures of quality of life address patients' abilities to work, to take care of themselves, to maintain their households, and to participate in society. Patients often are also asked about their abilities to take care of business, to get around, and to participate in the recreational activities of their choice.

Calculating Effects on Overall Quality of Life

There are two distinct approaches to calculating the effects of a particular treatment on a patient overall health-related quality of life. One way is to describe the results separately with respect to each component. Under this approach, the patient responses to questions in the instrument measuring quality of life might suggest, for example, that for a particular treatment the patient's physical functioning improved but that his or her perceived health did not change.

Another approach is to combine the ratings for all the various components into a single number that serves as a summary of the overall quality of life. Researchers following this approach must first determine how much weight to give to function, psychological distress, and measurements of other aspects of patients' lives, so that ratings for those different components can be combined quantitatively. The methods used to assign weights to different aspects of quality of life include statistical models, ratings by physicians, average ratings by patients, and ratings by samples of the general public. Perhaps the most obvious method is to ask people how they value their quality of life overall (4,35,36,37,38,59,64).

Measuring Condition-Specific vs. General Effects on Quality of Life

As described above, studies of medical outcomes usually require condition-specific measures aimed at describing the status of the patients' conditions, complications of common treatments, and perceptions of how the conditions and treatments have affected patients' lives. In addition, most studies now include general measures of health-related quality of life that provide comprehensive views of the patients' health at various points during the course of the treatment.[1]

[1] The many strategies for measuring health status and health-related quality of life have been extensively described and reviewed. McDowell and Newell (58) describe and review 50 measurement schemes based on subjective judgments and ratings. Patrick and Erickson (64) provide an excellent discussion of the conceptual underpinnings of the major efforts to measure perceptions of health. as well as a more detailed description of the development, uses, and limits of some of the most important approaches. Froberg and Kane (35,36,37,38) and Stewart and Ware (7 I) also provide excellent reviews of issues related to various aspects of the measurement of functioning, well-being, and health status.

The question of how much a patient is limited because of a particular condition, such as lower back pain, contains two components: to what degree is the patient limited, and to what extent is the limitation tied to the lower back pain? If the patient has only one condition that affects functioning, any limitation may be attributable to that condition. A person who has multiple conditions, however, may have difficulty attributing any particular effect to a specific condition or health problem. Indeed, as people age, many of their physical and intellectual capabilities decline, which may make it increasingly hard to report on the effects of each specific health condition. As a result, the effect of treatments may be ascertained more accurately by asking patients to assess their functioning and well-being over time, with and without treatments, than by asking them to attribute their deviations from perfect health to particular health problems.

General measures of health-related quality of life have another advantage as well. Fixing one condition, even a troublesome condition, may do only a little to benefit the overall quality of life of a patient with multiple conditions. Measuring overall quality of life in a way that reflects the effects of all the patient's health problems can demonstrate the true value of the treatment to the patient. Overall measures of quality of life also enable *re*searchers to take into account both the benefits and the downsides of treatments for a particular condition.

Patients' Satisfaction with Care

Patients' satisfaction with care is often mentioned as part of assessing medical outcomes (18,87). Satisfaction with the results of treatment reflects how patients rate their post-treatment states of health. Satisfaction with the process of care, however, depends on physicians' personal styles and how patients have been treated. A patient's assessment of the quality of care, therefore, doesn't necessarily indicate whether the treatment improved a medical condition (17).

Nonetheless, satisfaction with care is sometimes important for assessing medical services. Tests or examinations may be used, for example, simply to assuage patients' fears and worries. In such cases, the patients' satisfaction with the fact that procedures have been performed may be important. In assessing how a treatment has affected a medical condition or health status, however, patients' satisfaction with how the process itself was carried out is usually irrelevant.

I Instruments for Measuring Health-Related Quality of Life

Instruments to measure peoples' health status have been in use for decades (box 1-l). Attempts to measure health-related quality of life in a broader sense using survey instruments that ask detailed questions of the patients themselves, however, is a much newer development.

There are now numerous instruments used around the world to measure health-related quality of life, although not all of them rely on patients' reports. The Nottingham Health Profile is widely used in the United Kingdom (41), for example, and the EuroQol has been used in a 14-country study in Europe (25). The Arthritis Input Measurement Scale (57) and the OARS* Multi-dimensional Functional Assessment Questionnaire (27) are two of the more frequently cited instruments that rely on self-reporting.

In the United States, several programs for developing general measures of patients' well-being for use in clinical studies have been particularly important in influencing research on the outcomes of medical care. These programs include the Sickness Impact Profile research, the Medical Outcomes Study, and the Quality of Well-Being Scale.

[2] OARS IS the abbreviation for the Older American's Resources and Services Schedule.

BOX 1-1: The Historical Development of Instruments To Measure Health Status

Early efforts to clinically measure patients' functioning and the severity of conditions include the development of the widely used Karnofsky Index for patients with cancer in the 1940s and the development of scales for the activities of daily living in the 1950s (22). Neither of these approaches based its ratings on the patients' own reports, however, and neither attempted to assess health status across wide ranges of patients or the general population.

Still, these early measures greatly influenced later survey instruments. The early rating schemes for how well people could take care of the basic activities of daily living (ADLs) (such as bathing, dressing, eating, and toileting) and the instrumental activities of daily living (IADLs) (such as housekeeping, getting around, and participating in social events) have been the basis for numerous scales using reports from patients and experts (e.g., 46,51 ,58,63). Moreover, ADLs and IADLs usually are part of more comprehensive strategies for assessing health status.

Another important influence on current strategies for measuring health has been the National Health Interview Survey (NHIS), which was established in the late 1950s to characterize and monitor the health of the nation (97). The NHIS pioneered the concept of asking people to rate their own health, using three main approaches. First, to detect the presence of health conditions, interviewers read lists of diagnoses to respondents and ask them whether they have or have had the conditions. Second, to ascertain the effects of illnesses, the NHIS asks respondents about the extent to which illness has caused any loss of work, absences from school, or days in which normal activities have been restricted. Third, since its inception the NHIS has asked respondents the following widely used health status question, which has proven valuable for many purposes: "Overall how *would you* rate your health--exce//ertt, very *good, good, fair, or poor?"*

Measuring the effects of illnesses by measuring the resulting disabilities or restrictions in activities has allowed researchers to evaluate the costs and other consequences of illness at a population level. The extent to which illness causes people to restrict their activities is also a functional measure that shows up in many studies of the outcomes of medical care. Although the NHIS was not designed for such studies, it is one of the most pervasive sources of questions used to assess health status and medical treatment.

SOURCE: **F.J. Fowler, 1995**

Sickness Impact Profile

In the 1970s, the National Center for Health Services Research (the predecessor of the Agency for Health Care Policy and Research) funded a program to develop a comprehensive instrument to measure the effect of sickness on people (8). This instrument, the Sickness Impact Profile (SIP), includes 136 statements about people's functioning and activities, such as:

= "I am not doing heavy work around the house."
= "I laugh or cry suddenly."
= "I walk shorter distances or stop to rest often."

These questions are grouped into three broad categories ("indices"), each of which has several subcategories. Peoples' responses to these statements thus produce measures of how illness affects 12 different aspects of patients' lives (box 1-2).

The entire SIP takes about 30 minutes to administer and is perhaps the most comprehensive and detailed inventory in common use. It has been subjected to extensive psychometric evaluation to assess its reliability, its stability over time, its ability to differentiate well people from sick people,

BOX 1-2: Aspects of Health-Related Quality of Life Measured by the SIP and the SF-36

The Sickness Impact Profile

The SIP measures 12 aspects of health-related quality of life, grouped in three categories, as follows:

/dependent categories

- sleep
- eating
- work
- home management
- recreation and pastimes

Physical

- ambulation
- mobility
- body care and movement

Psychosocial/

- social interaction
- alertness behavior
- emotional behavior
- communication

RAND **36-Item Health Survey**

The aspects of health-related quality of life covered in the SF-36 include:

current perception of health

- psychological well-being
- role limitations due to physical health problems
- role limitations due to mental health problems
- physical function
- social relations
- pain
- fatigue

SOURCE M Bergner, R A Bobbttt, W B Carter, etal,, "The Sickness Impact Profile: Development and Final Revision of a Health Status Measure," Medical Cara 19(8) 787-805, 1981, and J,E. Ware and C.D. Sherbourne, "The MOS 36-item Short-Form Health Survey (SF-36)–1 Conceptual Framework and Item Selection," Medical Care 30(6):473-483, 1992

and its capacity to reflect positive effects of treatment, as well as to verify the internal consistency of its scales. The aspects of living reflected in the profile's 12 subindices tend to mirror those in current assessments of medical outcomes and patients' functioning. The basic approach developed in the SIP has had a major influence on subsequent efforts to develop better methods to evaluate medical outcomes. Moreover, all or part of the SIP is often used today in studies of medical outcomes.

The RAND 36-Item Health Survey

The SF-36 survey is probably the nation's most widely used generic instrument for measuring patients' assessments of health-related quality of life. The origins of the SF-36 lie in a health survey developed for the Health Insurance Experiment (HIE),[3] one of the major health research efforts of the 1970s (13,84). The 20-item questionnaire that emerged from the HIE later became a key instrument for collecting data in the Medical Outcomes Study, undertaken in the 1980s by some of the researchers who had worked on the HIE. The primary goal of the Medical Outcomes Study was to describe the health status of patients before and after medical treatment. The questionnaire later evolved into a 36-item, eight-index set of questions measuring various aspects of health, functioning, and quality of life (86) (box 1-2).

A primary goal in the development of the SF-36 was to identify a minimum set of health status dimensions that would cover most of the gen-

[3] The HIE focused chiefly on how various insurance packages with different deductibles affected utilization and cost. An important part of the study was an assessment of how people's health and well-being were affected by the different programs. Participants in the HIE filled out numerous questionnaires, which contained items from most of the major health survey measures available at the time, including the SIP. The results were analyzed to identify redundancy and independent dimensions of health status and functioning (13).

eral medical outcomes that researchers would want to measure, and to ask the minimum number of questions that would reliably and validly measure each dimension. The measure of psychological well-being, for example, consists of five items that have proved to be the measurement equal (for assessing aggregate outcomes) of as many as 30 of the items frequently used in other instruments to assess mental distress (9). Like the SIP, the SF-36 was envisioned as a generic instrument that would be appropriate for use in studying the treatment of virtually any health condition.

Although the developers of the SF-36 encourage researchers to use it as a complete package, the individual indices included in the SF-36 may be used by themselves, as can subsets of the SIP (71).

Quality of We//-Being Scale (QWB)

The Quality of Well-Being Scale (QWB), which emerged from the work of James Bush and his associates (44), uses a different approach. The SIP and the SF-36 rely on patients' reporting alone and were designed to be analyzed by looking at scores on individual subscales, which produce markedly different profiles depending on the type of illness. A summary SIP score can be calculated for overall functioning and for each of the three subdomains, and work is underway to derive a total score for the SF-36, but neither questionnaire was designed primarily to produce a single summary of well-being. In designing the QWB, however, Bush and his associates focused specifically on producing a quantitative measure of overall well-being.

To do so, these researchers created a list of deviations from perfect health. The list includes symptoms (such as headaches, sore throats, and trouble sleeping), conditions (such as hernias, overweight, and blindness), and activity or role limitations (such as missing work and being unable to drive a car). The respondent is asked whether any of these problems occurred during the preceding four days, and he or she rates each problem numerically according to the degree to which each of the problems reduced his or her well-being (64). These ratings are then combined by the researchers to produce a single number representing the overall well-being of the person.

In a variation on this approach, Torrance has developed the Health Utilities Index (73), which identifies nine health domains (vision, hearing, speech, the ability to get around, the use of hands and fingers, feelings, memory, thinking, and pain and discomfort). As with the QWB Scale, the Health Utilities Index entails calculating a weighted score that reflects the existence and seriousness of the problems reported in each domain.

RELIABILITY AND VALIDITY

For those schooled in the physical and biological sciences, the notion that good measurement can come from asking people questions seems somewhat implausible. Nonetheless, the criteria for evaluating questions as measures of health status are the same as those for evaluating measures used in laboratories, physicians' offices, or anywhere else. When consistent standards are applied, measurements based on asking people questions stand up very well.

I Reliability

Reliability means that measurement is consistent: when two patients are in the same situation, their answers to the questions should be the same. To the extent that there is inconsistency among patients, or at different times with respect to the same patient (when the patient's circumstances have not changed), the measurement is unreliable and imprecise.

The most commonly used measure of reliability in medical science is test-retest reliability, in which researchers compare two readings from the same person at different points in time. When no change in the patient's condition is thought to have occurred, the readings should be consistent. Researchers assessing the subscales used in the SIP, the SF-36, and other similar questionnaires routinely reported test-retest reliabilities of 0.85 and above.[4]

Validity

Assessing the validity of patients' reports for the purpose of evaluating medical outcomes is often difficult. Where there is a standard, a measure that everyone agrees is an accurate measure, validity can be assessed simply by comparing the results to the standard. Because no generally accepted standard for measuring patients' functioning or well-being exists, however, the evidence for the validity of patients' reports must come from the predictability of relationships.

Clinical measures are evaluated by examining the extent to which they discriminate between known groups and the extent to which they are responsive to treatments thought to be effective (47). Thus, a valid measure of symptoms of prostate disease, for example, should show higher levels among patients diagnosed with BPH than among the general population and higher levels in patients before they are surgically treated than after they are treated. The general approach of looking at patterns of association, how well the measures correlate with things with which they ought to be correlated, is the primary basis on which validity is assessed.

Many survey instruments used in clinical work ask patients multiple questions that cover the same general area. The study of associations between the answers to similar questions constitutes an important strategy for validating questions as measures. Questions about pain should correlate positively with other measures of discomfort, and they should be less correlated with measures of fatigue. The measurement can be strengthened by combining the answers to several questions to form an index. The reasons for using multi-item scales is that, all things being equal, multi-question scales are better than a single question at measuring what those questions have in common. (The extent to which multi-item scales provide a consistent and reliable measure of what they have in common is calculated by a statistic called Cronbach's alpha [20].)

Another issue *is face validity,* which means that the answers to questions mean what a reader of the wording of the question would most likely think they would mean. On the one hand, having questions that clinicians agree adequately cover what needs to be covered is critical to the acceptance of the results. On the other hand, questions cannot be presumed to be good measures just because they sound like the right questions. A requirement for any scientific enterprise is that the quality of measurement be documented through experiment and observation.

A good example was set by the researchers responsible for developing the SIP and those developing the SF-36 and related measures (8, 13,71). In the course of these programs of research, the investigators uniformly reported the ability of the measures to discriminate among clinical groups, the internal consistency of multi-index measures, the responsiveness to treatment, and the patterns of association with other measures with which they should be correlated. In all these respects, measures of the subscales in the SIP and of the various scales used in the Medical Outcomes Study meet high standards. The Cronbach's alpha rates routinely exceed 0.80, and correlations among related concepts are also very high.

The same kind of standards can be applied to more specific measures aimed at particular conditions and symptoms. A recent effort by the Measurement Committee of the American Urological Association to measure symptoms of BPH, comparing alternative measures of symptoms, demonstrates the high quality of measurement based on patients' reports (6). The committee compared and contrasted four different sets of questions about symptoms of BPH (7). Samples of patients and nonpatients answered questions

[4] A reliability of 1.0 would mean that the instrument yielded identical answers every time. A score of 0.85 is generally considered acceptably high (58).

twice, one week apart. The test-retest reliabilities of all four scores exceeded 0.75. The internal-consistency measure, Cronbach's alpha, for all four indices exceeded 0.80. The intercorrelations among the four indices, which partly reflected the overlap of items, were all above 0.75. These statistics show that BPH has meaningful symptoms that sets of questions can measure in a consistent and apparently valid way. Furthermore, when the answers of patients diagnosed as having BPH were compared with the sample of healthy individuals, 85 percent of the people would have been correctly classified as BPH patients or nonpatients based on their answers to those questions.

Comparison with Other Tools

Although the results of diagnostic tests and results based on patients' reports are difficult to compare directly, studies of diagnostic tests and of measurements taken in clinical settings almost always reveal considerable error. Blood pressure readings, for example, are often inaccurate: using the wrong cuff size is common and produces serious overestimates of blood pressure (30,53); and in a phenomenon known as "white coat" response, 20 to 40 percent of people who have elevated blood pressure readings in doctors' offices have normal blood-pressure readings in other settings (49,65). Thus, even though the measurement of blood pressure is considered an important procedure upon which important diagnoses and treatment decisions are based, the measurement process is fraught with potential error.

The lack of correspondence between the results of image studies and the symptoms of people with lower back pain provides another example of a traditional medical test that is not a consistently reliable or valid indicator of a health condition (12,67,89). Similar problems have arisen with the use of image studies to diagnose arthritis (23,81).

To help evaluate BPH, urologists have traditionally used a measure of the residual urine left in the bladder after voiding and have also begun using a measure of the rate of urine flow to assess obstruction. These measures correlate poorly with patients' symptoms, however (5,14). Although factors that have nothing to do with BPH status (such as recent fluid intake and patients' anxiety) apparently affect the measures, they continue to be a common part of the urologic diagnostic process.

There are at least three reasons why clinical tests may not be valid measures.

- First, the variable state being measured may not be a reliable indicator of the condition of a patient. (Because blood pressures go up and down in response to circumstances, the reading at any point in time may not be a good indicator of the usual state of a person's blood pressure.)
- Second, the measurement may be performed inconsistently or incorrectly, affecting the results. (Using the wrong cuff to measure blood pressure yields an erroneous reading.)
- Third, what can be measured may not be informative about the condition of interest. (In the case of back pain, some of the things that affect the nerves coming out of the spine are apparently not visible in image studies.)

Measuring clinical or medical states by asking people questions is subject to the same kinds of problems. Whether people can answer questions that provide valid measurements of a clinical state is an empirical question to be tested, and the validity of patients' reports can vary from condition to condition. (No matter how well they can describe their pain or functioning, for instance, patients cannot say what their blood pressures are based on feelings alone.) Aspects of the data collection procedures, such as the quality of interviewing in those cases where interviewers are used, also can affect the results (33).

Patients' reports cannot substitute for other strategies of clinical observation and diagnosis, nor are medical tests inherently unreliable. The reliability and validity of clinical and laboratory tests vary, as do those of measures based on patients' reports. For measuring what patients observe and experience, however, properly designed questions can produce measures that compare favorably in reliability and validity with traditional clinical measures of health.

ISSUES

Broad-based agreement on how to conduct good studies of the outcomes of medical care is emerging, but consensus on the details is lacking. Many studies are now designed to collect data about the treated condition and complications of treatment using patients' questionnaires or interviews. Ideally, for the sake of simplicity and comparability, all studies of a particular condition would use the same measures. There are very few conditions, however, for which a specific set of questions is widely accepted. In general, researchers are still developing and revising questions to meet their perceptions of what is best.

The lack of consensus on specific measures of health states does not mean that studies cannot be compared. Questions that validly measure the same underlying conditions will produce similar results, even if the wording is different. The scores of the resulting indices of four recently evaluated series of questions that have been used in published studies to measure the severity of BPH, for example, intercorrelated very highly (6). Consequently, studies using any of the four measures are likely to produce similar conclusions about the effect of treatment.

Medical outcomes studies now routinely include general, as well as condition-specific, measures of functioning and perceived well-being. The domains (i.e., the aspects of health and well-being) covered in the general measures are similar, drawing on those covered in the SIP and SF-36, but the particular indices and questions vary.

Some of the diversity in the choice of measures reflects the characteristics of the condition or the populations being studied. The range of functioning to be measured in studies of stroke victims, for example, is very different from that in studies of women who have had Cesarean sections, both because of the patients' ages and because of the way the conditions affect people.

Most of the outcomes studies being done by the PORTS, funded by the Agency for Health Care Policy and Research, are using some of the indices from the SF-36. In addition, various PORTS are using all or part of the SIP, asking specific questions about activities of daily living and instrumental activities of daily living,[5] and inquiring about disability days or restricted-activity days, as part of their protocols to assess the effects of treatment comprehensively. One argument for using the entire SIP or SF-36 is that these measures produce comprehensive profiles of the patients. Researchers differ, however, in how much they value measures of domains that are not likely to be affected by a particular treatment.

Thus, although there is virtually complete agreement on the need for measures of patients' self-reported health, there is diversity in the questions chosen by different researchers. The differences reflect the conditions being studied, the populations of patients, the burdens deemed appropriate for respondents in particular projects, and the personal convictions of the researchers about which specific measures are best suited for studying particular treatments.

Some convergence will probably occur as researchers gain experience. More systematic evaluation of questions is needed, however. Questions need to be tested with cultural minorities, for example, to ensure the questions really mean the same thing to everyone. Optimal questions about role limitations-questions that apply equally well to all age groups, including children and retired persons—have yet to be found.

Although these issues are important and need to be addressed, they are relatively minor problems that should not detract from the agreement about the need for general measures of health-related quality of life and about the advances in developing good measures of the major aspects of quality of life. Nonetheless, at least three major methodological challenges remain. If medical

[5] See box 1-1.

outcomes studies are to live up to their promise, each of the following issues must be resolved:

- How should prospective and retrospective designs be modified to ensure accurate measurements of the effects of treatment?
- How should researchers collect information about the results that would have been expected had a particular treatment not been given?
- How should the effects of treatment be calibrated to facilitate comparisons across conditions?

Prospective vs. Retrospective Designs

Both prospective and retrospective designs are used to assess treatment effects with patient surveys. In a prospective study, patients are asked the question *"How are you doing?"* before treatment and again at a later point in time. The effect of the treatment is assessed by comparing the two answers. In a retrospective approach, people who have already been treated are asked to compare their present state with how they felt prior to the treatment: *"DO you think you are doing better now, worse now, or about the same?"*

The two methods do not always yield the same results. Some people report that they feel better after treatment even though comparisons of their reported symptoms before and after treatment indicate no changes in their conditions (39,54). Some studies of medical outcomes are most easily done retrospectively: one cannot easily identify (or collect data from) the individuals who will later have heart attacks or suffer accidental injuries, whereas surgical patients are relatively easy to identify after they have had surgery. Therefore, it is important to develop an understanding of how best to conduct both prospective and retrospective studies. The measurement implications of the two kinds of designs should also be considered.

1 Better-Than-Expected Results

Assessing whether results are better than expected is another methodological problem that needs work. Although ascertaining whether patients change for the better by virtue of being treated may seem a good way to assess treatment results, considerable medical care is intended merely to keep patients from getting worse. The management of a patient who has had an acute myocardial infarction (AMI),[6] for example, is designed to make the recovery process as good as possible. Because people who have suffered AMIs cannot reasonably be expected to be better off than they were before the AMIs, their health status must be compared with what it would have been had they been treated differently. By the same token, although measuring the reduction in symptoms may be a good way to assess the value of the treatment (where symptom relief was the primary goal of the treatment), some people improve without treatment.

These examples underscore the fact that all treatment or outcomes studies require controls or comparisons and for those, studies of untreated people (or some other "control" group) are necessary. The traditional standard in clinical research is a randomized controlled trial (RCT). The design is good when it is feasible, but such studies often have not been done and sometimes cannot be done.

In the absence of good RCTS, researchers have been trying to do better descriptive studies of patients who undergo particular treatments, but valid conclusions about the treatments' effects are difficult to reach without good data about what would have happened to the patients had they received no treatment or alternative treatments. Such data are scarce. One critical gap is the relative lack of natural history studies. Patients who present themselves to physicians and meet criteria for surgical treatment are 1ikel y to get the surgical treatment, particularly in the United States. There is a dearth of studies that systematically follow candidates for surgery or hospitalization who do not actually receive the surgery or hospitalization.

[6] An acute myocardial infarction is a type of heart attack.

On a related issue, cohorts who are not given an extreme treatment, such as surgery, tend to be different from the aggressively treated group. As a result, appropriate data about the symptom status, comorbidities, and general health of both cohorts must be collected so that appropriate controls can be used in analytic comparisons of the outcomes. The Medical Outcomes Study used this kind of design. Its approach was a major advance over having no comparison group at ali, but researchers often have difficulty making adjustments to ensure that the comparisons are appropriate (43). Agreed-upon methods for cohort studies of people who receive different treatments or no treatments must be developed to provide data that will enable researchers to reach valid conclusions about treatment effects.

Calibrating Measurements of Treatment Effect

Measuring patients' views of the significance of particular clinical states, including complications of treatment, is central to the problem of how to measure the value of a treatment. In the past, there were few good studies of the overall health status of patients before and after treatments. As more such studies are conducted, however, the question of how to calibrate benefits will become much more salient.

A single summary measure of the net significance or value of medical treatment would be useful for decision analysis,[7] for ranking the value of performing various medical procedures, or for deciding whether a particular treatment is one for which we are willing to pay.

In clinical practice at the individual patient level, a single summary measure can be obtained simply by describing the various possible results to the patient, who then makes his or her own choice based on personal preferences and values. Some problems, however, raise social questions of cost, ethics, or best medical practice (box 1-3). Producing good statistical descriptions of the results of treatments might improve judgments and social choices in these cases.

The QWB Scale and the Health Utilities Index seek to address this need by asking groups of people to rate quantitatively how they value vari-

BOX 1-3: Questions That Might Benefit from a Summary Measure of Treatment Effect

- Suppose two treatments are available: one has an 80-percent chance of relieving the symptoms and a 20-percent chance of producing certain side effects; the alternative is less effective but has fewer side effects. Which is the best treatment?
- Suppose the costs of two treatments are significantly different. ɪs the more expensive treatment justified?
- Suppose a treatment is found that can make a measurable improvement in the cognitive functioning of mentally impaired elderly patients, but the treated patients remain substantially impaired after treatment How should the value, if any, of such a treatment be calculated?
- Suppose a treatment will prevent 30 premature deaths for every 1,000 people treated, but most of the treated people will have significant short-term side effects, a few will have long-term quality-of-life loss, and the treatment is expensive. Should the treatment be used?

SOURCE F.J. Fowler, 1995

[7] In a decision analysis, the analyst considers the variety of possible treatment options, associates each treatment option with a set of probabilities for good and bad outcomes, and tries to put them together to illuminate the implications of each treatment (45). Critical components of any decision analysis are the values assigned to the various health states in which patients may find themselves, with or without treatment. These measures of significance-numbers assigned to describe how good or bad patients' states **are--can be derived from each individual patient, from the average ratings of a group of patients, or from independent ratings (59).**

ous states of health. In doing so, instruments like these raise issues about the method by which the ratings of health status are derived—issues that do not arise with instruments such as the SIP and SF-36, which do not attempt to come up with a summary measure of health-related quality of life. Two issues are especially central:

1. What questions should be asked to rate health status?
2. Who should be the raters?

One way to measure what significance a condition holds for people is to ask them to rate it numerically: on a scale from O to 100, where O is death and 100 is perfect health, what number would you give to, for example, lower back pain? This is the approach used by Kaplan and his associates (42). Other researchers, however, believe that the valid measurement of the significance or importance of a condition requires asking people how much they are willing to pay, to risk, or to lose in order to get rid of a condition—an approach that leads to an entirely different set of questions (26,35,36,37,38,73). These researchers prefer the standard reference gamble, which goes something like this:

> Option A is to have no treatment at all and stay in your current state of health. Option B involves accepting a treatment. If the treatment is successful, it will cure your condition and return you to perfect health. If it is unsuccessful, you will die. With what chance of success would you choose Option B over Option A?

A variation is called the time tradeoff. It also trades off life against health, quantity versus quality of life, but in a different way:

> Consider the possibility that you will live 10 years with your health just the way it is now. Suppose I could offer you a treatment that would return you to perfect health, without the condition, but you would live fewer years. How many years of perfect health would you consider to be the same as 10 years in your current health state?

These approaches presume that the greater the risks people are willing to take or the more of their lives they are willing to give up to improve their current health, the worse the states of health in which they find themselves.

Studies assessing the significance of health conditions have used all of these approaches: asking patients to rate how they think they are affected by various health conditions, asking people to rate how they think they would feel if they were in various health states, and asking expert raters (such as physicians) to say how they think patients would feel if they got into various states. The QWB and Health Utilities Index use ratings by samples of people to assign weights and produce a summaries of well-being. Both have been used in clinical studies of medical outcomes. In addition, a variation of QWB was used in Oregon to set priorities for proposed revisions in the Medicaid payments system (34,77), and the Health Utilities Index was used by Statistics Canada to assess well-being in a general population survey in Ontario.

Researchers disagree about whether scale- or risk-based approaches are better. Many researchers believe that questions based on the standard gamble or time tradeoff approach are by far the best way to measure how significant particular health states are to people (72). Others point out that these are very hard questions to answer, and that the answers may not have the meaning the researchers hoped for. Moreover, questions based on gambles and tradeoffs reflect not only the value of health states but also the individuals' attitudes about trading quality and quantity of life and toward taking risks, and thus they have been criticized as producing confounded—rather than better—measures of the value of health states. Research using both approaches continues.

As for the issue of whose values should be reflected in the ratings, the answer depends in part on the purposes for which data are being collected. If a physician is treating an individual patient, the patient's preferences should have priority. For managed health care, however, the values of the average patient might be the most relevant (59). A different set of priorities might be appropriate for an insurance company. In that context, the perspectives of the people who are paying the premiums might be most appropriate. Applying that logic to government-funded health care might entail using the values of a cross-section of the general public to determine the ratings (64). But in the context of government, where the question of whose values matter is a political as well as an academic one, there is no unambiguous answer.

Thus, although the SIP, the SF-36, and similar survey instruments can yield summary measures, their strength lies in producing profiles of the various ways a health condition affects people's lives. The QWB and Health Utilities Index researchers address the problem more directly, but they have not resolved the perplexing issues of which questions to ask, whose values to measure, and how to create an overall summary of the quality of life. Describing patients' post-treatment status on various indices may actually be the best form in which to convey information to patients and physicians, but those who want a simple summary number—for decision analysis, for ranking the value of hysterectomies and fixing broken legs, or for deciding whether to pay for a particular treatment-do not yet agree about how to proceed.

CONCLUSION

It is not accidental that researchers' recent interest in developing measures of health-related quality of life has coincided with widespread interest in better assessing the value of current medical treatments. Patients' reports about their perceptions of their symptoms, about the significance of their conditions, and about their general functioning and quality of life are essential to documenting what benefits, if any, patients derive from treatments.

One of the contributions of the PORT concept has been to emphasize the patients' perspective in the evaluation of medical treatments. Although some very good work was done in the 1970s and became the foundation of current work, the focus on how patients fare after treatment is mainly arecent phenomenon. That patients' reports can provide valid and reliable measures of their health status has been clearly demonstrated. Indeed, measures from patients' reports often prove better than those from commonly used clinical and laboratory tests, and studies that include patients' perspectives have produced sound results that sometimes raise questions about standard medical practice.

Work remains to be done in developing and improving measures. Researchers need to increase their understanding of how best to conduct these studies to reach valid conclusions and how best to assess the significance of the results. Nonetheless, in a comparatively short time, an appreciation for patient-oriented outcomes studies and how to do them has developed a great deal. They can be done, and they produce considerable knowledge that neither patients, clinicians, nor researchers have had before.

REFERENCES

1. Abrams, P. H., and Greffeths, D.J., "The Assessment of Prostatic Obstruction from Urodynamic Measurements and from Residue Urine," *British Journal of Urology 51:129, 1979.*
2. American College of Cardiology/American Heart Association, "Special Report: A Report of the American College of Cardiology/ American Heart Association Task Force on Assessment of Diagnostic and Therapeutic Cardiovascular Procedures (Subcommittee on Coronary Artery Bypass Graft Surgery)," *Circulation 83(3) :1125-1172, 1991.*
3. Andersen, J.T., Nordling, J., and Walter, S., "Prostatism-I: The Correlation Between Symptoms, Cytometric, and Urodynamic Findings," *Scandinavian Journal of Urolology and Nephrology 13;229, 1979.*
4. Andrews, F. M., and Withey, S. B., *Social Indicators of Well-Being: Americans' Perceptions of Life Quality (New* York, NY: Plenum, 1976).
5. Barry, M.J., Cockett, A. T. K., Holtgrewe, H.L., et al., "Relationship of Symptoms of Prostatism to Commonly-Used Physiological and Anatomical Measures of the Severity of Benign Prostatic Hyperplasia," *Journal of Urology,* in press.

6. Barry, M.J., Fowler, F.J., O'Leary, M. P., et al., "The American Urological Association Symptom Index for Benign Prostatic Hyperplasia," *Journal of Urology 148: 1549-1557, 1992.*

7. Barry, M.J., Fowler, F.J., O'Leary, M. P., et al., "Correlation of the American Urological Association Symptom Index with Self-Administered Versions of the Madsen-Iversen, Boyarsky, and Maine Medical Assessment Program Symptom Indexes," *Journal of Urology 148:1558-1563, 1992.*

8. Bergner, M., Bobbitt, R. A., Carter, W. B., et al., "The Sickness Impact Profile: Development and Final Revision of a Health Status Measure," *Medical Care 19(8):787-805, 1981.*

9. Berwick, D. M., Murphy, J. M., Goldman, P. A., et al., "Performance of a Five-Item Mental Health Screening Test," *Medical Care 29(2): 169-176, 1991.*

10. Bigos, S.J., Battie, M. C., Fisher, L. D., et al., "A Prospective Evaluation of Preemploy - ment Screening Methods for Acute Industrial Back Pain," *Spine 17(8):922-926, 1992.*

11. Black, N., Petticrew, M., Ginzler, M., et al., "Do Doctors and Patients Disagree? Views of the Outcome of Transurethral Resection of the Prostate," *International Journal of Technology Assessment in Health Care 7(4):533-54-4,*

12. Boden, S. D., Davis, D. O., Dina, T. S., et al., "Abnormal Magnetic-Resonance Scans of the Lumbar Spine in Asymptomatic Subjects," *Journal of Bone and Joint Surgery* 72(3): 403-408, 1990.

13. Brook, R. H., Ware, Jr., J. E., Davies-Avery, A., et al., *Conceptualization and Measurement of Health for Adults in the Health Insurance Study: Vol. VIII—Overview,* Publication No. R-1987/8-HEW (Santa Monica, CA: RAND Corp., 1987).

14. Bruskewitz, R. C., Iversen, P., and Madsen, P. O., "Value of Postvoid Residual Urine Determination in Evaluation of Prostatism," *Urology* 20:260, 1982.

15. Carlson, K.J., Miller, B.A., and Fowler, Jr., F.J., "The Maine Women's Health Study: I-Outcomes of Hysterectomy," *American Journal of Obstetrics and Gynecology 83:556-565,* 1994a.

16. Carlson, K. J., Miller, B. A., and Fowler, Jr., F.J., "The Maine Women's Health Study: II—Outcomes of Nonsurgical Treatment for Fibroids, Abnormal Bleeding, and Chronic Pelvic Pain," *American Journal of Obstetrics and Gynecology 83:566-572,* 1994b.

17. Cleary, P. D., Edgman-Levitan, S., Roberts, M., et al., "Patients Evaluate Their Hospital Care: A National Study," *Health Affairs 10(4):254-267, 1991.*

18. Cleary, P. D., and McNeil, B.J., "Patient Satisfaction as an Indicator of Quality of Care," *Inquiry 25:25-36, 1988.*

19. Collins, R., Pete, R., MacMahon, S., et al., "Blood Pressure, Stroke, and Coronary Heart Disease," *Lancet 335:827-838, 1990.*

20. Cronbach, L.J., "Coefficient Alpha and the Internal Structure of Tests," *Psychometrika 16:297, 1951.*

21. 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.*

22. Deyo, R. A., "The Quality of Life, Research and Care" (editorial), *Annals of Internal Medicine* 114:695-697, 1991.

23. Deyo, R. A., "Comparative Validity of the Sickness Impact Profile and Shorter Scales for Functional Assessment in Low Back Pain," *Spine* 11:951-54, 1986.

24. Epstein, A. M., Hall, J. A., Tognetti, J., et al., "Using Proxies To Evaluate Quality of Life: Can They Provide Information About Patients' Health Status and Satisfaction with Medical Care?" *Medical Care 27(3):S91-98, 1989.*

25. EuroQol Group, "EuroQol-A New Facility for the Measurement of Health-Related Quality of Life," *Health Policy 16: 199-208, 1990.*

26. Feeny, D. H., and Torrance, G. W., "Incorporating Utility-Based Quality-of-Life Assessment Measures in Clinical Trials: Two Examples," *Medical Care* 27(3)(suppl.):S *190-S204, 1989.*

27. Fillenbaum, G.G., and Smyer, M. A., "The Development, Validity, and Reliability of the OARS Multidimensional Functional Assessment Questionnaire," *Journal of Gerontology* 36:428-434, *1981.*

28. Fisher, B., Redmond, C., Poisson, R., et al., "Eight-Year Results of a Randomized Clinical Trial Comparing Total Mastectomy and Lumpectomy With or Without Irradiation in the Treatment of Breast Cancer," New *England Journal of Medicine* 320:822-828, *1989.*

29. Fleming, C., Wasson, J. H., Albertsen, P. C., et al., "A Decision Analysis of Alternative Treatment Strategies for Clinically Localized Prostate Cancer," *Journal of the American Medical Association* 269:2650-2658, *1993.*

30. Floras, J. S., Hassan, M. O., Osikowska, B., et al., "Cuff and Ambulatory Blood Pressure in Subjects with Essential Hypertension," *Lancet* 2(8238): 107-109, 1981.

31. Fowler, F.J., "Patient Reports of Symptoms and Quality of Life Following Prostate Surgery," *European Journal of Urology 20* (suppl. 2):44-49, 1991.

32. Fowler, F. J., and Mangione, T. W., *Standardized Survey Interviewing, Minimizing Interviewer-Related Error* (Newbury Park, CA: Sage Publications, 1990).

33. 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(20): 3018-3022, 1988.*

34. Fox, D. M., and Leichter, H. M., "Rationing Care in Oregon: The New Accountability," *Health Affairs* 10(2):728, *1991.*

35. Froberg, D.G., and Kane, R. L., "Methodology for Measuring Health-State Preferences—1: Measurement *Strategies, ''Journal of Clinical Epidemiology* 42(4):345-354,1989.

36. Froberg, D. G., and Kane, R. L., "Methodology for Measuring Health-State Preferences—II: Scaling *Methods, ''Journal of Clinical Epidemiology 42(5):459-471, 1989.*

37. Froberg, D. G., and Kane, R. L., "Methodology for Measuring Health-State Preferences—III: Population and Context Effects," *Journal of Clinical Epidemiology* 42(6):585-592, *1989.*

38. Froberg, D. G., and Kane, R. L., "Methodology for Measuring Health-State Preferences—IV: Progress and a Research Agenda," *Journal of Clinical Epidemiology* 42(7):675-685, *1989.*

39. Guyatt, G. H., Townsend, M., Keller, J. L., et al., "Should Study Subjects See Their Previous Responses? Data from a Randomized Control Trial," *Journal of Clinical Epidemiology* 42(9):913-920, *1989.*

40. Herron, L. D., and Turner, J., "Patient Selection for Lumbar Laminectomy and Discectomy with a Revised Objective Rating S ystem," *Clinical Orthopedics and Related Research* 199:145-152, *1985.*

41. Hunt, S. M., McEwen, J., and McKenna, S. P., *Measuring Health Status* (Dover, NH: Croom Helm, 1986).

42. Kaplan, R. M., Anderson, J. P., Wu, A. W., et al., "The Quality of Well-Being Scale: Applications in AIDS, Cystic Fibrosis, and Arthritis," *Medical Care* 27(3)(suppl.):S27-S43, *1989.*

43. Kaplan, R. M., and Berry, C. C., "Adjusting for Confounding Variables," *Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data--Conference Proceedings,* (Rockville, MD: Agency for Health Care Policy and Research, May 1990)

44. Kaplan, R. M., and Bush, J. W., "Health-Related Quality of Life Measurement for Evaluation and Research and Policy Analysis," *Health Psychology* 1:61, *1982.*

45. Kassirer, J.P., Moscowitz, A. J., Law, J., et al., "Decision Analysis: A Progress Report," *Annals of Internal Medicine 106:275-291,* 1987.

46. Katz, S., Ford A. B., Moskowitz, R. W., et al., "Studies of Illness in the Aged. The Index of ADL: A Standardized Measure of Biological and Psychosocial Function," *Journal of the American Medical Association* 185:914-919, *1963.*

47. Kirshner, B., and Guyatt, G., "A Methodologic Framework for Assessing Health Indices," *Journal of Chronic Disease* 38:27, *1985.*
48. Lee, N. C., Dicker, R.C., Rubin, G. L., et al., "Confirmation of the Preoperative Diagnoses for Hysterectomy," *American Journal of Obstetrics and Gynecology,* 150:283-287, *1984.*
49. Lerman, C. E., Brody, D. S., Hui, T., et al., "The White-Coat Hypertension Response: Prevalence and Predictors," *Journal of General Internal Medicine* 4:226-231,1989.
50. Magaziner, J., Simonsick, E. M., Kashner, T. M., et al., "Patient-Proxy Response Comparability on Measures of Patient Health and Functional Status," *Journal of Clinical Epidemiology 41(1* 1):1065-1074, *1988.*
51. Mahoney, F. I., and Barthel, D. W., "Functional Evaluation: The Barthel Index," *Maryland State Medical Journal* 14:61-65,1965.
52. Manning, W.G., Newhouse, J. P., and Ware, Jr., J. E., "The Status of Health in Demand Estimation, or Beyond Excellent, Good, Fair, and Poor," *Economic Aspects of Health,* V.R. Fuchs (cd.) (Chicago, IL: University of Chicago Press, 1982).
53. Manning, D. M., Kuchirka, C., and Kaminski, J., "Miscuffing: Inappropriate Blood Pressure Cuff Application," *Circulation* 68(4):763-766, *1983.*
54. MacKenzie, R. C., Charlson, M. E., DiGioia, D., et al., "Can the Sickness Impact Profile Measure Change? An Example of Scale Assessment," *Journal of Chronic Disease* 39(6):429-438, *1986.*
55. Mebust, W. K., Holtgrewe, H. L., Cockett, A.T., et al., "Transurethral Prostatectomy: Immediate and Postoperative Complications. A Cooperative Study of 13 Participating Institutions Evaluating 3,885 Patients, ''Journal *of Urology* 141:243-247,1989.
56. Medical Research Council Working Party, "Stroke and Coronary Heart Disease in Mild Hypertension: Risk Factors and the Value of Treatment," *British Medical Journal 296: 1565-1570, 1988.*
57. Meenan, R.F., "The AIMS Approach to Health Status Measurement: Conceptual Background and Measurement Properties," *Journal of Rheumatology* 9:785-788,1982.
58. McDowell, I., and Newell, C., *Measuring Health, A Guide to Rating Scales and Questionnaires (New* York, NY: Oxford University Press, 1987).
59. Mulley, A.G., "Assessing Patients' Utilities: Can the Ends Justify the Means?" *Medical Care 27(suppl.):S269-S281, 1989.*
60. Neal, D.E., Styles, R. A., Ng, T., et al., "Relationship Between Voiding Pressures, Symptoms, and Urodynamic Findings in 253 Men Undergoing Prostatectomy," *British Journal of Urology* 60:554,1987.
61. Paajamen, H., Erkintalo, M., Dahlstrom, S., et al., "Disc Degeneration and Lumbar Instability. Magnetic-Resonance Examination of 16 Patients," *Acta Orthopaedica Scandinavia* 60(4):375-378, *1989.*
62. Parliament, M. B., Danjoux, C. E., and Clayton, T., "Is Cancer Treatment Toxicity Accurately Reported?" *International Journal of Radiation Oncology, Biology, Physics* 11:603-608, 1985.
63. Patrick, D. L., Darby, S.C., Green, S., et al., "Screening for Disability in the Inner City," *Journal of Epidemiology and Community Health* 35:65-70, *1981.*
64. Patrick, D. L., and Erickson, P., *Health Status and Heath Policy, Allocating Resources to Health Care (New* York, NY: Oxford University Press, 1993).
65. Pickering, T. G., James, G. D., Boddie, C., et al, "How Common Is White Coat Hypertension?" *Journal of the American Medical Association* 259:225-228, *1988.*
66. Rothman, M. L., Hedrich, S. C., Bulcrot, K. A., et al., "The Validity of Proxy-Generated Scores as Measures of Patient Health Status," *Medical Care 29(2):* 115-124, 1991.
67. Spangfort, E. V., "Lumbar Disc Herniation: A Computer-Aided Analysis of 2,504 Opera-

tions," *Acta Orthopaedica Scandinavia, 142(suppl.):l-95, 1972.*

68. Spengler, D. M., Ouellette, E. A., Battie, M., et al., "Elective Discectomy for Herniation of Lumbar Disc. Additional Experience with an Objective Method," *Journal of Bone and Joint Surgery* 72(2):230-237, *1990.*

69. 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(7):907-913, 1989.*

70. 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(7):724-735, *1988.*

71. Stewart, A. L., and Ware, J.E. (eds.), *Measuring Functioning and Well-Being, the Medical Outcomes Study Approach* (Durham, NC: Duke University Press, 1992).

72. Torrance, G.W., "Measurement of Health State Utilities for Economic Appraisal," *Journal of Health Economics* 5:1-30, *1986.*

73. Torrance, G. W., "Utility Approach to Measuring Health-Related Quality of Life," *Journal of Chronic Disease* 40(6):593-600, *1987.*

74. Turner J. A., Ersek, M., Herron, L., et al., "Surgery for Lumbar Spinal Stenosis: Attempted Meta-Analysis of the Literature," *Spine* 17(1):1-8, *1992.*

75. Turner, C., and Martin, E., *Surveying Subjective Phenomena (New* York, NY: Russell Sage, 1984).

76. Uhlmann, R. F., Pearlman, R. A., and Cain, K. C., "Physicians' and Spouses' Predictions of Elderly Patients' Resuscitation Preferences," *Journal of Gerontology* 43(5):M115-M121, 1988.

77. U.S. Congress, Office of Technology Assessment, *Evaluation of the Oregon Medicaid Proposal,* OTA-H-531 (Washington, DC: U.S. Government Printing Office, May 1992),

78. U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, National Center for Health Statistics, "Hysterectomies in the United States, 1965 -1984," *Vital and Health Statistics,* Series 13, No. 155, DHHS Pub. No. (PHS)87-1753 (Hyattsville, MD: 1987).

79. U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, National Center for Health Statistics, "The National Ambulatory Medical Care Survey: United States, 1975-1981 and 1985 Trends," *Vital and Health Statistics, Series* 13, No. 93, DHHS Pub. No. (PHS)88-1754 (Hyattsville, MD: 1988).

80. U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, National Center for Health Statistics, "Data Systems of the National Center for Health Statistics," *Vital and Health Statistics, Series* 1, No. 23, (Hyattsville, MD: 1989).

81. Verbrugge, L. M., "Scientific and Professional Allies in Validity Studies," *Health Survey Research Methods-Conference Proceedings* (Rockville, MD: National Center for Health Services Research and Health Care Technology Assessment, September, 1989).

82. Verbrugge, L. M., and Balaban, D.J., "Patterns of Change in Disability and Well-Being," *Medical Care 27(3)(* suppl.):S128-S147, *1989.*

83. Veronesi, U., Banfi, A., Salvadori, B., et al., "Breast Conservation Is the Treatment of Choice in Small Breast Cancer: Long-Term Results of a Randomized Trial," *European Journal of Cancer* 26:668-670,1990.

84. Ware, Jr., J. E., Brook, R. H., Davies, A. R., et al., *Conceptualization and Measurement of Health for Adults in the Health Insurance Study: Volume I—Model of Health and Methodology,* Publication No. R-1987/ 1-HEW. (Santa Monica, CA: RAND Corp., 1987)

85. Ware, J., and Davies, A., *Scoring the Short-Form Mental Health Inventory* (MHI-5) (Santa Monica, CA: RAND Corp., 1983).

86. Ware, Jr., J.E., and Sherboume, C. D., ‘^{b}The MOS 36-Item Short-Form Health Survey (SF-36)—I: Conceptual Framework and Item Selection, ''Medical *Care* 30(6):473-483, 1992.

87. Ware, J. E., Snyder, M. R., Wright, R., et al., "Defining and Measuring Patient Satisfaction with Medical Care," *Evaluation and Program Planning* 6:247-263,1983.

88. Wasson, J. H., Cushman, C. C., Bruskewitz, R. C., et al., "A Structured Literature Review of Treatment for Localized Prostate Cancer," *Journal of the American Medical Association, 1993.*

89. Weisel, S. W., Tsourmas, N., Feffer, H. L., et al., "A Study of Computer-Assisted Tomography—1: The Incidence of Positive CAT Scans in an Asymptomatic Group of Patients," *Spine* 9(6):549-551, *1984.*

90. Wilson, R.G., Hart, A., and Dawes, P. J. D. K., "Mastectomy or Conservation: The Patient's Choice," *British Medical Journal 297(6657):* 1167-1169, 1988.

Large Administrative Database Analysis

Background Paper 2

SUMMARY

Large health administrative databases are used in three different ways to assess the effectiveness of medical treatments: in descriptive studies, in comparative studies, and as adjuncts to other research methods.

In descriptive studies, administrative databases can be used to provide estimates of the rates at which medical treatments are used. The degree to which these rates vary across population subgroups, time periods, and geographic areas can be contrasted. Administrative databases can also be used to provide general assessments of important clinical and economic outcomes experienced by individuals who receive the treatments. Such assessments can sometimes provide surprising results that raise questions about how medical treatments-even well-established treatments-are used.

Some researchers have also used administrative databases in comparative studies, to identiy populations that receive competing types of medical treatment. The populations' health outcomes -e.g., rates of mortality, rehospitalization, or reoperatio-are then compared. These comparative studies, howevever are rarely sufficient themselves to draw definitive conclusions about relative effectiveness, because like other nonrandomized studies their results are susceptible to unrecognized underlying biases that can render the conclusions invalid. Moreovever the quality and quantity of data in existing databases often limit the researchers' ability to use the adjustment techniques employed in other observational research.

This technique is most likely to produce valid results if the medical condition al issue and associated risk factors have been well-

by
Jeff Whittle
Veterans Administration
Medical Center
Pittsburgh, PA

studied, if the treatment is applied in a standardized way, and if the data needs are defined prospectively, so that the database is assembled with the research question in mind.

Although administrative databases are severely limited in their ability to be the basis for valid comparisons between technologies, they may substantially increase the weight of evidence about a treatment. They are also useful as adjuncts to other research methods. Using administrative databases as sampling frames, for example, allows researchers to identifiy populations of particular interest for further study. The ability to retrieve data from the medical records of all members of a representative population in a database, or to contact the individuals directly, could be extremely valuable, although it raises privacy issues that must not be dismissed. Another approach entails using a claims database to enhance the followup for a population carefully characterized by a study that entailed primary data collection.

Linking administrative databases with other medical information sources (e.g., cancer registry data), augmenting administrative data with additional elements (e.g., information on health status and functioning), and making other improvements in the availability and accuracy of data could also expand the usefulness of this tool.

One of the earliest demonstrations of the potential power of using routinely collected data on health services was an analysis published in 1938 by Glover (58), who described how the rates at which tonsillectomies were performed on British schoolchildren varied among school districts. Over the course of the past two or three decades, the prominence and volume of such analyses have soared (27), following several developments that facilitated the use of data from claims and discharge abstracts.

Advances in computer systems have allowed large numbers of records to be manipulated at reasonable cost, which has resulted in computers being used to store huge amounts of fiscal and administrative data and has facilitated the development of large registries of diseases and procedures. At the same time, processing these large databases has become less time consuming and expensive, making them more accessible for health services research.

In addition, new computer software (167) allows analysts to use sophisticated statistical techniques. For data sets as large as those now being examined, some statistical tests—such as Cox's proportional hazard analysis of survival data (1 15) and multiple logistic regression (200)—would have been nearly impossible to perform with manual techniques.

Another significant development occurred in 1965, when the United States established Medicare, the nationwide health insurance plan for individuals 65 years of age and older[1] (87,157). Because the vast majority of Americans in this age group are eligible and choose to participate, a sample of Medicare beneficiaries approximates a sample from the U.S. population aged 65 and over (47,65,1 14). Medicare gathers data on most of the health care that is provided to beneficiaries, and Social Security data on mortality can be used to ascertain a beneficiary's vital status after treatment (9). Because each Medicare beneficiary has a unique identification number, his or her use of health services over time can be traced (with some limitations) (114).

The interest in database analysis for assessing medical care burgeoned in response to several factors. First, the aggregate costs of medical care in the United States continued to climb (98). Second, demonstrations of wide variations in the use of common treatments for common conditions

[1] Some of the first claims-based technology assessment research was conducted with data from the Canadian health care system. Canada's national health insurance was introduced in various provinces between 196 I and 1971.

(118,195) suggested that traditional research methods had not defined the best courses of action for many common clinical problems (34,191). Third, when the outcomes of some procedures differed in practice from what had been suggested by the medical literature, analysts recognized the need for more research into how well medical technologies perform in real-world medical practice (175,198,204), as opposed to more limited settings (e.g., academic medical centers). And fourth, the federal government, primarily through the Agency for Health Care Policy and Research (AHCPR), significantly increased funding for this type of research.

AHCPR was established in 1989 to “enhance the quality, appropriateness, and effectiveness of health care services through a broad program of scientific research and dissemination” (184). Prominent among its sponsored research activities are the Patient Outcomes Research Teams (PORTS). These multidisciplinary, multiyear research teams each focus their studies on a particular health condition, and large administrative database analysis has been one of the research tools emphasized by the PORTS.

In addition to being a prime funding source for research into the effectiveness of medical care, AHCPR has encouraged database analysis by facilitating professional communications about its limitations and potential, by carrying out assessments of the existing databases, and by developing databases for use by researchers (184,185). A major function of one branch of AHCPR, the Office of Science and Data Development, is the development of databases as research tools (184).

Types of Databases

This paper focuses on four types of health care databases:

= Claims databases, which consist of claims to third-party payers for reimbursement for medical services provided to covered individuals. The claims can be made for prescription drugs, hospital care, outpatient care, medical equipment, and so on. Claims databases are maintained by third-party payers.

▪ Discharge abstract databases, which compile summaries of information regarding hospital stays. Each abstract generally includes information regarding the patient’s age, sex, and race, the conditions treated during the hospitalization, the procedures performed, and other aspects of the hospital stay, such as the dates of admission and discharge. Discharge abstract databases generally contain information similar to that submitted in claims to third-party payers for reimbursement of hospitalization expenses, although additional data elements may also be included.

Disease and procedure registries, which contain data regarding individuals who have specific diseases or undergo specific procedures. Disease and procedure registries include all the individuals in a defined population who have the disease of interest or undergo the procedure of interest.

- Practice databases, which contain data accumulated in the course of providing clinical care to patients. All patients receiving care in a particular setting are included, regardless of their diseases or the procedures they undergo. Generally, a practice database consists of a room full of patients’ medical records.

Table 2-1 describes these and some other types of large databases that are of potential use for assessing medical care, including some addressed in this paper chiefly with regard to their use with claims and discharge abstract databases. The latter include databases generated as part of large epidemiologic studies such as the Framingham Heart Study (150). (Population surveys, such as those conducted by the National Center for Health Statistics (187), also have a role in medical evaluation but are not discussed in any detail here.)

Role in Evaluating Medical Technologies

The analysis of large health administrative databases has three related but distinct potential applications in efforts to evaluate medical technologies and services.

TABLE 2-1: Some Types of Databases That Could Be Used for Evaluating Medical Technologies

Database type	Description of population included	Data elements typically available	Examples
Claims for insurance payment	All individuals covered by an insurance plan.	Provider, service provided (e.g., a procedure code), reason for service (e.g., diagnosis), charge, payment, patient demographics, patient identifier.	Medicare, Medicaid, private insurance claims databases.
Discharges abstract registries	A defined set of hospital admissions-e. g., all those occurring in a state or all those in hospitals participating in a voluntary registry	Descriptions of hospitalization (including patient characteristics, discharge status, procedures performed, admission and discharge dates), hospital identifier. Certain registries may collect additional data elements (e.g., in New York State, detailed data on catheterization results for patients undergoing CABG.	State discharge registry, VA Patient Treatment File, CPHA database.
Disease registries	All people with certain disease(s) who meet specific criteria (e.g., seen at a participating hospital, resident of a geographic area).	Detailed disease-specific information, patient demographics, patient vital status. May include information on initial treatment. Since these are often gathered at a single site, care received offsite may be poorly recorded; for example, outpatient chemotherapy maybe missed by a hospital-based tumor registry.	Cancer registries, communicable disease reporting systems.
Procedure registries	All people undergoing certain procedures who meet specific criteria (e.g., have procedure done by a provider that is participating in the registry). Thus, these registries are typically not population-based, since not all providers in a region are participating.	Details of procedure (e.g., results of cardiac catheterization, complications of procedure), demographics, vital status in followup Since focus of registry is on the procedure, very detailed data relating to the procedure may be available.	CASS cardiac catheterizatlon registry.
Databases gathered as part of a separate research project	The population identified for study in the original research project.	Patient characteristics, data collected for the original project (frequently quite detailed in the original area of interest, especially if the population included is relatively small), followup data regarding study endpoints. Patient identifiers may not be available because they are destroyed to preserve patient confidentially. Thus, added data regarding the cohort may be difficult to obtain.	Framingham Heart Study, Multiple Risk Factor Intervention Trial screening cohort.
Practice databases	All patients in a given practice setting.	Data gathered in the course of practice (e.g., laboratory tests, physical exam results, diagnoses), demographics, charges payments. These data typically do not conform to a predefine set of data that is gathered on each patient. This disadvantage is weighed against the fact that all of the data available to the clinicians managing the patient are available to the researcher using the database.	Traditional medical records, Duke Database for Cardiovascular Diseases.

KEY CABG = coronary artery bypass graft surgery; CASS = coronary artery surgery study; CPHA = commission on professional and hospital activities; VA = Veterans Administration

SOURCE: Jeff Whittle, 1994.

First, large databases have come to be a staple tool for descriptive *studies* of medical behavior and clinical practice. These include research to describe the variation across areas or populations in the rates with which procedures are done, studies that describe the outcomes associated with a particular procedure or practice, and studies whose goal is to describe the current state of clinical practice. This use of large database analysis is well-established, is used widely in research associated with the federal government medical effectiveness initiative, and is relatively noncontroversial.

Second, large databases have been used to conduct *comparative studies:* studies in which the outcomes of two or more interventions are compared in an attempt to determine which is the most effective. This application has also been promoted under the federal medical effectiveness initiative. However, it is much more controversial than descriptive studies, because of the difficulties in conducting valid comparative studies using observational rather than experimental designs.

Third, large health administrative databases have been used as *adjunct methods* to enhance other research techniques. This set of applications is potentially wide and is only beginning to be explored.

Each of these applications is described below in more detail, along with a discussion of some of the issues and caveats each entails. Many of these issues, such as problems of incomplete or incorrect coding, have been debated and investigated primarily in the context of the use of these databases in descriptive studies. They are discussed here in that context, although they often apply to other applications as well. Other issues, however, are unique to a particular application. This is especially true of comparative uses of large administrative database analysis. Because this use has featured prominently in many of the research projects sponsored by AHCPR, it is discussed in detail below.

DESCRIPTIVE STUDIES

Applications

Variations in Clinical Practice

One important step in assessing medical care is to determine who receives it. Analyzing databases allows researchers to describe the population of patients and to contrast the rates at which subgroups defined by geography, race, sex, or other characteristics undergo particular procedures. The rate at which the use of a treatment changes over time can also be of interest, especially when the increased use of one treatment may be related to the decreased use of an alternative treatment.

Geographic variation

Many demonstrations of variation using claims and administrative databases have been published since Glover reported the variation in tonsillectomy rates across different areas of England (58). Research combining discharge abstract databases with other databases has shown that geographic variation cannot be satisfactorily explained by differences in population characteristics, availability of services, or other structural factors (23,106,130, 158,192,197) (although there is a relationship between the number of providers and the amount of service provided (1 18,194)).

Wennberg has hypothesized that the variation reflects a lack of professional consensus about when treatments are appropriate (19). Even after joint discussions and reviews of the literature, expert clinicians have widely varying opinions regarding whether certain clinical scenarios are appropriate indications for a variety of procedures (13,142). The uncertainty implied by the unexplained variations in the provision of treatment (21 ,193) was a major impetus for the Health Care Financing Administration’s (HCFA’S) effectiveness initiative (162) and the formation of AHCPR (97).

The demonstrations of variation have affected practice. After practitioners learned that hysterectomy rates were highly variable among regions in Saskatchewan, Canada, the number of hysterectomies performed by high users decreased (33). Similar results occurred with tonsillectomies in Vermont (190,195). Orthopedic surgeons in Maine are examining indications for a number of orthopedic procedures, whose use has been shown to vary geographically (103).

Variations among specific populations

Claims and discharge abstract databases have also been used to study variations in the provision of certain treatments to population subgroups, including elderly and poor people (201,209), racial minorities (51,62,200), and residents of rural areas (156). Many of these studies have found that some segments of the population are treated less frequently than others, which impels researchers and society to consider the reasons for the variation (72,202).

Although the variations seen in database studies often seem to be important, a precise understanding of the roots of the variations cannot be gleaned from the current databases. For example, the relatively low rates at which invasive procedures are performed on African Americans, compared with whites, could result from differences in coronary anatomy, baseline comorbidity, or patients' preferences—information attainable in prospective studies but not in current databases (162,200,202).

Variation over time

Claims databases have also been analyzed to detect temporal changes in the provision of treatment. One study, for example, showed that the performance of radical prostatectomy had increased nearly sixfold between 1984 and 1990. This implies a significant change in how the procedure is used to treat prostate cancer—a change that is undergoing further evaluation (1 22). Database analyses of changes in the provision of medical care have been used to assess the compliance with consensus recommendations (39,140,169), the introduction of new treatments (155), and the effects of Medicare's change to case-based prospective payment as a means of paying for hospital care (100,172). By demonstrating poor adherence and weak response to the recommendations of the National Institutes of Health (NIH), database analyses have contributed to the growing recognition of the need for research into the dissemination of new technology and information.

Technology substitution

A special type of variation occurs when the use of one procedure decreases the use of another-a phenomenon that is likely to be of particular interest when one of the procedures is new. Unfortunately, very new procedures are often hard to detect in administrative databases. In a recent study of substitution of angioplasty for bypass surgery in the treatment of peripheral vascular disease of the leg, for example, no specific code existed for angioplasty of the arteries to the legs. At a cost of considerable time and money, the researchers had to design and test an algorithm using a combination of diagnosis and procedure codes to identify the patients who had received angioplasties (79).

Analyses of claims and discharge abstract databases can address the question of whether increases in the use of a procedure were associated with decreases in the use of its alternative in a particular population, but not whether the new procedure caused the decrease. In the study noted above, the researchers found no decrease in the rate of peripheral artery bypass surgery as the rate of angioplasty increased overtime, but they could not determine whether the rates of surgery would have been greater had angioplasty not been available. Thus, the actual question—whether the use of angioplasty reduces the need for surgery—was not directly answered. Certain data available in prospective studies but not in the database (information regarding angiograms, clinical conditions, and the like) would have helped researchers answer the actual question.

Outcomes Assessment

Another aspect of evaluating medical technology entails determining the effects of putting a procedure into practice. These effects include both economic and clinical outcomes, whether good (pain relief or improved functioning) or bad (rehospitalization, complications, or deaths). Because the outcomes are likely to differ for patients with different characteristics, an ideal assessment would describe all the relevant outcomes and explain how they vary among patients defined by such characteristics as age, sex, clinical condition, and the setting in which they were treated. These data might then be used to identify groups of patients for whom the treatment's effects were good or were bad.

Databases that can be linked to reliable sources of information about death provide a powerful means of looking at mortality in a defined population. In addition, the rates of hospitalization, reoperation, and certain complications can be determined. Studies of the outcomes of surgical treatment for benign prostatic hyperplasia (BPH) (161,198,199), for instance, were influential in bringing about a recognition of the need for the evaluation of common procedures (162). Furthermore, a decision analysis combining these data with primary data regarding symptomatic outcomes clarified the importance of patients' preferences in selecting management options (53).

Similarly, reports that short-term morbidity and mortality following carotid endarterectomy were higher than expected in the Medicare population may have contributed to a trend toward lower rates of the treatment nationwide (8 1,205) (and to a decline in the enrollment rates in a randomized trial of the treatment (1 l)). Studies of Medicare patients' outcomes during hospitalization have been another kind of influential (though controversial) assessment.[2]

Issues and Limitations

Coding Issues

Much of the concern about claims database analyses has focused on the coding system used to represent diagnostic and procedural information (40).[3] Hospital discharge data in the United States are coded using the *International Classification of Diseases, 9th Revision-Clinical Modification (ICD-9-CM) (181). The ICD-9-CM* includes more than 10,000 numeric codes, with as many as five digits apiece. All five-digit codes are subsets of four-digit codes, which are subsets of three-digit codes. The three-digit codes were organized into 17 chapters representing broad disease categories, which range from "neoplasms" to "symptoms, signs and ill-defined conditions"4 (40).

The information represented by these codes is the basis for hospital payment by Medicare, many national health statistics, and other uses. The coding system is updated periodically in order to meet reimbursement needs, to allow more precise identification of diseases and procedures that have grown in significance, or to clarify how certain diagnoses or procedures should be coded.[s]

The time lag before the implementation of the coding changes that are needed to identify new diseases or procedures causes problems for database researchers. The acquired immunodeficiency syndrome (AIDS) first received a specific diagnosis code in 1986, eight years after the first case reports of AIDS were published and three years after the etiologic agent had been identified. Similarly, new procedures may be part of practice for some time before new ICD-9-CM codes are devised to describe them specifically. For example, percutaneous transluminal coronary angioplasty (PTCA)

[2] These assessments had been a part of the annual HCFA hospital mortality reports, which were suspended in 1993.

[3] For more detail on this issue, see any of the several reviews that have been published (40,84,175).

[4] For example, the chapter "Diseases of the Digestive System" includes the code 532 for "duodenal ulcer." Code 532.0 refers to an acute, ›leeding, duodenal ulcer, and 532.01 refers to an acute, bleeding, duodenal ulcer with obstruction.

[5] An ICD-9-CM Coordination and Maintenance Committee was established in 1985 to advise HCFA and the National Center for Health Statistics about the need for changes in the coding system.

was initially coded to a procedure category that was also used for open-heart surgical procedures. The use of this code caused the patients who underwent PTCA to be assigned to a Medicare payment category that was reimbursed at a level much higher than that of the usual costs (173). New procedure codes created in 1986 placed the procedure in a more specific (and less generously reimbursed) category.

To limit the number of codes in the system, the ICD-9-CM lumps certain entities, which can obscure important differences. Codes for patients with renal dysfunction, for example, distinguish between acute or chronic cases but not among levels of dysfunction, which range from a slight change in a biochemical test that has only minor functional effects to complete cessation of kidney function (35). Moreover, the ICD-9-CM does not systematically include the sidedness (left or right) of a disease or a procedure. Consequently, researchers conducting database studies of procedures that can be done on either side (e.g., cataract surgery) may have difficulty interpreting whether certain procedures or diagnoses that occur after the procedure of interest are related to it (92).

Another limitation in the ICD-9-CM coding system is that different codes can sometimes be used to describe the same condition. For example, a code for a symptom (angina), a disease process (myocardial ischemia), or an anatomic abnormality (coronary atherosclerosis) can all be correctly and legitimately used to describe a patient with narrowing of the coronary arteries that causes chest discomfort with exertion (175).

A limitation of hospital discharge data is that they do not reveal whether coded conditions were present at the time of admission (preexisting conditions) or developed during the hospitalization (possible complications). To address this problem in New York, coders were asked to indicate whether conditions were present at the start of the hospitalization, but initial studies show that the coders have been slow to implement the change (61). An alternative approach is to exclude those conditions that could develop as complications from being considered as comorbidities. Unfortunately, many of the most important factors affecting an individual's baseline condition fall into that category.

In addition to dealing with coding problems in individual databases, researchers analyzing more than one database for a particular study may have to resolve differences in the coding systems. Procedure codes in the ICD-9-CM, for example, do not correspond to the coding system used for most professional service claims[6] (2).

Inaccurate Data

The accuracy of a database's coding depends on how often the codes entered into the database are the codes prescribed by the rules of the coding system. According to studies of coding accuracy that were conducted before Medicare's prospective payment system (PPS) for hospitals was introduced (in 1983), patients' age, sex, admission date, and discharge date were generally accurate, but the diagnosis and procedure codes were not (31,89,90). Even at the three-digit level, more than 25 percent of the principal diagnosis codes were different from those assigned by expert reviewers (90).

Financial incentives for complete coding were introduced with the Medicare PPS, which linked the amount of payment to a patient's diagnosis, and coding accuracy did improve. For example, at the three-digit level, overall agreement on the principal diagnoses increased from 73 percent in 1977 to 78 percent in 1985 (47). Subsequent data suggest that accuracy has continued to improve (79,80).

[6] HCFA mandates the use of the Current Procedural Terminology (CPT) coding system for professional service claims and supplemented the CPT with additional codes for nonphysicians' services, which it did not include. This augmented CPT is known as the HCFA Common Procedure Coding System (HCPCS). In addition, local intermediaries (insurers) can create codes with the approval of HCFA. This complicates analyses that cross states, as well as those that combine inpatient and outpatient data.

Accuracy apparently varies significantly among diagnostic and procedure codes. In the 1985 Medicare discharge data, 90 percent of the patients coded as having lung cancer had actually been diagnosed with lung cancer, and only 7 percent of patients who had actually been diagnosed as having lung cancer had not been coded as such. By contrast, peripheral vascular disease was coded for fewer than 60 percent of the patients who had the condition, and only 53 percent of those coded as having the disease actually did (47). Major procedures, which often affect Medicare payment, are quite accurately coded. For example, 96 percent of the patients who underwent coronary artery bypass graft (CABG) surgery were coded, and when the procedure was coded as having been done, it had always been done. Minor procedures, however, are much less reliably recorded. One study identified far fewer individuals coded as receiving total parenteral nutrition (a form of specialized intravenous feeding) than were known (from other data sources) to be receiving the treatment (120).

It is unclear whether improvements in coding that accompanied the implementation of the Medicare PPS are reflected in codes not used for Medicare payment. Although many other databases have been subjected to review (25,1 17), the results are seldom published (101).

Financial incentives can sometimes result in one-sided errors. For example, the Medicare PPS pays more for more severely ill patients, and the study of coding accuracy in 1985 discharge data showed that errors in how severity of illness was coded systematically tended to overstate severity, increasing hospitals' reimbursements (80). (A study of 1988 discharge data did not reveal the same tendency, perhaps because of the strict laws that now require attending physicians to certify the accuracy of the designated diagnoses (79)).

The accuracy of the coding on claims for professional reimbursement may be somewhat higher, because professionals tend to perform the same procedures (and use the same codes) repeatedly. In a recent study of carotid endarterectomy, in which professional claims were used to identify patients (205), codes for the procedure were verified for more than 95 percent of the patients. Diagnostic information is seldom available for professional claims, however, and has very rarely been used in research.

Studies of temporal and geographic variations may be subject to bias because of disparities in the quality of coding overtime or among regions (44). If coding is unusually complete in one area, the rates at which procedures are performed in that area may appear to be unusually high. Both selective coding to maximize reimbursement(171) and the trend to move some procedures from inpatient to outpatient settings (163) could create a false appearance of temporal trends in treatment rates if inpatient databases are used.

In addition to errors in coding the diagnoses that are recorded in patients' charts, physicians themselves sometimes make diagnostic errors, and one doctor's diagnoses are likely to differ to some extent from those of another. The lack of precise definitions leads to wide variations in the reported incidence rates of diseases, depending on the criteria used for making the diagnoses (40). In a discharge abstract database, a correctly coded diagnosis means only that the attending physician made that diagnosis, irrespective of whether the appropriate diagnostic criteria were used.

Differences among physicians in how completely they choose to evaluate their patients can lead to bias, because patients who have more complete evaluations are more likely to be found to have signs that indicate poor prognoses, for example, the spread of cancer from the site of origin. The prognosis for patients whose spread of cancer has been found only after extensive testing falls between the prognosis for patients in whom the spread of cancer is obvious and that for patients whose cancer has not spread. Moving this intermediate group from the good prognosis (no spread) group to the poor prognosis group improves the prognoses of both groups[7] (41).

[7] Alvan Feinstein named this the Will Rogers effect, after the humorist observation that the move of many Oklahomans to California was increasing the average intelligence in both places.

Groups defined by their clinical characteristics or treatments may receive systematically different evaluations. For example, testing to rule out the spread of lung cancer may be more extensive for patients treated with surgery than for patients treated with radiation therapy. If so, a comparison of how the two groups fared could be biased.

Identifying the Relevant Population

Another important issue with which researchers must contend involves being able to identify correctly the universe of people of interest. For example, when dealing with the rates at which treatment is provided, researchers must not only know the denominator (the population under investigation) but also be able to identify the numerator (the people who receive the treatment). The numerator in a claims study is usually defined as those individuals who are coded for the treatment of interest. The appropriate denominator is not always clear. One simple denominator comprises all the people whose receipt of a treatment could appear in the database. Such a denominator usually includes everyone in the population, unless the treatment of interest is only performed on members of one sex. If a procedure can be done only once, however, anyone who has already had the procedure should not be included in the denominator. Thus, if hysterectomy rates are to be described, the denominator should include only women and should exclude any woman who no longer has a uterus (because of prior hysterectomy). The fact that as many as half of some female populations have had hysterectomies might bean important source of variation (158).

Often, a different denominator—people with the medical condition for which the treatment is provided—may be more appropriate. In a study of variation in rates of radical prostatectomy,[8] for example, a useful denominator would be men with prostate cancer (122). Similarly, because CABG surgery is performed only on patients with coronary artery disease, the analysis ideally would cover only such patients (202).

Defining the appropriate denominator can be difficult. Researchers conducting claims-based analyses of variation in the rates of treatment have often assumed that similar proportions of different populations are at risk, perhaps after accounting for differences in the age and sex distributions. This assumption allows the total population, which is easy to quantify, to be used as the denominator, but the assumption is not always valid.

Although some analyses of variation have used patients hospitalized with conditions of interest as the denominator, many similarly afflicted patients are not hospitalized. The interpretation of the denominator, therefore, becomes difficult. Nonetheless, such conditions as myocardial infarction (180), childbirth (119), and hip fracture (45)-which are almost always treated in the hospital if they are recognized-can define more complete denominators. Otherwise, the differences in hospitalization practices across time, regions, or population subgroups could cause spurious apparent variations in the rates.

For many procedures, appropriate databases are hard to find. Outpatient procedures, for example, are not included in statewide discharge databases. HCFA now retains all professional claims under Medicare, including bills for procedures performed in physicians' offices, but this is a recent development, and the data go back only to 1991.

Other procedures that are difficult to assess are those performed on hospitalized patients but not reliably recorded (e.g., total parenteral nutrition). New procedures are especially problematic, because until they have been assigned unique codes, they cannot reliably be identified in claims databases.

Even a well-defined population can receive some care that is missed if a researcher uses only a single database. Some individuals, for example, may be covered by both their own insurance and

[8] Radical prostatectomy is a surgical procedure used to treat localized prostate cancer. It is significantly more dangerous than the procedures ısed to treat benign prostatic hypertrophy.

that of their spouses. Others may lose their coverage or never have any at all. Restricting a study to the consistently covered portion of the population raises the possibility of selection bias (164). Medicare data are particularly useful because nearly all the recipients continue to be enrolled until they die. Even then, however, the beneficiaries may receive treatment from providers (e.g., the Department of Veterans Affairs (VA)) whose services do not appear in the Medicare database (49).

Obtaining Outcomes Information

Although outcomes assessments have been important products of database analyses, a number of limitations are obvious. In addition to problems with coding and with services provided by sources not covered in the databases, many important outcomes other than death are not included in claims and discharge abstract databases. When such outcomes are the appropriate measures for comparing the benefits of different treatments, database analyses are difficult.

For instance, total joint replacement--considered the most important advance in the management of arthritis in the past 20 years (55)—almost certainly does not increase the survival rates of patients with arthritis and might even decrease the rates slightly (because the disease is not fatal, whereas surgery carries some risk of death) (204). The objective of the procedure is to improve the patient's quality of life by relieving pain and increasing mobility (63). Unfortunately, neither claims nor discharge abstract databases include measures of these outcomes. Among the other important outcomes not available in such databases are the relief of such symptoms as incontinence and diarrhea, the ability to function socially, and a sense of well-being.

Alternative Approaches

There are alternative methods for examining patterns in the use of medical services. One approach, suitable for procedures that require a single expensive piece of equipment (such as an artificial hip joint), is to survey the manufacturers of the equipment. If there are only a few suppliers, relatively good estimates of overall rates of treatment seem possible (74). Another approach is to survey a sample of providers regarding the frequency with which they provide the treatment (32,74). Unlike database studies, these approaches do not require coding conventions to identify the treatment. A third alternative is to sample a population to identify individuals who have been treated with a particular procedure. This allows researchers to collect precisely the variables that are of interest in characterizing both the numerator and the denominator.

These approaches have their own drawbacks, however. Studies of the use of a procedure at a single facility or a few hospitals, for example, might include relatively few patients who have undergone the treatment. The denominator population from which the patients were drawn would be hard to define, and the treatment rates at the participating sites might not be representative. Several of the health surveys conducted by the National Center for Health Statistics (e.g., the National Medical Care Utilization Survey) could be used to study treatment rates, but the surveys are relatively expensive and time-consuming, include relatively few individuals, and provide only limited details about which medical services are used. Primary data collection to address these concerns would probably be prohibitively expensive. Moreover, the data collection would have to be continued if temporal trends were of interest.

COMPARATIVE STUDIES

Frequently, when researchers assess a treatment, the most difficult question to answer is whether the outcomes (mortality, morbidity, cost) for patients treated with therapy A are better than those for similar patients treated with therapy B (or for patients who do not undergo treatment). The use of claims databases to address this question has generated much controversy (7,15,16,66,70,1 10, 124,125. 178).

I Rationale

For researchers attempting to assess the comparative effectiveness of health technologies, database analysis offers a number of potential advantages over other methods. These advantages helped generate the enthusiasm for using database analysis in effectiveness research. Studies comparing the outcomes of transurethral prostatectomy (TURP) and open prostatectomy on men with BPH, for example, stimulated a reexamination of the question of which treatment is the most appropriate (46,73). A recent study of outcomes associated with management of cataract surgery patients showed that the rate of retinal detachment was several times higher among patients who had undergone posterior capsulotomy[9] than among patients who had not undergone the procedure (93)--information that, if confirmed, could enhance decisionmaking regarding the timing of the procedure.

Large Size

Probably the most obvious advantage of using large preexisting databases to conduct comparative effectiveness studies is that the databases are large. More than 25 million patients are represented in the Medicare claims database, and about 10 percent of the U.S. population lives in the areas covered by the National Cancer Institute's Surveillance, Epidemiology and End Results (SEER) cancer registries. Analyses of such large databases can provide estimates of the rates of even relatively uncommon events (e.g., the adverse effects of particular drugs (147,176) or the complications that can occur after certain types of surgery (92,154)) or can identify cohorts of patients with rare conditions (e.g., endocarditis (7)). The size of the databases also allows researchers to subdivide groups of patients by age, race, or sex and still have a significant number of subjects whose experiences with treatment can be studied (180).

Representative Samples

Because they are generated routinely in the course of providing care to patients, claims and discharge abstract databases may include everyone with the condition of interest in the populations for which the databases are maintained. For example, Medicare claims data cover more than 95 percent of Americans over the age of 64 (1 14). State discharge abstract databases generally cover nearly all the hospitalizations that occur in the state. The people or hospitalizations covered in such databases are generally much more representative than the populations studied at individual facilities or at a few academic medical centers (175). The factors that cause patients to enter tertiary medical centers for care are often related to the results of the treatment they undergo (164). Thus, studies using Medicare claims data to determine the short-term mortality rates following pneumonectomy for lung cancer (203), carotid endarterectomy to prevent stroke (204), and transurethral prostatectomy (TURP) for BPH (161) have found higher mortality rates than those found in studies of patients who were treated at medical centers that had particular interests in the diseases (50,57,132).

Opportunity for Followup

"Long-Term Follow-Up is a Problem" is the title of a 1983 editorial in the *American Journal of Public Health (5).* This statement holds true for a number of outcomes of great interest in assessing medical care. Mortality, cost of care, health status, and rehospitalization long after treatment are important in defining a treatment's utility. Because individuals who are not followed up may differ systematically from the rest of the population being studied, a major portion of research efforts are directed toward assuring complete followup (54). The concern is so great that most investigators try to exclude patients who are unlikely to follow up

[9] Posterior capsulotomy is frequently done following extracapsular cataract extraction as a treatment for posterior capsular opacity, a common complication following extracapsular cataract extraction. Since posterior capsular opacity can vary in severity, the decision to perform posterior capsulotomy relies in part on a subjective assessment of the opacity's effect on vision.

reliably, despite the possibility that their exclusion could affect the representativeness of the population (83).

In some cases, insurance databases can provide more complete followup at considerable y less cost. For example, the Medicare claims database includes data, gathered by the. Social Security Administration (SSA), regarding the vital status of beneficiaries (199). Because vital status is important for determining Social Security payments, considerable care is given to ascertaining whether and when beneficiaries have died. The expense of this ascertainment is borne by the SSA, which frees researchers from a heavy burden.

Less "Attention Bias"

Many aspects of medical care are difficult to study, because physicians and patients may change their behavior if they know that they are being studied. Physicians may practice more cost-effective medicine, schedule more timely follow-up visits, or provide more preventive services. Patients may be more likely to stop smoking, adhere to a recommended diet, or comply with a complex medical regimen (83). This has been called "attention bias" (164) or the "Hawthorne effect" (after the Chicago industrial site where research in the 1920s showed that productivity improved when workers were being observed) (151).

By using data gathered in the course of routine clinical practice, researchers can avoid this phenomenon. The providers and patients involved do not know that they are the subjects of a study. Indeed, at the time the care was delivered, the patients were not in a study, inasmuch as the study began sometime later.

Timeliness

Because claims data are needed immediately for payment purposes, they become available for research fairly quickly. Thus, Medicare claims data from 1991 and VA data from fiscal year 1992 were available for research in 1993. Because data regarding several years' worth of patient followup are included in the Medicare and VA databases, researchers can relatively easily conceive of and carry out studies to address questions that have been newly recognized, even if they involve events that have occurred long after the treatment (e.g., long-term survival following lung cancer resection). In contrast, a study using primary data collection may have to wait many years for a sufficient number of patients to be identified and to experience the outcomes of interest.

Another way in which database analyses can be timely derives from their power. Treatments can change rapidly, particularly when they are new. By the time a randomized controlled trial (RCT) of a treatment is complete, the treatment under evaluation may not be acceptably close to the treatment that has become state-of-the-art since the RCT began. The results of coronary artery bypass graft surgery, for example, changed significantly as the technology evolved. The first major randomized trial of CABG found excessively high rates of surgical mortality, but the results of that study do not seem to apply to surgery done with modem techniques (146). By contrast, the numbers of patients covered by databases are so great that researchers can use data from a short period when the technology is likely to be relatively stable. Thus, database analysis can address the moving-target problem that plagues other approaches to assessing health care.

Validity and Reliability

The relative ease of database analysis makes it an increasingly attractive method to address clinical questions, particularly when the outcomes of interest are rare but are likely to result in events that the databases can reveal (92). The important question, however, is whether database analysis can provide valid answers to these questions (1 5).

Using Observational Data for Comparative Studies

Differences in the outcomes of alternative treatments do not necessarily mean that one approach is superior, unless the way the treatments are provided and the composition of the populations receiving them are comparable. A sick population provided with a superior treatment might well fare worse than a healthy population provided with an inferior treatment.

Much scientific discussion has focused on the validity of comparing the outcomes experienced by apparently similar groups of patients. For example, researchers often compare the results for patients receiving a new treatment at a particular center with those for historical controls-patients who had the same condition and received the standard treatment at the same center in the past (56,165). The outcomes experienced by a series of patients treated at a center can also be compared with the outcomes experienced by similar patients as reported in the literature or by contemporaneous patients who have undergone the alternative therapy.

In each of these examples, the use of the new treatment is not the only thing that truly differs for the groups being compared. Historical controls were diagnosed with older tests, which might not have discovered their disease until it had reached more advanced stages. The historical controls also received older supportive therapy, which means that improved outcomes might simply reflect the general improvement in medical care over time, rather than derive from the treatment being evaluated. Similar concerns exist about controls selected from the literature (they reflect the experience at different centers, often from earlier times) and contemporaneous controls from the same institution (the decision to use a new treatment may by influenced by the severity of the patient's condition).

Experimental studies—specifically, RCTs—address these problems of comparability by ensuring that the treatments are performed on groups of patients who are only as different as the random play of chance would allow. This approach permits any eventual differences in outcomes to be assessed by the question of how likely it is that the set of outcomes would have occurred by chance alone.

Although current opinion holds that an RCT provides the most scientifically rigorous way to compare treatments (8,1 12), it has a number of disadvantages. These include the practical (177) and ethical (67,144) difficulties of enrolling patients when there are strong theoretical reasons or clinical suspicions that one treatment is better than another (56), the long delay before the results are available, and the limited numbers of patients often included. Moreover, such studies can be expensive to carry out due to the need for screening many candidates to identify a few eligible patients, the need for extensive quality control activity to make sure that the interventions are applied in a standard fashion, and the practice of obtaining extensive data regarding each patient.

Just as the study's design ensures that the participants are relatively homogeneous, it also limits the researchers' ability to generalize an RCT'S results to patients who do not meet the eligibility criteria. When the eligibility criteria do permit participation by identifiably different subjects (both men and women, for instance, or patients whose disease differs in severity), the relatively small number of patients often makes it difficult to determine whether the results apply equally to all the subgroups.

As an example, a recent trial of lowering cholesterol with cholestyramine showed that the treatment reduced rates of myocardial infarction among 3,806 men between the ages of 35 and 59, the vast majority of whom were white (150). A similar trial in an elderly, female, or African American population would be hard to justify—both on ethical and economic grounds—although some researchers have questioned whether the trials can be extrapolated to populations that were not studied (151). These limitations have spurred interest in designing RCTS that keep the benefits of randomization while incorporating some of the generalizability of database research (1 7). (See J. Burning, M. Jonas, and C. Hennekens, "Large and Simple Randomized Trials," background paper no. 3 in this volume).

When RCTS have not been carried out, what role should nonexperimental data play in the assessments? Should the data be used only to extend the results of RCTS to population subgroups not included in the original trials? Should observational studies never be used to make comparative judgments? Or can fair comparisons of treatments be based on the outcomes experienced by populations "assigned" to different treatments in nonrandom ways?

The problems with using observational data in general for comparisons offer insights into the problems with using claims databases—a particular source of observational data—for comparative analyses. The most weighty concern about using observational data to compare therapies is summarized by Byar: "In medicine, the doctor chooses the therapy precisely in order to affect outcomes" (123). The choices generally differ for different patients. One way to address the problem is to carefully note all the important differences between the groups receiving the treatments being compared and then to use statistical techniques to correct for the differences. This presupposes that the important differences can be identified and are recorded when the data are collected. Researchers using historical controls, literature controls, or contemporaneous controls have taken this approach when comparing treatments.

At least three more potential problems with comparisons based on observational data have been noted. First, the ability to characterize the treatments being compared is limited, because providers may vary in the way they apply the same treatment, whereas both the new and the control treatments are administered in standard fashions in traditional RCTS.

Second, patients are assigned to one group or another at the time of randomization in RCTS, but group assignments may be ambiguous in observational studies. For example, if a patient is treated medically for coronary artery disease and then undergoes surgery because the medical therapy was ineffective, the patient's subsequent death might be attributed either to the medical group or to the surgical group. Approaches to this problem have been investigated by researchers using the Coronary Artery Surgery Study registry of patients undergoing catheterization for possible coronary artery disease (20), but no course is completely satisfactory. One feasible approach in a prospective study would be to ask the physician to outline a plan of action at the time followup begins. Researchers would then use the plan in assigning the patient to a treatment group.

Third, the ability to characterize subsets is limited by the data that are collected: if new, important risk factors are discovered, they cannot be presumed to be present to the same degree in the two comparison groups. Similarly, the level of detail in which the data are obtained—both at the baseline assessment and in the followup-is likely to be inconsistent unless the investigator prescribes, in advance, what data are to be collected. Worse yet, the level of detail is likely to be inconsistent in a nonrandom way. It is quite plausible that more tests or more followup visits, through which complications could be discovered, would be ordered for patients receiving the new treatment. Whether differences in the baseline characteristics result from biased assessments is difficult to ascertain.

Randomized vs. Nonrandomized Studies

One way to gauge whether valid comparisons of treatments can be made without randomization is to look at instances of comparisons made with nonrandomized and randomized study designs and try to draw conclusions about whether they are equally valid.

There is considerable agreement that spectacular effects (137) do not require randomized assessment. Most of the major advances in cancer chemotherapy, for example, were made without the benefit of randomized trials (56). Similarly, the treatment of endocarditis and tuberculous meningitis with antibiotics could be recognized as major advances without RCTS (70,133). What these cases have in common is that before the availability of the treatment, the patients uniformly fared badly and that the treatment considerably improved their chances of survival.

This does not mean that the comparisons were fair. In fact, people diagnosed with Hodgkin's disease today almost certainly have better prognoses, on average, than did individuals diagnosed with the disease before the advent of chemotherapy, but the availability of chemotherapy is not the sole reason for the improvement. Modem imaging techniques enable the diagnosis to be made earlier, the treatment of infectious complications has improved, and patients are more likely to seek medical care earlier in the course of the disease.

Nonrandomized comparisons have shown treatments to be beneficial, only to have subsequent trials demonstrate them to be useless. Gastric freezing for peptic ulcer disease (13S) and internal mammary artery ligation for coronary artery disease (24,59) are frequently cited examples. Recent studies by careful investigators have revealed similar patterns of unreliability as well. Major trials involving patients with advanced cases of non-Hodgkin's lymphoma have concluded that a first-generation chemotherapy regimen, CHOP[1o] (19), is as good as, or better than, more complex and toxic second- and third-generation regimens (48,60). Previously, the results of the newer regimens were compared with the results of CHOP reported in the literature. These reports, and accompanying editorials, had strongly implied that the newer regimens were superior (26,105,170). Despite an acknowledged need for randomized trials, practitioners began using the more complex regimens. A 10-year lag occurred between the early reports and the publication of the studies showing that the supposed benefits of the more complex regimens were not real (48,60,121).[11]

Another modem example of unreliable observational data is particularly interesting because of the methodologic care that was taken to avoid any identifiable biases. Researchers evaluated whether administering lidocaine prophylactically to patients with acute myocardial infarction helped prevent arrhythmia deaths (76). The observational study used stringent entry criteria, a well-defined endpoint, and an adjustment for the differences in the risks associated with the endpoint. The data were collected by trained researchers who were blinded to the study's hypothesis. Nonetheless, the finding that lidocaine had a beneficial effect was not borne out by subsequent RCTS or a meta-analysis of all the available RCTS (68,207). Although neither the RCTS nor the meta-amdysis had sufficient statistical power to identify arrhythmia deaths, the standard interpretation of the available data has been that lidocaine usually should not be administered prophylactically in the treatment of acute myocardial infarction (10).

On the other hand, in several instances nonrandomized comparisons have yielded results similar to those of randomized comparisons. A study that addressed the use of tonsillectomies for children with recurrent sore throats (141) provides an especially good example because the random and nonrandom comparisons were carried out simultaneously at a single institution. When the parents of children who were eligible for the randomized comparison decided against randomization, the children received the therapy their parents requested. The initial evaluations and data collection processes, however, were identical for both the randomized and nonrandomized patients. Moreover, the two groups were followed up in the same manner, including the frequency of visits, the definitions of the endpoints, and the management of throat infections. The researchers compared subgroups matched for important predictors, such as age and frequency of episodes in the preceding two years, and found that the outcomes experienced by the randomized and nonrandomized patients were indistinguishable.

Another study, which used the Duke Database for Cardiovascular Disease (71), was explicitly designed to test the hypothesis that observational data could be used to make fair comparisons between groups assigned to different treatments in a nonrandom fashion. The treatments compared were CABG and medical management for patients with coronary artery disease. These treat-

[10] CHOP, one of the first combination chemotherapy regimens introduced for lymphoma in the mid-1970s, includes cyclophosphamide, vincristine, doxorubicin, and prednisone.

[11] It should be acknowledged that many oncologists do not believe the results of the RCT comparisons. Even with RCTs, purportedly definitive results do not always prove definitive.

ments had been compared in three separate RCTS, whose results could be compared with those of observational studies.

The researchers used the Duke database to identify individuals who would have been eligible for each of the RCTS. The predicted survival rates of these patients, first assuming that all of them had received medical therapy and then assuming that surgery had been performed on them, were calculated using a statistical model derived from the overall Duke database. The survival curves were then compared with the actual survival rates of the participants in each arm of the trials. Differences between the Duke database analysis and each of the RCTS were within the limits of random variation. In addition, the five-year mortality rate of nearly every subgroup of the Duke patients with varying severities of the disease, whether managed medically or surgically, differed from the rates in the RCTS by no more than would have been expected by chance alone.

Still another study assessed the use of beta-blockers after heart attacks, an intervention that has been shown to be beneficial in a number of RCTS (77). The observational study was modeled directly on a specific RCT, the Beta-Blocker Heart Attack Trial (BHAT) (6). Patients in the observational study were patients consecutively admitted to the Yale-New Haven hospital for acute myocardial infarction between 1978 and 1982, while the BHAT enrolled patients from June 1978 to October 1980. The authors of the observational study explicitly compared their results to those of the BHAT (77) and found no significant differences between them, after adjusting for age and severity of disease in the groups who received beta-blockers and those who did not. Each study showed reductions in both 24- and 36-month mortality rates. Moreover, the magnitude of the difference was very similar in each trial.

Lessons

The fact that fair comparisons were made using observational data does not guarantee that similar designs would lead to fair comparisons in other studies. Without RCTS for comparison, how can we determine whether observational data are reliable?

The examples provide some clues. The tonsillectomy study (141) points out the need for good followup, particularly when the outcome of interest is not death. In claims and discharge abstract databases, outcomes other than death are detectable only if the patients or physicians have taken some action. An outcome like “need for repeat surgery” may be biased by a greater level of surveillance for one of the study groups (164). The methods for determining outcomes should be identical for groups receiving alternative treatments. The tonsillectomy study also had the advantage of patients with few comorbidities, and the risk factors thought to predict the outcomes of interest were precisely determined at the outset of the trial. The clinicians involved had little bias toward one treatment or the other, and parents were advised that the randomized trial was an appropriate option. Therefore, there is little chance that the clinicians would have encouraged patients with better prognoses to opt for a particular treatment.

Although the salient risk factors for death from coronary artery disease do not predict even half of the variability in who lives and who dies, they are well studied (69,70,7 1). Clinicians making decisions are unlikely to consider any factors that do not appear in the model used by the researchers who analyzed the Duke database. In other words, the other factors that predict outcomes are unlikely to vary among groups assigned to different treatments because no one knows what those factors are. Their distribution should be random. It is difficult to disprove the contention that clinicians can detect and interpret subtle differences among patients that cannot be captured as concrete data and incorporated in a model, but the Duke researchers have shown that their risk-adjustment model is better than expert clinicians at predicting the prognoses for patients with coronary artery disease (107,1 16). It would be unreasonable to think that these same clinicians could select the patients with the best prognoses for the group that is to receive CABG.

The researchers in the beta-blocker study emphasized several aspects of their method (77). The eligibility criteria in their observational study, and particularly their definition of myocardial infarction, were identical to those used in the BHAT. The researchers excluded any patient who would have been excluded from the BHAT, and they carefully considered how to assign zero time—the time at which each patient's baseline was established. After zero time, followup events were assigned according to their treatment groups. The researchers emphasized that this selection should duplicate, as much as possible, the time at which the random assignment would occur in an RCT. As in the BHAT, adjustments were made for differences in known confounding variables, such as age and baseline prognoses.

The studies in these examples have several important features in common:

1. Each of the clinical entities (tonsillitis, coronary artery disease, myocardial infarction) is well studied and understood.
2. The treatment under study was likely to be applied in a fairly standard fashion. The control group was also likely to receive state-of-the art management, because the control patients and the experimental patients were treated in the same sophisticated settings (although this does not guarantee standard care).
3. The data needed to define the baseline prognoses were collected prospectively and purposefully for each of the observational cohorts. Thus, from the beginning of the data-collection process, the database was oriented toward the type of study that was eventually performed.

Despite the fact that the study of lidocaine incorporated many of these positive features, the results differed from those of the RCTS. The least hopeful interpretation is that nonrandomized studies sometimes provide the right answers and sometimes provide the wrong answers, and that there is no way to tell the difference without an RCT to determine the true answer (15,16).

This interpretation suggests that the usefulness of databases is determined by the answer to a simple question: "Can valid comparisons be made with observational data alone?" Phrasing the question this way, however, ignores the fact that such comparisons are based on careful assessments of patterns of data. In general, the results of a single RCT are not definitive. Indeed, the validity of the comparison of CHOP with the advanced chemotherapy regimens has been challenged by a number of respected experts (121). Thus, many authors would argue that rigorous comparisons using observational techniques have a role in assessing medical treatments, because they can contribute data to the pattern, even if the results of observational analyses alone are not definitive (76,77).

Several differences between the data in these observational studies and the data in claims and discharge abstract databases bear emphasis. Each of the researchers used careful, quality- controlled methods of collecting data elements that had been defined prospectively as appropriate for the problem at hand, but database researchers must use whatever data are available. In general, even databases supplemented with clinical data have difficulty yielding answers to questions that were not formulated carefully before the data were collected (69).

For example, using a dataset that included detailed clinical data from the MedisGroups medical severity classification system in addition to routine hospital-discharge data, researchers examined how patients whose coronary artery disease was treated with percutaneous transluminal coronary angioplasty[12] fared in comparison with patients who were treated with CABG surgery (64). After being adjusted for differences between the groups in the age of patients and the presence of a number of clinical variables, the data showed

[12] Angioplasty is the procedure in which a balloon is inflated in an artery that has been narrowed or closed by a pathologic process (usually ıtherosclerosis). It usually decreases the degree of narrowing. PTCA is angioplasty of the arteries to the heart.

that the patients who were at low risk of dying did better with PTCA than with CABG. Despite the size and detail of this database, however, the lack of prospective collection of important data relevant to outcomes (e.g., the number of coronary arteries that were blocked) makes it difficult to draw conclusions.

The current claims databases' limitations for making comparisons have been well documented. There is evidence that discharge data restricted to ICD-9-CM codes may not contain adequate detail to allow for valid comparisons of outcomes across treatments (28) or hospitals (111). For example, the five-year survival rates were lower for men whose BPH was treated with TURP than for men treated with open prostate surgery, and an adjustment based on ICD-9-CM-coded discharge data did not explain the difference. This suggested that the open procedure was superior, but an adjustment using more detailed information obtained by reviewing charts showed that the differences probably derived from the higher numbers of important comorbid conditions in the population undergoing TURP (28). Another study found that case-mix adjustment using ICD-9-CM codes explained much less of the variation in hospital mortality than did case-mix adjustment using additional data abstracted from the clinical records (11 1).

Other studies have suggested that using ICD-9-CM codes in making adjustments for differences in groups might sometimes actually produce misleading results. Researchers found that the presence of a number of comorbid conditions actually improved the survival rates of patients admitted for several serious illnesses (95). The researchers interpreted their results to mean that certain diagnoses were usually coded only when there were no more important diagnoses to be coded. At the time of that study, the discharge abstract contained only a limited number of fields in which diagnostic codes could be recorded, but a later study found that the same thing occurred when the number of available fields was not restrictive (85).

Comparisons after adjustments for differences in baseline risks of poor outcomes are almost certainly more valid than comparisons made without such adjustments. Improved methodologies, including hierarchical modeling and instrumental variables techniques, increase the believability of the adjustments[13] (129). Comparisons are also more believable when the medical condition under study is well understood, when the variables are objectively defined, and when the data collection is complete and accurate. Unfortunately, most claims and discharge abstract databases do not currently meet that standard. Moreover, enhancements that would make the data in these databases more similar to those in good observational studies would eliminate the advantages of speed and cost (and therefore size) that make database research attractive.

Data-dredging—taking advantage of the convenience of large databases to test multiple hypotheses—is frequently raised as an issue in large database analysis (box 2-l). The issue is not unique to the analysis of administrative data (it can occur with RCTS as well, for instance), but it is of particular concern in this context.

The contribution of database analysis to comparisons of treatments maybe more appropriately assessed in terms of how it can contribute, rather than whether it is definitive (43a). Cross-design synthesis, a formal mechanism for incorporating the results of database research into a comparative assessment of treatments, was described in a General Accounting Office (GAO) report to Congress (188). This methodology attempts to formalize the use of database analysis as a complement to RCTS in comparing medical technologies.

[13] The use of newer statistical techniques is daunting to the average clinician. One concern that has been raised is that as statistical adjustment becomes more sophisticated, it will be poorly accepted by the providers whom researchers would like to influence. Researchers studying acute myocardial infarction note that the instrumental variable and hierarchical modeling techniques being developed there have been perceived as intuitive by a panel of clinicians who provide clinical support (127).

BOX 2-1: Testing Multiple Hypotheses

Large preexisting health databases are often convenient for testing multiple hypotheses about whether various characteristics of the patients or providers are related to particular outcomes. If enough combinations are examined, however, some of the characteristics will have closer-than-expected relationships to some of the outcomes by chance alone-i. e., they will be statistically significant. These relationships might be interpreted as important, even though the odds of finding at least one statistically significant relationship in every 14 such combinations are better than 50-50, even if there are no real relationships among the variables.

The practice of testing multiple hypotheses in search of one that is statistically significant is known as data-dredging. Whether this process has taken place may not be readily apparent to the reader, particularly inasmuch as the immensity of the databases invites the use of complex multivariable statistical techniques (138). Although multiple analyses of research datasets are common and often appropriate, the potential magnitude of the problem in large administrative databases grants it special importance (124).

Fortunately, the size of the databases makes statistical approaches—such as developing hypotheses in one half of a dataset and testing them in the other—feasible for addressing the problem (7,138). Moreover, because many similar databases are available, interesting results can be subsequently retested in independent databases. In addition, experienced database researchers develop analytical plans that focus on relationships suggested by previous research or theory, which decreases the likelihood of spurious results. Finally, as with the findings of any other form of research, the results of database analyses should be examined in the context of a much larger body of research.

SOURCE: Jeff Whittle, 1995.

ADJUNCTS TO OTHER RESEARCH METHODS

Applications

Research that combines primary data collection with the analysis of claims and discharge abstract databases reaps the advantages of both methods. The use of large, population-based databases as sampling frames, for example, facilitates the identification of representative samples. One of the original purposes of the National Cancer Institute's SEER program was to provide researchers with a tool: the case-finding capability of the SEER network (210) of cancer registries, each of which lists all cancers diagnosed in residents of a particular area. Data from the Professional Activity Study conducted by the Commission on Professional and Hospital Activities (CPHA) have been analyzed to identify potential cases for case-control studies of unusual occurrences, such as myocardial infarction following the use of oral contraceptives (96,97). Because myocardial infarction in women of childbearing age is rare, a single center could not have accumulated an adequate number of cases for a study. CPHA, however, collects data from thousands of North American hospitals and has a database of more than 150 million discharge abstracts (96).

Medicare eligibility files have been used to define a representative cohort of elderly individuals. The Medicare hospitalization file has been used to identify representative samples of discharge abstracts for studies of coding (79,80), for assessments of the quality of care before and after the introduction of Medicare's prospective payment system for hospital inpatient care (100), and for studies of the outcomes experienced by patients who had suffered myocardial infarction, stroke, pneumonia, or congestive heart failure (30). Before studying the appropriateness of various procedures, researchers have used Medicare files for professional claims to identify cohorts of patients who have undergone various procedures (22). Claims databases can also be used as sampling frames for pseudorandomized trials that take advantage of the varied treatment assignments created by regional differences in the treatment of common conditions (120).

Issues

One concern about using databases as sampling frames is that researchers may have difficulty obtaining data about the patients whose cases have been identified. The researchers who studied the effects of the Medicare prospective payment system on the care provided to the program's beneficiaries, however, managed to obtain the medical charts for 96.2 percent of the patients in a sample identified from a Medicare claims database (100). The analysts who conducted the coding studies of 1985 and 1988 obtained 99.6 percent and 91.8 percent, respectively, of the charts of the patients they had identified (79,80).

Other researchers have used Medicare data to identify representative samples of hospitalized patients and then contacted them or their health care providers to obtain additional information. Researchers studying cataract surgery are using the Medicare database to find a sample of cases for further study of how posterior capsulotomy following the surgery is associated with retinal detachment (174). The researchers will contact providers to obtain information regarding factors (e.g., the length of the eyeball) that place patients at risk for retinal detachment but that cannot be determined from the claims data. These data will permit the researchers to control for differences in those variables, providing stronger observational evidence that the posterior capsulotomy itself increases the risk of retinal detachment. PORT researchers studying prostate disease and total knee replacement have contacted patients to ask about their levels of functioning after certain procedures have been performed on them (52,145).

Another major concern is that the patients' inclusion in the databases is involuntary. Consequently, a request to participate in a study can be an unexpected imposition. Nonetheless, researchers who have taken this approach have found that the individuals are generally willing to participate. These researchers believe that the privacy mechanisms currently in place are adequate to protect patients' confidentiality and their freedom to choose whether to participate in a study. (The Institute of Medicine has recently issued a report with recommendations regarding national policy on the conflict between patients' privacy and databases' usefulness.)

If patients are willing participants, many health care providers are not. Some providers will not comply with requests for records or participate in studies of their decisionmaking processes (101). The fact that providers who decline to participate may be systematically different corrupts the very generalizability that makes the use of databases as sampling frames so attractive. Researchers at the RAND Corporation have described methods to increase participation (108), but the methods are costly and do not result in participation by 100 percent of the providers.

Large databases can also be used in conjunction with primary data collection to provide followup

[14] Each patient receives a letter from the Health Care Financing Administration several weeks before the researchers plan to contact them. The letter tells patients about the study and allows them to indicate that they prefer not to be contacted. A phone number is provided for patients who have questions. Of 1,750 patients who received letters as part of one PORT project, two called and expressed grave concern that this step had been taken without their consent (145).

for populations that are well characterized by the primary data collection. Researchers at the National Center for Health Statistics, for example, have used personal identifiers to link individuals who are included in several surveys (e.g., the National Health Interview Survey Supplement on Aging) (94,187) with information about the individuals in Medicare databases. Using personal identifiers that allow linkage between RCTS and the National Death Index is an inexpensive way to improve long-term followup of the participants' vital status (29). Linkage to the Medicare claims database could also provide information regarding the need for hospitalization for specific diagnoses, should yield estimates of the costs of subsequent care, and might improve researchers' chances of contacting patients directly. Unfortunately for researchers, people under the age of 65 have no identifiers other than their Social Security numbers and have no population-wide insurance system that could (like Medicare) be used to track medical events over time.

Variant approaches that use claims databases both for identifying samples and for conducting followup are surveys of the Medicare population. The Medicare Current Beneficiary Survey gathers personal data in interviews with samples of beneficiaries, then obtains followup information from the Medicare claims data. More than 12,000 beneficiaries were surveyed in 1991, the first year in which data were collected (149,1 82). The Medicare Beneficiary Health Status Registry uses a mailed survey to obtain data on a spectrum of issues affecting health, including lifestyle risk factors, functional status, medical history prior to Medicare eligibility, sociodemographics, and quality of life (1 26). Neither the Medicare Current Beneficiary Survey nor the Medicare Beneficiary Health Status Registry has existed long enough for an assessment of its utility.

OTHER ISSUES

The issues in database research vary with the technology under consideration, the database, and the focus of the assessment. Thus, coding inaccuracies are much more important for a study of myocardial infarction (where the error rates are high, perhaps in a biased pattern) than for a study of lung cancer (where the error rates are low). Similarly, the inability to distinguish the reason for a procedure in outpatient Medicare data might affect a study of mammography (a procedure that could just as easily be done for diagnosis or screening as for a workup for known cancer), but would not affect a study of cataract extraction (a procedure that is generally performed only for cataracts related to existing or anticipated visual impairment).

Enhancing Databases

As the use of databases for assessing health care has grown, so has the realization that the existing databases are often inadequate for the proposed uses. This has stimulated interest in designing databases that are better suited for the analyses. Strategies for doing so include not only improving and augmenting the data but, in some cases, even designing entirely new databases (box 2-2).

Collecting More Data

A number of studies have shown that clinical data beyond the ICD-9-CM-coded discharge abstract data can explain the differences in the resources used for patients in identical diagnosis-related groups (the clinical grouping categories used as a basis for Medicare payment) (18,36,75,139). By collecting this data in multiple hospitals at multiple locations, proponents of these systems have shown that it is feasible for the data to be collected on a large scale (3,88). The addition of just three clinical variables not reflected in the discharge abstract data markedly improves the ability of a model based on ICD-9-CM data to predict mortality following CABG surgery (61).

The costs of such data collection, however, can be high. The experience of the SEER program is instructive. Early data collected in SEER registries included detailed data regarding the extent to which each patient's cancer had spread. Quality-control activities disclosed that the reliability of the data regarding fine gradations in stage was limited, but the data were very accurate for distin-

BOX 2-2: Creating a New Kind of Database: Community Health Management Information Systems

An ambitious method of addressing the problem of limited data is to create an entirely new database that both links and augments existing data sources. The Hartford Foundation's Community Health Management Information System project is an example of an effort that has begun the process of developing a communitywide database to enhance both the quality of care and the ability to assess the effects of medical treatment.

The foundation has organized community leaders to consider the advantages of improved health data systems in several cities around the country, with varying degrees of support from local governments. Which elements the database should include, how the data should be gathered, and how the quality of the data should be maintained are all decisions that will be made explicitly as the database is designed.

SOURCE: Jeff Whittle, 1995.

guishing coarser gradations. The coding system was designed to provide optimally useful scientific information on stage, but practical difficulties in obtaining accurate coding limited the use of the data, which were gathered at considerable cost (91). The difficulty of precise coding in this research-oriented, single-disease database should give pause to those who want to initiate more detailed collection in other large databases.

Pennsylvania now requires that all discharge abstracts include the key elements necessary for determining the MedisGroups severity levels (12,86). Researchers in the pneumonia PORT have used the nationwide MedisGroups database to analyze predictors of length-of-stay and mortality among patients admitted with pneumonia (42,43); they will use the statewide Pennsylvania MedisGroups database in the future.

HCFA has had an active interest in a similar set of clinical data, the Uniform Clinical Data Set (UCDS), for several years (109). The UCDS began as a set of more than 1,700 clinical data elements that could be collected by reviewing the charts associated with a subset of Medicare discharge abstracts. As time has gone by, the UCDS concept has evolved to a more flexible model that entails collecting different sets of data for different clinical entities (37).

A careful assessment of the experience of those who use the databases with supplemental clinical data may provide future guidance regarding the overall usefulness of a number of variants of this technique.

Including a Health Status Measure

Health status is an important outcome that is unavailable in current databases. Because most treatments are intended to improve the patients' health rather than to prolong their lives, death as an outcome of treatment is likely to be inapplicable in many situations.

Some observers have suggested incorporating a health status measure in the claims records maintained by Medicare. Selecting a measure appropriate for all patients, however, is problematic. The best measure for assessing the benefits of total hip replacement is certainly not the best measure for assessing the benefits of cataract surgery. Moreover, the important outcomes of surgery often cannot be determined at the time of hospital discharge. The reduction in pain following a hip replacement is best assessed after the patient has been discharged, and patients who undergo radical prostatectomy cannot be expected to assess their sexual functioning before leaving the hospital. One approach would be to allow researchers to contact the patients later to obtain the specific functional data of interest. This has been done by research groups studying the effects of prostatectomy and total knee replacement.

The experience of the Medicare Beneficiary Health Status Registry may provide guidance regarding whether and how health-status outcomes might be added to claims and discharge abstract databases.

Linking Databases

Where the data in a single database are limited, researchers can sometimes combine two databases. In a study comparing open prostatectomy with TURP in the treatment of BPH, the only clinical detail available from the claims database was coded with ICD-9-CM, but some of these patients could be linked to a list of patients for whom an anesthesiologist had carried out preoperative risk assessments (161).

This approach can sometimes provide useful additional data. Epidemiologists link databases when they determine vital status by comparing the names of study participants with information from death registries, driver's-license agencies, and telephone books. By carefully reviewing each linked piece of data, the researchers can be reasonably certain that the data refer to the same individuals (9).

Database linkages can provide information beyond vital status. Researchers have linked state discharge data and cancer registry data from New Jersey to study differences in the results obtained by breast cancer patients with varying insurance coverage (4). A linked Medicare-SEER database, currently under development, will provide information about patients' treatment and the costs of their care from Medicare and detailed information about their cancer from SEER. Linkages between Medicare and Medicaid allow researchers to obtain information about the use of prescription drugs from Medicaid and longitudinal followup from Medicare (147,148).

The experience of several investigators demonstrates the feasibility and potential usefulness of this approach. The SEER-Medicare link—using each patient's name, sex, Social Security number, date of birth, and date of death—has identified the Medicare claims record for nearly 95 percent of the individuals over the age of 64 who are in the SEER cancer registry. In another case, Medicare records were identified for 85 percent of the men over 64 who used VA facilities in the Northeast over a four-year period. The researchers were able to study the degree to which this cohort used medical services provided by entities other than the VA—an accomplishment that has facilitated more accurate interpretations of studies of the VA administrative database (49).

Database linkages are not without problems, however. Most practically, linked studies require access to two (or more) databases, which doubles (or more) the cost of acquisition, the potential for violations of privacy, and the amount of data cleaning that is needed. In addition, different databases may use different definitions for similar concepts. For example, the coding for cancer surgeries in the SEER database differs from the ICD-9-CM system in the Medicare hospital database and from the Current Procedural Terminology (CPT) **system in** the professional claims database.[15]

Other problems of linkage are more technical. To a greater or lesser extent, all linkages are probabilistic—that is, the researcher identifies pairs of members of each database that have a certain probability of being the same persons (two records that share the same sex, the same last name, the same date of birth, and the same maternal last name represent the same person with a certain likelihood). The use of unique identifiers, such as social security numbers, can allow very high confidence about a match (160), but they often fail to be truly unique. For example, women who are eligible for Medicare because of their husbands' eligibility use their husbands' social security num-

[15] The Agency for Health Care Policy and Research is working to address these issues (25).

bers (although each beneficiary's own number will be included in the Medicare claims database in the future). In addition, many older individuals have been assigned more than one social security number (104). 1[6]

Methods of increasing linkage, and the confidence with which probabilistic matches can be regarded as true matches, are the subjects of active research. Issues include increasing linkage rates, enhancing the accuracy of links, providing estimates of the accuracy of links, dealing with uncertain linkages, and minimizing the computational burden of matching records between large databases {78,159,1 81,1 85).

I The Electronic Medical Record

As expanded databases become more complete, researchers will have access to more of the data that are in patients' charts. Taken to its extreme, this concept would result in the complete computerization of medical records, which could provide researchers with access to all the information that is generated during the patients' hospital stays.

Researchers at Beth Israel Hospital in Boston have used a computerized hospital database for more than 10 years and have demonstrated its ability to easily supplement the information found in the coded discharge abstracts. This does not eliminate the concern that risk adjustments made with these data may be biased because the data are obtained for nonrandom reasons (166). The lack of uniformly applied diagnostic criteria remains a potential source of error in any database that does not impose definitions for the diagnoses of interest, whether the medical record is coded or electronic. Moreover, certain data that a researcher might be anxious to have (e.g., scores on a particular functional status scale) are unlikely to appear anywhere in the medical record. Of course, many of these concerns apply whether the data about patients are obtained from a readily used, computerized practice database or by a laborious review of charts (the traditional method).

The electronic medical record has many potential uses for research into the effectiveness of health care. Inasmuch as more data are available in an electronic format, however, researchers should find it easier to retrieve information about samples of patients who have been identified through claims databases. The electronic medical record, therefore, will probably complement, rather than replace, large databases as tools for evaluating medical technologies.

Retrieving Primary Data

One alternative to adding information to the database is to allow researchers to retrieve the required data directly. Access to primary data allows theresearcher more flexibility in choosing data elements, as well as direct control over the quality of the data collection. With the ability to contact patients or providers, researchers could even identify data that are not recorded in the charts. This approach raises difficult questions of privacy (inasmuch as the patients may not be asked for permission to use their charts), logistics (because some databases reflect admissions throughout the country), and selection bias (because some charts-which are likely to be the unusual ones—will not be found) (168).

Several studies have demonstrated high rates of record retrieval when Medicare databases were used. Moreover, both the prostate and total-knee-replacement PORT teams have found that when individuals who are identified through database analyses are contacted directly, they cooperate with research efforts. In addition, HCFA and individual researchers have developed methods that facilitate research with appropriate concern for privacy rights. There is little experience with similar research using private or state databases, perhaps because concerns about the potential for law-

[16] The vast majority of these people use only one of the numbers; they acquired the others early in the history of Social Security, when some people thought a different number was needed for each job.

suits provide strong incentives for private insurers to keep researchers from contacting the individuals in the insurers' databases.

Other Possibilities for Improvement

Unique Identifiers

Probably the most common concern of database researchers is their inability to get a complete picture of medical services used over time, except from the Medicare claims data. The root of the problem is the lack of unique identifiers for patients in most other databases. The identifiers in insurance databases are unique, but patients move into and out of the systems relatively frequently, which hinders followup. Moreover, insurance databases generally do not include representative samples of the population.

Unique identifiers would provide several benefits:

- Researchers could more easily follow patients through time, between providers, and in various settings.
- All health care services covered by a database could be linked.
- Linkage would be simplified.
- A representative sample could be generated by starting with the list of all active unique identifiers.

Improving Data Accuracy

Improving the accuracy, completeness, and reliability of medical data is widely believed to be a good thing, for obvious reasons.

There are several potential mechanisms to facilitate improvements. For example, the computers used in the data entry systems could be programmed to reject out-of-range values, inconsistent codes, and nonspecific codes as the data are entered. Moreover, the programs could be altered to prevent the submission of incomplete data, to incorporate prompts regarding common errors, and to minimize transcription errors. Precise guidelines on appropriate coding could also help by making coding more consistent. Although the coding manuals used by all coders contain identical codes, the ambiguities in the ICD-9-CM system lead to wide variations in how coders in different sites apply the codes. Another useful mechanism would be to provide financial incentives for accurate and complete coding. Unfortunately, it is more difficult to envision a reward system than a system that imposes payment delays or other financial penalties for coding that fails to meet standards.

Regardless of any improvements, the data will always contain errors. As with research involving primary data collection, assessing the quality of data is important to understanding the results of database research. Ideally, researchers should be able to assess the quality of specific data elements in the specific database they use. With the exception of reports on Medicare hospital-claims data, however, few studies regarding the quality of data have been published.

Improving the Coding Systems

Inasmuch as the ICD-9-CM coding system itself has been blamed for much of the difficulty with using claims data, changing the coding system might enhance the databases' usefulness. Obvious gaps that could be filled include information about whether diseases were present or procedures were performed on the patient's right side, left side, or both; whether conditions existed at the times of admission or developed during the hospitalization; and how severe conditions are. Like adding new fields of data to be collected, however, increasing the complexity of the coding system adds to the data-collection burden. In New York, a binary code to designate whether a condition was present at admission has been introduced but has not been reliably implemented (61).

The ICD-9-CM is under continuous revision. In addition to HCFA, the entities most concerned are special-interest groups that wish to make certain conditions more precisely identifiable. Reliable, regular communication with persons knowledgeable about the data needs of medical researchers could provide guidance about how proposed coding changes might affect the usefulness of coded data for their research. For example, AHCPR staff made suggestions regarding changes that could make the coding of incontinence and prostatic diseases more useful for evaluating the effects of treatment (131).

National standardization of HCFA'S coding system for professional claims might facilitate the use of professional service claims in the Medicare database.

Exchanging Information About Database Research Methods

Information about how the data in a database are collected, verified, and stored is crucial for understanding and using a database. Although the accuracy and completeness of specific data elements in particular databases may have been studied, the results often are not published and summary documents are unavailable, which means that other researchers must spend time and money rediscovering or recreating vital pieces of information. Increasing the availability of such information is a major focus of AHCPR'S Office of Science and Data Development (184), but researchers are reluctant to write summaries of their findings about the databases, because the subject is too arcane for publication and the writing is too time-consuming to undertake without strong incentives.

Although there are many published and unpublished studies of the quality of the information in various databases, the need for current information about the accuracy of data never ceases. Continuous changes in coding incentives and conventions can affect coding accuracy in ways that are unpredictable in magnitude, if not direction. Because the accuracy of databases varies considerably, depending on which conditions are being studied, information on the specific codes of interest must be used. This usually means that the researchers must carry out coding studies that specifically target the issue at hand, because even the large (more than 7,000 charts) validation studies carried out in connection with the change in Medicare payment included fewer (usually much fewer) than 100 patients with any one condition. Ideally, the studies should be published, but they frequently are not.

Methodologic Advances

Certain aspects of methodology are particularly important for research using claims and discharge abstract data. For example, better case-mix adjusters based only on ICD-9-CM data would be extremely useful. This area has been active, with several proprietary and open systems available. At present, there is no consensus as to the most useful method, although different methods will probably prove best for different applications.

Newer statistical methodologies, including hierarchical modeling (129), are being explored for use in adjusting. Of course, improvements in statistical techniques do not eliminate the requirement for accurate, reliable data on important potential confounders.

Analytical Costs

Compared with research that entails primary data collection, database analysis is inexpensive (92,175). Researchers incur fewer expenses for collecting claims data, for example, which has already been gathered for billing purposes. Moreover, because the patients are receiving routine clinical care and the clinicians are paid as part of routine practice, the study does not need to fund the patients' care (202). Nonetheless, a number of the costs involved in preparing the data for use in research should be considered.

First, the data must be acquired from whomever collects them. The charges are generally low compared with those for primary data collection, but they are not insignificant. The costs of the public use tapes that include 100 percent samples of Medicare hospital discharge data with linkable identifiers, for example, are $6,120 for each year of data (182). Pennsylvania provides state discharge abstracts, supplemented with the admission and followup MedisGroups severity scores,1[7] for one cent per discharge plus the computer costs of se-

[17] MedisGroups is the trade name for a system that uses clinical data from hospital records to generate scores that predict individual patients' costs and outcomes. These data are also used to detect clinical deterioration that occurred during the patients' hospitalization and that might signal problems with the quality of the care being delivered.

lecting the desired population. The cost of acquiring discharge abstract data enhanced with the more than 100 clinical data elements that go into the MedisGroups scoring system was 15 cents per record for the pneumonia PORT.

Elements of particular interest to researchers are sometimes poorly documented, and the researchers incur additional costs in time and computer resources while discovering the weaknesses in the data and validating the data elements. Researchers who wish to link two or more databases must pay the costs of acquiring and examining each database separately and then performing the linkage. Furthermore, the statistical analyses are generally more expensive to perform in studies using large databases than in studies using primary data collection from relatively small numbers of individuals.

As efforts are made to enhance the databases, the costs of data collection itself may become a significant factor. Gathering data solely for a database is not inexpensive. The Connecticut Tumor Registry, for example, lists nearly 15,000 new cancer cases each year (and follows them up) at a cost of approximately $1 million (170). If elements are added to routine billing data for research purposes, the added costs should be assigned to the research.[18] As part of an effort to provide risk-adjusted data on costs and outcomes in state hospitals, the Pennsylvania Health Care Cost Containment Commission (HCCCC) has, since 1986, required most of the state's hospitals to gather the clinical data (Key Clinical Findings) needed for determining the MedisGroups severity scores when patients are admitted and again during their hospital stays (12,86).

Pennsylvania's experience illustrates the potential costs of such data collection. The HCCCC has a budget of $2 million, somewhat over half of which is allocated to data analysis (38). The added cost to the hospitals is significant: researchers calculated the cost of collecting these dataat$13.90 per discharge for the first year (88), which—in a state with approximately 2 million discharges annually—meant that the total cost of compliance approached $25 million. Anecdotal evidence suggests that the costs decrease by $1 or $2 per discharge as the process becomes routine, but a statewide cost in excess of $20 million annually appears likely. MediQual, which markets the MedisGroups system, estimates the time commitment at 20 minutes per medical chart and 30 minutes per surgical discharge abstract (136).

The potential usefulness of the MedisGroups data for quality assurance had led 15 percent of the hospitals in Pennsylvania to collect this data prior to the HCCCC'S mandate. Moreover, the value of the data for nonresearch activities should be considered when evaluating the cost. Nonetheless, if data requirements are imposed on hospitals to support research activities, a careful consideration of the financial implications for all parties concerned is warranted.

Researchers working with the PORTS report varying commitments of time and resources to database analyses, but all agree that the amount of time is both greater than expected and substantial. An estimate of the actual cost is difficult to obtain, because time allotted by salaried investigators is a major variable. Depending on the PORT, between 10 percent and half of the resources committed to the overall project have been devoted to database analyses (102).

Several things are likely to reduce the costs of database analysis in the future. Information about which elements in databases are reliable will become available from the researchers who are currently exploring them, which will save future re-

[18] Because health maintenance organizations (HMOs) are not now required to generate claims as part of the billing process, any costs that HMOs incur because of requirements that they generate similar data might be considered research costs as well. Anecdotally, many HMOs gather fairly extensive utilization data as part of their internal quality control activities, so this concern may be only theoretical.

searchers a lot of time.[19] Databases may become available in already linked form, allowing the cost of linkage to be paid just once. After two databases have been linked, the cost of updating them will probably decrease. AHCPR has been active in trying to identify and remove the obstacles to the efficient use of databases (185).

REFERENCES

1. Abrams, H. B., Detsky, A. S., Roos, Jr., L.L., et al., "Is There a Role for Surgery in the Acute Management of Infective Endocarditis? A Decision Analysis and Medical Claims Database Approach," *Medical Decision Making 8: 165-174,* 1988.
2. American Medical Association, ***CPT Physicians' Current Procedural Terminology*** (Chicago, IL: American Medical Association, 1991).
3. Averill, R. F., McGuire, T. E., Manning, B. E., et al, "A Study of the Relationship Between Severity of Illness and Hospital Cost in New Jersey Hospitals," *Health Services Research* 27:587-606; *1992.*
4. Ayanian, J. Z., Kohler, B. A., Abe, T., et al., "The Relation Between Health Insurance Coverage and Clinical Outcomes Among Women with Breast Cancer," New *England Journal of Medicine* 329:326-331, *1993.*
5. Beebe, G. W., "Long-Term Follow-Up Is a Problem," *American Journal of Public Health* 73:245-246, 1983.
6. Beta-Blocker Heart Attack Trial Research Group, "A Randomized Trial of Propranolol in Patients with Acute Myocardial Infarction I.: Mortality Results," *Journal of/he American Medical Association* 247: *1707-1714, 1982.*
7. Blumberg, M. S., "Potentials and Limitations of Database Research Illustrated by the QMMP AMI Medicare Mortality Study," *Statistics in Medicine* 10:637-646, *1991.*
8. Boissel, J. P., "Impact of Randomized Clinical Trials on Medical Practices," *Controlled Clinical Trials* 10 (suppl.):120S-134S, 1989.
9. Boyle, C. A., and Decoufle, P., "National Sources of Vital Status Information: Extent of Coverage and Possible Selectivity of Reporting, ''*American Journal of Epidemiology* 131:160-168, *1990.*
10. Braunwald, E., *Heart Disease: A Textbook of Cardiovascular Medicine* (Philadelphia, PA: W.B. Saunders, 1992).
11. Brem, H., John Hopkins Hospital, Baltimore, MD, personal communication, January 1990.
12. Brewster, A. C., Karlin, G. B., Hyde, L. A., et al., "MEDISGROUPS: A Clinically Based Approach to Classifying Hospital Patients at Admission," *Inquiry* 22:377-387, *1985.*
13. Brook, R. H., Kosecoff, J. B., Park, R. E., et al., "Diagnosis and Treatment of Coronary Disease: Comparisons of Doctors' Attitudes in the USA and the UK," *Lancet* 1:750-753, 1988.
14. Bunker, J. P., "Surgical Manpower: A Comparison of Operations and Surgeons in the United States and in England and Wales," New *England Journal of Medicine 282:* 135-*144, 1970.*
15. Byar, D. P., "Why Data Bases Should Not Replace Randomized Clinical Trials," *Biometrics* 36:337-342, *1980.*
16. Byar, D. P., "Problems with Using Observational Databases to Compare Treatments," *Statistics in Medicine* 10:663-666, *1991.*

[19] A, this time, there is no reliable source for such information. Because it is of interest to only the few people carrying out database research, it is unlikely to be published in widely circulated medical or health services research journals. AHCPR has facilitated the exchange of such information by sponsoring conferences that provide opportunities for database researchers to exchange information (217). The products of these conferences serve as valuable resources. AHCPR has also established the Use of Claims Data Work Group, which includes many of the PORT investigators who analyze claims.

17. Byar, D.P., Schoenfeld, D.A., Green, S. B., et al., "Design Considerations for AIDS Trials," New *England Journal of Medicine* 323:1343-1348, *1990.*
18. Calore, K.A., and Iezzoni, L. I., "Disease Staging and PMCs: Can they Improve DRGs?" *Medical Care* 25:724-737, *1987.*
19. Devita, Jr., V.T., Hellman, S. and Rosenberg, S.A. (eds.), *Cancer: Principles and Practice of Oncology* (Philadelphia, PA: J.B. Lippincot Company, 1993).
20. CASS Principal Investigators and Their Associates, "Coronary Artery Surgery Study (CASS): A Randomized Trial of Coronary Artery Bypass Surgery. Comparability of Entry Characteristics and Survival in Randomized Patients and Nonrandomized Patients Meeting Randomization Criteria," *Journal of the American College of Cardiology 3:* 114-128, 1984.
21. Chassin, M.R., Brook, R. H., Park, R. E., et al, "Variations in the Use of Medical and Surgical Services by the Medicare Population," New *England Journal of Medicine* 314 (30):285-290, *1986.*
22. Chassin, M. R., Kosecoff, J., Park, R.E., et al., "Does Inappropriate Use Explain Geographic Variations in the Use of Health Care Services? A Study of Three Procedures," *Journal of the American Medical Association 258* (18):2533-2537, *1987.*
23. Clearly, P. D., Greenfield, S., Mulley, A.G. et al., "Variations in Length of Stay and Outcomes for Six Medical and Surgical Conditions in Massachusetts and California," *Journal of the American Medical Association* 26:73-79, 1991.
24. Cobb, L. A., Thomas, G. I., Dillard, D. H., et al., "An Evaluation of Internal-Mammary - Artery Ligation by a Double-Blind Technique," New *England Journal of Medicine* 260:1115-1118, *1959.*
25. Cohen, E., Bemier, D., Tam, S., et al., "Data Quality and DRGs: An Assessment of the Reliability of Federal Beneficiary Discharge Data in Selected Manhattan Hospitals," *Journal of Community Health* 10:238-246, 1985.
26. Coleman, M., "Chemotherapy for Large-Cell Lymphoma: Optimism and Caution," *Annals of Internal Medicine* 103:140-142, *1985.*
27. Cohen, M. F., "Clinical Research Databases—A Historical Review," *Journal of Medical Systems* 14:323-344, *1990.*
28. Concato, J., Horwitz, R. I., Feinstein, A. R., et al., "Problems of Comorbidity in Mortality After Prostatectomy," *Journal of the American Medical Association 267:1077-1082, 1992.*
29. Curb, J. D., Ford, C.E., Pressel, S., et al., "Ascertainment of Vital Status Through the National Death Index and the Social Security Administration," *American Journal of Epidemiology* 121:754-766, *1985.*
30. Daley, J., Jencks, S., Draper, D., et al., "Predicting Hospital-Associated Mortality for Medicare Patients: A Method for Patients with Stroke, Pneumonia, Acute Myocardial Infarction, and Congestive Heart Failure," *Journal of the American Medical Association* 260:3617-3624,1988.
31. Demlo, L. K., and Campbell, P. M., "Improving Hospital Discharge Data: Lessons from the National Hospital Discharge Survey," *Medical Care 19: 1030-1040, 1981.*
32. Doughty, A., Nash, S.1., and Gift, D. A., "Deployment and Utilization of MR Imaging in Michigan: Observations of a Statewide Data Base," *Radiology* 185:53-61, *1992.*
33. Dyck, F.J., Murphy, F.A., Murphy, J. K., et al, "Effect of Surveillance on the Number of Hysterectomies in the Province of Saskatchewan," New *England Journal of Medicine* 296:1326-1328, *1977.*
34. Eddy, D. M., and Billings, J., "The Quality of Medical Evidence: Implications for Quality Care," *Health Affairs* 7(1):19-32, spring 1988.
35. Eggers, P.W., Connerton, R., and McMullan, M., "The Medicare Experience with End-Stage Renal Disease: Trends in Incidence,

Prevalence, and Survival," *Health Care Financing Review* 5:69-88, 1984.

36. Eisenberg, B.S., "Diagnosis-Related Groups, Severity of Illness and Equitable Reimbursement Under Medicare," *Journal of the American Medical Association* 251(5): *645-646, 1984.*

37. Ellerbeck, E., U.S. Department of Health and Human Services, Health Care Financing Administration, Office of Quality Assurance, Baltimore, MD, personal communication, June 1993.

38. Essa, E.J., Health Care Cost Containment Commission, Harrisburg, PA, personal communication, June 1993.

39. Farow, D. C., Hunt, W. C., and Samet, J. M., "Geographic Variations in the Treatment of Localized Breast Cancer," New *England Journal of Medicine* 326(17):1097-1101, *1992.*

40. Feinstein, A.R., "ICD, POR, and DRG: Unsolved Scientific Problems in the Nosology of Clinical Medicine," *Archives of Internal Medicine* 148:2269-2274, *1988.*

41. Feinstein, A.R., Sosin, D. M., and Well, C. K., "The Will Rogers Phenomenon: Stage Migration and New Diagnostic Techniques as a Source of Misleading Statistics for Survival in Cancer," New *England Journal of Medicine* 312:1604-1608, *1985.*

42. Fine, M.J., Hanusa, B. H., Lave, J.R., et al., "Variation in Length of Stay in Patients Hospitalized with Community-Acquired Pneumonia" (abstract), *Clinical Research 39: 142A, 1991.*

43. Fine, M.J., Hanusa, B. H., Singer, D.E., et al., "Validation of a Pneumonia Mortality Risk Index Using the MedisGroups Comparative Hospital Database" (abstract), Clinical Research 39: 142A, 1991.

43a. Fisher, E. S., Dartmouth Medical School, Hanover, NH, personal communication, 1993.

#. Fisher, E. S., Baron, J. A., Malenka, D.J., et al., "Overcoming Potential Pitfalls in the Use of Medicare Data for Epidemiologic Research," *American Journal of Public Health* 80:1487-1490, *1990.*

45. Fisher, E. S., Baron, J.A., Malenka, D.J., et al., "Hip Fracture Incidence and Mortality in New England," *Epidemiology* 2:116-122, *1991.*

46. Fisher, E. S., Malenka, D.J., Wennberg, J.E., et al., "Technology Assessment Using Insurance Claims. Example of Prostatectomy," *International Journal of Technology Assessment in Health Care 6: 194-202, 1990.*

47. Fisher, E. S., Whaley, F. S., Krushat, W. M., et al., "The Accuracy of Medicare's Hospital Claims Data: Progress Has Been Made but Problems Remain," *American Journal of Public Health* 82:243-248, *1992.*

48. Fisher, R. I., Gaynor, E.R., Dahlberg, S., et al., "Comparison of a Standard Regimen (CHOP) with Three Intensive Chemotherapy Regimens for Advanced Non-Hodgkin's Lymphoma," New *EnglandJournal ofMedicine 328:1002-1006, 1993.*

49. Fleming, C., Fisher, E. S., Chang, C.H., Bubolz, T. A., Malenka, D.J., "Studying Outcomes and Hospital Utilization in the Elderly: The Advantages of a Merged Database for Medicare and Veterans Affairs Hospitals," *Medical Care* 30:377-391, *1992.*

50. Fode, N. C., Sundt, T. M., Robertson, J.T., et al., "Multicenter Retrospective Review of Results and Complications of Carotid Endarterectomy in 1981," *Stroke* 17:370-376, *1986.*

51. Ford, E., Cooper, R., Castaner, A., et al., "Coronary Arteriography and Coronary Bypass Survey Among Whites and Other Racial Groups Relative to Hospital-Based Incidence Rates for Coronary Artery Disease: Findings From NHDS," *American Journal of Public Health* 79:437-440, *1989.*

52. Fowler, F. J., University of Massachusetts, Boston, MA, personal communication, October 1992.

53. 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.*

54. Friedman, L. M., Furberg, C.D., and DeMets, D. L., *Fundamentals of Clinical Trials* (Littleton, MA: PSG Publishing Company, Inc., 1985).

55. Fries, J.F., "Advances in the Management of Rheumatic Disease: 1965 to *1985,'' Archives of Internal Medicine 149:1002-1011,* 1989.

56. Gehan, E.A., "The Evaluation of Therapies: Historical Control Studies," *Statistics in Medicine* 3:315-324, *1984.*

57. Ginsberg, R.J., Hill, L.D., Eagan, R.T., et al., "ModemThirty-Day Operative Mortality for Surgical Resections in Lung Cancer," *Journal of Thoracic and Cardiovascular Surgery* 86:654-658, *1983.*

58. Glover, J.A., "The Incidence of Tonsillectomy in School Children," *Proceeding of the Royal Society of Medicine* 31:1219-1235, *1938.*

59. Glover, R.P., Davila, J. C., Kyle, R. H., et al., "Ligation of the Internal Mammary Arteries as a Means of Increasing Blood Supply To the Myocardium," *Journal of Thoracic and Cardiovascular Surgery* 34:661-678, *1957.*

60. Gordon, L. I., Barrington, D., Andersen, J., et al., "Comparison of a Second-Generation Combination Chemotherapeutic Regimen (m-BACOD) with a Standard Regimen (CHOP) for Advanced Non-Hopkin's Lymphoma," New *England Journal of Medicine* 327:1342-1349, *1992.*

61. Hannan, E.L., Kilburn, Jr., H., Lindsey, M. L., et al., "Clinical Versus Administrative Data Bases for CABG Surgery: Does It Matter?" *Medical Care* 30:892-907, *1992.*

62. Hannan, E.L., Kilburn, Jr., H., O'Donnell, J.F., et al., "Interracial Access to Selected Cardiac Procedures for Patients Hospitalized with Coronary Artery Disease in New York State," *Medical Care* 29:430-441, *1991.*

63. Harris, W. H., and Sledge, C. B., "Medical Progress: Total Hip and Total Knee Replacement (First of Two Parts)," New *England Journal of Medicine* 323:725-731,1990.

64. Hartz, A.J., Kuhn, E. M., Pryor, D. B., et al., "Mortality After Coronary Angioplasty and Coronary Artery Bypass Surgery (The National Medicare Experience)," *American Journal of Cardiology* 70: *179-185,* 1992.

65. Hatten, J., "Medicare's Common Denominator: The Covered Population," *Health Care Financing Review,* fall:53-64, 1980.

66. Held, P.J., Levin, N. W., Bovbjerg, R. R., et al., "Mortality and Duration of Hemodialysis Treatment," *Journal of the American Medical Association* 265(7):871-875, *1991.*

67. Hellman, S., and Hellman, D. S., "Of Mice But Not Men. Problems of the Randomized Clinical Trial," New *England Journal of Medicine* 324:1585-1589, *1991.*

68. Hine, L.K., Laird, N., Hewitt, P., et al., "Meta-Analytic Evidence Against Prophylactic Use of Lidocaine in Acute Myocardial Infarction," *Archives of Internal Medicine* 149:2694-2698, *1989.*

69. Hlatky, M.A., Stanford University, Palo Alto, CA, personal communication, July 1993.

70. Hlatky, M.A., "Using Databases to Evaluate Therapy," *Statistics in Medicine* 10:647-652, *1991.*

71. Hlatky, M. A., Califf, R. M., Harrell, F. E., et al., "Comparison of Predictions Based on Observational Data with the Results of Randomized Controlled Clinical Trials of Coronary Artery Bypass Surgery," *Journal of the American College of Cardiology* 11(2):237-*245, 1988.*

72. Holmes, M. D., Hodges, D., and Rich, J., "Racial Inequalities in the Use of Procedures for Ischemic Heart Disease," *Journal of the American Medical Association 261:3242-3243, 1989.*

73. Holtgrewe, H. L., "An American Urological Association Prospective, Randomized Clinical Trial in the Treatment of Benign Prostatic Hyperplasia," *Cancer* 70:351-354, *1992.*

74. Hori, R. Y., Lewis, J.L., Zimmerman, J.R., et al., "The Number of Total Joint Replacements in the United States," *Clinical Orthopaedics and Related Research* 132:46-52, *1978.*

75. Horn, S.D, Ashworth, M.A., "Adjusting DRGs for Severity of Illness," *Socioeconomics of Surgery,* I.M. Rutkow (cd) (St. Louis, MO: C.V. Moseby, 1989).

76. Horwitz, R. I., Viscoli, C. M., Clemens, J. D., et al., "Improved Observational Method for Studying Therapeutic Efficacy: Suggestive Evidence That Lidocaine Prophylaxis Prevents Death in Acute Myocardial Infarction," *Journal of the American Medical Association 246* (21):2455-2459, *1981.*

77. Horwitz, R. I., Viscoli, C. M., Clemens, J. D., et al., "Developing Improved Observational Methods for Evaluating Therapeutic Effectiveness," *American Journal of Medicine 89* (5):630-638, *1990.*

78. Howe, G. R., and Lindsay, J., "A Generalized Iterative Record Linkage Computer System for Use in Medical Follow-Up Studies," *Computers and Biomedical Research* 14:327-340, *1981.*

79. Hsia, D. C., Ahem, C. A., Ritchie, B.P., et al., "Medicare Reimbursement Accuracy Under the Prospective Payment System, 1985 to 1988," *Journal of the American Medical Association* 268:896-899, *1992.*

80. Hsia, D. C., Krushat, W. M., Fagan, A. B., et al., "Accuracy of Diagnostic Coding for Medicare Patients Under the Prospective Payment System," *New England Journal of Medicine* 318:352-355, *1988.*

81. Hsia, D. C., Krushat, W. M., and Moscoe, L. M., "Epidemiology of Carotid Endarterectomies Among Medicare Beneficiaries," *Journal of Vascular Surgery* 16:201-208, *1992.*

82. Hughes, R. G., Hunt, S. S., Luft, H. S., "Effects of Surgeon Volume and Hospital Volume on Quality of Care in Hospitals," *Medical Care* 25:489-503, *1987.*

83. Hulley, S. B., Cummings, S.R., Browner, W. S., et al., *Designing Clinical Research: An Epidemiologic Approach* (Baltimore, MD: Williams and Wilkins, 1988).

84. Iezzoni, L. I., "Using Administrative Diagnostic Data To Assess the Quality of Hospital Care. Pitfalls and Potential of ICD-9-CM," *International Journal of Technology Assessment in Health Care* 6:272-281, *1990.*

85. Iezzoni, L. I., Foley, S. M., Daly, J., et al., "Comorbidities, Complications, and Coding Bias: Does the Number of Diagnosis Codes Matter in Predicting In-Hospital Mortality?" *Journal of the American Medical Association* 260:3159-3163, *1988.*

86. Iezzoni, L.I., Moskowitz, M. A., "A Clinical Assessment of MedisGroups," *Journal of the American Medical Association 260:3159-3163, 1988.*

87. Iglehart, J. K., "Canada's Health Care System" (part 1), New *EnglandJournal ofMedicine* 315:202-208, *1986.*

88. Institute of Policy Research & Evaluation, "Hospitals' Response to Pennsylvania's Mandated Cost and Quality Data Reporting; Center for Health Policy Research: CHPR Research Report," The Pennsylvania State University, May 1990.

89. Institute of Medicine, *Reliability of Hospital Discharge Abstracts* (Washington, DC: National Academy of Sciences, 1977).

90. Institute of Medicine, *Reliability of Medicare Hospital Discharge Records,* (Washington, DC: National Academy of Sciences, 1977).

91. Jablon, S., Garfinkel, L., Gordis, L., et al., *Jablon Report,* 1981.

92. Javitt, J. C., "Outcomes of Eye Care from Medicare Data" (editorial), *Archives of Ophthalmology 109: 1079-1080, 1991.*

93. Javitt, J. C., Tielsch, J. M., Canner, J. K., et al., "National Outcomes of Cataract Extraction. Increased Risk of Retinal Complications Associated with Nd:YAG Laser Capsulotomy. The Cataract Patient Outcomes Re-

search Team," *Ophthalmology 99: 1487-97, 1992.*

94. Jekel, J.F., "Rainbow Reviews V: Recent Publications of the National Center for Health Statistics,' 'Journal *of Clinical Epidemiology 45:* 1185-1190, 1992.

95. Jencks, S.F., Williams, D. K., 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.*

96. Jick, H., "The Commission on Professional and Hospital Activities-Professional Activity Study: A National Resource for the Study of Rare Illnesses," *American Journal of Epidemiology* 109:625-627, *1979.*

97. Jick, H., Dinan, B., and Rothman, K.J., "Oral Contraceptives and Nonfatal Myocardial Infarction," *Journal of the American Medical Association* 239:1403-1406, *1978.*

98. Kahn, C. N., "Policy Implications of Outcomes Research," *International Journal of Technology Assessment in Health Care* 6:295-296, *1990.*

99. Kahn, H.A., *An Introduction of Epidemiologic Methods (New* York, NY: Oxford University Press, 1983).

100. Kahn, K. L., Rogers, W. H., Rubenstein, L. V., et al., "Measuring Quality of Care with Explicit Process Criteria Before and After Implementation of the DRG-Based Prospective Payment System," *Journal of the American Medical Association* 264: *1969-1973,* 1990.

101. Kallich, J., RAND Corp., Santa Monica, CA, personal communication, May 1993.

102. Kapoor, W., University of Pittsburgh, Pittsburgh, PA personal communication, May 1993.

103. Keller, R. B., Soule, D. N., Wennberg, J. E., et al., "Dealing with Geographic Variations in the Use of Hospitals. The Experience of the Maine Medical Assessment Foundation Orthopaedic Study Group," *Journal of Bone and Joint Surgery* 72: *1286-1293, 1990.*

104. Kestenbaum, B., U.S. Department of Health and Human Services, Social Security Administration, Office of the Actuary, Baltimore, MD, persona! communication, July 1992.

105. Klimo, P., and Connor, J. M., "MACOP-B Chemotherapy for the Treatment of Diffuse Large-Cell Lymphoma," *Annals of Internal Medicine 149:* 1177-1181, 1989.

106. Knickman, J. R., and Foltz, A., "Regional Differences in Hospital Utilization: How Much Can Be Traced to Population Differences?" *Medical Care* 22:971-986, *1984.*

107. Kong, D. F., Lee, K. L., Harrell, Jr., F. E., et al., "Clinical Experience and Predicting Survival in Coronary Disease," *Archives of Internal Medicine* 102:596-602, *1985.*

108. Kosecoff, J., Chassin, M. R., Fink, A., et al., "Obtaining Clinical Data on Appropriateness of Medical Care in Community Practice," *Journal of the American Medical Associations* 258:2538-2542,1987.

109. Krakauer, H., "The Uniform Clinical Data Set," *Electiveness and Outcomes in Health Care,* K.A. Heithoff and K.N. Lohr (eds.) (Washington, DC: National Academy Press, 1990).

110. Krakauer, H., and Bailey, R. C., "Epidemiologic Oversight of the Medical Care Provided to Medicare Beneficiaries," *Statistics in Medicine* 10:521-540, *1991.*

111. Krakauer, H., Bailey, R. C., Skellan, K.J., et al., "Evaluation of the HCFA Model for the Analysis of Mortality Following Hospitalization," *Health Services Research 27:3:317-335, 1992.*

112. Kramer, M. S., and Shapiro, S. H., "Scientific Challenges in the Application of Randomized Trials," *Journal of the American Medical Association* 252:2739-2745,1984.

113. Kreuter, W., University of Washington, Seattle, WA, personal communication, July 1993.

114. Lave, J., Dobson, A., and Walton, C., "The Potential Use of Health Care Financing Ad-

ministration Data Sets for Health Care Services Research," *Health Care Financing Review* 5(1):93-98, *1983.*

115. Lawless, J. F., *Statistical Models and Methods for Lifetime Data (New* York, NY: John Wiley and Sons, 1982).

116. Lee, K. L., Pryor, D. B., Harrell, Jr., F.E., "Predicting Outcomes in Coronary Disease: Statistical Models versus Expert Clinicians," *American Journal of Medicine* 80:553-560, *1986.*

117. Lessler, J.T., Harris, B. S. H., *Medicaid Data as a Source for* ***Postmarketing*** *Surveillance Information: Final Report* (Chapel Hill, NC: Research Triangle Institute, 1984).

118. Lewis, C. E., "Variations in the Incidence of Surgery," New *England Journal of Medicine* 281:880-884, *1969.*

119. Localio, A. R., Lawthers, A.G., Bengtson, J. M., et al., "Relationship Between Malpractice Claims and Cesarean Delivery," *Journal of the American Medical Association* 269:366-373, *1993.*

120. Loft, A., Anderson, T.F., and Madsen, M., "A Quasi-Experiment Design Based on Regional Variations: Discussion of a Method for Evaluating Outcomes of Medical Practice," *Social Science and Medicine* 28: *147-154, 1989.*

121. Longo, D. L., deVita, Jr., V.T., and Young, R. C., "CHOP versus Intensive Regimens in Non-Hodgkin's Lymphoma," New ***England*** *Journal of Medicine* 329:580-581, *1993.*

122. Lu-Yao, G. L., McLerran, D., Wasson, J., et al., "An Assessment of Radical Prostatectomy: Time Trends, Geographic Variation and Outcomes," *Journal of the American Medical Association* 269:2633-2636,1993.

123, Malenka, D.J., Roos, N., Fisher, E. S., et al, "Further Study of the Increased Mortality Following Transurethral Prostatectomy: A Chart-Based Analysis," *Journal of Urology* 144:224-227, *1990.*

124. Mantel, N., "Cautions on the Use of Medical Databases," *Statistics in Medicine* 2:355-362, *1983.*

125. McDonald, C.J., and Hui, S. L., "The Analysis of Humongous Databases: Problems and Promises," *Statistics in Medicine* 10:51 1-518, 1991.

126. McBean, M. A., U.S. Department of Health and Human Services, Health Care Financing Administration, Office of Research and Demonstrations, Baltimore, MD, personal communication, September 1993.

127. McNeil, B.J., Harvard University, Boston, MA, personal communication, August 1993.

128. McNeil, B.J., "Claims Data and Effectiveness: Acute Myocardial Infarction and Other Examples," *Electiveness and Outcomes in* ***Health*** *Care* K.A. Heithoff and K.N. Lohr (eds.) (Washington, DC: National Academy Press, 1990).

129. McNeil, B.J., Pederson, S. H., Gatsonis, C., "Current Issues in Profiling Quality of Care," *Inquiry* 29:298-307, *1992.*

130. 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.

131. Meads, S., U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, MD, personal communication, June 1993.

132. Mebust, W. K., Holtgrewe, H. L., Cockett, A.T., et al., "Transurethral Prostatectomy: Immediate and Postoperative Complications. A Cooperative Study of 13 Participating Institutions Evaluating 3,885 Patients," *Journal of Urology* 141:243-247, *1989.*

133. Medical Research Council, "Streptomycin Treatment of Tuberculous Meningitis," ***Lancet*** *1:582-596,* 1948.

134. Melsey, K., Indiana University, Indianapolis, IN, personal communication, May 1993.

135. Miao, L. L., "Gastric Freezing: An Example of the Evaluation of Medical Therapy by Randomized Clinical Trials," *Costs, Risks and Benefits of Surgery,* J.P. Bunker, B.A.

Barnes, F. Mosteller (eds.) (New York, NY: Oxford University Press, 1977).

136. Monaghan, J., MediQual Systems, Westborough, MA, personal communications, July 1993.

137. Moses, L.E., "Framework for Considering the Role of Data Bases in Technology Assessment," *International Journal of Technology Assessment in Health Care* 6:183-193, *1990.*

138. Moses, L.E., "Innovative Methodologies for Research Using Databases," *Statistics in Medicine* 10:629-633, *1991.*

139. Munoz, E., Lausghlin, A., Regan, D. M., et al., "The Financial Effects of Emergency Department-Generated Admissions Under Prospective Payment Systems," *Journal of the American Medical Association 254:1763-1771, 1985.*

140. Nattinger, A. B., Gottlieb, M. S., Veum, J.,et al., "Geographic Variation in the Use of Breast-Conserving Treatment for Breast Cancer," New *England Journal of Medicine* 326(17):1 *102-1107, 1992.*

141. Paradise, J. L., Bluestone, C. D., Bachman, R.Z., et al, "Efficacy of Tonsillectomy for Recurrent Throat Infection in Severely Affected Children: Results of Parallel Randomized and Non-Randomized Clinical Trials," New *England Journal of Medicine* 310:674-683, *1984.*

142. Park, R. E., Fink, A., Brook, R. H., et al., "Physician Ratings of Appropriate Indications for Six Medical and Surgical Procedures," *American Journal of Public Health* 76:766-772, *1986.*

143. Pashos, C. L., Harvard University, Boston, MA, personal communication, August 1993.

144. Passamani, E., "Clinical Trials—Are They Ethical?" New *England Journal of Medicine* 324:1589-1592, *1991.*

145. Paul, J. E., Burroughs Wellcome, Research Triangle Park, NC, personal communications, August 1993.

146. Rahimtoola, S. H., "A Perspective on the Three Large Multicenter Randomized Clinical Trials of Coronary Bypass Surgery for Chronic Stable Angina," *Circulation 72* (suppl. V): V123-V135, 1985.

147. Ray, W.A., Griffin, M.R., "Use of Medicaid Data for Pharmacoepidemiology," *American Journal of Epidemiology* 129:837-849,1989.

148. Ray, W. A., Griffin, M. R., and Baugh, D. K., "Mortality Following Hip Fracture Before and After Implementation of the Prospective Payment *System, ''Archives of Internal Medicine* 150:2109-21 *14, 1990.*

149. Rimes, C., U.S. Department of Health and Human Services, Health Care Financing Administration, Office of National Health Statistics, Baltimore, MD, personal communication, September 1993.

150. Roberts, M. S., D'Agostino, R. B., Dillon, M., et al., "Technology Assessment in the Framingham Heart Study," *International Journal of Technology Assessment in Health Care 7: 156-170, 1991.*

151. Roethlisberger, F.J., and Dickson, W.J., *Management and the Worker* (Cambridge, MA: Harvard University Press, 1939).

152. Romano, P. S., and Luft, H. S., "Getting the Most of Out of the Messy Data: Problems and Approaches for Dealing with Large Administrative Data Bases," *Medical Effectiveness Research: Data Methods (Conference Proceedings, Summary Report,* M.L. Grady and H.A. Schwartz (eds.) (Rockville, MD: Agency for Health Care Policy and Research, 1992).

153. Roos, N. P., "Hysterectomy: Variations in Rates Across Small Areas and Across Physicians' Practice s," *American Journal of Public Health* 74:327-335,1984.

154. 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.*

155. Roos, N. P., and Lyttle, D., "Hip Arthroplasty Surgery in Manitoba: 1973 -1978," *Clinical Orthopedics and Related Research* 199:248-255, *1985.*

156. Roos, N. P., and Lyttle, D., "The Centralizations of Operations and Access to Treatment: Total Hip Replacement in Manitoba," *Amer-*

ican Journal of Public Health 75: *130-133, 1985.*

157. Roos, L. L., and Nicol, J. P., Johnson, C. F., et al., "Using Administrative Data Banks for Research and Evaluation: A Case Study," *Evaluation Quarterly* 3:236-255,1979.

158. Roos, N.P., and Roos, L. L., "High and Low Surgical Rates: Risk Factors for Area Residents," *American Journal of Public Health* 71:591-600, *1981.*

159. Roos, L. L., Wajda, A., and Nicol, J. P., "The Art and Science of Record Linkage: Methods That Work with Few Identifiers," *Computers in Biology and Medicine* 16(1):45-57, *1986.*

160. Roos, L. L., Wajda, A., Nicol, J. P., Roberts, J., "Record Linkage: An Overview," *Medical Effectiveness Research: Data Methods,* M.L. Grady and H.A. Schwartz (eds) (Rockville, MD: Agency for Health Care Policy and Research, 1992).

161. 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.

162. 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.

163. Rutkow, I. M., "Determining the Rates of Surgical Operations," *Socioeconomic of Surgery,* I.M. Rutkow (cd.) (St. Louis, MO: C.V. Moseby, 1989).

164. Sackett, D. L., "Bias in Analytic Research," *Journal of Chronic Disease* 32:51-63,1979.

165. Sacks, H., "Randomized versus Historical Controls for Clinical Trials," *American Journal of Medicine* 72:233-240, *1982.*

166. Safran, C., "Using Routinely Collected Data for Clinical Research," *Statistics in Medicine* 10:*559-564, 1991.*

167. SAS Institute Inc., *SAS User's Guide: Statistics: Version 5 Edition* (Cary, NC: SAS Institute Inc., 1985).

168. Savitz, D.A., and Grace, C., "Determinants of Medical Record Access for an Epidemiologic Study," *American Journal of Public Health* 75: *1425-1426,* 1985.

169. Sherman, C. R., Potosky, A. L., Weis, K. A., et al., "The Consensus Development Program. Detecting Changes in Medical Practice Following a Consensus Conference on the Treatment of Prostate Cancer," *International Journal of Technology Assessment in Health Care* 8:683-693,1992.

170. Shipp, M. A., Yeap, B. Y., Barrington, D. P., et al, "The m-BACOD Combination Chemotherapy Regimen in Large-Cell Lymphoma: Analysis of the Completed Trial and Comparison with the M-BACOD Regimen," *Journal of Clinical Oncology* 8:84-93,1990.

171. Simborg, D. W., "DRG Creep: A New Hospital-Acquired Disease," New *England Journal of Medicine 304: 1602-1604, 1981.*

172. Sloan, F. A., Morrisey, M. A., and Valvona, J., "Medicare Prospective Payment and the Use of Medical Technologies in Hospitals," *Medical Care* 26(9):837-853, *1988.*

173. Smits, H. L., and Watson, R. E., "DRGs and the Future of Surgical Practice," New *England Journal of Medicine 311:* 1612-1615, 1984.

174. Steinberg, E. P., John Hopkins Medical Institutions, Baltimore, MD, personal communication, May 1993.

175. Steinberg, E.P., Whittle, J., and Anderson, G. F., "Impact of Claims Data Research on Clinical Practice," *International Journal of Technology Assessment in Health Care* 6:282-287, *1990.*

176. Strom, B. L., Carson, J. L., Halpem, A. C., et al., "Using a Claims Database To Investigate Drug-Induced Stevens-Johnson Syndrome," *Statistics in Medicine* 10:565-576, *1991.*

177. Taylor, K. M., Margolese, R. G., and Soskolne, C. L., "Physicians' Reasons for Not Entering Eligible Patients in a Randomized Clinical Trial of Surgery for Breast Cancer," New *England Journal of Medicine 310:* 1363-*1367, 1984.*

178. Tierney, W. M., and McDonald, C.J., "Practice Databases and Their Uses in Clinical Research," *Statistics in Medicine* 10:541-557, *1991.*

179. Tunis, S. R., Bass, E. B., and Steinberg, E.P., "The Use of Angioplasty, Bypass Surgery, and Amputation in the Management of Peripheral Vascular Disease," New *England Journal of Medicine* 325:556-562,1991.

180. Udvarhelyi, 1. S., Gatsonis, C., Epstein, A. M., et al., "Acute Myocardial Infarction in the Medicare Population: Process of Care and Clinical Outcomes," *Journal of the American Medical Association 268:2530-2536, 1992.*

181. U.S. Department of Health and Human Services, ***ICD-9-CM:** The International Classification of Diseases, 9th Revision, Clinical Modification,* PHS Pub. No. 80-1260 (Washington, DC: U.S. Department of Health and Human Services, 1980).

182. U.S. Department of Health and Human Services, Health Care Financing Administration, Office of Statistics and Data Management, *Public Use Files Price List (July 1, 1993)* (Baltimore, MD: 1993).

183. U.S. Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research, *Conference Proceedings: Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data,* (Washington, DC: Agency for Health Care Policy and Research, 1990).

184. U.S. Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research, *Report to Congress: Progress of Research on Outcomes of Health Care Services and Procedures* (Rockville, MD: Public Health Service, 1991).

185. U.S. Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research: *Report to Congress: The Feasibility of Linking Research-Related Data Bases to Federal and Non-Federal Medical Administrative Data Bases,* AHCPR Pub. No. 91-0003 (Rockville, MD: Public Health Service, 1991).

186. U.S. Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research, *Medical Electiveness Research Data Methods,* (Rockville, MD: Agency for Health Care Policy and Research, 1992).

187. U.S. General Accounting Office, *Cross Design Synthesis: A New Strategy for Medical Effectiveness Research* (Washington, DC: U.S. Government Printing Office, 1992).

188. U.S. General Accounting Office, *Cross Design Synthesis: A New Strategy for Medical Effectiveness Research* (Washington, DC: U.S. Government Printing Office, 1992).

189. Wajda, A., Roos, L. L., "Simplifying Record Linkage: Software and Strategy," *Computers in Biology and Medicine* 17(4):239-248, *1987.*

190. Wennberg, J. E., "Physician Uncertainty, Specialty Ideology, and a Second Opinion Prior to Tonsillectomy," *Pediatrics* 59:952, *1977.*

191. Wennberg, J. E., "On Patient Need, Equity, Supplier-Induced Demand, and the Need To Assess the Outcome of Common Medical Practices," *Medical Care* 23:512-520, *1985.*

192. Wennberg, J. E., "Population Illness Rates Do Not Explain Population Hospitalization Rates: A Comment on Mark Blumberg's Thesis That Morbidity Adjusters Are Needed To Interpret Small Area Variations," *Medical Care* 25:354-359,1987.

193. Wennberg, J. E., "The Paradox of Appropriate Care," *Journal of the American Medical Association* 258:2568-2569, *1987.*

194. Wennberg, J.E., Barnes, B. A., and Zubkoff, M., "Professional Uncertainty and the Problem of Supplier-Induced Demand," *Social Science in Medicine* 16:81 *1-824, 1982.*

195. Wennberg, J. E., Blowers, L., Parker, R., et al., "Changes in Tonsillectomy Rates Associated With Feedback and Review," *Pediatrics* 59:821-826, *1977.*

196. Wennberg, J., and Gittlesohn, "A., "Small Area Variations in Health Care Delivery," *Science* 182:1 *102-1108, 1973.*

197. Wennberg, J. E., Gittlesohn, A., and Soule, D., "Health Care Delivery in Maine II: Conditions Explaining Hospital Admission," *Journal of the Maine Medical Association* 66(10):255-261, *269, 1975.*

198. Wennberg, J.E., Mulley, A.G. Jr., Hanley, D., et al., "An Assessment of Prostatectomy for Benign Urinary Tract Obstruction: Geographic Variations and the Evaluations of Medical Care Outcomes," *Journal of the American Medical Association 259:3027-3030, 1988.*

199. Wennberg, J. E., Roos, N., Sola, L., et al., "Use of Claims Data Systems To Evaluate Health Care Outcomes: Mortality and Reoperation Following Prostatectomy," *Journal of the American Medical Associations* 257:933-936, *1987.*

200. Wenneker, M.B., and Epstein, A. M., "Racial Inequalities in the Use of Procedures for Patients with Ischemic Heart Disease in Massachusetts," *Journal of the American Medical Association* 261:253-257,1989.

201. Wenneker, M. B., Weissman, J. S., and Epstein, A. M., "The Association of Payer with Utilization of Cardiac Procedures in Massachusetts," *Journal of the American Medical Association* 264:1255-1260, *1990.*

202. Whittle, J., Conigliaro, J., Good, C. B., et al., "Racial Differences in the Use of Invasive Cardiovascular Procedures in the Department of Veterans Affairs Medical System," New *England Journal of Medicine* 329:621-627, *1993.*

203. Whittle, J., Steinberg, E. P., Anderson, G. F., et al., "Use of Medicare Claims Data To Evaluate Outcomes in Elderly Patients Undergoing Lung Resection for Lung Cancer," *Chest* 100:729-734, 1991.

204. Whittle, J., Steinberg, E. P., Anderson, G. F., et al, "Mortality After Elective Total Hip Arthroplasty in Elderly Americans: Age, Gender, and Indication for Surgery Predict Survival," *Clinical Orthopedics and Related Research 295:* 119-126, 1993.

205. Winslow, C. M., Solomon, D. H., Chassin, M.R., et al., "The Appropriateness of Carotid Endarterectomy," New *England Journal of Medicine* 318:721-727, *1988.*

206. Wittes, R.E., "Paying for Patient Care in Treatment Research-Who Is Responsible?" *Cancer Treatment Reports* 71:107-1 *13, 1987.*

207. Wyse, D.G., Kellen, J., and Rademaker, A. W., "Prophylactic versus Selective Lidocaine for Early Ventricular Arrhythmias of Myocardial Infarction, ''Journal *of the American College of Cardiology* 12:507-513,1988.

208. Wyshak, G., Burdick, E., and Mosteller, F., "Technology Assessment in the Connecticut Tumor Registry," *International Journal of Technology Assessment in Health Care* 7:129-133, *1991.*

209. Yancik, R., Ries, L.G., and Yates, J. W., "Breast Cancer in Aging Women: A Population-Based Study of Contrasts in Stage, Surgery and Survival," *Cancer* 63:976-981, *1989.*

210. Young, J.L. Jr., Percy, C. L., and Asire, A.J., *Surveillance, Epidemiology, and End Results: Incidence and Mortality Data, 1973-1977* (Washington, DC, National Cancer Institute, 1981).

Large and Simple Randomized Trials

Background Paper 3

SUMMARY

For addressing many important research questions, randomized trials are neither necessary nor desirable. However, if the effects of a hypothesized intervention are likely to be only small to moderate in size, a randomized trial with a large sample size will be necessary to provide a definitive test of such research questions. Large trials, if properly designed, can be conducted using relatively simple protocols in which minimal screening or data collection is required.

Large and simple trials are characterized by their emphasis on enrolling large numbers of participants; testing an intervention's effect on a readily ascertained, clinically important outcome; and collecting a relatively limited amount of baseline and followup data. Such trials are particulary appropriate for addressing questions about the relative effectiveness of treatments with wide potential applicability. Because they enroll such a broad range of participants, their results are directly relevant to the wide range of patients seen in clinical practice. Because such trials often involve nonacademic as well as research-oriented clinicians and health care institutions, their results also may be more rapidly incorporated into standard care of patients.

Not all areas of medical research are suitable for large, simple trials. Nevertheless, many questions could be tested using far simpler protocols than those that have been used in most randomized clinical trials. Where appropriate, large and simple trials can provide more reliable tests of an intervention than can other feasible research approaches, and do so at very low cost per patient randomized.

by
Julie E. Buring,
Michael A. Jonas, and
Charles H. Hennekens
Division of Preventive
Medicine,
Department of Medicine
Brigham and Women's
Hospital and Department
of Ambulatory Care and
Prevention
Harvard Medical School
Boston, MA

Two types of epidemiologic studies can be used to test hypotheses: observational studies (case-control or cohort investigations) and randomized trials. Because of the methodologic limitations inherent in observational studies, randomized controlled trials (in which the investigators allocate the treatment to participants at random) represent the type of analytic study in humans that most closely resembles the highly controlled experiments possible in the laboratory (29).

Randomized controlled trials are particularly useful for detecting small to moderate effects of treatments or interventions-effects that are likely to change outcomes by 10 to 50 percent. In such circumstances, observational studies, which evaluate self-selected exposures and the subsequent occurrence of disease, are particularly vulnerable to the effects of unmeasured or unmeasurable confounding factors that may account for all or part of any observed association. For example, in an observational study that reports a 30-percent lower risk of cancer among individuals with high dietary intake of the antioxidant vitamin beta-carotene, the participants with greater intake of this micronutrient might have other dietary or lifestyle practices not fully accounted for in the study analysis that might be partially or entirely responsible for the observed benefit.

Even when an observational study reports a large effect, the amount of uncontrolled confounding may affect the magnitude of the estimated relative risk. Confounding factors, for example, could mean that the reported 15- to 20-fold higher risk of lung cancer among lifelong smokers than among nonsmokers could actually be as high as 25 or as low as 10. It is unlikely, however, that the confounding factors would change the conclusion that a strong relationship exists between smoking and lung cancer. In the case of current smoking and coronary heart disease, if the true effect of smoking is about an 80-percent increased risk of heart disease, uncontrolled confounding may mean that the observed effect is as small as 60 percent or as large as 100 percent. Again, however, this uncertainty does not materially affect the conclusion that current cigarette smoking increases the risk of coronary heart disease.

Thus, when the most plausible effects of an intervention or exposure are relatively large, they can be easily detected through observational studies.1 But when the most plausible effect size is between 10 and 50 percent, as is the case with many promising interventions, a small amount of uncontrolled confounding could mean the difference between a 20-percent decreased risk, no effect, or even a 20-percent increased risk. While such modest effects are difficult to detect reliably, they can have tremendous public health impact for a common or serious condition. Reliably detecting modest effects of a treatment, however, can only be done through randomized trials.

If such trials are sufficiently large, they eliminate the residual confounding that cannot be controlled in observational studies, by randomly allocating participants to the exposure of interest. For example, if a randomized trial is conducted to test whether beta-carotene reduces the risk of cancer, some participants would be assigned at random to take beta-carotene supplements, while others would serve as the comparison group by receiving no beta-carotene supplements. Such a strategy eliminates the self-selection of exposure that occurs in observational studies, and the impact of other variables that might be more prevalent among those who choose to eat diets high or low in beta-carotene. The unique strength of randomization is that, if the sample is large enough, the two study groups will usually be comparable with respect to all confounding variables, known and unknown, that might independently be related to risk

[1] For an exposure hypothesized to confer harm rather than benefit (e.g., cigarette smoking), randomized trials cannot be justified, because it would be unethical to assign study participants to such an exposure. In such cases, observational studies remain the only epidemiologic study design available, even when the likely effect of the exposure is modest.

of the disease. Randomized trials thus achieve a degree of control over bias and confounding that is not possible with any other epidemiologic design strategy.

Recognizing that small to moderate treatment effects can be reliably detected only with large samples, some researchers have focused on large and simple trials to answer important medical questions. The size of such trials, which generally involve several thousand participants, and the simplicity of their study protocol and streamlined collection of followup data distinguish these investigations from most randomized trials conducted to date.

PRINCIPLES OF LARGE AND SIMPLE TRIALS

The basic principles of clinical trial methodology must be considered in the design and analysis of any randomized trial, regardless of size. However, the design and conduct of large and simple trials rest on several additional principles and considerations (49):

- the need for *Large sample sizes* in order to reliably detect the most plausible small to moderate effects of particular treatments or to exclude with statistical certainty the possibility of such effects,
- the importance of testing widely *practicable treatments* that could have broad application if demonstrated to be effective,
- the use of *broad entry criteria* to determine eligibility for inclusion in trials,
- the use of *streamlined protocols,* and
- the use of a *clinically important outcome measure* to assess the effects of treatments.

Need for Large Samples Sizes

Through the random assignment of treatment, trials maximize the probability that both known and unknown confounding variables will be distributed equally among the treatment groups. Because this phenomenon works “on average,” equal distribution is more likely to occur if the trials are large. Moreover, large samples also enhance the statistical power of trials—i.e., the likelihood that a trial will detect an effect if one is truly present.

A fundamental aim of any randomized trial should be to assemble a sample size that is adequate to permit the researchers to definitively detect an effect if it exists, or to clearly demonstrate the lack of an effect if there isn’t one. Many randomized trials have failed to provide definitive tests of research hypotheses simply because they were too small to rule out the play of chance as a plausible alternative explanation for any findings that emerged. Such trials can actually do scientific harm if their results are interpreted as providing clear evidence of no effects when the trials simply had inadequate statistical power to answer the research questions with certainty. Null findings have emerged from a number of small trials testing treatments that were later shown unequivocally in investigations with adequate samples to confer clear net benefits.

Two examples of the importance of large samples to definitively evaluate a hypothesis involve the testing of promising treatments for acute heart attacks, or myocardial infarction (MI). The International Studies of Infarct Survival (ISIS) is a set of studies on treatment for MI, conducted through a worldwide collaboration of hospitals, that began in the early 1980s. The first ISIS trial was designed to test the effects of the beta-blocker drug atenolol. More than 16,000 patients in the acute phase of a suspected heart attack were enrolled into ISIS- 1 and assigned at random to receive atenolol (5 to 10 mg intravenously and then 100 mg per day orally for 7 days) or to serve as controls (32). Another trial of beta-blocker therapy, the Metoprolol in Acute Myocardial Infarction (MIAMI) trial, enrolled approximately 6,000 patients to test this treatment (39).

When the two trials were completed, the estimates of the effects of treatment were very similar, with the study participants who received beta-blocker therapy experiencing reductions in vascular mortality of approximately 13 percent in the MIAMI trial and 15 percent in ISIS-1. Though the estimates of effect in the two trials were virtually identical, the ISIS- 1 result achieved statistical significance, whereas the MIAMI result did not. This difference in the strength of the conclusions that could be drawn from the two trials resulted almost wholly from their respective sample sizes.

Another promising area of research in the treatment of acute MI in the early 1980s was the use of thrombolytic drugs, agents given during the acute phase of a heart attack to dissolve the clots in the coronary artery that had precipitated the attack. By restoring blood flow to areas of the heart muscle that have been starved of oxygen-rich blood by the blockage, these drugs can spare the heart from permanent damage.

By the mid- 1980s, 24 separate trials had tested the hypothesis that the use of an intravenous thrombolytic agent (primarily streptokinase) would decrease the risk of mortality in patients with acute MI. Of these trials, five reported a statistically significant benefit on mortality from use of a thrombolytic drug, 11 suggested a benefit but were not statistically significant, and eight reported a harmful trend but were also not statistically significant (50). The discrepancies in the trials' findings most likely derived from the fact that the effect of such agents was anticipated to be modest (on the order of a 10- to 30-percent decrease in mortality), and the majority of the individual trials were simply too small to detect such a benefit accurately (none enrolled more than 750 patients).

The uncertainties left by these trials led directly to ISIS-2, in which more than 17,000 patients were randomized to the thrombolytic drug streptokinase or placebo as well as to a month-long regimen of daily low-dose aspirin or placebo (33). With respect to vascular mortality, patients who received streptokinase experienced a statistically significant 25-percent reduction in risk, those receiving aspirin experienced a statistically significant 23-percent decrease, and those who received both treatments experienced a significant 42-percent decrease in vascular death. Thus, this large and simple trial was able to detect definitively the modest but clinically meaningful benefits of thrombolytic therapy in the treatment of acute MI.

The reason that larger trials are better able tore-liably detect modest treatment effects derives not just from the numbers of randomized participants but rather from the number of events they experience. For example, whereas a trial of aspirin in the primary prevention of heart disease might require a sample of 22,000 men over the age of 40 in order to detect a 20-percent reduction in risk, a sample of 40,000 women over the age of 45 would be required to detect the same effect, because women have a lower baseline rate of heart disease than men do. Thus, trials must be large enough to accrue sufficient numbers of outcome events to demonstrate either definitive positive results or truly informative null findings.

The identification of effective treatments for a condition also affects the sample size requirements of future investigations. As the efficacy of thrombolysis and aspirin has been demonstrated in large trials, these therapies have become more common components of the routine management of MI patients (35,40). Any new therapies, then, must be shown to confer additional benefits beyond those of an expanding regimen of effective standard treatments. As a result, the absolute magnitudes of any further benefits are likely to be progressively smaller. Such benefits may be very worthwhile, since MI is a common and serious condition, but detecting them will become increasingly difficult and will require trials with even larger samples.

A second circumstance that affects a trial's sample size requirements is the need to compare directly two or more treatments to determine whether one has clear advantages. It was just such a question that led to ISIS-3 (34). Randomized trials had suggested that, in addition to streptokinase, two other thrombolytic agents—tPA (tissue plasminogen activator) and APSAC (anisoylated plasminogen-streptokinase activator complex)—were effective in dissolving clots in acute MI and reducing subsequent mortality. Although thrombolytic therapy was clearly a valuable treatment, it was unclear whether there were any important differences in the benefits and risks of the three principal thrombolytic drugs, so a head-to-head comparison of the agents was carried out.

All patients in ISIS-3 received thrombolytic drugs, with one-third of the study participants randomly assigned to each agent. To detect meaningful differences among the treatments, all of which were expected to confer roughly comparable benefits, 1S1S-3 randomized more than 41,000 patients. The trial provided statistically conclusive evidence that there were no significant differences between the three thrombolytic drugs in reducing mortality following acute MI. Moreover, in terms of the most serious adverse effects associated with thrombolytic drugs, tPA and APSAC were shown in ISIS-3 to be associated with significantly more cerebral hemorrhages than streptokinase. The three drugs differ substantially in cost, which ranges from roughly $300 per dose for streptokinase to approximately $1,700 for APSAC and $2,200 for tPA.

A subsequently reported trial, GUSTO (Global Utilization of Streptokinase and Tissue Plasminogen Activator for Occluded Coronary Arteries), which included comparisons of streptokinase and tPA, suggested that a newer method of administering tPA very rapidly conferred a slight advantage in reducing mortality over streptokinase (26). Again, however, tPA was associated with a higher rate of cerebral hemorrhage. Considerable controversy has surrounded specific issues in the interpretation of the GUSTO findings (41).

An issue raised by the findings from GUSTO and other trials of thrombolytic agents is the need to distinguish between differences in treatment effects that are statistically significant, based on comparisons of tens of thousands of patients, and those that are clinically meaningful in treating patients. In the case of streptokinase and tPA, currently available evidence from large-scale trials suggests that emphasizing the differences in thrombolytic agents' efficacy and safety is far less important than encouraging their wider use, since all of them confer clear benefits in a large proportion of acute MI patients (41).

I Testing Widely Practicable Treatments

The need to test widely practicable treatments is another principle of large and simple trials. From a public policy standpoint, a treatment is likely to have a greater effect on public health if it can be readily administered at most community hospitals than if it is very complicated or expensive (or requires specialized training or resources available only at tertiary care facilities), even if the two treatments confer the same degree of benefit.

For example, three recent small randomized trials of treatments for acute MI patients compared the effects of a clot-dissolving thrombolytic agent with those of coronary angioplasty, a procedure in which a balloon-tipped catheter is guided into the blocked coronary artery and briefly inflated to reopen the occluded vessel (22,23,53). In two of the three trials, patients receiving angioplasty experienced lower rates of mortal it y or recurrent MI than did those receiving thrombolytic therapy (23,53). The third trial found no clear evidence of a difference in the effects of the two treatment strategies (22).

These results suggest that the two approaches may be equally effective, or perhaps even that angioplasty has a short-term advantage. Of far more significance from a public health perspective, however, is the fact that only 18 percent of U.S. hospitals are capable of performing angioplasty, with even fewer equipped to conduct emergency coronary bypass surgery (which is necessary in the small number of cases where a vessel abruptly closes following angioplasty). Many acute MI patients in the United States probably live reasonably near hospitals equipped to perform angioplasties as well as emergency coronary bypass surgery, but the widespread use of angioplasty instead of thrombolytic therapy would greatly increase the demands on such facilities and would have tremendous implications for the level of coronary care services required in U.S. hospitals. Consequently, the editorial accompanying the three trial reports concluded that "the strategy of immediate angioplasty for acute myocardial infarction has limited applicability because of the severely restricted accessibility of the procedure" (37). At present, therefore, thrombolytic therapy, which can be administered at most emergency care facilities—and in prehospital settings in some areas-can have a far greater overall public health impact on mortality following acute heart attack.

I Use of Broad Entry Criteria

Many randomized trials have studied relatively homogeneous, narrowly defined groups of patients, thereby meeting the scientific urge for precision in knowing exactly which types of patients will benefit from particular interventions. By contrast, large, simple trials generally have used very broad and flexible entry criteria. On a practical level, the use of broad entry criteria aids the recruitment of large numbers of patients and minimizes costs by eliminating the need for elaborate screening procedures. In addition, however, there is a compelling scientific rationale for such a practice (52).

The goal of randomized trials is to provide reliable evidence of treatment effects that can be used to improve clinical practice. Large, simple trials have used very wide entry criteria so that the heterogeneous population under study will more closely mirror the broad population of patients to whom the results can be generalized. A basic premise underlying the use of broad eligibility criteria is that the direction, though not necessarily the magnitude, of the net effect of a treatment is likely to be similar for many subcategories of patients. In other words, the magnitude of any benefit or harm may well differ according to certain patient characteristics, but such quantitative differences in the size of the effect are much more likely than unanticipated qualitative differences, in which one group of participants benefits from a treatment while another either does not benefit or is harmed.

The use of narrow eligibility criteria can unnecessarily limit the generalizability of findings. For example, from the results of animal experiments, researchers thought that thrombolytic therapy would be ineffective or perhaps even harmful if initiated more than six hours after the onset of symptoms. Some early trials of thrombolytic therapy, therefore, restricted participation to patients with symptoms of less than six hours' duration. The rationale for this limitation was that little benefit would accrue to patients whose symptoms were of longer duration but that the drugs' known risks (e.g., cerebral hemorrhage) would still exist. However, even if the results of such trials suggested a benefit of thrombolysis, they could not answer whether the treatment might also benefit patients who arrived at hospitals more than six hours after their symptoms began.

ISIS-2 adopted much wider eligibility criteria, enrolling patients up to 24 hours after the onset of MI symptoms. Large-scale trials can, and indeed should, collect data on key variables that may define clinically important subcategories of patients in whom treatment effects may substantially differ. Therefore, the time that had elapsed since the onset of symptoms was one of the select variables in ISIS-2 for which information was gathered at baseline. The collection of such data allowed for the analysis of trial results according to duration of symptoms prior to treatment, an analysis that demonstrated that the benefit of streptokinase, although greatest for patients treated early, extends to those treated up to 24 hours after the onset of symptoms. Overall, a 25-percent reduction in cardiovascular death was associated with streptokinase treatment given within 24 hours of the onset of symptoms. The reduction was 35 percent for those treated within four hours and 17 percent for those treated within five to 24 hours.

Although the ISIS-2 results demonstrated the advantages of wide eligibility criteria, precise definition of the eligibility criteria for a trial is a matter of scientific judgment, based on the particular question being asked. Randomizing patients up to one week following the onset of MI symptoms, for example, makes little biological sense, in view of the known properties of thrombolytic drugs and the pathophysiology of MI over such a period. Not only would such broad eligibility criteria unnecessarily expose late-treated patients to the possible risks of thrombolysis, but they would also dilute any benefit of treatment to such an extent that the overall finding from the trial might be null, even if analyses restricted to early treated patients suggested a clear benefit. In fact, while some early trials of thrombolysis used unduly restrictive entry criteria, others cast too wide a net, randomizing patients up to 72 hours after initial symptoms (50). Thus, reasonable judgments must be made not only in identifying the population at risk for the outcome under study, but in defining the group of individuals in whom an effect of the intervention is biologically plausible.

One criticism of the use of broad entry criteria is that even though the study's overall results may apply to a wide population of patients with a particular disease, they do not offer much guidance about how to treat individual patients with specific medical profiles. This tension—between the broadly relevant data available from large, simple trials and the highly detailed information upon which practicing clinicians might ideally wish to base individual treatment recommendations—may never be fully resolved. However, several factors support the use of wide, rather than narrow, entry criteria in many trials evaluating promising medical interventions. First is the belief that unless there is a clear reason to believe otherwise, a beneficial treatment is likely to be effective across a broad spectrum of patients. Results from trials using broad entry criteria, therefore, are directly relevant to the wide spectrum of patients to whom the results will be generalized in actual clinical medicine. Second, if the effect of an intervention differs among categories of patients, a large-scale trial enrolling a broad range of patients might be the only way to detect the differences. Even in large trials, however, the statistical power to detect treatment effects among subcategories of patients may be inadequate. Further, if many subcategories are analyzed, it becomes increasingly likely that an erroneous finding will emerge simply from the play of chance. Therefore, for research questions that require the enrollment of large numbers of participants, the main finding will be one that answers whether, on average, the study intervention confers a net benefit compared with no treatment (or the alternative treatment). More precise evidence may emerge from analyses of select subcategories, but applying trial results to medical practice will always involve making individual clinical judgments based on each patient's medical profile. A decade ago, a paper describing the principles of large, simple trials addressed these issues succinctly:

> Trials are at least a practical way of making some solid progress, and it would be unfortunate if desire for the perfect (i.e., knowledge of exactly who will benefit from treatment) were to become the enemy of the possible (i.e., knowledge of the direction and approximate size of the effects of the treatment of wide categories of patient) (49).

I Use of Streamlined Protocols

The use of streamlined study protocols has very practical advantages in the design of a large-scale trial. If a trial requires many thousands of patients in order to answer a question reliably, the trial organizers usually must reach beyond the confines of the academic medical centers (where most research is conducted) to involve general-care community hospitals or even medical settings in a number of countries. This can be accomplished only if treatments can be administered in a wide range of settings, as is the case for thrombolytic therapy. Furthermore, to secure the cooperation of busy physicians and nurses (whose primary mission is to care for their patients, not to conduct research), trial treatments must be relatively simple to administer, and the added burdens of participation must be minimized whenever possible by using streamlined screening procedures and collecting only the most important followup data needed for assessing the efficacy and side effects of the treatment.

The cost of research is also an important factor in the move toward simple trial protocols. Particularly during an era of shrinking research budgets and increased competition for funding, efficient study designs are imperative if large trials are to be funded to any significant extent. For example, the Beta-Blocker Heart Attack Trial (BHAT), which began in 1977, randomized 3,837 patients with prior heart attacks in order to test whether the beta-blocker drug propranolol hydrochloride reduced total mortality, at a total cost of $20 million (3). In contrast, a trial testing the drug digitalis among patients with congestive heart failure, which was begun in 1991 and is employing a streamlined trial protocol, randomized 7,790 patients and will have a total budget of $16 million (21). After adjustment for inflation, the earlier BHAT investigation cost approximately $11,350 per participant, while the ongoing digitalis trial will incur costs of approximately $2,050 per participant.

Similar efficiencies are possible in studies of preventive interventions in apparently healthy participants. Most such investigations have collected extensive baseline and followup data and required regular clinic visits, with costs generally ranging from $3,000 to $15,000 per randomized participant for a five-year trial. In contrast, the Physicians' Health Study, a trial testing aspirin and beta-carotene in the prevention of cardiovascular disease and cancer, has been conducted entirely by mail among 22,071 U.S. male physicians at a cost of approximately $80 per participant per year (4).

Practical considerations underscore the need for simple trial protocols, but in addition, the most widely practicable treatments are often those that are simple. And for interventions where the outcome of interest is a straightforward, easily ascertainable event such as mortality, most of the crucial information needed for future clinical decisionmaking and public health policy is available from the streamlined data collected in large, simple trials.

In ISIS-2, for example, virtually every patient entering a participating hospital within 24 hours of the onset of symptoms of suspected MI was considered eligible to participate. If there were no clear indication for or against the trial treatments, the patient was eligible to be randomized. If informed consent were obtained, a 24-hour toll-free randomization telephone line was dialed, and the physician or nurse collaborator provided basic identifying data on the patient as well as information about a very few select medical variables, such as time since the onset of symptoms. A randomization code was then obtained and matched against one of the treatment packs stored in the hospital, and the contents of the pack administered to the patient. At the time of the patient's hospital discharge, the clinician completed a simple one-page followup form, providing information on vital status (i.e., whether the patient was alive or dead) as well as major in-hospital events, such as reinfarction, stroke, or significant bleeding episodes. The clinician then sent this form, along with the results from a pre-randomization electrocardiogram, to the international coordinating center in England. At that point, the clinician responsibilities to the trial were over.

An important assumption underlying the use of a simple protocol with streamlined followup is that the areas of chief concern regarding adverse effects of the intervention have been reliably identified. Although the balance between the benefits and the risks of a treatment is unknown-and, indeed, is the principal question being asked in most large trials—preliminary testing or knowledge of biological mechanisms should have allowed the researchers to identify the most serious potential side effects so that the collection of followup data could be confined to a few key variables. Trials of agents or procedures for which there is little prior knowledge concerning safety may require much more detailed data collection and thus will more closely resemble traditional randomized controlled trials.

I Use of Clinically Important Outcome Measures

Small and more complex trials may be important early in the development of a treatment. Such investigations may collect data on scores of variables to assess their response to treatment. This may, in turn, provide important information about the action of the drug, its side effects, or features of the disease itself. When an intervention is sufficiently promising to warrant testing for efficacy in a large-scale trial, however, the fundamental goal is to obtain information that can inform clinical practice and public health policy. For this reason, the primary outcome in a large, simple trial should be a clinically meaningful event, not an intermediate marker whose clinical significance is unknown. In most trials of serious diseases, the fundamental question is whether a treatment increases patients' chances of survival. Major morbidity events, such as nonfatal heart attacks, may also be suitable endpoints in some trials, but the use of subclinical or intermediate markers as surrogates for clinical endpoints can lead to spurious conclusions.

Reliance on an intermediate endpoint in studies of the effect of thrombolytic drugs in the treatment of acute MI, for example, may have led to erroneous conclusions about the relative benefit of different agents. Many physicians believed that the thrombolytic drug tPA was superior to streptokinase because it appeared to be faster at dissolving the clots in the coronary artery that precipitated the attack. This conclusion was based on angiographic studies demonstrating that 90 minutes after treatment, blood flow was restored through the previously occluded artery in 70 percent of patients receiving tPA compared with 50 to 55 percent of patients receiving streptokinase (46). However, further studies indicated that coronary patency rates for tPA and streptokinase become equal over the next several hours. Moreover, for the primary clinical endpoint of mortality, the results of large-scale trials demonstrated identical 35-day vascular mortality rates for patients given tPA and those given streptokinase (25,34).

In addition to making a clinically important outcome the primary focus, a large and simple trial must also have a main outcome event that can be fairly readily ascertained without extensive, specialized testing or frequent in-person followup visits. In this regard, mortality is the most straightforward outcome event, inasmuch as its occurrence is not subject to dispute and can even be tracked by searching death certificate databases or using other indirect methods of followup. Nonfatal medical events may also be suitable endpoints for large, simple trials. For example, most nonfatal heart attacks or cancer diagnoses can be verified using existing medical record information that would be available regardless of whether an individual was part of a trial protocol.

OTHER APPLICATIONS OF LARGE, SIMPLE TRIALS

Acute MI has been the clinical context in which the principles of large, simple trials have been most widely applied to date, as discussed above. Because it is an easily defined, common, and serious clinical event—and one for which the fundamental measure of a treatment's efficacy can be made over a relatively short time frame—acute MI is particularly well-suited to this research approach. In addition, however, trials employing these principles have been conducted and proposed for a wide range of treatments and health conditions, including longer-term trials of chronic heart disease, the management of women with high-risk pregnancies, treatments for patients with human immunodeficiency virus (HIV) infection or acquired immune deficiency syndrome (AIDS), and the surgical treatment of cancer, as well as the testing of promising interventions in the primary prevention of cancer and heart disease among apparently healthy participants.

Polio Vaccine Field Trial: An Early Example

Perhaps the first large and simple randomized trial was carried out 40 years ago, when the National Foundation for Infantile Paralysis recruited a team of physicians and public health researchers to mount a massive randomized trial to test the efficacy of the Salk polio vaccine (38). More than 400,000 U.S. school children took part in this experiment in the spring and summer of 1954. The polio vaccine trial randomly assigned half of the participants to receive the vaccine, while half received a placebo injection. The incidence of disease in the two groups was then tracked by simply monitoring the hospitalizations for polio in the areas where the field trial was carried out. Over the course of several months, the effectiveness of the vaccine in preventing this serious, disabling childhood disease became clear (20).

In many respects, the large and simple design of the massive polio trial was a response to the urgency of the problem, in which there was tremendous pressure to provide a quick and reliable test of the newly developed vaccine in a single polio season. While the polio trial may have been the first example of a large and simple randomized trial, it was not until several decades later—in the late 1970s and early 1980s—that the principles of this approach to answering health questions were more formally described and its methods more widely used to evaluate clinical questions (38).

Digitalis in the Treatment of Congestive Heart Failure

Researchers have recently begun testing the drug digitalis in the treatment of congestive heart failure in a long-term trial that has incorporated many features of large, simple trials (12). Although overall rates of cardiovascular disease have declined significantly in the United States, over the past two decades the incidence and prevalence of congestive heart failure (CHF) have increased significantly, a pattern that is expected to continue as the population ages. CHF, a cardiac syndrome characterized by a weakening of the contractions of the heart muscle, is estimated to be a primary or contributing cause of 250,000 deaths in the United States each year.

Digitalis preparations, which have been available for more than 200 years, are one of the most commonly prescribed treatments for CHF. In 1986, more than 12 million prescriptions for this drug were written in the United States (12). Despite its widespread use, the net effect of this drug on mortality in patients with CHF remains uncertain. Although a number of small trials of digitalis have been conducted, the results of the trials are inconsistent (51). Digitalis may improve the output of blood by the heart (ejection fraction) and thereby slow the progression of CHF and decrease mortality, but the drug has other biochemical properties that, in theory, may increase the risk of dangerous changes in cardiac rhythm. In view of the continued uncertainty regarding the net effect of digitalis, a large trial of the drug was initiated in 1991 under the direction of the National Heart, Lung and Blood Institute (NHLBI).

Because any large benefit of digitalis would probably have been clear from the smaller studies already conducted to date, its true benefit-if any-in reducing mortality is likely to be on the order of 10 to 15 percent. However, as with thrombolysis or aspirin therapy of acute MI, even a modest mortality benefit for such a common condition could be of great public health value. For researchers to detect a benefit of digitalis in this range, about 2,000 deaths would need to occur in a trial population. To enable researchers to observe this number of events over a relatively short period, the trial has enrolled nearly 8,000 patients at more than 300 hospitals throughout the United States and Canada.

The digitalis trial will involve treatment and followup of patients for three years. The principal entry criterion for the trial will be moderate or severe CHF (ejection fraction< 0.45). All patients in the trial must have had a chest x-ray within the past six months and cardiac ejection fraction documented by either angiography or echocardiogram. Patients will be randomized via telephone calls to a central coordinating center, with key baseline data given directly by phone for entry in the study's database. Each randomized patient must return for a followup visit in four weeks, and every four months thereafter. Because digitalis has a relatively narrow therapeutic window, with high toxicity at elevated doses, blood will be drawn during followup visits to monitor the serum levels of digitalis as well as those of potassium, creatinine and magnesium. In addition, since the appropriate dose of digitalis depends on patient characteristics such as age, weight, and sex, four different dose regimens will be used.

Despite these considerations, which add complexity to the trial protocol, the digitalis trial retains two chief characteristics of large, simple trials: the collection of followup data is limited (a one-page questionnaire at each visit), and the clinic visits and many of the monitoring tests of dose level, as well as the required radiologic studies for eligibility, would be carried out anyway as part of the standard clinical management of CHF patients. The principal outcome measurement in the study will be mortality, and the trial should provide clear evidence of digitalis's net effect on mortality in patients with CHF.

Aspirin in Treatment of High-Risk Pregnant Women

Pre-eclampsia, a condition caused by high blood pressure, is a common and serious complication of the second half of pregnancy. It can lead to intrauterine growth retardation and fetal death as well as to complications of prematurity, because early delivery of the baby is the only effective approach to the condition.

Several trials of aspirin have suggested that treatment in high-risk pregnant women is beneficial, but the small samples of patients in most of the trials have left a great deal of uncertainty concerning the treatment's effects. To address this problem, the Collaborative Low-Dose Aspirin Study in Pregnancy (CLASP) randomized 9,364 women in 16 countries to either 60 mg of aspirin or a placebo daily (8). According to the trial's broad entry criteria, women were eligible if they were between their 12th and 32nd weeks of pregnancy and were judged by their treating clinicians to be at sufficient risk of pre-eclampsia to consider aspirin treatment. Randomization was carried out by having clinic staff telephone a 24-hour randomization service. For each patient, data were collected on several key variables at entry, and a single-page followup form was completed following hospital discharge at the end of the pregnancy, recording information on treatment compliance, use of other drugs, and major clinical events occurring after randomization.

Overall, those assigned to receive aspirin experienced a 12-percent reduction in the development of pre-eclampsia, but this difference was not statistically significant. Aspirin-allocated women did experience a modest but significant lower rate of delivery before 37 weeks estimated gestation. However, there were no significant differences between treatment groups in the proportion of stillbirths, neonatal deaths, or babies with intrauterine growth retardation. Because of the possibility that the benefits of aspirin might be restricted to certain subgroups of women, the CLASP protocol called for separate analyses of data based on several entry characteristics. There were no subgroups in whom the reduction in pre-eclampsia was as large as that reported in the earlier small trials. The authors concluded that currently available data do not support the widespread use of aspirin in women at high risk for pre-eclampsia. Nonetheless, among women with preterm deliveries, there was a significant trend toward greater reductions in the development of pre-eclampsia in the group that received aspirin. The authors suggest that aspirin may have effects in women who are susceptible to early pre-eclarnpsia that it does not have among women who develop this condition in the late stages of gestation. Aspirin may, therefore, be justified in those at particularly high risk of early-onset (before 32 weeks) pre-eclampsia, but inasmuch as these women are difficult to identify prospectively, the clinical implications of the CLASP findings may be restricted to high-risk women with prior histories of early-onset pre-eclampsia.

Treatments for Patients with HIV or AIDS

Several investigators have suggested that large, simple trials could be used for the efficient testing of potential treatments for those infected with HIV or with diagnosed AIDS (6,7,13,14,42). Very detailed studies in specialized centers are clearly crucial to gain more knowledge about this disease. Indeed, it is from the intensive study of patients and potential treatments that promising hypotheses will emerge. To reliably answer the broader question of a treatment's net clinical effect, however, will require collaborative trials using the principles of large, simple trials, because most of the promising therapies are likely to have only small to moderate effects. In addition, as with the treatment of acute MI, trials will need to be designed to detect the equal or superior efficacy of new treatments in relation to an expanding array of standard therapies.

Both the National Institutes of Health and the American Foundation for AIDS Research have established networks of community physicians for research studies. Such consortia could form the organizational basis for the implementation of large, simple trial protocols (13, 14). Another avenue that has been suggested for the development of large trials is the enrollment of patients now receiving treatments as part of the system known as the treatment IND (investigational new drug) or *parallel track.* This expanded-access program was approved by the Food and Drug Administration to make treatments still undergoing experimental evaluation available to a broad population of patients who have life-threatening diseases and who are no longer able to tolerate or benefit from the standard available treatments. Although these programs are providing many patients with experimental AIDS drugs and uncontrolled observational followup, direct comparison of such treatments could be carried out as part of simple, randomized treatment protocols, without undue requirements for additional work by busy clinicians, but with systematic coordination by a data center.

Because large-scale community-based trials would collect uniform data on only a small number of important variables, more extensive data could be gathered at selected participating sites, such as academic research centers. This strategy, which has been used in other large trials, may be particularly appropriate for AIDS treatment, where the rapid development of new experimental therapies means that there is frequently much less long-term experience with a drug's toxicity orother effects than is often the case with agents being tested in large-scale trials. Such trials might, therefore, more appropriately be considered hybrid trials, with a large component that uses a simple trial protocol and a small subgroup for whom a more detailed randomized clinical trial protocol is implemented.

Since a major goal of AIDS treatments is to prolong survival, large numbers of patients must be enrolled in trials if any net benefit of these treatments on mortality is to become known relatively quickly. New antiretroviral agents, for example, could be compared with current standard therapies in large-scale trials to assess their survival benefits (42). Other important questions that could be answered through large trials include determining the optimal doses of available treatments (13). Many currently available AIDS treatments have significant toxicity. Randomized trials comparing different doses of a particular drug could determine whether lower, less toxic doses confer a similar survival benefit. Studies of zidovudine (AZT), for example, have already shown that daily doses of 600 mg are as effective at prolonging survival as 1200-mg doses. Even lower doses might work equally well, and such a finding could significantly improve the quality of life for many AIDS patients (13).

At present, AIDS treatments differ qualitatively from those used in other conditions. In the case of acute MI, where treatment with thrombolysis or aspirin saves several lives for every 100 patients treated, these individuals are, in some sense, considered "cured" because they avoided death during the high-risk period immediately following their attack. Such patients remain at higher risk for cardiovascular death, but they could live for decades and then die from nonvascular causes. There is no comparable life-saving effect of current AIDS treatments, which confer only short-term survival benefits.

This suggests some justification for using other clinical endpoints besides mortality, such as quality-of-life measures or the development of opportunistic infections, to determine the benefits of some agents. It is crucial to keep in mind, however, that an observed improvement in such outcomes may not translate into longer patient survival. The demonstrated benefits of any treatments approved on such a basis must be clearly identified to avoid overstating their known effects.

Biological markers, such as CD4 cell counts, have also been proposed for use as endpoints in AIDS trials. Because drugs may affect these surrogate endpoints much sooner than clinical outcomes (e.g., opportunistic infections or death), the use of such endpoint markers can reduce the needed size and duration of a trial. This approach is clearly attractive in the face of a fatal epidemic. Unfortunately, none of the biological markers currently measured in AIDS patients has been shown to predict clinical course or survival reliably enough for use as a firm endpoint (7).

After the initial demonstration that AZT confers a short-term reduction in the mortality of symptomatic patients (19), a trial was conducted to test whether early treatment would delay the onset of AIDS in asymptomatic individuals infected with HIV. A clear delay in disease onset was observed and the trial was stopped prematurely (47). However, it was not at all clear whether the early use of AZT in asymptomatic patients would extend their survival beyond what would be achieved by initiating therapy at the onset of the disease. Moreover, early use of an antiretroviral agent may render the drug less effective later, during the actual disease phase, thereby potentially shortening the survival time after the development of full-blown AIDS (42). Because of the debilitating and fatal nature of the disease, this may be an acceptable choice to patients, who face limited life expectancies regardless of which treatment course they pursue. Information on this question should be available, however, so that patients can make informed choices.

To provide further data on the relative merits of immediate versus delayed treatment with AZT, a randomized controlled trial in Europe compared how the two treatment approaches affect mortality (1,9). The Concorde trial was a multicentered trial carried out in England, Ireland, and France among 1,749 HIV-infected individuals who were symptom-free at baseline. Half of the participants were randomized to begin immediate treatment with AZT; the others were randomized to deferred treatment, which entailed taking inert placebo pills that resembled AZT. Once patients exhibited symptoms of AIDS or AIDS-related complex (ARC), or had persistently low CD4 cell counts that led their physicians to believe treatment was indicated, their assignments were unblinded and those who were receiving placebos began AZT therapy.

Throughout the trial, the patients randomized to immediate AZT treatment had significantly higher CD4 cell counts. It has been postulated that higher levels of these disease-fighting cells indicate the efficacy of an AIDS treatment, and that decreases in CD4 cell counts signal the progression of HIV disease. Despite the favorable effect of immediate treatment on CD4 cell counts in asymptomatic patients, the three-year survival rates in the two treatment groups were virtually identical (92 percent in the immediate treatment group vs. 94 percent in the deferred therapy group). Even more surprising, the Concorde results indicated that early treatment of HIV did not appear to slow the rate of progression of asymptomatic HIV disease to ARC, AIDS, or death—a finding in marked contrast to previous studies, which had indicated a benefit of early treatment of asymptomatic patients. However, these tria~s were stopped much earlier than the Concorde trial. Short-term followup data from the Concorde trial were also compatible with the finding of a benefit from early treatment, but the apparent advantage of immediate AZT therapy disappeared with longer-term treatment and followup.

Although the Concorde findings appear to rule out any large benefit from early treatment with AZT in asymptomatic individuals, the trial was not large enough to rule out the possibility of a small advantage of such treatment.

The Concorde results raise important questions about the ultimate public health benefit of the rapid approval of AIDS drugs in the United States. The highly organized activities of individuals with HIV mark an unprecedented degree of direct involvement by affected patients in the quest for advances in treatment of their condition. This activism has led to many positive changes in what some have regarded as an often cumbersome and unduly bureaucratic drug approval process. At the same time, however, the pressure to speed drug approval may also lead to rapid decisions made without full benefit of the optimal quality or quantity of randomized trial data.

The guiding principle of broad entry criteria in large, simple trials has particular relevance to the study of AIDS treatments. Many AIDS patients are interested in participating in treatment protocols, but have been excluded because of stringent entry criteria. Because most of the treatments that are found to be effective will be made available to most patients, it is reasonable—and indeed desirable—to include a broad range of patients in trials (6).

As in the large trials of patients with MI or CHF, several key baseline variables should be collected to allow for the assessment of any differing effects of treatments among subgroups.

Breast Cancer Treatments

Some of the most significant advances in the treatment of breast cancer have resulted from the collaboration of a large number of hospitals in the National Surgical Adjuvant Breast Project (NSABP), which has coordinated multicentered randomized trials comparing different treatment approaches to breast cancer (17, 18). For many decades, the standard treatment was radical mastectomy, which entails removal of the breast, axillary lymph nodes, and pectoral muscles. Anecdotal evidence suggested that less disfiguring approaches might be as effective as more extensive surgery, but definitive evidence was not available to settle the debate.

In a study to determine whether alternative treatments to radical mastectomy increased the risk of cancer recurrence or death (18), a total of 34 institutions in the United States and Canada randomized 1,665 women with operable breast cancer. Women judged to be free of cancer in axillary nodes were randomly assigned to undergo radical mastectomy, total mastectomy with regional radiation treatments, or total mastectomy alone. Those judged to have cancer in the axillary nodes were assigned randomly to either radical mastectomy or total mastectomy with radiation treatment. The overall rates of survival and cancer recurrence were similar for all three groups of patients with clinically negative axillary nodes. The overall survival at 10 years for patients with positive axillary nodes was similar for those who underwent radical mastectomy and those who had total mastectomy with accompanying radiation treatment. This trial provided clear evidence that surgery less extensive than radical mastectomy could be safely performed with no decrease in long-term survival.

A second trial coordinated by the NSABP sought to determine whether even greater breast conservation could be safely achieved through segmental mastectomy (also referred to as lumpectomy), in which only the tumor and immediately surrounding tissue are removed (17). This trial randomized 1,843 women who had breast tumors no more than 4 cm at the largest dimension. The three types of treatment tested were: total mastectomy, segmental mastectomy (lumpectomy), or segmental mastectomy with accompanying radiation treatments. All of the women underwent removal of their axillary nodes, and those patients found to have evidence of nodal cancer underwent chemotherapy. After five years, the overall rates of survival were better for the women who had received segmental mastectomy, with or without radiation, than for those who had undergone total mastectomy, and the rates of survival with no recurrence of the disease recurrence were better for those who had undergone segmental mastectomy with radiation treatment than for those who had undergone total mastectomy.[2]

Because the NSABP trials enrolled a broad range of patients, their results are clinically relevant to a large proportion of women, and the outcomes measures-disease-free survival and overall survival-were easily ascertained, clinically important events. The trial treatments and follow-up monitoring were clearly more complex than those of nonsurgical large-scale trials, but most of the procedures are part of the standard manage-

[2] The NSABP has made enormous contributions to the field of breast cancer treatment, which would not have been possible without such large-scale collaborative trials. Recently reported evidence of misconduct by one of the NSABP clinical collaborators has raised concern about the scientific integrity of the findings concerning segmental mastectomy. However, the principal findings were unchanged in a re-analysis that was performed excluding data on patients who may have been inappropriately entered into the trial (16).

ment of breast cancer patients and would have been followed anyway.

Promising Therapies in Primary Prevention

For a common and serious condition such as acute MI, even modest reductions *in* mortality can have a significant public health impact, saving tens of thousands of lives per year. At the same time, effective means of preventing such a disease could, in theory, have a far greater impact, preventing perhaps hundreds of thousands of deaths each year. The conduct of large-scale trials of promising therapies in primary prevention presents unique challenges not faced by those conducting trials to test treatments in a population with a specific disease.

One primary prevention trial that has employed many of the principles of large and simple trials is the Physicians' Health Study, an ongoing, randomized, double-blind, placebo-controlled trial begun in 1982 to test the effect of low-dose aspirin on cardiovascular disease and beta-carotene on cancer risk among a population of 22,071 apparently healthy U.S. male physicians, ages 40 to 84 (28,29).

Because the rates of disease and death among an apparently healthy population "at usual risk" are much lower than rates among a comparable group of individuals with a serious condition such as MI, a primary prevention trial must not only enroll a large number of participants, but also follow them for an extended period in order to allow for valid tests of the study's hypotheses. The participants must also remain compliant with their assigned treatment regimens and be conscientious in maintaining contact with the researchers to report their health experiences. Significant noncompliance with the assigned study regimens or losses to followup will weaken the study's ability to generate valid results. The choice of a study population for such a trial, therefore, is particularly important. Because of their clear interest in health issues, physicians were considered a group who would be motivated participants willing to follow daily pill-taking regimens for an extended period.

A guiding principle of large, simple trials is to minimize the necessity for procedures or clinic visits beyond those that would take place in the standard management of a condition (49). But participants in a primary prevention trial are, by definition, free of major disease and therefore have no regular clinic visits or procedures. Because physicians were deemed capable of reporting on their own health with a high degree of accuracy, the trial could be conducted entirely by mail. Annual supplies of study medications are sent in convenient monthly calendar packs, and brief, followup questionnaires are mailed to collect data on compliance and relevant outcomes at 12-month intervals. Reports of study outcomes are verified by seeking permission to obtain copies of confirmatory medical records.

The Physicians' Health Study also implemented a prerandomization run-in period, in which the participants took their study pills for approximately 18 weeks before their official randomization into the study took place. After this run-in period, the doctors were sent a brief followup questionnaire, and only those who reported that they took their study pills at least two-thirds of the time were randomized into the trial. Participants were randomly assigned to one of four treatments: aspirin alone, beta-carotene alone, both active agents, or both placebos. This efficient design (known as a 2x2 factorial design) has allowed the trial to test two agents simultaneously, with little increased cost over that of a study testing one agent alone. The choice of study population, the enrollment procedures, and the prerandomization run-in period were designed to assemble a group of proven compliers who would be likely to follow the study regimen and report accurately on their health experience for the extended period of trial treatment and followup. Compliance with the assigned regimen is an important factor in determining a trial's statistical power to answer a research question. The run-in, therefore, increased the trial's power. Indeed, a study- population of 22,071 men who remain compliant with the regimen will have greater statistical power to answer a question than would a group of 33,000, one-third of whom become noncompliant (36).

The initial assembly of the study population involved much more effort than that of many large, simple trials, but once the participants were randomized, the trial procedures and followup in the Physicians' Health Study have proved remarkably streamlined and simple. Moreover, compliance rates after 10 years remain at over 80 percent, morbidity followup is over 95 percent, and every death among participants has been recorded.

In 1988, the external Data Monitoring Board terminated the aspirin component of the trial early because the group taking aspirin had experienced a statistically extreme 44 percent reduction in the risk of a first MI. The beta-carotene component of the trial has continued uninterrupted and is scheduled to end in 1995.

In addition to performing added work to assemble its study population, the Physicians' Health Study differed from many other large, simple trials in that it enrolled a relatively homogeneous study population. From the standpoint of generalizability of the study findings, testing these hypotheses in a more heterogeneous group might have seemed preferable. From the standpoint of validity, however, in view of the need to conduct the trial efficiently by mail, to maintain the followup of all participants, and to rely upon their own reporting of health outcomes and motivation to adhere to the treatment regimen over an extended period, a group of health professionals was considered an ideal population in which to conduct the trial. The increased validity and efficiency derived from choosing physicians for the study population were judged, overall, a greater asset to the generalizability of the trial results than a more representative study population unable to maintain adequate compliance for the duration of the study.

Although it may be reasonable to assume that the direction of any net effects seen in the Physicians' Health Study would be similar for other groups, the balance of benefits and risks may well differ for populations with different risk profiles. Because of the desirability of obtaining direct evidence of aspirin's primary prevention effect in women, a large-scale trial of aspirin was begun in 1992 among apparently healthy U.S. female health professionals. The Women's Health Study plans to enroll approximately 40,000 women, ages 45 and older, and will test the effects of low-dose aspirin, beta-carotene, and vitamin Eon the risks of cardiovascular disease and cancer (5).

IMPACT OF LARGE, SIMPLE TRIALS ON CLINICAL PRACTICE

Although the immediate goal of any large-scale randomized trial is to provide a reliable test of the intervention, the ultimate goal is to provide information that can be incorporated into the clinical management of patients in general medical practice. It is difficult to draw broad generalizations about the effects of large, simple trials on medical practice. Nevertheless, because large trials are generally better than small trials at answering research questions, their degree of scientific reliability is high, an important factor that probably influences clinicians' receptivity to research findings. The results of large trials also tend to be published in prominent journals, making physicians and the public more aware of them.

Several reports suggest that the recent large, simple trials of the treatment of acute MI have indeed had a measurable impact on medical practice. For example, in the mid- 1980s, about 25 percent of acute MI patients in the United States received thrombolytic therapy. By 1989, following publication of the ISIS-2 results, this figure rose to just under 40 percent (40). Aspirin was administrated to 39 percent of all acute MI patients before the ISIS-2 report, and approximately 72 percent of all patients following the report (35). Significantly more patients could benefit from thrombolysis and aspirin in acute heart attacks than are presently receiving these treatments (31). Nevertheless, the data indicate that practitioners are adopting these treatment regimens, which have been demonstrated to confer clear benefit.

Although findings from a trial maybe clear and unequivocal, in terms of their public health impact, it is equally important that they address questions considered important by clinicians. For this reason, the ability to provide reliable data on important possible modifying factors, such as age or time delay to treatment for thrombolysis, may also be a determinant of the impact of the findings for a large, simple trial on medical practice.

Large and simple trials may also affect clinical practice more directly and immediately than small, academic-based investigations because the large trials involve the collaboration of the practicing physicians who ultimately decide how to incorporate research results into patient care (48). Perhaps the best example of this is a series of trials of acute MI treatment in Italy. The GISSI trials (Gruppo Italiano per 10 Studio della Streptochinasi nell'Infarto Miocardico) have tested questions similar to those addressed in the ISIS trials. The first GISSI trial paralleled ISIS-2, testing intravenous streptokinase versus standard treatment among 11,806 patients at 176 coronary care units throughout Italy (24). Three-fourths of Italy's coronary care units collaborated in the trial. Within a year after the publication of the trial finding demonstrating a clear benefit of streptokinase, the drug had become routine treatment for acute MI in 99 percent of the country's coronary care units, suggesting a remarkably rapid and complete incorporation of the trial results into clinical practice (45). The GISSI investigators also held symposia for collaborating clinicians in order to provide scientific background on the trial treatments as well as information about aspects of randomized trial methodology.

LIMITATIONS OF LARGE, SIMPLE TRIALS

Testing a promising intervention by assembling a large study population and using a streamlined trial protocol can be an extremely effective approach to answer a wide range of important health questions. However, there are many questions for which large, simple trials are either not necessary or not feasible.

The anticipated number of outcome events is the primary determinant of how big a sample must be in order for a trial to reliably detect a moderate treatment effect. In a primary prevention trial, the sample must be quite large. In the Physicians' Health Study, where the anticipated rate of cardiovascular disease was very low, more than 22,000 participants were randomized in order to detect meaningful differences between the group receiving aspirin and the group receiving a placebo. Fewer than 400 heart attacks occurred during a five-year period (139 in the aspirin group versus 239 in the placebo group). The 44-percent decrease in the risk of a first MI among those assigned aspirin was highly statistically significant, with probability of less than 1 in 100,000 that the finding was the result of the play of chance (p-value < 0.00001) (44). A trial one-fourth the size of the Physicians' Health Study, with 5,500 participants (which would still be substantially larger than most randomized controlled trials), would have had inadequate statistical power to detect with certainty the 44-percent reduction in mortality.

By contrast, in a trial testing a chemotherapeutic agent inpatients with advanced metastatic cancer, half of whom were expected to die within six months, a small trial of several hundred patients could reliably detect the fact that six-month survival could be achieved in three-quarters rather than half of the patients. Large-scale trials, therefore, are not needed to detect even modest treatment effects in a population where the expected outcome rate is extremely high.

With respect to the trial treatment and study protocol, large, simple trials may not be feasible for studying interventions that are complex to administer and that require frequent clinic visits. Examples include trials testing new physical therapy treatments or drug trials in which frequent blood tests are necessary to maintain proper dosing levels and to monitor potential toxicity.

Also impractical for study in large, simple trials are interventions for which information on intermediate biological markers is deemed important for understanding the treatment's actions or aspects of the pathophysiology of the disease. For example, in a trial of a potential AIDS treatment, detailed laboratory studies at regular intervals may be important, because there maybe comparatively limited knowledge concerning the drug's side effects as well as the postulated mechanism for its benefit. An alternative to a highly complex investigation may be to carry out more intensive laboratory studies among a subset of participants, while the overall trial design incorporates features of the large, simple trial approach.

Finally, the principle of evaluating a treatment's effect on a clinically important outcome that is easily ascertained also limits the types of questions that can be addressed by large, simple trial designs. Examples in which the outcome of interest may be clinically important but difficult to assess in a streamlined trial include tests of promising treatments for arthritis, where objective measures of mobility must be made; mental illness, where detailed clinical evaluations are needed; visual acuity; and many quality of life outcomes, which are measured in terms of competence in carrying out activities of daily living.

ADDITIONAL ISSUES

Timing of Randomized Trials

Timing is an important issue in the initiation of all randomized trials. For ethical reasons, there must be sufficient belief in the potential benefit of a drug or procedure to justify exposing half the individuals to it, while at the same time, there must be sufficient doubt about its efficacy to justify withholding the intervention from the other half. Ideally, therefore, randomized trials should be conducted as soon as there is a belief that a treatment might confer a net benefit.

The danger in waiting too long to conduct a trial of an intervention that is potentially effective, but as yet unproved, is that it may be adopted into widespread clinical use and become accepted as standard therapy, even without firm evidence supporting its efficacy. Not only does this expose individuals to medical interventions that might not be beneficial (or might even be harmful), but it also increases the logistical difficulty of conducting subsequent randomized trials to evaluate the intervention. Such trials may not be feasible if potential participants or health care providers become reluctant to be part of a trial in which some participants will not receive a treatment that has come to represent standard medical care. It may be difficult to find a sufficiently large population of individuals willing to forego a treatment or practice believed to be beneficial for the duration of the trial, even if there is no sound evidence to support this view.

For example, radical mastectomy for breast cancer gained wide acceptance as the standard of care after its introduction in the early 20th century by William Halsted. Halsted's clinical impression was that removal of the surrounding lymph nodes and muscle, in addition to the breast itself, would decrease the risk of recurrence or spread of the cancer and subsequent mortality. By the 1970s, however, questions that had been raised concerning the necessity of the radical mastectomy (11) prompted randomized trials comparing this procedure to less extensive surgery. Many physicians resisted the call to randomize patients into such a trial, because the radical mastectomy was so entrenched as the standard of care (2).

Eventually, however, a large number of women were enrolled in two multicenter trials. The trials clearly demonstrated that for women with localized tumors and no evidence of spread beyond the breast itself, five-year mortality rates as well as overall rates of recurrence were similar in those undergoing radical mastectomy, those undergoing simple mastectomy, and those undergoing a lumpectomy followed by radiation therapy (17,18).

Use of Factorial Designs

In view of the cost and feasibility issues of large, simple clinical trials, one technique to improve efficiency is to test two or more hypotheses simultaneously in a single trial, using a factorial design. In a 2x2 factorial design, participants are first randomized to treatments A or B to address one hypothesis, and then within each treatment group there is further randomization to treatments C or D to evaluate a second question. In a 2x2x2 factorial design, each of these subgroups would be further randomized into two additional intervention groups to address a third hypothesis, and so on.

For example, ISIS-2 used a 2x2 factorial design to evaluate streptokinase as well as aspirin in the treatment of acute MI (33). Patients were randomized to 1.5 million units of streptokinase or a placebo given intravenously, as well as 160 mg of aspirin or a placebo daily for 30 days, for a total of four treatment groups: participants receiving either streptokinase alone, aspirin alone, both active agents, or both placebos.

The principal advantage of the factorial design is that it allows the simultaneous testing of more than one question in a single trial, while costing little more than a trial of one of the questions alone. Ideally, of course, the additional treatments in a factorial design should not complicate trial operations, materially affect eligibility requirements, or cause side effects that could lead to poor compliance or losses to followup. In addition, the possibility of an interaction between the treatment regimens must be considered. Although the possibility of such interactions is considered by some to be a limitation of a factorial design, only through such a design can any combined effects of trial treatments be detected (43). ISIS-2 showed that streptokinase alone and aspirin alone clearly reduced 35-day vascular mortality, but that the participants who received both drugs experienced the greatest reduction in risk. The factorial design allowed this interaction to be assessed, which would not have been possible in a single-factor study.

Subgroup Analyses from Randomized Trials

Looking at the effect of an intervention among specific subgroups of participants might appear to be a way to address the question of whether the findings of a trial conducted in a wide group of patients are also applicable to patients with particular characteristics. It is important to keep in mind, however, that the most valid comparison in a randomized trial is between the originally allocated treatment groups. It is only in this comparison that randomization, in trials of adequate sample size, assures nearly even distribution of all the potential confounding variables, both known and unknown. In an analysis of any subgroup, whether defined on the basis of compliance or any other baseline characteristic, the comparison is no longer randomized and the potential role of confounding must be evaluated and controlled to the extent possible, just as in any observational study.

This point is illustrated by the experience of the Coronary Drug Project trial, a study testing the effect of the cholesterol-lowering drug clofibrate in the reduction of mortality following MI (10). In that trial, the five-year mortality rates in the groups receiving the clofibrate and the placebo were very similar (18.0 percent versus 19.5 percent). Because there was substantial noncompliance with the clofibrate regimen, the investigators attempted to more clearly evaluate the efficacy of the drug by also analyzing the mortality rates within the clofibrate group. They found that patients whose compliance was at least 80 percent had a mortality rate of 15 percent, compared with a rate of 24.6 percent among those who were less compliant. Such a finding might be erroneously interpreted to indicate that clofibrate reduces mortality. An analysis within the placebo group, however, found a similar disparity in mortality between compliers and noncompliers, with rates of 15.1 percent and 28.2 percent, respectively. Even after controlling for 40 known possible confounding variables, researchers still found a difference between the mortality rates of compliers and noncompliers in the placebo group. These data indicate that subgroup analyses of compliers did not provide valid results, because of the inability to control for the confounding effects of differences between compliers and noncompliers that independently affected their prognosis.

Another problem in the interpretation of findings from subgroup analyses relates to the meaning of testing for statistical significance, or the p-value. In medical research, the conventional level of statistical significance is p=0.05. This means that once in 20 times, a finding will be said to be statistically significant by chance alone, even though no difference between the treatment groups actually exists. This implies that if enough comparisons were made in a trial, as would occur in an evaluation of the treatment's effect among a large number of subgroups, one in 20 would be statistically significant, even if the intervention actually had no effect. The interpretation of this finding, however, greatly depends on whether the subgroup was previously defined as being of interest or was found by "fishing" or "data-dredging."

To illustrate the potential pitfalls of analyzing many subgroups, the ISIS-2 investigators carried out analyses according to patients' astrological signs. The researchers found that those born under the birth signs Gemini and Libra experienced a nonsignificant adverse effect of aspirin in acute MI, whereas aspirin significantly reduced the mortality rate of those born under all other zodiac signs (33). This was not an a priori hypothesis, as there was no clinical basis for the belief that aspirin would differentially affect those born under certain astrological signs, and the analysis clearly demonstrates the caution that must be used in interpreting the results of subgroup analyses.

The Decision To Terminate a Trial Early

In the design phase of a trial, the researchers need to develop guidelines for deciding whether the trial should be modified or terminated before its originally scheduled conclusion. To assure that the welfare of the participants is protected, the unblinded data should be monitored by a group that is independent of the investigators who are conducting the trial. If the data indicate that the intervention has a clear and extreme benefit on the primary endpoint, or if a treatment is clearly harmful, the modification or early termination of the trial must be considered.

A decision to terminate a study early is based on a number of complex issues and must be made with a great deal of caution. It is critical that a trial not be stopped prematurely based solely on emerging trends from a small number of patients, because these findings might well be transient and disappear or even reverse after data have accumulated from a larger sample. As a general rule, the first requirement for even considering the modification or early termination of an ongoing trial is the observation of a sustained statistical association that is so extreme, and so highly statistically significant, that its emergence by chance alone would be virtually impossible. The observed association must then be considered in the context of the totality of evidence. A number of specific guidelines have been used in various studies, but the aim is to achieve an equitable balance between, on the one hand, protecting randomized participants from real harm and, on the other hand, minimizing the risks of mistakenly modifying or stopping the trial prematurely.

Whenever a trial is ended prematurely because of findings related to one endpoint, the ability to answer other, often equally important, questions may be lost. The Physicians' Health Study is a case in point. The study was designed to evaluate two primary prevention hypotheses: whether low-dose aspirin reduces cardiovascular mortality and whether beta-carotene decreases cancer incidence. In early 1988, the trial's Data Monitoring Board prematurely terminated the randomized aspirin component of the trial (7a). This decision was based on all the available evidence, including three major considerations: the presence of a statistically significant ($p<0.00001$) reduction in the risk of MI among those in the group receiving aspirin; the fact that no effect of aspirin on cardiovascular mortality could be detected in the trial until the year 2000 or later, because of the exceptionally low cardiovascular death rates among the participating physicians; and the fact that aspirin was subsequently prescribed for more than 85 percent of the participants who experienced nonfatal MIs, which would render any later findings about aspirin and cardiovascular mortality particularly difficult to interpret.

Two other significant outcomes of interest in relation to aspirin were stroke and cardiovascular death, both of which occurred less frequently than MI. As a result of the early termination of the aspirin component, participants experienced inadequate numbers of strokes and cardiovascular deaths to permit a reliable assessment of aspirin's effect. There was an apparent increased risk of stroke—primarily in the subgroup of hemorrhagic strokes—but it was not statistically significant. No reduction in the risk of cardiovascular mortality was associated with aspirin. Although a number of explanations have been proposed, the primary consideration must be that the number of cardiovascular deaths in the trial at the time the aspirin component was terminated was simply too small to reliably evaluate the endpoint. Thus, two major pieces of the benefit-to-risk equation for the use of aspirin in the primary prevention of cardiovascular disease could not be determined, because of the ethical and practical considerations that prompted the early termination of the aspirin component of the trial.

Role of Meta-Analyses in Randomized Trials

The sample size of a trial and its resultant statistical power determine the extent to which chance may have influenced the study findings. If a study's sample size is inadequate, then a finding of no statistically significant association between the intervention and the outcome (a so-called null finding) may well be uninformative, because a true lack of association would be difficult or impossible to distinguish from a true association that simply could not be detected because of inadequate statistical power.

The ambiguity of the results from individual trials with small samples provides a strong rationale for much larger trials that could reliably detect modest treatment effects. Some investigators have argued that small trials should be pursued first, with larger investigations undertaken only if shown to be necessary. Because uninformative null results from small trials may erroneously suggest no effect, however, it would appear far preferable to mount a large trial once there is sufficient belief in a treatment's potential. If the effect is far greater than anticipated, a large trial can always be terminated earlier than scheduled.

Although a single well-designed and -conducted trial of sufficient size to detect the true effects of an intervention is usually considered optimal, in the absence of a definitive study, statistical overviews or meta-analyses that consider in aggregate the data from several small trials can provide useful information by minimizing the role of chance as an explanation for the findings (30) (see M.P. Longnecker, *Meta-Analysis,* background paper no. 4). However, meta-analysis cannot overcome the effects of bias or confounding present in the individual trial results.

One of the most important uses of meta-analyses of small trials may be not to provide a definitive answer to a question, but to provide a reliable estimate of the most likely effect of an intervention. That estimate then can be used in planning a future trial with adequate power to detect such an effect if it truly exists. With respect to estimating the size of any reduction in risk, the results from a pilot study or even a single small trial are likely to be quite unstable due to sampling variability. By contrast, the unique strength of meta-analysis of data from all randomized trials is to minimize the variability of the overall estimate that is obtained from each individual study. Thus, meta-analyses provide the most reliable risk estimates that can be obtained in the absence of individual trials of adequate statistical power.

The pitfalls of estimating the likely effects of an intervention from a single small trial instead of an overview can be illustrated by comparing two trials that tested the effects of beta-blockers, drugs given in the early acute phase of a heart attack: the Metoprolol in Acute Myocardial Infarction (MIAMI) trial and the First International Study of Infarct Survival (ISIS- 1). Before the initiation of the MIAMI trial, the investigators conducted a pilot study of approximately 1,400 participants. Based on an observed 36-percent reduction in total mortality, approximately 6,000 individuals were enrolled in the full-scale trial. By contrast, the sample size for ISIS- 1 was calculated from an overview of 21 previously conducted trials of beta-blocker therapy, which indicated an approximate 10-percent reduction in total mortality. On the basis of this estimate, more than 16,000 patients were enrolled in ISIS-1. When the two trials were completed, the estimates of the effects of treatment were similar, with vascular mortality reduced by approximately 13 percent in MIAMI and 15 percent in ISIS- 1, but the results from the MIAMI trial did not achieve statistical significance, while the results from ISIS-1 did (32,39).

CONCLUSIONS

For addressing many important research questions, randomized trials are neither necessary nor desirable. However, if the effects of a hypothesized intervention are likely to be only small to moderate in size, a trial with a large sample will be necessary to provide a definitive test of such research questions. Large trials can, if properly designed, be conducted using relatively simple protocols in which minimal screening or data collection is required.

There must be flexibility in the application of these principles to suit the particular circumstances of each research question. Although trials of in-hospital treatment of acute MI may be extremely simple in design and collect minimal data, those of chronic disease treatments, such as the digitalis trial in congestive heart failure, may require somewhat more involved protocols. Trials of primary prevention, in turn, may necessitate more prolonged screening phases to enroll populations of willing and eligible participants, and longer treatment and followup will be needed to accrue sufficient numbers of endpoints to permit valid tests of the trial hypotheses.

All these trials are characterized by their emphasis on enrolling large numbers of participants; testing an intervention's effect on a readily ascertained, clinically important outcome; and collecting a relatively limited amount of baseline and followup data. Many areas of medical research are suitable for large, simple trial protocols. Some types of questions, however, do not lend themselves to testing using such an approach, and are therefore best evaluated in more traditional trials with complex protocols and extensive data collection.

Nevertheless, many questions could be tested using far simpler protocols than those that have been used in most randomized controlled trials. Physicians are by training-and perhaps by temperament-oriented toward gathering detailed information on individual patients. This has very likely contributed to the use of very complex trial protocols with small samples in which hundreds of variables are collected on a few participants. Although an extensive history on a particular patient is correctly viewed as crucial to rendering individually appropriate clinical decisions, the same is not necessarily true for attempting to answer fundamental questions regarding the effectiveness of a promising medical intervention. It is far preferable, in the case of many trials, to have data on a few variables for hundreds or thousands of participants than to collect information on scores of variables from only a few study participants.

Where appropriate, therefore, large and simple trials can provide more reliable tests of an intervention than can other feasible research approaches. As their broad contributions to medicine become more widely understood, such investigations may play an increasing role in answering important research questions, and in providing a sound basis for formulating rational clinical decisions for individual patients and public health policy for the general population.

REFERENCES

1. **Aboulker,** J.P., and Swart, A. M., "Preliminary Analysis of the Concorde Trial," *Lancet* **341:889-890,** 1993.
2. **Altman,** L., "Fall of a Man Pivotal in Breast Cancer Research," *The New York Times,* p. B1O, Apr. 4, 1994.
3. **Boissel,** J. P., "Impact of Randomized **Clinical** Trials on Medical Practices," *Controlled Clinical Trials* **10:12S-134S,** *1989.*
4. **Buring,** J.E., and Hennekens, C. H., "Cost and **Efficiency** in Clinical Trials: The U.S. Physicians' Health Study," *Statistics in Medicine* **9:29-33,** *1990.*
5. **Buring,** J. and Hennekens, C. (for the Women's Health Study Research Group), "The Women's Health Study: Summary of the Study Design," *Journal of **Myocardial Ishcemia*** **4:27-29, 1992.**
6. **Byar,** D., "Design Considerations for AIDS Trials," ***Journal** of AIDS* **3(suppl.** *2): S16-S19,* **1990.**
7. Byar, D. P., **Schoefeld,** D. A., Green, S. B., et al., "Design Considerations for AIDS

Trials," *New England of Medicine 323: 1343-1348, 1990.*

7a. Cairns, J. Cohen, L., Colton, T., et al., "Issues in the Early Termination of the Aspirin Component of the Physicians' Health Study, *Annals of Epidemiology* 1:345-405, *1991.*

8. CLASP (Collaborative Low-Dose Aspirin Study in Pregnancy) Collaborative Group, "CLASP: A Randomised Trial of Low-Dose Aspiring for the Prevention and Treatment of Pre-Eclampsia Among 9,364 Pregnant Women," *Lancet* 343:619-629, *1994.*

9. Concorde Coordinating Committee, "Concorde: MRC/ANRS Randomised Double-Blind Controlled Trial of Immediate and Deferred Zidovudine in Symptom-Free HIV Infection, " *Lancet* 343:871-881, 1994.

10. Coronary Drug Project Research Group, "Influence of Adherence to Treatment and Response of Cholesterol on Mortality in the Coronary Drug Project, New *England Journal of Medicine 303: 1038, 1980.*

11. Crile, G., "Treatment of Breast Cancer by Local Excision," *American Journal of Surgery* 109:400-403, *1965.*

12. Digitalis Investigation Group, Trial To Evaluate the Effect of Digitalis on Mortality in Heart Failure, final protocol, U.S. Department of Health and Human Services, National Institutes of Health, National Heart, Lung and Blood Institute, Bethesda, MD, unpublished report, 1992.

13. Ellenberg, S. S., "Do Large, Simple Trials Have a Place in the Evaluation of AIDS Therapies? *Oncology* 6:55-59, 1992.

14. Ellenberg, S. S., and Foulkes, M. A., "The Utility of Large, Simple Trials in the Evaluation of AIDS Treatment Strategies, *Statistics in Medicine* 13:405-413, 1994.

15. Ellenberg, S. S., Cooper, E., Eigo, J., et al., "Studying Treatments for AIDS: New Challenges for Clinical Trials-A Panel Discussion at the 1990 Annual Meeting of the Society for Clinical Trials," *Controlled Clinical Trials* 13:272-292,1992.

16. EMMES Corporation, "Reanalysis of NSABP Protocol B-06—Final Report," prepared by D.M. Stablein and staff of the EMMES Corporation, Potomac, MD, Apr. 11, 1994.

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. Fisher, B., Redmond, C., Fisher, E.R., et al., "Ten-Year Results of a Randomized Clinical Trial Comparing Radical Mastectomy and Total Mastectomy with or without Radiation," New *England Journal of Medicine* 312:674-681, *1985.*

19. Fishl, M.A., Richman, D. D., Grieco, M. H., et al., and the AZT Collaborative Working Group, "The Efficacy of Azidothymidine (AZT) in the Treatment of Patients with AIDS and AIDS-Related Complex: A Double-Blind, Placebo-Controlled Trial," New *England Journal of Medicine 317: 185-191, 1987.*

20. Francis, T., Kerns, F.T., Voight, R. B., et al., "An Evaluation of the 1954 Poliomyelitis Vaccine Trials: Summary *Report, ''American Journal of Public Health* 45:1-49, *1955.*

21. *Garg,* R., Scientific Project Officer, U.S. Department of Health and Human Services, National Institutes of Health, National Heart, Lung and Blood Institute, Bethesda, MD, personal communication, Jan. 6, 1992.

22. Gibbons, R.J., Holmes, D.R., Reeder, G. S., et al., "Immediate Angioplasty Compared with the Administration of a Thrombolytic Agent Followed by Conservative Treatment for Myocardial Infarction," New *England Journal of Medicine* 328:684-691, *1993.*

23. Grines, C.L., Browne, K.F., Marco, J., et al., "A Comparison of Immediate Angioplasty with Thrombolytic Therapy for Acute Myocardial Infarction," New *England Journal of Medicine* 328:673-679, *1993.*

24. Gruppo Italiano por 10 Studio della Streptochinasi nell'Infarto Miocardico (GISSI),

"Effectiveness of Intravenous Thrombolytic Treatment in Acute Myocardial Infarction," *Lancet* i:397-402, 1986.

25. Gruppo Italiano por 10 Studio della Sopravvivenza nell'Infarto Miocardico, "GISSI-2: A Factorial Randomized Trial of Alteplase versus Streptokinase and Heparin versus no Heparin among 12,490 Patients with Acute Myocardial Infarction," *Lancet* 336:65-71, 1990.
26. GUSTO Investigators, "An International Randomized Trial Comparing Four Thrombolytic Strategies for Acute Myocardial Infarction," New *England Journal of Medicine* 329:673-682, *1993.*
27. Hennekens, C. H., and Buring, J. E., *Epidemiology in Medicine* (Boston, MA: Little, Brown, and Company, 1987).
28. Hennekens, C. H., and Buring, J. E., "Methodologic Considerations in the Design and Conduct of Randomized Trials: The U.S. Physicians' Health Study," *Controlled Clinical Trials* 10(suppl.): 142S-150S, 1989.
29. Hennekens, C. H., and Eberlein, K., "A Randomized Trial of Aspirin and Beta-Carotene Among U.S. Physicians, *Preventive Medicine 14: 165-168, 1985.*
30. Hennekens, Buring, J. E., and Hebert, P., "Implications of Overviews for Randomized Trials," *Statistics in Medicine* 6:397-402, *1987.*
31. Hennekens, C. H., Jonas, M. A., and Buring, J. E., "The Benefits of Aspirin in Acute Myocardial Infarction: Still a Well-Kept Secret in the U.S.," *Archives of Internal Medicine* 154:37-39, *1994.*
32. ISIS-1 Collaborative Group, "ISIS-l: A Randomised Trial of Intravenous Atenolol Among 16,027 Cases of Suspected Acute Myocardial Infarction," *Lancet* 2:57-66, 1986.
33. ISIS-2 Collaborative Group, "ISIS-2: Randomised Trial of Intravenous Streptokinase, Oral Aspirin, Both or Neither Among 17,187 Cases of Suspected Acute Myocardial Infarction," *Lancet* 2:349-360, *1988.*
34. *ISIS-3* Collaborative Group, "ISIS-3: A Randomised Trial Comparing SK vs. tPA vs. APSAC and Comparing Aspiring Plus Heparin vs. Aspirin Alone Among 41,299 Suspected Acute Myocardial Infarction," *Lancet* 339:1-18, 1992.
35. Lamas, G. A., Pfeffer, M.A., Hamrn, P., et al. (for the SAVE Investigators), "Do the Results of Randomized Clinical Trials of Cardiovascular Drugs Influence Medical Practice?" New *England Journal of Medicine* 327:241-247, *1992.*
36. Lang, J. M., Buring, J. E., Rosner, B., et al., "Estimating the Effect of the Run-In on the Power of the Physicians' Health Study," *Statistics in Medicine* 10: 1585-1593, 1991.
37. Lange, R. A., and Hillis, L. D., "Immediate Angioplasty for Acute Myocardial Infarction," New *England Journal of Medicine* 328:726-728, *1993.*
38. Meier, P., "Polio Trial: An Early Efficient Clinical Trial," *Statistics in Medicine* 9:13-16, *1990.*
39. The MIAMI Trial Research Group, "Metoprolol in Acute Myocardial Infarction (MIAMI): A Randomised Placebo-Controlled International Trial," *European Heart Journal 6: 199-226, 1985.*
40. Pfeffer, M.A., Moye, L. A., Braunwald, E., et al., "Selection Bias in the Use of Thrombolytic Therapy in Acute Myocardial Infarction," *Journal of the American Medical Association* 266:528-532, *1991.*
41. Ridker, P. M., O'Donnell, C., Marder, V., et al., "Large-Scale Trials of Thrombolytic Therapy for Acute Myocardial Infarction: GISSI-2, ISIS-3, GUSTO- l," *Annals of Internal Medicine* 119:530-532, 1993.
42. Smith, R. P., Meier, P., Abrams, D., et al., "Time for Some Really Large, Simple Randomized Trials in HIV Disease," *Oncology* 6:63-64, 1992.

43. Stampfer, M., Buring, J. E., Willet, W., et al., “The 2x2 Factorial Design: Its Application to a Randomized Trial of Aspirin and Carotene in U.S. Physicians,” *Statistics in Medicine 4:11* 1-116, 1985.

44. The Steering Committee of the Physicians’ Health Study Research Group, “Final Report on the Aspiring Component of the Ongoing Physicians’ Health Study,” New *England Journal of Medicine* 321: *129-135,* 1989.

45. Tognoni, G., Franzonsi, M. G., Garattini, S., et al., “The Case of GISSI in Changing the Attitudes and Practices of Italian Cardiologists,” *Statistics in Medicine 9: 17-27, 1990.*

46. Topol, E.J., “Thrombolytic Intervention,” *Textbook of Interventional Cardiology,* E.J. Topol (cd.) (Philadelphia, PA: W.B. Saunders, 1989).

47. Volberding, P. A., Lagakos, S.W., Koch, M. A., et al., and the AIDS Clinical Trials Group of the National Institute of Allergy and Infectious Diseases, “Zidovudine in Asymptomatic Human Immunodeficiency Virus Infection: A Controlled Trial in Persons with Fewer Than 500 CD4-Positive Cells Per Cubic Millimeter,” New *England Journal of Medicine* 322:941-949, *1990.*

48. Wilhelmsen, L., “The Applicability of the Results of Streamlined Trials to Clinical Practice,” *Statistics in Medicine 9: 185-191, 1990.*

49. Yusuf, S., Collins, R., and Pete, R., ’’Why Do We Need Some Large, Simple Randomized Trials?” *Statistics in Medicine* 3:409-420, *1984.*

50. Yusuf, S., Collins, R., Pete, R., et al., “Intravenous and Intracoronary Fibrinolytic Therapy in Acute Myocardial Infarction: Overview of Results on Mortality, Reinfarction, and Side Effects from 33 Randomised Controlled Trials,” *European Heart Journal* 6:556-585, *1985.*

51. Yusuf, S., Garg, R., Held, P., et al., “Need for a Large Randomized Trial To Evaluate the Effects of Digitalis on Morbidity and Mortality in Congestive Heart Failure,” *American Journal of Cardiology* 69:64G-70G, *1992.*

52. Yusuf, S., Held, P., Tee, K. K., et al., “Selection of Patients for Randomized Controlled Trials: Implications of Wide or Narrow Eligibility Criteria,” *Statistics in Medicine* 9:73-86, *1990.*

53. Zijlstra, F., de Boer, M. J., Hoomtje, J.C.A., et al., “A Comparison of Immediate Coronary Angioplasty with Intravenous Streptokinase in Acute Myocardial Infarction, ’’New *England Journal of Medicine* 328:680-684, *1993.*

Meta-Analysis

Background Paper 4

by
Matthew P. Longnecker
University of California
Los Angeles, CA

SUMMARY

A meta-analysis is a systematic, quantitative review of a subject. Using very explicit procedures, the analyst reviews the existing studies of a subject and re-analyzes their results to arrive at a more robust and comprehensive result. Three major features distinguish this method from a traditional narrative literature review:

- *the formal and comprehensive search for relevant data;*
- *the explicit, objective criteria for selecting studies to be included; and*
- *the quantitative statistical analysis of the studies' results.*

The justification for analyzing studies' results together in a meta-analysis is that all the component studies provide results that address the same research question.[1] *Where all the results come from similar randomized controlled experiments, meta-analysis is widely recognized as a powerful technique for evaluating the effectiveness of health technologies. Where the existing studies are less ideal for a meta-analysis, using the principles of this technique (e.g., making one's criteria for selecting and reviewing studies formal and explicit) can still improve the analyst's ability to undertake an objective, comprehensive review.*

Several issues regarding the appropriateness and methodological rigor of meta-analyses are still matters ofdiscussion and

[1] The definition of the term ***same*** in this context depends on the goal of the specific investigation. It may mean that the studies were virtually identical but carried out indifferent places or that the studies were quite different but addressed a similar problem.

debate. These include:

1. *Issues relating to the combinability of studies. Whether to use meta-analysis for nonexperimental or dissimilar studies is controversial and best evaluated on a case-by-case basis. The approach can, however, sometimes provide important insights that might not be evident with traditional narrative review methods.*
2. *Issues relating to publication bias. Results from unpublished studies can be different from published study results. Thus, not including all available studies can lead to bias. Meta-ana - lysts differ in how they attempt to overcome this problem.*
3. *Issues relating to the procedure for conducting a meta-analysis. Meta-analysts also differ in the specific procedures they follow to try to ensure that the review is unbiased and to recognize differences in the quality of the studies being reviewed.*

Despite the continuing discussion of these issues, and their importance for readers to consider when evaluating the quality and validity of any particular meta-analysis, the general technique is now well-established, and its applications continue to grow.

The practice of combining the results of different studies to obtain a more powerful and conclusive result has along history. In 1904, Pearson summarized the relation between mortality and inoculation against enteric fever by calculating the average correlation between mortality and inoculation across five communities (75). Statistical methods for combining the results of agricultural experiments were developed in the 1930s (42). Several applications of statistical methods for combining results across studies appeared in the medical literature in the 1950s (2,64), but it was the application of meta-analysis in the social sciences in the 1970s (37,59) that led to its frequent use in medicine today.

Applied to medical care, meta-analysis can be used to evaluate a treatment effect on any sort of outcome (e.g., to assess a treatment that is supposed to reduce the level of serum cholesterol) or to describe other characteristics when no treatment is involved (e.g., to calculate across studies the average sensitivity of a screening test, the average level of cholesterol indifferent populations, or the average correlation between sex and height). This paper focuses on the use of meta-analysis for assessing the effect of a treatment[2] on a health outcome, such as the risk of death, and assumes that the effect of a treatment is compared with an alternative-no treatment, a placebo, or an accepted treatment.

Rationale

The traditional method of combining the results of previous studies is the narrative review of a subject. Narrative reviews have generally been considered an acceptable evaluation and synthesis of data, but they have several well-known drawbacks. The method of identifying and selecting information is rarely defined, the information may be reviewed haphazardly, and the quality of the data is rarely assessed systematically (71). Because the narrative review approach is not quantitative, nor formally and explicitly systematic in its procedures, a traditional literature review may fail to include important studies and (because of its nonquantitative approach) may fail to make full use of the available data (91). The reviewer's biases may influence the assessment of the data, and directly comparing results across studies can

[2] *Treatment* is used here in a broad sense to include not only medical therapies aimed at improving one's condition (e.g., a drug or a surgical procedure) but also other health interventions and health behaviors (e.g., alcoholic-beverage consumption or cigarette smoking).

be difficult when the treatment effects are expressed differently.[3] In addition, in traditional reviews, authors often assess evidence by "vote counting" (tallying the number of studies that provide evidence for and against the presence of a given treatment effect (40)) without considering that some studies are larger or better than others.

In contrast, in a meta-analysis, the existing studies of the subject of interest are reviewed systematically and quantitatively, using formal and explicit procedures (box 4-1). The advantages of meta-analysis stem from two factors:

1. The use of explicit procedures for identifying and processing the study results. The comprehensive search for relevant studies minimizes the possibility y that available data are omitted. Explicit procedures for evaluating and handling the study results assure, to the extent possible, an unbiased assessment of the data in each study. These explicit procedures also help the reader to assess the competency and appropriateness of the meta-analysis.
2. The expression of the results of individual studies in comparable quantitative terms. A meta-analysis expresses the results of each study in a uniform way, facilitating comparisons of the results and their relation to the size of the individual studies. The uniform expression of results in a meta-analysis allows the analyst to calculate a summary number representing the average effect of a treatment across studies (if such a summary is of interest). A treatment effect is more easily detected when the results of several studies are considered together than when the results are examined individually; a related benefit is that a treatment effect within a subgroup of participants may become clear in the huge sample that is formed when study results are combined. The meta-analytic method facilitates objectivity and reliability, and the use of statistical methods can help researchers identify reasons for any variation in the studies' results (39,54,62,63,84). Identification of patterns in the variation of the treatment effect may contribute to the understanding of the generalizability of the result (31) and may suggest new hypotheses (34).

The astute traditional narrative reviewer may take the size and quality of studies into account, but such steps are key features of the meta-analytic approach. In a recent comparison of conclusions from traditional reviews and meta-analyses, Antman and his associates found that traditional reviews had often failed to recognize important treatment effects that were clearly evident from meta-analyses (l).

Although some authors have asserted that traditional narrative reviews are no longer useful (90), that view seems extreme. If the resources to support a meta-analysis are unavailable or the data are too different for statistical combination, a well-conducted review may be the only alternative. The review, however, will be most useful if the principles of meta-analysis are incorporated to the extent feasible.

The theoretical justification of meta-analysis rests on the assumption that the component studies all address the same research question. If the populations, the treatment, the study design, and the outcomes measured in each study are virtually identical, the studies essentially replicate the same protocol. Consequently, any differences in the treatment effect across studies can be presumed to occur by chance. Under these circumstances, a meta-analysis and an analysis of data from a multicenter clinical trial differ only slightly, and even a skeptic is likely to view a meta-analysis as appropriate.

In practice, however, the studies being combined in a meta-analysis are seldom virtually identical. As the component studies of a meta-analysis become less similar, the appropriateness of ana-

[3] If, for example, one author expressed the treatment effect as the difference in the observed and expected number of cases in the treatment group, and another author expressed the treatment effect as the ratio of the mortality rate in the treated and untreated groups, the results would not be directly comparable.

BOX 4-1: Definitions of Meta-Analysis

The term *meta-anatysis* **was** coined in 1976 by Glass, a social scientist (37), who defined it as “the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings.”1 Other broad definitions in frequent use in the medical and health care literature include:

- “the practice of using statistical methods to combine the outcome of a series of different experiments or investigations” (54);
- “a quantitative summary of research in a particular area” (31); and
- “[the use of the results of collections of research papers to answer specific questions, usually in a quantitative manner” (63).

The National Library of Medicine has developed a detailed definition that actually specifies the procedures to be followed in a meta-analysis (92). It defines this analytic tool as:

> “aquantitative metho dof combining the results of independent studies (usually drawn from the published literature) and synthesizing summaries and conclusions which may be used to evaluate therapeutic effectiveness, plan new studies, etc., with application chiefly in the areas of research and medicine. The method consists of four steps: a thorough literature review, calculation of an effect size for each study, determination of a composite effect size from the weighted combination of individual effect sizes, and calculation of a fail-safe number (number ofunpublished studies with opposing conclusions needed to negate the published literature) to assess the certainty of the composite size.”

The elements common to most definitions of meta-analysis are that the analysis is quantitative, that it is based on observations in independent studies, and that the results of the independent observations are summarized across studies.[2] The most prominent difference among the definitions of meta-analysis relates to the kinds of studies that may be included in the analysis. Some definitions stipulate that only results from randomized trials should be analyzed (1 1,98).

The exact systematic procedures used (such as literature searches and quantitative analyses) vary somewhat among published meta-analyses. Thus, the meta-analytic approach is more a set of general principles than a set of standard rules invariably followed. Nonetheless, it is note worthy that at least for the meta-analysis of randomized clinical trials, “meta-analysis has matured as a scientific discipline, with well-documented standards and methods” (57).

Some authors (78) use the term *meta-analysis* to refer to a combined analysis (47) or a pooled analysis (17,60), Unlike other meta-analyses, however, in a combined or pooled analysis the data for the individual participants in different studies are combined into one data set and analyzed as if they were from a multicenter study with a common protocol.[3] In contrast, in most meta-analyses, the studies’ results—rather than the original data—are combined. In practice, the results of a combined or pooled analysis usually are virtually identical to those from other meta-analyses. Pooling can be more difficult to perform than other meta-analyses, because It often requires the cooperation of many scientists (to obtain their raw data), but it has the advantage of facilitating the analysis of treatment effects in subgroups of participants.

[1] For a scholarty discussion of the etymology of the term, see Dickersin(24).

[2] other terms sometimes used to describe meta-anafyses include systematic overviews, pooling, data syntheses, and quantitative syntheses (24) and, less frequently, integrative research reviews, research integrations, research consolidations, research syntheses, quantitative assessments, surveys, re-analyses, and quantitative reviews (50,69).

[3] A study protocol defines the characteristics of people who are eligible to be in the study, describes the nature and duration of the treatment, discusses how the effect of treatment is assessed, and provides other details of how the study is conducted.

SOURCE: Matthew Longnecker, 1995

lyzing them jointly becomes a matter of judgment and, therefore, subject to debate. Yet even when the results that are combined come from somewhat dissimilar studies, meta-analysis may be useful—not so much for calculating a summary treatment effect as for allowing the analyst to examine how the treatment effect varies according to study characteristics or across subgroups of participants (72).

Current Applications: An Example

A good illustration of the manner in which meta-analyses are being used can be found in the work of Yusuf and his associates (101), who assessed whether a drug that dissolves blood clots (fibrinolytic therapy) decreased mortality in patients who had heart attacks (myocardial infarctions). The motivation for their assessment was that the results of individual clinical trials addressing this topic appeared contradictory and unreliable. The analysts examined data from studies in which patients were assigned at random to receive either treatment or no treatment (randomized clinical trials).

Using a computerized literature search, reviewing abstracts from scientific meetings, and contacting investigators who had completed trials but not published the results, the analysts located relevant studies and identified 24 eligible trials. For each trial, the number of patients treated with the fibrinolytic therapy, the number not treated with the therapy (the control group), and the number of deaths occurring in each of these two groups were noted. The analysts then used a relatively simple statistical method to calculate across all 24 studies the average effect of the treatment on mortality.

When the data from all the studies were considered together, 51 fewer deaths were found in the group treated with the fibrinolytic therapy than would have been expected if the treatment had no effect on mortality rates. This reduction was found to be statistically significant (see box 4-2). In contrast, just five of the 24 studies had individually shown a statistically significant beneficial effect of treatment. Using the quantitative methods of meta-analysis to consider the results of all the trials simultaneously demonstrated that the treatment was effective in reducing mortality, in a way that a simple narrative review of the results of individual studies would not. (For a detailed discussion of the quantitative methods used in this and other meta-analyses, see appendix 4-A.)

CONDUCTING A META-ANALYSIS

Before conducting a meta-analysis, the analyst should evaluate its utility and desirability and the combinability of the studies. Questions to be asked include:

Are there any good studies that address the research question?

If so, are the study designs similar enough that combining them makes sense?

Given the available data, are the results of the meta-analysis likely to make an important contribution to knowledge?

If the answers to these questions are positive, the analyst then proceeds.

Conducting a meta-analysis is a systematic process (30,50,53,85) that entails the following steps:

1. defining the research question,
2. defining the admissibility criteria for studies,
3. searching for relevant data,
4. reviewing the retrieved data to determine admissibility,
5. assessing the quality of the eligible studies,
6. correcting for bias,
7. performing the data analysis (including sensitivity analysis and influence analysis),
8. assessing the publication bias, and
9. interpreting the results.

Defining the Research Question

In defining the research question, the analyst specifies the treatment under investigation, the treatment's alternative, the outcome, the study populations, and the quantitative measure of the effect in which the analyst is interested.

BOX 4-2: Statistical Significance

"Statistical significance" is a phrase that traditionally has been used to indicate the researcher's belief that the effect observed in an experiment represents a real phenomenon and is unlikely to be due entirely to chance. It is sometimes contrasted with "clinical significance, " which indicates that the effect is not only real but is large enough or important enough to have a meaningful impact.

In a typical medical experiment to determine whether a new treatment has a beneficial effect (compared with plausible alternatives), the researchers begin by assuming that it does not (the "null hypothesis") and then attempting to disprove that assumption. If the study is a well-designed, randomized trial, it is unlikely that an observed apparent treatment effect will be due to chance alone. In statistics, tradition holds that if an observed effect has less than a 5-percent probability of being observed where no treatment effect exists (i.e., p<.05), the treatment effect is most likely not zero. In that case, the treatment effect is said to be statistically different from zero (statistically significant).

The use of the 5-percent cutoff level is a popular scientific convention that is somewhat arbitrary; one could also justify choosing 1 percent, or 0.1 percent, or some other low value, as the cutoff for significance. Thus, whether a result is considered statistically significant depends, in part, on the chosen significance level.

An alternative approach, which is growing in favor with researchers and analysts alike, is to place confidence limits around an observed treatment effect, concerning oneself more with the size of the treatment effect and one's certainty about how big it is than with an absolute answer to whether it exists. A 95-percent confidence interval, for example, indicates that in 95 of 100 hypothetical repetitions of the experiment, the true treatment effect would fall within the range of estimated treatment effects included in the confidence interval. (For a complete discussion of confidence intervals, see Rothman (83)).

SOURCE: Matthew Longnecker, 1995.

Defining the Admissibility Criteria

The next step is to define formal admissibility criteria for the component studies. The research question is expressed in specific terms that facilitate the decisions about whether potentially eligible studies should be included. The criteria might require, for example, that to be admissible a study must:

- be double-blinded,[4]
- have a placebo as the alternative treatment,
- have the dose of the treatment be in a certain range,
- have study participants whose ages are within a certain range,
- present its results in a manner that permits the relevant effect to be calculated,
- evaluate the effect of treatment on the outcome within a specific length of time, and
- be written in English.[5]

Expertise on the research topic is indispensable at this stage (33).

Searching for Relevant Data

The computerized literature search (45), the most important part of the formal search for study re-

[4] In a double-blinded study, neither the patient nor the clinician administering the treatment know which treatment the patient is actually receiving.

[5] Note that this criterion tight eliminate many otherwise eligible studies.

suits, requires special training and is often done in consultation with a qualified librarian. The librarian performs an over-inclusive search using the admissibility criteria for the meta-analysis. Searching at least two different computer databases for studies to include in the meta-analysis increases the number of eligible studies found (14,86).

Because computerized searches can miss important references (23), meta-analysts usually supplement the searches by perusing the reference lists of the identified articles and by consulting experts in the field, abstracts of conferences where relevant papers are likely to have been presented, and any other informally identified sources.

Reviewing the Data for Admissibility

The articles and papers resulting from the literature search are then reviewed to determine whether they meet the criteria for admissibility. Careful documentation of the rejected studies has been advocated (85). The relevant information is abstracted from the admissible articles, and the study results are re-expressed in a standard fashion, if necessary, for subsequent statistical analysis. The characteristics of the individual studies are recorded for use in the data analysis. Ensuring that the analysis does not include multiple studies based on the same participants prevents the inclusion of redundant data (18).

Some authors recommend that the information in the admissible studies be re-abstracted by a second researcher as well, and the extracted data double-checked (35,100). This process is time-consuming but is believed to improve the quality and objectivity of the analysis.

Assessing the Quality of Studies

Many meta-analysts assess the quality of the eligible studies with the aid of standard, published criteria (10,1 4,21) or with criteria specially tailored to the research question under investigation (6,60). Subjective methods of assessing quality have also been employed (6). Examples of criteria used to assess the quality of a randomized clinical trial are:

- whether the participants knew what they received (the treatment or the placebo),
- whether the investigators knew which participants received the treatment and which received the placebo during the trial,
- whether the presentation of the data was appropriate, and
- whether the statistical analyses were appropriate.

Meta-analysts often try to quantify the quality of the studies by awarding points that reflect how well each study approached the ideal for each criterion; the sum of these points for a given study is then used as its summary quality score (16).

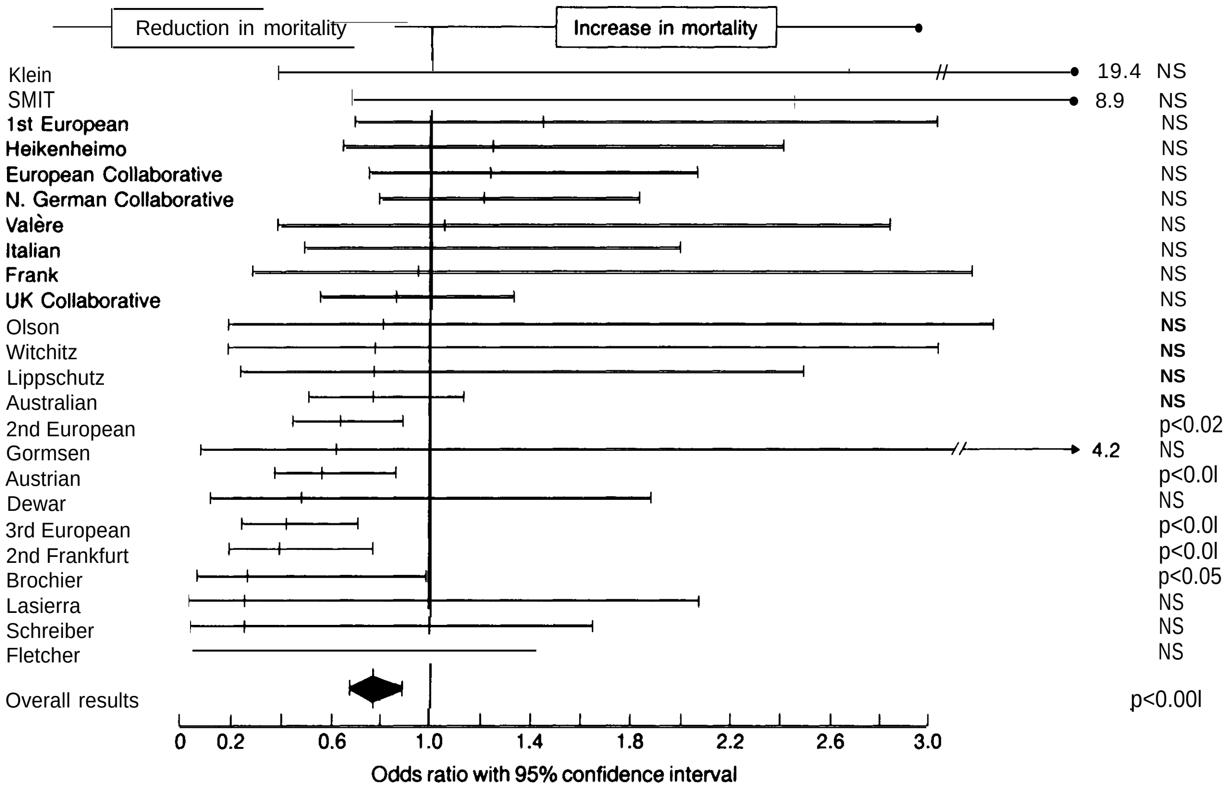
Correcting for Bias

The manner in which a study was designed, conducted, or analyzed can cause the observed effect of the treatment to differ from what would have been observed if the study had been done better. For example, investigators who are aware of what the participants received during a trial tend to find larger treatment effects than do investigators who are blinded to the participants' treatment or lack thereof (19). This may occur because of the investigators' desire to find the new therapy efficacious, which interferes with their ability to make equally accurate assessments of the outcomes in the treatment and control groups. High-quality studies are presumed to provide better estimates of the true effect of the treatment.

If the treatment effect observed in a given study is not an accurate measure of the true effect of the treatment, the result of the study is biased. The amount of bias is reflected by the difference between the observed effect and the true effect.

In some studies, the size of the bias is known with enough certainty that the observed treatment effect can be adjusted for the bias (28,39,93). The adjustment entails taking the treatment effect observed in a study and making it larger or smaller by an amount proportional to the bias (before the study's result is included in the meta-analysis). Correcting the results of studies for bias has not been a frequent practice, however, because it often is nearly impossible to determine whether a given

FIGURE 4-1: Summary of Study Results on Fibrinolytic Agents and Survival After Myocardial Infarction



NOTE: These study results are shown in a different form in tables 4-A-1 and 4-A-2. Here, the studies are listed according to the size of the odds ratio, which has a 95-percent confidence level.

SOURCE: S. Yusuf, R. Collins, R. Peto, et al., "Intravenous and Intracoronary Fibrinolytic Therapy in Acute Myocardial Infarction: Overview of Results on Mortality, Reinfarction, and Side-effects from 33 Randomized Controlled Trials," *European Heart Journal* 6:556-558, 1985

type of bias occurred in a study or not. Even if a bias is known to have occurred, the degree to which the bias reduced or increased the observed treatment effect is difficult to estimate with certainty.

▮ Analyzing the Data

The quantitative data analysis sets meta-analyses apart from other systematically conducted reviews. In the data analysis, the meta-analyst first examines the results of the component studies. Graphical representation of the studies' results are well-suited to this purpose. For example, figure 4-1 represents the data from the meta-analysis conducted by Yusuf and his associates described above. This figure demonstrates that most of the studies found a beneficial effect of treatment (i.e., an odds ratio less than 1), that the variation in study-specific treatment effects appeared rather large (suggesting heterogeneity), and that many of the individual studies were imprecise.

The reasons for variation among studies' results may be identified by analyzing subgroups of studies separately or by using regression analysis. The degree of variation in the studies' results is assessed with a formal statistical calculation. If a summary estimate of the treatment effect is appropriate, the effects from the different studies are combined (see appendix 4-A).[6]

Sensitivity analyses are then conducted to determine the extent to which the findings of the meta-analysis depend on assumptions made by the analysts. If, for example, the authors of a meta-analysis excluded several studies on grounds that others might challenge (e.g., the authors excluded studies not published in English), the meta-analysis could be repeated after including those studies to determine whether the overall results were sensitive to those exclusion criteria. If the second meta-analysis yields essentially the same results as the first one did, the authors' findings can more readily withstand criticism.

Influence analyses are another way of testing the robustness of the results. They examine whether the findings of the meta-analysis depend on the inclusion of the results of any particular study, such as a single large study or a study in which the treatment effect is extreme. In an influence analysis, the analyst recalculates the results of the meta-analysis after excluding the particular study of interest (e.g., the study with the unusual treatment effect) to determine whether the new results support the conclusions that were reached when all the data were included. If the results of the meta-analysis do not depend on the inclusion of such studies, the analyst can be more confident of the results.

Information regarding the quality of the studies may be considered in the data analysis. For example, summary quality scores may be included in the calculation of the overall result, in a regression analysis, or in the sensitivity or influence analyses.

Assessing Publication Bias

In the data analysis, the effect of publication bias on the result of the meta-analysis is assessed. Publication bias occurs when the published studies are not representative of the results of all the studies that have been conducted on the research question. Publication bias reflects the preference for publishing studies that have statistically significant findings or that support popular ideas (23). The analyst evaluates the potential effect of publication bias when graphically representing the individual studies' results (3,58) or after estimating the summary treatment effect (48,81).

Interpreting the Results

Like other analysts, meta-analysts conclude the process by interpreting the results so that their generalizability and their implications for practitioners and researchers are clear.

RELIABILITY AND VALIDITY

Reliability

A reliable meta-analysis is one that gives the same result when used again to assess the same research question using the same set of studies. If the identical studies were available to two meta-analysts and both were addressing precisely the same research question, the comparison of the results of the two meta-analyses would be a direct reflection of the reliability of the method. In practice, however, the research questions in replicate meta-analyses usually differ slightly, or the meta-analyses are done at different times, when different studies are available. Thus, some differences between replicate meta-analyses are expected.

[6] Meta-analyses often require specialized statistical methods, because the units of observation in meta-analyses differ from those used in traditional statistical analyses. The units of observation in meta-analyses are the results of independent studies, whereas the units of observation in clinical trials are the data for individual participants. Several reviews of statistical methods in meta-analysis are available (39,42,54) for readers interested in the technical details.

Meta-analyses appear to be at least moderately reliable. In an investigation of 20 replicated meta-analyses done by others, Chalmers and his associates (13) found that the differences among meta-analyses of the same research question were "almost always of degree rather than direction." The authors' interpretations of the findings in the replicate meta-analyses differed more than did the estimates of the summary treatment effects.

In a more recent examination of the reliability of meta-analyses (44), 20 more research questions were examined in replicate meta-analyses. These meta-analyses appeared to be more reliable than the ones studied by Chalmers and his colleagues (13). Henry and Wilson attributed the disagreements between the meta-analyses to differences in the research questions addressed by the analysts. In meta-analyses of progestins to prevent early pregnancy failure, for example, one group found no effect, whereas another group-which focused on studies whose subjects were women with histories of recurrent miscarriages-found evidence that the treatment was effective.

Inasmuch as the quality of meta-analyses before 1987 was found to be highly variable (85), the greater reliability of recent meta-analyses may reflect their improved quality.

Validity

Assessing the validity of meta-analysis requires the comparison of the results of applying this technique with the treatment's true effect, which is rarely known. As a substitute, investigators often use the results from a large clinical trial (one that is not part of the meta-analysis) as an estimate of the true effect. Where the true effect is unknown, this practice may be the most reasonable method for assessing validity.

Two teams of investigators have compared the results of meta-analyses with the results of single, large randomized clinical trials (15,44). Comparing three meta-analyses with the results of their respective clinical trials, Chalmers and his associates found that only one pair clearly agreed on the treatment effect (15). The researchers offered no explanation for the disagreement between another meta-analysis and its large trial but suggested that the third meta-analysis was based on such a small number of subjects that its result might have been greatly influenced by publication bias. Henry and Wilson (44) compared a large trial of oral anticoagulants with a meta-analysis addressing the same question and" found their results comparable. Although similar examples are easily identified (57), their importance is unclear. There has been no comprehensive survey to test the validity of meta-analysis. (Box 4-3 discusses what such a survey might look like if it were conducted.)

The reliability and validity of meta-analyses are likely to improve as the quality of meta-analyses improve. Several investigators have proposed guidelines for assessing the quality of meta-analyses (69,74,79,85). Whether these guidelines succeed in identifying meta-analyses of greater reliability and validity has not been established.

ISSUES AND CONTROVERSIES

Meta-analytic results can be controversial (66) because of concerns regarding the combinability of results, publication bias, or the meta-analytic protocol.

Combinability

Important questions regarding the combinability of studies (72) include the following:

- What types of studies should be included in a meta-analysis?
- Are the study protocols similar enough to warrant combining the results?
- What should be done if the treatment effects vary widely across studies?

Including Studies with Different Designs

Although evidence from good randomized clinical trials is widely accepted as valid, the validity of results from nonrandomized trials is less clear. Because the quality of randomized studies is related to the size of the treatment effect that is observed (19), some researchers believe that nonexperimental data—which presumably are particularly susceptible to bias—have no place in a

BOX 4-3: Designing a Survey To Test the Validity of Meta-analysis

Existing studies that have examined the validity of meta-analyses on particular subjects are few and have somewhat conflicting results. One possible way to examine the question more conclusively would be to conduct a comprehensive survey to address this topic.

Such a survey would begin with the definition of the broad research area about which an investigation of validity is desired. If the area were defined as, for example, the effects of all drugs on total mortality, the investigator would enumerate all the specific drugs for which meta-analyses of the effects on mortality have been done, select at random several specific drug-mortality meta-analyses to evaluate. For each drug-mortality meta-analysis, the investigation would perform a new meta-analysis in the following way: the investigator would order the original studies according to their dates of publication, then take the first five studies and compare the inverse variance of their combined meta-analytic estimate of treatment effect to the inverse variance of the treatment effect of the next published report. If the inverse variance of the next report is at least 50 percent of the inverse variance of the meta-analytic estimate, a comparison of the estimates of the treatment effect will provide information about validity If the inverse variance is less than 50 percent, the meta-analytic estimate should be recalculated to include the result of the sixth study. This result should be compared with that of the seventh study, and the process should be repeated, if necessary.

Some refinement might be required to make this approach work. If it could be carried out, however, its results would enable users of meta-analyses to be more confident of the validity of their results.

SOURCE Matthew Longnecker, 1995.

meta-analysis, and these results are excluded from many meta-analyses (76). Some researchers even define meta-analysis as including only data from randomized trials (1 1,98).

For many research questions, however, data from such trials either are unavailable (4) or not currently possible to collect (e.g., for logistical or ethical reasons). If data from randomized clinical trials (experimental data) are not available, a meta-analysis of observational (nonexperimental data) may still provide a more useful summary of data than a traditional narrative review would provide. The analyst's interpretation of the results of a meta-analysis of observational data should be appropriately conservative, as should the interpretations of the underlying individual observational studies.

Some types of observational studies are more susceptible to bias than others (box 4-4), a fact which meta-analysts must take into account. In a meta-analysis of alcohol consumption in relation to the risk of breast cancer, for example, the analysts examined the results of followup studies and case-control studies separately (60). The treatment effect in the followup studies was found to be larger than that in the case-control studies, and the results of the two types of studies were not combined, because the analysts felt that the results of the followup studies were more likely to represent an unbiased estimate of the treatment effect. This represents an empirical approach to decisions regarding combinability. A consensus regarding the appropriateness of combining the results of observational studies with different designs (regardless of their results) has not been reached (85). At this time, it is common in meta-analyses of observational studies to present there-sults separately according to the types of study design.

BOX 4-4: Bias in Study Designs

Study designs can be classified according to the methods of obtaining data: simple observation (nonexperimental studies) or observation after some type of intervention (experimental studies). The designs can be further classified according to the units upon which the observations are made: a population (for example, ecologic studies) or individuals (e.g., followup and case-control studies). Study designs vary with respect to the type of bias most likely to occur and the likelihood that the bias would materially affect the result.

In theory, experimental studies (randomized controlled trials) are the best method by which to assess the effect of a treatment. Data obtained from clinical trials are the most powerful for making causal inferences and are the least likely to be biased. Unfortunately, clinical trials are sometimes infeasible for practical, financial, or ethical reasons, When clinical trials cannot be performed, followup and case-control studies are the two nonexperimental study designs most commonly used. In a followup study, the occurrence of disease among the individuals who have and have not undergone the treatment of interest is compared. In case-control studies, the prior treatment experience of persons who already have the disease is contrasted with that in nondiseased (control) subjects, who represent samples of the population in which the cases occur. In general, the results of followup studies are considered less likely to be biased than are the results of case-control studies.

An example of an ecologic study is an examination of the rate of death from breast cancer in relation to per capita sales of fat in different countries, Ecologic studies such as this provide only weak evidence for causal inference, because it is not known whether the subjects who ate fat are the ones who got breast cancer, and because some factor (other than fat intake) correlated with fat sales may be the true reason for the variation in rates of breast cancer across countries.

SOURCE: Matthew Lortgnecker, 1995.

Different Protocols

Another combinability issue is whether the protocols of the studies (e.g., a group of randomized trials) are similar enough to warrant the combining of the studies' results. The answer depends on the particular research question, and the decision to combine the results depends on the judgment of the meta-analysts and their audience.

Many of the criticisms of meta-analysis have been related to decisions regarding combinability in a specific meta-analysis rather than to the method itself (9,34,36,46,52). In conducting a meta-analysis of nonmedical treatments for chronic pain, for example, Malone and Strube (65) calculated the average effect of one treatment on several different kinds of pain, including headache pain and cancer pain. Holyrod and Penizen (46) criticized the meta-analysis because the treatment effect might have been very different for headache and cancer pain, and the summarization across different types of pain might have obscured a treatment effect.

In another example, Held and associates (43) performed a meta-analysis in which they sought to summarize the effect of a general class of drugs (calcium antagonists) on preventing death among persons who had had heart attacks. The meta-analysis was criticized (9) because specific drugs within this class differed in their treatment efficacy. One subclass of drugs (which lowered the heart rate) reduced morbidity and mortality, whereas another subclass increased these outcomes. By analyzing all these drugs together, Held's team had come to the potentially misleading conclusion that the general class of drugs was not effective in reducing mortality and morbidity.

BOX 4-5: Fixed-Effects vs. Random-Effects Models—A Hypothetical Example

Fleiss and Gross (34) have presented an interesting hypothetical example that illustrates some of the issues regarding the choice between the fixed-effects model and the random-effects model (described in greater detail in appendix 4-A). One meta-analysis includes two published studies with odds ratios[1] of 1.0 and 6.0, and another includes two published studies with odds ratios of 2.0 and 3.0. All four odds ratios have the same variance (of the logarithm of the odds ratio): 0.01. When a fixed-effects model is used for both meta-analyses, the summary odds ratio is 2.45, and the 95-percent confidence intervals extend from 2.13 to 2.81. In both of these meta-analyses, the fixed-effects model's confidence intervals do not even include the values upon which they are based.

If a random-effects model is used to analyze the same data, however, the 95-percent confidence intervals are 0.5 to 10.0 in the first meta-analysis and 1.65 to 3.64 in the second. In both cases, the random-effects confidence intervals include the values upon which they are based, and the width of the confidence intervals reflects the amount of variation in treatment effect between the studies.

To take the example further, note that the random-effects summary in the first meta-analysis (2.45) gives the impression that, on average, a treatment effect exists, even though one of the two studies showed no treatment effect at all (an odds ratio of 1). Although the confidence interval is wide, an observer might look at the summary estimate and fail to appreciate that some studies showed no effect. Despite the disadvantages of summaries from random-effects models, however, the disadvantages of fixed-effects models are often even greater. As a result, random-effects models are gaining widespread acceptance (57).

[1] The odds ratio is the ratio of the odds of an event occurring under one set of circumstances to the odds of the event occurring under another set of circumstances

SOURCE Matthew Longnecker, 1995

Heterogeneous Results

A third combinability issue arises where the treatment effects in the component studies vary markedly—for example, when several studies show a large beneficial effect but other studies show a harmful effect. Large variation (heterogeneity) in the results of individual studies, when present, is usually evident when the study-specific results are represented graphically (see figure 4-1). The analyst can also assess the degree of variability to applying formal statistical tests of this characteristic. Summarizing a treatment effect across studies even when the study results are heterogeneous has been common in meta-analysis, although the practice is a subject of debate (39,77).

Also debated is the appropriate statistical procedure for summarizing the treatment effect in such cases (68,77). The two methods used most frequently to summarize treatment effects across studies are the fixed-effect model and the random-effects model (described in greater detail in appendix 4-A). In practice, if the treatment effects found in the component studies vary greatly, the results of meta-analyses using the two approaches maybe somewhat different (box 4-5). If the results of the studies are homogeneous, however, the two approaches give the same result.

The assumptions underlying the fixed-effects model are that all studies are estimating the same treatment effect and that the difference in effect observed across studies results by chance. The assumptions underlying the random-effects model are that the treatment effect truly differs across studies and that the goal is to determine the average of the different effects. Fixed-effects models have been used frequently in the past and still have some strong advocates (39,77), although the use of random-effects models to summarize the treatment effect has been favored recently (20,34,72).

Critics of random-effects models (39) question the assumption underlying the model: that the studies were sampled from a hypothetical universe of studies where the true treatment effect varies. They also note that the meaning of random-effects summaries are often misinterpreted. (The correct interpretation is that the random-effects treatment effect is an estimate of the average treatment effect in the universe of hypothetical studies with differing treatment effects.) Proponents of random-effects models argue that they are appropriately imprecise when heterogeneity is present (34).

Publication Bias

Publication bias refers to the fact that results are more likely to be published if they are statistically significant than if they are not (3,23,27). The likelihood of publication is also greater for results from large studies or results that are perceived as important (27). The exclusion of unpublished study results thus can cause the results of a meta-analysis (or any literature review) to be misleading (3).

Informal graphic methods of detecting publication bias have been proposed (58). These methods are easy to use and are widely employed, although their sensitivity and specificity are unknown.

Formal statistical methods for detecting and assessing publication bias have also been proposed (5,48,81), but experts disagree about which formal statistical approach is best (49). Some of these methods are more easily implemented (81) than others (5,48). An advantage of the more computationally intensive methods (5,48) is that they can be used to estimate the true effects of treatment (what would have been observed if there were no publication bias). Estimating the true effects of treatment or determining the number of negative studies necessary for canceling out a positive finding in a meta-analysis is possible, however, only if assumptions are made, and these assumptions may be untestable or not entirely reasonable. The Iyengar and Greenhouse (48) approach, for example, relies on the assumption that the often inappropriate fixed-effects model is used for summarizing treatment effect and that only results significant in one direction[7] are published, which apparently is not entirely true (82). Berlin's approach relies on the assumption that a study's size is unrelated to the size of the treatment effect, which may be incorrect, as well (5). Thus, the formal approaches to assessing publication bias are useful but imperfect solutions to the problem.

The definitive method of correcting publication bias is to include all unpublished results in a meta-analysis, subjecting them to the same inclusion criteria and quality scoring methods as published studies. Accordingly, some authors suggest that analysts routinely attempt to include all the relevant unpublished data in meta-analyses (35, 100).

However, tracing unpublished studies can be difficult. Registries of studies undertaken in a given field are becoming more common (26), but where registries are unavailable, the inclusion of every unpublished study may be impossible for lack of information (22) or may be infeasible on practical grounds (97). When unpublished results can be easily obtained, their inclusion in a meta-analysis, at least in a sensitivity analysis, is reasonable.

Because of a concern that unpublished results may be less reliable than published ones, including unpublished results to combat publication bias is not universally accepted (4).[8] This concern is

[7] Treatments, if they have an effect, can work in two directions: they can either benefit or harm patients. "Statistically significant in one direction" means, for example, that only these studies are published that show the treatment decreased deaths.

[8] This would cause publication bias if unpublished studies tended not only to show no treatment effect but also gave biased estimates of the treatment effect.

probably unwarranted: there is no strong evidence that a study's quality is related to its publication (22,27), and the quality of the component studies may be assessed and considered in a meta-analysis.

Protocol Controversies

The third major category of issues regarding meta-analysis concerns the details of the meta-analytic protocol—the specific procedures followed when conducting the meta-analysis.

Variations in the Standard Meta-Analytic Protocol

Chalmers (12) has long advocated the use of blinding in evaluating the studies for a meta-analysis: to minimize bias in the evaluation process, identifying information is obscured on each article that is potentially eligible for inclusion in the meta-analysis. Thus, the names of the authors, where they did their study, whether they found an effect of treatment, and other pieces of information that might bias an assessment are not available to the person who determines whether to include the study in the meta-analysis. Blinding also helps ensure an unbiased evaluation of the study's quality. Nonetheless, the added assurance that studies are chosen and evaluated in an unbiased fashion comes at a price. If a large number of studies must be reviewed, the blinding process can substantially increase the cost of the meta-analysis. Although the theoretical justification for such blinding is understandable, its actual benefit has not been studied.

Chalmers also recommends that two persons independently evaluate the quality of the studies in a meta-analysis (12). This practice serves as a form of quality control, but its cost-effectiveness has not been documented.

Some statisticians who have spent considerable time thinking about the analytic issues in meta-analysis have recommended that procedures to correct bias be implemented (39,70,84). In practice, however, few investigators have done so. This fact stems partly from a lack of the information necessary for correcting the bias. More important, as mentioned previously, the validity of bias-correction procedures depends on strong assumptions that may be untestable, wrong, or controversial. Furthermore, bias-correction procedures complicate the analysis and may decrease the understandability of its results. Nonetheless, such procedures may be the only sensible approach when optimal analysis of flawed data is a priority.

How best to incorporate assessments of the quality of the component studies in a meta-analysis has not been resolved. One issue is whether a summary measure of a study's quality should be used. Another issue is how to use the summary measure.

The debate as to whether a summary measure of a study's quality has a use in meta-analysis stems from uncertainty about whether such measures reliably and validly identify biased studies. The specific weaknesses that bias a study's observation of the treatment effect are often unknown. A summary score that reflects what the meta-analyst suspects are specific flaws in a study thus may exclude information that is, in fact, related to the findings. If certain studies in a meta-analysis are given less weight because of the analyst's impression that an irrelevant aspect of the study was not ideally conducted, the results of the meta-analysis may be misleading (92). Also, the summary score of quality may contain too much "noise" to adequately reflect the problem of interest.[9] Furthermore, because of space limitations in publications, authors may not present enough information for the quality of the study to be fully assessed; this problem is particularly acute for those attempting to assess observational studies.

Rubin (84) has suggested that individual features of the quality of the studies be examined in relation to the treatment effect, so that important features can be identified. One possible compo-

[9] Noise refers (to the random variation that may obscure the general trend or characteristics of the Item of interest.

nent of a summary measure of quality, for example, is whether the study participants were blinded to the treatment they received. Analysts could examine whether the treatment effect found in studies with blinding differed from that found in studies where blinding was not used. Although this approach is sensible, it is of limited use in practice: the attributes of different studies are often so highly correlated and the numbers of studies so limited that analysts have dificulty linking specific aspects of the quality of a study to the size of the treatment effect observed.

How best to incorporate a summary measure of the quality of a study into a meta-analysis is also unclear. Detsky (21) has outlined the major options for using the information. The first option is to exclude poor studies from the analysis. The second option is to weight a given study not only according to its statistical precision, but also according to its quality. Finally, the quality scores may be included as terms in statistical models or serve as the basis for sensitivity or subgroup analyses. A consensus about which of these methods is theoretically preferable has not emerged in the literature on meta-analysis.

Bayesian Meta-Analysis

A Bayesian approach to meta-analysis (box 4-6) is strongly supported by some investigators (29). Few meta-analysts have used Bayesian methods and few empirical comparisons between the results from the Bayesian and traditional methods have been presented (79).

Bayesian methods have three potential advantages. First, the statistical results are more easily interpreted than are those from the traditional fre-

BOX 4-6: Frequentist and Bayesian Approaches to Data Analysis

The frequentist and Bayesian approaches to data analysis are two different ways to use data to make inferences about the treatment effect, The frequentist approach is more prevalent throughout the sciences, though the use of the Bayesian approach is growing.[1]

The frequentist approach assumes that a given study could hypothetically be repeated an infinite number of times, and that the particular treatment effect observed in the study actually done is just one of all possible observations, selected at random.

In the Bayesian approach to statistics, the analyst specifies in quantitative terms his or her belief (and certainty in that belief) about the size of the treatment effect under investigation, and the observations made in a particular study are used to modify the analyst's belief.

A frequentist, at the end of the data analysis, specifies an estimate of the size of the treatment effect (based only on the data in the study performed) and also presents a p-value or an equivalent 95-percent confidence interval (see box 4-2). This confidence interval describes statistically the interval within which the true effect of the treatment would lie in 95 of 100 hypothetical repetitions of the experiment. Because most studies cannot be repeated multiple times, the assumptions upon which the statistics are based cannot be verified directly,

A Bayesian, at the end of the data analysis, specifies an estimate of the size of the treatment effect and an interval in which he or she believes with 95-percent certainty the true treatment effect lies. This approach gives validity to the analyst's subjective beliefs, a controversial issue behind some of the resistance to broader use of Bayesian statistics in the sciences,

[1] For an in-depth discussion of these two approaches, see **Oakes (73).**

SOURCE: Matthew Longnecker, 1995.

quentist approach (box 4-6) (38). Second, when adjusting treatment effects for bias, the analysts may incorporate their degrees of certainty or uncertainty about the adjustment into the analysis. Third, greater flexibility is possible when combining different types of information about the treatment effect.

The Bayesian approach also has three disadvantages. First, even fewer people understand Bayesian methods than understand the frequentist approach. Second, performing Bayesian analyses requires specialized computer software that has only recently become widely available (29). Third, because Bayesian analyses can be based on even more assumptions than can frequentist analyses, the Bayesian results maybe subject to more debate. Once empirical comparisons of the two methods are available and more investigators have experience with Bayesian methods, the relative merits of the approach will be easier to assess.

FUTURE APPLICATIONS

Meta-analysis is gaining in popularity, especially in the medical field. The tool has been used frequently for assessing technology and promises to be useful for improving assessments of risk and, by strengthening the estimates of the effects of treatments, for increasing the accuracy of cost-effectiveness analyses.

The number of meta-analyses conducted each year is growing. Dickersin and her associates (24), in their examination of the literature, found three meta-analyses published between 1966 and 1969, nine published between 1976 and 1978, and 44 published between 1985 and 1987. Seventy percent of these meta-analyses were on medical topics. The computerized database of the National Library of Medicine began formally identifying meta-analyses and related work in 1989. A computerized search for articles relating to meta-analysis in that database resulted in 232 articles for the year 1989,297 articles for 1990, and 368 articles for 1991. Although only a portion of these articles are themselves meta-analyses (many are merely *about* meta-analysis), the increasing prominence of this tool in the medical literature is evident.

The use of meta-analyses is also growing. For example:

- Influential medical professionals use evidence from meta-analyses to evaluate treatment efficacy (57,67,91).
- The Food and Drug Administration allows the results of meta-analyses to support New Drug Applications (34).
- The U.S. General Accounting Office (GAO) has endorsed meta-analysis as a method of assessing treatment efficacy (93).
- The Agency for Health Care Policy and Research is using meta-analyses to guide policy regarding medical procedures that will be reimbursable under Medicare (23).

GAO (93) has proposed that meta-analyses combining results from randomized clinical trials and "database analyses" be conducted. 10 The justification for combining results among studies conducted using different designs is that randomized clinical trials tend to measure the treatment effects in only small subsets of all the types of subjects who might receive the treatments in practice. Database analyses provide an estimate of the treatment effect in a much more diverse group of subjects, and they reflect the effect of treatment as administered by physicians in general, not just those specialists conducting clinical trials. They are also observational studies, however, and an estimated treatment effect based on observational data alone is often not reliable (see J. Whittle, "Analysis of Large Administrative Databases," background paper #2).

GAO has named this type of data synthesis "cross-design synthesis." The technique entails combining the results from studies with different

[10] In "database analyses" (as the term is used by GAO (91)), information about patients—including the treatments they have received and he outcomes of those treatments—from computer records routinely kept for accounting purposes is analyzed to provide estimates of the treatment effects.

designs and analyzing raw data from the database(s) as part of the analysis, whereas meta-analysis entails simply analyzing the results of studies. The validity of cross-design synthesis will be even more difficult to establish than the validity of traditional meta-analysis has been. Still, like meta-analysis in general, cross-design synthesis might sometimes facilitate more efficient use of existing data than is possible with the traditional narrative approach to evaluating the effects of treatments.

The potential for meta-analysis to improve risk assessments [11] has been recognized by several observers (32,87). Meta-analysis may improve the accuracy of cost-effectiveness analyses and can identify effective therapies, gauge the treatment effect, and estimate other quantities that influence cost-effectiveness (88).

Because the meta-analytic approach can be applied to virtually any problem in the evaluation of medical technology that has been previously studied (28), its use in the health care field can be expected to increase, but the benefit of a given meta-analysis in relation to its cost deserves critical evaluation. The cost of a meta-analysis depends on the number of potentially eligible studies, the number of admissible studies, the use of blinding (or lack there of), the usability of the format in which the data have been presented, the experience of the analysts, the number of decisions that the analysts must make, and other factors.

For meta-analysis to be beneficial, its results must be persuasive. The results of a meta-analysis are most likely to be persuasive where there is little controversy about how it was done or how its results should be interpreted. The credibility of a meta-analysis is likely to be greatest when the approach is applied to clearly combinable, homogeneous results from methodologically strong randomized clinical trials that were identified through a registry of all trials conducted on a given research question. As the circumstances of a meta-analysis depart from this ideal, the validity of its results will be less clear and increasingly difficult to assess. Even when the results of a meta-analysis are controversial, however, they may provide insights into data not attainable with traditional review methods. Several authors have suggested criteria to be used in evaluating the results of a meta-analysis (53,85), and these may assist an evaluation of a meta-analytic result. With or without such guidelines, the evaluator must have subject-matter expertise to fully appreciate the worth of a given meta-analysis.

CONCLUSION

Despite the controversies, meta-analysis appears to be generally accepted as a useful tool for analyzing data from, at least, randomized clinical trials (51,89). Yet unquestioning reliance on the results of meta-analysis (67) has been criticized (55), because despite the advantages of meta-analysis' explicit, formal approach, the results of meta-analyses are still influenced by the sometimes fallible judgments of their authors (93).

The usefulness of meta-analysis may best be considered on a case-by-case basis. Where the setting is ideal for a careful meta-analysis, the method may accelerate and aid the evaluation of health care technologies and practices. Where the setting is less than ideal, the method may help investigators to identify the combination of treatment and participant characteristics where the efficacy is greatest or the circumstances under which more or better data regarding a treatment effect are needed.

Although the results of meta-analyses may reduce the number of randomized controlled trials needed to evaluate a technology (57), meta-analy - sis should not eclipse the need for randomized trials. In fact, a meta-analysis may clarify the need for a trial when the meta-analytic result suggests that a treatment effect is present but the estimate of the effect is imprecise.

[11] Risk assessment, according to Last (56), is "the qualitative or quantitative estimation of the likelihood of adverse effects that may result from exposure to specified health hazards or from the absence of beneficial influences."

Meta-analysis, properly done, requires significant resources, including access to experts in the specific technique and in the subject being studied. Since identifying relevant studies is one of the most time-consuming steps, the systematic registration of randomized clinical trials (and other studies) could improve the efficiency of this technique.

REFERENCES

1. Antman, E. M., Lau, J., Kupelnick, B., et al., "A Comparison of Results of Meta-Analyses of Randomized Control Trials and Recommendations of Clinical Experts: Treatment for Myocardial Infarction," *Journal of the American Medical Association 268(2):240-248, 1992.*
2. Beecher, H. K., "The Powerful Placebo," *Journal of the American Medical Association 159(17): 1602-1606, 1955.*
3. Begg, C. B., and Berlin, I. A., "Publication Bias: A Problem in Interpreting Medical Data," *Journal of Royal Statistical Society* 151(3):419-463, *1988.*
4. Begg, C. B., and Berlin, J. A., "Publication Bias and Dissemination of Clinical Research," *Journal of the National Cancer Institute 81(2): 107-115, 1989.*
5. Berlin, J.A., Begg, C. B., and Louis, T. A., "An Assessment of Publication Bias Using a Sample of Published Clinical Trials," *Journal of the American Statistical Association* 84(406):381-392, *1989.*
6. Berlin, J. A., and Colditz, G. A., "A Meta-Analysis of Physical Activity in the Prevention of Coronary Heart Disease," *American Journal of Epidemiology* 132(4):612-628, *1990.*
7. Berlin, J. A., Laird, N. M., Sacks, H. S., et al., "A Comparison of Statistical Methods for Combining Event Rates from Clinical Trials," *Statistics in Medicine 8(2): 141-151,* 1989.
8. Berlin, J. A., Longnecker, M. P., and Greenland, S., "Meta-Analysis of Epidemiologic Dose-Response Data," *Epidemiology* 4:218-*288, 1993.*
9. Boden, W. E., "Meta-Analysis in Clinical Trials Reporting: Has a Tool Become a Weapon?" (editorial), *American Journal of Cardiology* 69(6):681-686, *1992.*
10. Brown, S. A., "Measurement of Quality of Primary Studies for Meta-Analysis," *Nursing Research* 40(6):352-355, *1991.*
11. Bulpitt, C.J., "Meta-Analysis," *Lancet* 2(8602):93-94, *1988.*
12. Chalmers, T. C., "Problems Induced by Meta-Analyses," *Statistics in Medicine* 10(6):971-979, *1991.*
13. Chalmers, T. C., Berrier, J., Sacks, H. S., et al., "Meta-Analysis of Clinical Trials as a Scientific Discipline. II: Replicate Variability and Comparison of Studies that Agree and Disagree," *Statistics in Medicine* 6:733-744, *1987.*
14. Chalmers T. C., Hewett P., Reitman D., et al., "Selection and Evaluation of Empirical Research in Technology Assessment," *International Journal of Technology Assessment in Health Care* 5(4):521-536, *1989.*
15. Chalmers, T. C., Levin, H., Sacks, H. S., et al., "Meta-Analysis of Clinical Trials as a Scientific Discipline. I: Control of Bias and Comparison with Large Cooperative Trials," *Statistics in Medicine* 6:315-325,1987.
16. Chalmers, T. C., Smith, Jr., H., Blackburn, B., et al., "A Method for Assessing the Quality of a Randomized Control Trial," Con*trolled Clinical Trials* 2(l):31-49, *1981.*
17. Checkoway, H., "Data Pooling in Occupational Studies," *Journal of Occupational Medicine* 33(12):1257-1260, *1991.*
18. Choi, B., "Meta-Analysis" (letter), *Annals of Internal Medicine* 109(l):83, *1988.*
19. Colditz, G. A., Miller, J. N., and Mosteller, F., "How Study Design Affects Outcomes in Comparisons of Therapy. I: Medical," *Statistics in Medicine* 8(4):441-454, *1989.*
20. DerSimonian, R., and Laird N., "Meta-Analysis in Clinical Trials," *Controlled Clinical Trials* 7:177-188, *1986.*

21. Detsky, A. S., Naylor, C.D., O'Rourke, K., et al., "Incorporating Variations in the Quality of Individual Randomized Trials into Meta-Analysis,'' *Journal of Clinical Epidemiology* 45(3):255-265, *1992.*

22. Dickersin, K., "The Existence of Publication Bias and Risk Factors for Its Occurrence," *Journal of the American Medical Association 263(10): 1385-1389, 1990.*

23. Dickersin, K., and Berlin, J. A., "Meta-Analysis: State-of-the-Science," *Epidemiologic Reviews* 14:154-176, *1992.*

24. Dickersin, K., Higgins, K., and Meinert, C. L., "Identification of Meta-Analyses: The Need for Standard Terminology," *Controlled Clinical Trials* 11(l):52-66, *1990.*

25. Durlak, J. A., and Lipsey, M. W., "A Practitioner's Guide to Meta-Analysis," *American Journal of Community Psychology* 19(3): *291-332, 1991.*

26. Easterbrook, P.J., "Directory of Registries of Clinical Trials," *Statistics in Medicine* 11(3):345-359, *1992.*

2'7. Easterbrook, P.J., Berlin, J. A., and Gopalan, R., et al., "Publication Bias in Clinical Research," *Lancet* 337(8746):867-872, *1991.*

28. Eddy, D. M., Hasselblad, V., and Shachter, R., "An Introduction to a Bayesian Method for Meta-Analysis: The Confidence Profile Method," *Medical Decision Making* 10(1): 15-23, 1990.

29. Eddy, D. M., Hasselblad, V., and Shachter, R., *Meta-Analysis by the Confidence Profile Method: The Statistical Synthesis of Evidence* (Boston, MA: Academic Press, 1992).

30. Einarson, T.R., Leeder, J. S., and Koren, G., "A Method for Meta-Analysis of Epidemiological Studies," *Drug Intelligence and Clinical Pharmacy* 22(10):81 *3-824, 1988.*

31. Ellenberg, S. S., "Meta-Analysis: The Quantitative Approach to Research Review," *Seminars in Oncology* 15(5):472-481,1988.

32. Farland, W. H., "The U.S. Environmental Protection Agency's Risk Assessment Guidelines: Current Status and Future Directions," *Toxicology and Industrial Health* 8(3):205-212, *1992.*

33. Felson, D.T., "Bias in Meta-Analytic Research," *Journal of Clinical Epidemiology* 45(8):885-892, *1992.*

34. Fleiss, J.L., and Gross, A.J., "Meta-Analysis in Epidemiology, with Special Reference to Studies of the Association Between Exposure to Environmental Tobacco Smoke and Lung Cancer: A Critique," *Journal of Clinical Epidemiology 44(2): 127-139, 1991.*

35. Furberg, C. D., and Morgan, T. M., "Lessons from Overviews of Cardiovascular Trials," *Statistics in Medicine* 6(3):295-306, *1987.*

36. Gelber, R. D., and Goldhirsch, A., "Meta-Analysis in Clinical Research" (letter), *Annals of Internal Medicine 108(1): 158-159, 1988.*

37. Glass, G. V., "Primary, Secondary, and Meta-Analysis of Research," *Educational Researcher* 5:3-8, *1976.*

38. Goodman, S. N., "Meta-Analysis and Evidence," *Controlled Clinical Trials 10:188-204, 1989.*

39. Greenland, S., "Quantitative Methods in the Review of Epidemiologic Literature," *Epidemiologic Reviews 9: 1-30, 1987.*

40. Greenland, S., and Longnecker, M. P., "Methods for Trend Estimation from Summarized Dose-Response Data, with Applications to Meta-Analysis," *American Journal of Epidemiology* 135: 1301-1309, 1992.

41. Greenland, S., and Salvan, A., "Bias in the One-Step Method for Pooling Study Results," *Statistics in Medicine* 9(3):247-252, *1990.*

42. Hedges, L. V., and Olkin I., *Statistical Methods for Meta-Analysis* (Orlando, FL: Academic Press, 1985).

43. Held, P. H., Yusuf, S., and Furberg, C. D., "Calcium Channel Blockers in Acute Myocardial Infarction and Unstable Angina: An Overview," *British Medical Journal 299* (6709):1187-1192, *1989.*

44. Henry, D.A., and Wilson, A., "Meta-Analysis Part 1: An Assessment of Its Aims, Validity and Reliability," *Medical Journal of Australia* 156(l):31-38, 1992.

45. Hewitt, P., and Chalmers, T.C., "Using MEDLINE To Peruse the Literature," *Controlled Clinical Trials* 6:75-83, *1985.*

46. Holroyd, K. A., and Penzien, D. B., "Meta-Analysis Minus the Analysis: A Prescription for Confusion," *Pain* 39(3):359-363, *1989.*

47. Howe, G., Rohan, T., Decarli, A., et al., "The Association Between Alcohol and Breast Cancer Risk: Evidence from the Combined Analysis of Six Dietary Case-Control Studies," *International Journal of Cancer* 47:707-710, *1991.*

48. Iyengar, S., and Greenhouse, J. B., "Selection Models and the File Drawer Problem," *Statistical Science 3(l): 109-117, 1988.*

49. Iyengar, S., and Greenhouse, J. B., "Rejoinder," *Statistical Science 3(l): 133-135, 1988.*

50. Jenicek, M., "Meta-Analysis in Medicine: Where We Are and Where We Want to Go," *Journal of Clinical Epidemiology* 42(l): *35-44, 1989.*

51. Kassirer, J. P., "Clinical Trials and Meta-Analysis: What Do They Do for Us?" (editorial), New *England Journal of Medicine 327(4): 273-274, 1992.*

52. Kriebel, D., Wegman, D. H., Moure-Erase, R., et al., "Limitations of Meta-Analysis: Cancer in the Petroleum Industry" (letter), *American Journal of Industrial Medicine* 17(2):269-271, *1990.*

53. L'Abbe, K. A., Detsky, A. S., and O'Rourke, K., "Meta-Analysis in Clinical Research," *Annals of Internal Medicine* 107:224-233, *1987.*

54. Laird, N. M., and Mosteller, F., "Some Statistical Methods for Combining Experimental Results," *International Journal of Technology Assessment in Health Care* 6(l):5-30, *1990.*

55. Lancet, "Cross Design Synthesis: A New Strategy for Studying Medical Outcomes" (editorial), *Lancet* 340:944-946, *1992.*

56. Last, J. M., *A Dictionary of Epidemiology (New* York, NY: Oxford University Press, 1988).

57. Lau, J., Antman, E. M., Jimenez-Silva, J., et al., "Cumulative Meta-Analysis of Therapeutic Trials for Myocardial Infarction," New *England Journal of Medicine 327(4): 248-254, 1992.*

58. Light R.J., and Pillemer D. B., *Summing Up* (Cambridge, MA: Harvard University Press, 1984).

59. Light, R. J., and Smith, P. V., "Accumulating Evidence: Procedures for Resolving Contradictions Among Different Research Studies," *Harvard Education Review 41(4):429-471, 1971.*

60. Longnecker, M. P., Berlin, J. A., Orza, M.J., et al., "A Meta-Analysis of Alcohol Consumption in Relation to Risk of Breast Cancer," *Journal of the American Medical Association* 260(5):652-665, *1988.*

61. Longnecker, M.P., Martin-Moreno, J. M., Knekt, P., et al., "Serum a-Tocopherol Concentration in Relation to Subsequent Colorectal Cancer: Pooled Data from Five Cohorts," *Journal of the National Cancer Institute* 84:430-435, *1992.*

62. Louis, T. A., "Assessing, Accommodating, and Interpreting the Influences of Heterogeneity," *Environmental Health Perspectives* 90:215-222, *1991.*

63. Louis, T.A., Fineberg, H. V., Mosteller, F., "Findings for Public Health from Meta-Analysis," *Annual Review of Public Health* 6:1-20, *1985.*

64. MacMahon, B., and Hutchison, G. B., "Prenatal X-ray and Children's Cancer: A Review," *Acta Unio Internationalism Contra Cancrum 20: 1172-1174, 1964.*

65. Malone, M. D., and Strube, M.J., "Meta-Analysis of Non-Medical Treatments for Chronic Pain," *Pain* 34(3):231-244, *1988.*

66. Mann, C., "Meta-Analysis in the Breech," *Science* 249:476-480, *1990.*

67. Manson, J. E., Tosteson, H., Ridker, P. M., et al., "The Primary Prevention of Myocardial

In farction," New *England Journal of* Medicine 326(21): 1406-1416, 1992.

68. Meier, P., "Commentary," *Statistics* ***in*** *Medicine* 6:329-331, *1987.*

69. Meinert, C. L., "Meta-Analysis: Science or Religion?," *Controlled Clinical Trials 10(4 suppl.):257S-263S, 1989.*

70. Mosteller, F., "Summing Up," *The Future of* ***Meta-Analysis,*** K.W. Wachter, and M.L. Straf (eds.) (New York, NY: Russell Sage Foundation, 1990).

71. Mulrow, C. D., "The Medical Review Article: State of the Science," *Annals of Internal Medicine* 106(3):485-488, *1987.*

72. National Research Council, *Combining Information: Statistical Issues and Opportunities for Research* (Washington, DC: National Academy Press, 1992).

73. Oakes, M., *Statistical Inference* (Chestnut Hill, MA: Epidemiology Resources, Inc., 1990).

74. Oxman A. D., and Guyatt G. H., "Validation of an Index of the Quality of Review Articles," *Journal of Clinical Epidemiology* 44(1 1):1271-1278, 1991.

75. Pearson, K., "Report on Certain Enteric Fever Inoculation Statistics," *British Medical Journal* 2:1243-1246, *1904.*

76. Pete, R., "Why Do We Need Systematic Overviews of Randomized Trials?" *Statistics in Medicine* 6(3):233-240, *1987.*

77. Pete, R., "Discussion," *Statistics in* ***Medicine*** 6(3):241-244, *1987.*

78. Pignon, J. P., Arriagada, R., Ihde, D. C., et al., "A Meta-Analysis of Thoracic Radiotherapy for Small-Cell Lung Cancer," New *England Journal of Medicine 327(23): 1618-1624, 1992.*

79. Pollard, W. E., ***Bayesian*** *Statistics for Evaluation Research: An Introduction* (Beverly Hills, CA: Sage Publications, 1986).

80. Robins, J., Breslow, N., and Greenland, S., "A Mantel-Haenszel Variance Consistent Under Both Large Strata and Sparse Data Limiting Models," *Biometrics* 42:31 *1-323, 1986.*

81. Rosenthal, R., "The File Drawer Problem, and Tolerance for Null Results," ***Psychological*** *Bulletin* 86(3):638-641, *1979.*

82. Rosenthal, R., and Rubin, D. B., "Comment: Assumptions and Procedures in the File Drawer Problem," *Statistical Science 3(l): 120-125, 1988.*

83. Rothman, K.J., *Modern Epidemiology* (Boston, MA: Little, Brown and Company, 1986).

84. Rubin, D. B., "A New Perspective," *The Future of* ***Meta-Analysis,*** K.W. Wachter, and M.L. Straf (eds.) (New York, NY: Russell Sage Foundation, 1990).

85. Sacks, H. S., Berrier, J., Reitman, D., et al., "Meta-Analysis of Randomized Controlled Trials," New *England Journal of Medicine* 316:450-455, *1987.*

86. Schoones, J.W., "Searching Publication Data Bases" (letter), ***Lancet*** 335(8687):481, 1990.

87. Shore, R. E., Lyer, V., Altshuler, B., et al., "Use of Human Data in Quantitative Risk Assessment of Carcinogens: Impact on Epidemiologic Practice and the Regulatory Process," *Regulatory Toxicology and Pharmacology 15: 180-221, 1992.*

88. Simes, J., "Meta-Analysis: Its Importance in Cost-Effectiveness Studies," *Medical Journal* ***of Australia*** *153* (suppl.):S13-S16, *1990.*

89. Spitzer, W.O., "Meta-Analysis: Unanswered Questions About Aggregating Data" (editorial), *Journal of Clinical Epidemiology* 44(2):103-107, *1991.*

90. Teagarden, J. R., "Meta-Analysis: Whither Narrative Review?" ***Pharmacotherapy*** *9(5): 274-281, 1989.*

91. Thacker, S. B., "Meta-Analysis. A Quantitative Approach to Research Integration," *Journal of the American Medical Association* 259(11):1685-1689, *1988.*

92. Thompson, S.G., and Pocock, S.J., "Can Meta-Analyses Be Trusted?" *Lancet 338* (8775):1 *127-1130, 1991.*

93. U.S. General Accounting Office, *Cross Design Synthesis: A New Strategy for Medical Effectiveness Research* (Washington, DC: U.S. Government Printing Office, 1992).

94. U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, National Library of Medicine, *Medical Subject Headings, Annotated Alphabetic List* (Bethesda, MD: 1989).

95. U.S. Department of Health and Human Services, Public Health Service, Office of the Assistant Secretary for Health, Office of Disease Prevention and Health Promotion, Preventive Services Task Force, *Guide to Clinical Preventive Services: An Assessment of the Effectiveness of 169 Interventions* (Baltimore, MD: Williams & Wilkins, 1989).

96. Wachter, K. W., "Disturbed by Meta-Analysis?" *Science* 241(4872):1407-1408, 1988.

97. Wachter, K. W., "Concepts Under Scrutiny: Discussion," *The Future of Meta-Analysis,* K.W. Wachter and M.L. Straf (eds.) (New York, NY: Russell Sage Foundation, 1990).

98. Whitehead, A., and Whitehead, J., "A General Parametric Approach to the Meta-Analysis of Randomized Clinical Trial s," *Statistics in Medicine* 10(1 1):1665-1677, 1991.

99. Wilson, A., and Henry, D. A., "Meta-Analysis Part 2: Assessing the Quality of Published Meta-Analyses," *Medical Journal of Australia 156(3): 173-174, 1992.*

100. Yusuf, S., "Obtaining Medically Meaningful Answers from an Overview of Randomized Clinical Trials," *Statistics in Medicine* 6(3):281-294, *1987.*

101. Yusuf, S., Collins, R., and Pete, R., et al., "Intravenous and Intracoronary Fibrinolytic Therapy in Acute Myocardial Infarction: Overview of Results on Mortality, Reinfarction and Side-Effects from 33 Randomized Controlled Trials," *European Heart Journal* 6:556-558, *1985.*

APPENDIX 4-A: QUANTITATIVE METHODS IN META-ANALYSIS

The quantitative methods appropriate for any analysis, including a meta-analysis, depend on the research question to be addressed. Although meta-analysis has many uses in health care-e. g., calculations across studies of the average value of a laboratory result, a disease rate, a population characteristic, or the sensitivity of a diagnostic test—the discussion here focuses on meta-analytic techniques for evaluating the effect of a medical treatment on an outcome.

Determining the Treatment Effect in a Single Study

Calculating the Size of the Treatment Effect

In evaluating a controlled trial of a medical treatment, the outcome measure in the treatment and control groups is usually expressed as a proportion, rate, or mean. One might be interested in a drug's effect, for example, on the proportion of participants who develop side effects, on the rate at which subjects die or develop a disease, * or on their mean level of cholesterol. In the meta-analysis of fibrolytic therapy performed by Yusuf and his associates, the outcome measure for the treatment and control groups in each of the component studies was the proportion of patients with myocardial infarction who died within a specified period of time.

The treatment effect in a study is the outcome measure in the treated group compared with that in the control group. The comparison between outcome measures may be a difference, a ratio, or a related measure. Where the outcome is a proportion, one might be interested, for example, in the difference in the proportion who died in the treated and control groups (the proportion dead in the treatment group minus the proportion dead in the control group); in the ratio of these proportions (the proportion dead in the treatment group divided by the proportion dead in the control group); or in the difference in the rates (rate difference) or the ratio of the rates (rate ratio) between the treatment and control groups. Where the outcome is a mean, the treatment effect usually examined is the difference of the means in the treated and control groups.

In practice, the treatment effect is often expressed as: 1) the difference between the observed and expected number of deaths in the treatment group, and 2) the odds of death in the treatment group divided by the odds of death in the control group (odds ratio). (Odds are related to proportions in that a proportion divided by one minus the proportion is the odds.) Neither expression is a simple example of a difference or a ratio of outcome measures, but they are worth explaining in detail because they are commonly used in the evaluation of medical therapies.

The first treatment effect commonly measured is the difference between the observed and expected numbers of deaths in the treatment group. The outcome in the Yusuf meta-analysis of fibrinolytic therapy was a proportion (the number of deaths in a treatment or control group divided by the number of subjects in that group) (table 4-A-l). The observed number of deaths is the numerator in the proportion. Thus, for the Fletcher study (the first study shown in table 4-A-l), the observed number of deaths in the treatment group was one, and the total number of study participants in the treatment group was 12. Thus, the proportion of observed deaths in the treated group was 1/12, or 8.3 percent. If the treatment had no effect, the proportions of deaths in the treatment and control groups could be expected to be the same.

The authors of the meta-analysis computed the proportion of deaths that would be expected in the treatment group if the treatment had no effect. This was accomplished by combining the number of deaths in the treatment and control groups, then dividing by the number of participants in both

[1] A death rate is calculated as the number of deaths per unit of person-time. If 100 study participants are observed for 2 years, for example, and one of them dies during that period, the death rate is 1/200 person-years. The proportion of participants who die in the 2-year period is 1/100.

TABLE 4-A-1: Summary of Study Results on Fibrinolytic Agents and Survival After Myocardial Infarction

study		Treated Group		Control Group			
No. *(1)*	Author (2)	Deaths *(3)*	Total (4)	Deaths *(5)*	Total (6)	0- E[a] *(7)*	V(O - E)[b] *(8)*
1	Fletcher	1	12	4	11	-1.6	1.0
2	Dewar	4	21	7	21	-1.5	2.1
3	1st European	20	83	15	84	2.6	7.0
4	Heikenheimo	22	219	17	207	2.0	8.9
5	Austrian	37	352	65	376	-1 2.3[c]	21.9
6	Italian	19	164	18	157	0.1	8.2
7	Australian	51	376	63	371	-6.4	24.2
8	NHLBI SMIT	7	53	3	54	2.0	2.3
9	Frank	6	55	6	53	-0.1	2.7
10	Valere	11	49	9	42	0.2	3.9
11	UK	48	302	52	293	-2.8	20.8
12	Witchitz	5	32	5	26	-0.5	2.1
13	Lasierra	1	13	3	11	-1.2	0.9
14	3rd European	25	156	50	159	-12.1C	14.3
15	Olson	5	28	5	24	-0.4	2.0
16	Schreiber	1	19	4	19	-1.5	1.1
17	2nd European	69	373	94	357	-1 4.3[c]	31.7
18	2nd Frankfurt	13	102	29	104	-7.8[c]	8.4
19	Klein	4	14	1	9	1.0	1.0
20	N. German	63	249	51	234	4.2	21.8
21	Lipshultz	6	43	7	41	-0.7	2.8
22	Gormsen	2	14	3	14	-0.5	1.1
23	Brochier	2	60	8	60	-3.0[c]	2.3
24	European	41	172	34	169	3.2	14.7
	Totals	463	2,961	553	2.896	-51.4[c]	207.1

[a] O - *E* refers to the difference between the observed and the expected number of deaths in the treated group (see main text).

[b] *V(O - E)* refers to the variance of O-E.

[c] p <0.05

SOURCE: Adapted from S. Yusuf, R. Collins, R. Pete, et al., "Intravenous and Intracoronary Fibrinolytic Therapy in Acute Myocardial Infarction: Overview of Results on Mortality and Side-effects from 33 Randomized Controlled Trials, " **European** *Heart Journal* 6:556-558, **1985. Only data from studies of the effect of intravenous streptokinase are shown,**

groups combined. For the Fletcher study, this number is 5/23 (i.e., (1 +4)/(12 + 1 l)), or 21.7 percent. If the treatment had no effect, 21.7 percent of the subjects in the treatment and control groups should have died. The expected number of deaths in the treated group is 21.7 percent of 12, or 2.6. Expressed as a formula,

$$E = rid/N,$$

where:

- *E* is the expected number of deaths in the treatment group if there were no treatment effect,
- N is the total number of participants in the trial,
- *n* is the number of treated participants, and
- ***d*** is the number of deaths in the treated and control groups combined.

Thus, among the treated participants, one case was observed and 2.6 were expected. The difference, -1.6 (i.e., 1- 2.6), is the treatment effect for the Fletcher study (see table 4-A-l); it suggests that treatment reduced (by 1.6)[13] the number of deaths that occurred in the treatment group. Calculating the difference in the proportions of deaths

[13] **One cannot, of course, actually** reduce a fraction of a real death. Statistically, however, one is assuming that **the 12 treated patients in the** study are representative of a larger population. If that population were sampled many times, drawing a sample of 12 people each time, on **average** the deaths in each sample would be reduced by 1.6.

TABLE 4-A-2: Results on Fibrinolytic Agents and Survival After Myocardial Infarction with Intermediate Calculations for Homogeneity Chi-Square and Fixed-Effects Model

		Treated		Control							
Study No. *(1)*	Author *(2)*	Deaths *(3)*	Total *(4)*	Deaths *(5)*	Total *(6)*	OR_i *(7)*	V_i *(8)*	W_i *(9)*	$(lnOR_i - lnOR_s)^2$ *(10)*	$(lnOR_i - lnOR_s)^2 * W_i$ *(11)*	$lnOR_i * W_i$ *(12)*
1	Fletcher	1	12	4	11	0.16	1.48	0.67	2.55	1.72	-1.24
2	Dewar	4	21	7	21	0.47	0.52	1.91	0.26	0.50	-1.44
3	1st European	20	83	15	84	1.46	0.15	6.80	0.38	2.61	2.58
4	Heikenheimo	22	219	17	207	1.25	0.11	8.72	0.21	1.87	1.93
5	Austrian	37	352	65	376	0.56	0.05	20.49	0.11	2.30	-11.81
6	Italian	19	164	18	157	1.01	**0.12**	8.18	0.06	0.52	0.10
7	Australian	51	376	63	371	0.77	**0.04**	23.92	0.00	0.01	-6.34
8	NHLBI SMIT	7	53	3	54	2.59	**0.52**	1.93	1.42	2.74	1.84
9	Frank	6	55	6	53	0.96	**0.38**	2.67	0.04	0.11	-0.11
10	Valere	11	49	9	42	1.06	**0.26**	3.87	0.09	0.35	0.23
11	UK	48	302	52	293	0.88	**0.05**	20.77	0.01	0.25	-2.75
12	Witchitz	5	32	5	26	0.78	**0.48**	2.06	0.00	0.00	-0.52
13	Lasierra	1	13	3	11	0.22	1.54	0.65	1.59	1.03	-0.98
14	3rd European	25	156	50	159	0.42	0.08	13.02	0.40	5.26	-11.42
15	Olson	5	28	5	24	0.83	0.50	2.02	0.00	0.01	-0.39
16	Schreiber	1	19	4	19	0.21	1.37	0.73	1.76	1.28	-1.14
17	2nd European	69	373	94	357	0.64	0.03	31.03	0.05	1.41	-14.09
18	2nd Frankfurt	13	102	29	104	0.38	0.14	7.35	0.54	3.94	-7.16
19	Klein	4	14	1	9	3.20	1.48	0.68	1.97	1.34	0.79
20	N. German	63	249	51	234	1.22	0.05	21.59	0.19	4.11	4.21
21	Lipschultz	6	43	7	41	0.79	0.37	2.73	0.00	0.00	-0.65
22	Gormsen	2	14	3	14	0.61	1.01	0.99	0.06	0.06	-0.49
23	Brochier	2	60	8	60	0.22	0.66	1.51	1.57	2.38	-2.26
24	European	41	172	34	169	1.24	0.07	14.53	0.21	3.05	3.16
	Totals	463	2,961	553	2,896			198.83		36.86	-47.96

Summary Treatment Effect

Fixed-Effects Model, odds ratio 0.79 (95-percent confidence interval, 0.69-0.91)

Random-Effects Model, odds ratio 0.79 (95 percent confidence interval, 0.64-1.00)

NOTE: *OR* is the odds ratio; *V* is the variance of the logarithm of the odds ratio; *W* is the weight (= 1/*V*); *lnOR* is the natural logarithm of the odds ratio; subscript *i* indexes the study; and OR_s is the summary treatment effect from the fixed-effect model. See text for details.

SOURCE: Adapted from S. Yusuf, R. Collins, R. Peto, et al., "Intravenous and Intracoronary Fibrinolytic Therapy in Acute Myocardial Infarction: Overview of Results on Mortality, Reinfarction and Side-Effects from 33 Randomized Controlled Trials," *European Heart Journal* 6:556-558, 1985. Only data from studies of the effect of intravenous streptokinase are shown.

in the treatment and control groups, -28.1 percent (i.e., 8.3 to 36.3 percent) would have yielded a similar conclusion.

The second measure of treatment effect in common use is the odds of death in the treatment group divided by the odds of death in the control group (odds ratio). An odds ratio is, under usual circumstances, an approximation of the ratio of the rate of disease in the treated group to the rate of disease in the control group (rate ratio or, in more generic terms, the relative risk). In the Fletcher study (see table 4-A-2), the proportion of deaths in the treatment group was 8.3 percent (1/1 2), and the odds of death were (8.3 percent/(100 percent -8.3 percent)), or 0.091. (Note that 1/(12 - 1) is another way to calculate the odds and is equal to 0.091.) The odds of death for the control group were 4/7, or 0.571. The ratio of these two odds is 0.091/0.571, or 0.16, the odds ratio (see table 4-A-2). If the odds of death were the same in the treatment and control groups, the odds ratio would be 1. In the Fletcher study, the odds ratio was much smaller than 1, which suggests that treatment decreased the odds of death. Calculating the proportion ratio-(1/12)/(4/l 1), or 0.23—would show that the proportion of cases in the treated group was about one-quarter of that in the control group. Note again the similarity in conclusion, regardless of the particular method of calculating a treatment effect.

Although there are various methods for expressing treatment effects, the choice of the type of treatment effect calculated is somewhat arbitrary and is often based on tradition and interpretability as well as practical and theoretical statistical considerations. As a general rule, so long as the treatment effect is correctly interpreted, the manner of expressing the treatment effect is not important.

Calculating fhe Precision of the Measured Effect

For each expression of the size of the treatment effect, there is an associated value (a variance) that reflects the precision with which the treatment effect has been measured. This measure of precision is similar to the concept of a standard deviation and is, in fact, calculated as the square of the standard deviation. If the variance of a treatment effect is large, the treatment effect has not been precisely measured.

The variance of a treatment effect reflects the amount of information in the study. The result of a small study is imprecise and offers little information about the treatment effect. Conversely, the result of a large study is precise and conveys much information about the treatment effect. As precision (information) increases, variance decreases, and vice versa. In other words, the inverse of the variance of the treatment effect reflects the informativeness of the study results.

The variance of a treatment effect is calculated from a simple formula. Understanding why the formulas for variances are constructed as they are is not important for understanding the basic concepts of meta-analysis. Nonetheless, the following examples show how variances are calculated.

The treatment effect for the Fletcher study (see table 4-A-1) is -1.6. The variance associated with this value is

$$E(l - n/N)(N - d)/(N - 1),$$

where:

- E is the expected number of deaths in the treatment group if there were no treatment effect,
- N is the total number of participants in the trial,
  n is the number of treated participants, and
- d is the number of deaths in the treated and control groups combined.

E is equal to rid/N, as explained earlier. Thus, for the Fletcher data, the variance is 2.6(1 - 12/23)(23 - 5)/(23 - 1), or 1.02 (see table 4-A-1). The square root of the variance (1.02) is the standard error (like a standard deviation) of the treatment effect (O - E), which in this case is 1.01.

Taking the ratio of the treatment effect to its standard error (-1 .6/1 .01) yields -1.58, a statistic that can be used to test the significance of the treatment effect. The ratio of an observed treatment effect to its standard error reflects the probability that an effect as large as the observed treatment effect would have been found if, in fact, no real treatment effect were present. The ratio is compared with values in a statistical table (of the Z distribution), which shows that if the absolute value of this ratio is <1.96, the probability of an effect of this size being observed by chance if no treatment effect existed is >0.05 (see box 4-2). If the probability of an observed treatment effect is >0.05, the analyst accepts the hypothesis that there was no evidence of a treatment effect in the study.[3] For the Fletcher data, with ratio-1.58, the treatment effect observed was not significantly different from the effect of no treatment.

The same example can be used to illustrate how the variance of an odds ratio is calculated. A popular variance formula for the odds ratio is complicated (78). In this example, because a much simpler formula (81) works nearly as well, it is presented instead. The variance can be estimated by the sum of the inverse of the number of deaths and nondeaths in the treatment and control group: 1/1 + 1/(12- 1)+ 1/4+ 1/(11 - 4), or 1.48 (see table 4-A-2). The standard error is 1.48, or 1.22. The statistic to test the significance of the odds ratio is obtained by dividing the natural logarithm of the odds ratio[4] by its standard error, which is ln(O. 16)/1.22, or -1.50. As before, because the absolute value of this ratio is <1.96, the probability of an effect of this size being observed by chance if no treatment effect existed is >0.05. Thus, the analyst would accept the hypothesis that there was no evidence of a treatment effect in this study. Note that the smallest value of the number of deaths and nondeaths in the treatment and control group-1 in this example—is the most important in determining the variance.

Summarizing the Treatment Effect Across Studies

The two methods used most frequently to summarize treatment effects across studies are the fixed-effects model and the random-effects model. The assumptions underlying the fixed-effects model are that all studies are estimating the same treatment effect and that the difference in the effects observed across studies results by chance. The assumptions underlying the random-effects model are that the treatment effect truly differs across studies and that the goal is to determine the average of the different effects. Although fixed-effects models were used frequently in the past, the use of random-effects models to summarize the treatment effect has been favored recently (70). In practice, if the results of the studies are homogeneous, the two approaches give the same result.

When the results from different studies are ready to be analyzed jointly, the analyst may choose either to search for different characteristics of the study designs or study populations that might account for the variation in results, or to evaluate the homogeneity of the results. If the results prove to be heterogeneous, varying considerably among studies, the analyst has two options:

1. to refrain from summarizing the results across studies (summarizing, instead, within groups that have similar results or not calculating a summary at all, if the results are markedly heterogeneous), or
2. to summarize the results across studies using a random-effects model.

One could also summarize heterogeneous results using a fixed-effects model, but although this has been a common practice in the past it is no longer recommended.

The search for characteristics of the studies or study populations that might account for variation in results can be undertaken by grouping the study results according to the characteristic under study and summarizing results within groups. The sum-

[3] In statistical terms, the null hypothesis (that there is no treatment effect) is being tested. If the probability of a given treatment effect's being observed by chance, if in fact the null hypothesis is true, is >0.05, then in formal statistical terms one would fail to reject the null hypothesis.

[4] The (natural) logarithmic scale is used with ratio measures of effect (such as odds ratios) so that treatment effects of the same size, but in opposite directions (like an odds ratio of 0.5 and 2; 1/0.5=2) will be of equal absolute size arithmetically. The natural logarithm of 0.5 is -0.6931; the natural logarithm of 2 is 0.6931. The sampling theories upon which the relevant statistics are based work only when such symmetry is present.

mary treatment effects are then compared across groups. Regression techniques that effectively accomplish the same goal can also be used.

Evaluating the Homogeneity of Results

Results sometimes vary greatly and inexplicably from study to study, which may influence how a meta-analysis is interpreted. If the study results are markedly heterogeneous, for example, one might have little confidence in one's ability to predict the effect of treatment in any future study.

The evaluation of the homogeneity of results across studies in a meta-analysis is based on the homogeneity chi-square statistic(81). This statistic is a sum across all studies of the square of the difference between the study-specific treatment effects and the summary treatment effect, multiplied by the inverse variance of the study-specific treatment effect. In statistical terms, the homogeneity chi-square statistic is as follows:

$$\chi^2=\Sigma w_i(te_i - te_s)^2,$$

where:

- te_i is a study-specific treatment effect,
- te_s is the summary treatment effect (described below),
- $w_i = 1/v_i$ (v_i is the study-specific variance), and
- i indexes the study.

Thus, the squared difference of each study's result from the overall average is weighted by the precision of the study. In this way, the deviation of a small study from the summary treatment effect contributes little to the homogeneity statistic, whereas the deviation of a large study from the summary treatment effect contributes much more. This makes sense intuitively, because smaller studies are more likely to deviate from an overall mean by virtue of sampling error alone (4). Deviation of a large study from the overall mean suggests that the studies in the meta-analysis may have been samples from populations in which the treatment effects differed. The expected size of the homogeneity chi-square statistic is based on the number of studies in the meta-analysis and is found in a statistical table for values of chi-square.

The statistical evaluation of homogeneity can be illustrated using the meta-analysis of fibrolytic therapy discussed previously. The calculation of the homogeneity chi-square statistic is based on several columns in table 4-A-2. Column 7 contains the study-specific odds ratio, and column 9 contains the study weight (l/vi). Column 10 contains the squared difference of the logarithm of the study-specific odds ratio from the logarithm of the summary odds ratio (described below). Column 11 contains the product of the study weight and the squared difference of the effects. At the bottom of column 11 is the sum of the contribution of each study, which is the homogeneity chi-square statistic. In this example, the chi-square statistic is 36.9. The value expected under homogeneity is 35.2 or less. (This is the value from a chi-square table for 23 degrees of freedom and p=O.05. The degrees of freedom are the number of studies minus 1.) Thus, the variation in study results is greater than expected, suggesting that something other than chance accounts for the differences in the findings. Perhaps the effect of the drug differs depending on the exact dose used, the patient population, the length of time the patients were studied, or some other factor.

One problem with the homogeneity chi-square statistic is that it may not detect a variation that has biologic or practical importance. Therefore, a search for factors related to study results is recommended, regardless of whether statistical homogeneity is present.

Combining Results Across Studies

Once homogeneity has been evaluated, the results across studies may be summarized, if deemed appropriate, using a fixed-effects model or a random-effects model. In a fixed-effects model, the contribution of each study within the meta-analy - sis to the summary treatment effect is inversely proportional to its variance. Thus, larger studies contribute more to the summary treatment effect because they have smaller variances. Random-effects models, however, weight the contribution of individual studies according to their inverse variance and according to a measure of the variability of results across studies. In random-effects models, as the degree of heterogeneity increases, the studies tend to be given more equal weight, as in a simple average. An advantage of the random-effects summary for heterogeneous results is that the estimate of the summary effect is less precise than that calculated in a fixed-effects model, reflecting the greater degree of variation in the study results.

The random-effects model is now generally considered preferable where substantial heterogeneity exists (70). The illustration below of the method of calculating a summary treatment effect across studies uses a fixed-effects model, however, because that procedure is more straightforward and reflects the essential points about how the results from different studies are combined in a meta-analysis.

To illustrate the fixed-effects model, assume that the treatment effects for the data from the Yusuf meta-analysis were homogeneous. The most straightforward method of combining results across studies is to calculate the simple average of the treatment effects, in which the results of each study carry equal weight. The fixed-effects model, however, is a weighted average in which each treatment effect is weighted by the inverse of the precision of the estimate (inverse variance weights). The previous section described how to calculate the observed deaths minus the expected number of deaths in the treatment group, O - E, and its variance, v(O - E), for an individual study. Summing the O - E across studies is a form of inverse-variance-weighted summary treatment effect. Note that if there were no treatment effect, the sum of O - E would be zero. If the treatment reduced the number of observed cases by 10 percent, the sum would change accordingly. The 0- E of a study with a large number of treated subjects would be larger than that of a study with a small number of treated subjects. In this way, the larger studies contribute more to the summary treatment effects

The sum of 0- E for all studies of fibrinolytic therapy is -51.4 (seethe bottom line of column 7 in table 4-A-l). (In statistical notation, this is $Z(O_i - E_i)$, where i indexes the study and the summation is across all studies.) In other words, there were 51.4 fewer deaths than expected in all the treatment groups combined. The variance of this summary treatment effect is the sum of the variances of each treatment effect, which is 207.1 (see the bottom line of column 8 in table 4-A- 1). In statistical notation, this is Zvi(Oi - E_i). The square root of 207.1 is 14.4, the standard error; taking the ratio of the summary treatment effect to its standard error (-51.4/14.4) yields -3.57, which has an absolute value greater than 1.96 and thus is statistically significant at the p<O.05 level. Therefore, the meta-analysis supports the conclusion that fibrinolytic therapy is effective in reducing death after myocardial infarction.

Using data from the same example, one can calculate the summary treatment effect as an odds ratio. This approach more directly illustrates the principle of the weighted average. The formula for the natural logarithm of the summary odds ratio is

$$Zwiln(ORi)/Zwi,$$

where $W_i = l/v_i$ (the inverse variance), *OR* is the odds ratio, and i indexes across study results. In other words, the weight for each study result is multiplied by the natural logarithm of the study's odds ratio. This quantity is then summed across all i studies (see the bottom line of column 12 in table 4-A-2) and divided by the sum of all the study weights (see the bottom line of column 9 in table 4-A-2). In the example, this yields an answer of -48.0/199, or -0.24, the natural logarithm of the summary odds ratio. Exponentiation of the logarithm of the summary odds ratio (e^{-024}) gives the

[5] Although the O - E method has been very popular in the past for meta-analysis of randomized trial data and probably provides the right answer in most cases, the method has been shown to be misleading in some situations (41). One advantage to the O - E method is that the results are relatively easy to understand and explain. Because of the problems with occasional misleading answers, however, meta-analysts recently have favored methods based on odds ratios.

odds ratio, 0.79. The variance of the summary log odds ratio is l/Xwl, or 0.005 (i.e., 1/199). The square root of the variance is the standard error of the logarithm of the odds ratio, or 0.071. Calculating the ratio of the summary logarithm of the odds ratio to its standard error (-0.24/0.071) yields -3.38, which has an absolute value greater than 1.96 and thus is significant at the $p<O.05$ level. This significance means that it is unlikely that a treatment effect this large would have been observed by chance if there truly were no treatment effect. Thus, the results suggest a benefit of treatment.

The standard error of the logarithm of the odds ratio can also be used to calculate the confidence limits of an odds ratio. The width of the confidence interval is proportional to the standard error of the odds ratio. Thus, a large confidence interval implies small precision, and vice versa. A confidence limit for an odds ratio that excludes 1 indicates that the treatment effect is statistically significant. The 95-percent confidence interval around the fixed-effects estimate for the example data is 0.69 to 0.91, which excludes 1 (see table 4-A-2).

The details of calculating a random-effects summary are beyond the scope of this document, although the calculation is not markedly more complicated than the procedures illustrated above. In this example, the random-effects model summary treatment effect is an odds ratio of 0.79 (95-percent confidence interval 0.64 to 1.00). Note that the confidence interval around this estimate is wider than the confidence interval around the fixed-effects estimate (see table 4-A-l). The greater width of the confidence interval reflects the fact that the study results were more variable than would be expected if chance were the only reason for variation. The heterogeneity in this example was relatively small; if more marked heterogeneity were present, the difference between the results of the fixed-effects and random-effects models would be greater (7).

Regression methods can also be used to combine studies' results (8,39,40,82). The regression approach allows the shape of dose-response curves to be estimated and provides a convenient method for identifying patterns in study results associated with characteristics of the study populations or study designs. Both fixed-effects and random-effects regression models can be constructed.

The frequentist approach, which is used routinely in medical meta-analysis, has been used for summarizing the treatment effects presented in this appendix. Another method of summarizing treatment effects is the Bayesian approach (see box 4-6 in main text). The Bayesian meta-analyst specifies his or her belief about the size of a treatment effect and the certainty about that belief prior to examining any of the results of the studies in a meta-analysis (28,29,77). In the absence of a strong prior belief, the Bayesian meta-analyst may find all possible values of the treatment effect equally likely (and thus has no certainty about its size). The results of the studies in the meta-analysis are then used to modify the analyst's belief about the size of the treatment effect. The result is an expression of the analyst's belief about the size of the treatment effect that primarily reflects the results of the studies and only minimally reflects the prior belief. The contribution of the prior belief (or data) relative to the contribution of the new data (the studies in the meta-analysis) depends on the strength of evidence from each source. For example, when there is much prior information about the size of a treatment effect and only a few small studies are in the meta-analysis, the result of including the new data in the synthesis may not much alter the estimate of the treatment effect. In practice, Bayesian methods give quantitative results that are similar to those from a random-effects model.

Clinical-Economic Trials

Background Paper 5

SUMMARY

One consequence of the desire for better information about the economic implications of medical technologies and practices has been the growing practice of collecting and analyzing economic data in clinical trials. This type of research entails incorporating economic measures into the prospective data collection activities of a clinical trial conducted to determine the safety and efficacy of a technology. Both the economic and clinical data from the trial are then analyzed to provide information about the treatment's cost-effectiveness. Clinical-economic trials-trials that include both clinical and economic components-can be initiated either early in a treatment's development (e.g., before its approval by the Food and Drug Administration) or after the treatment has been used in routine clinical practice.

The number of clinical-economic trials is still very small but is growing rapidly. Many of the methodological and practical issues that arise in these trials also arise in traditional clinical trials and in other methods of obtaining cost-effectiveness data. These include, for example, the importance of the economic perspective selected by the researcher and the great variations in methodological techniques that can affect the comparability of cost-effectiveness results. In addition, however clinical-economic trials raise some new issues, such as how to deal statistically with economic data that is more skewed or requires larger sample sizes for statistical significance than the clinical data from the same trial. Also, economic data from a clinical trial may reflect cost- efficacy (15) rather than cost-e festiveness in the real world, just as clinical trials may reflect clinical efficacy (under highly controlled and ideal circumstances) rather than clini-

by
Neil R. Powe and
Robert I. Griffiths
Johns Hopkins University
Baltimore, MD

cal effectiveness. The data from a study that has strict criteria for selecting patients and is performed at the best academic medical centers may contrast greatly with the data from a trial that is conducted in several community hospitals that are representative of the average U.S. hospital.

Nonetheless, clinical-economic trials offer the opportunity to examine the potential cost- effectiveness of a technology before it becomes irrevocably established in everyday practice, and they can provide sponsors with useful information. The demand for early information on the costs and effectiveness of new technologies is driven by health care policymakers who hope to improve medical care without increasing its costs; by providers who want to remain competitive in a cost-conscious environment; by insurers who must make decisions about coverage and reimbursement; and by manufacturers who adapt their research and marketing strategies in response to these concerns. In view of these demands, clinical-economic trials are likely to become increasingly common. The usefulness and validity of clinical-economic trials can be improved through the futher development of clinical-economic methodology and the establishment of some consensus standards for methods and dissemination of study results.

The cost of health care in the United States has risen rapidly in the past decade. The proportion of the gross national product (GNP) spent on health care rose from 9.1 percent in 1980 to 14 percent in 1992 (69). As a nation, we now spend more than $800 billion annually on health care, which is more per capita than is spent in any other country.

Concern over the rising costs and the deficiencies in our health care system has led to a widespread desire to increase the availability and quality of care while containing or reducing the costs. Policies have been implemented to contain costs and to promote a more rational allocation of resources for health care services. The federal government has established a per-case prospective payment system (Public Law 98-21) to control the cost of hospital care for Medicare patients and a resource-based relative value scale (56) to control the costs of physicians' services. State Medicaid programs have also imposed severe constraints on payments for medical services. Managed care has become an important alternative for employers who are struggling to keep costs down and was incorporated into the Clinton Administration's proposals for health care reform.

Expensive and new medical technologies and practices (e.g., magnetic resonance imaging (63) and recombinant erythropoietin (52)) have received considerable attention as one factor that has contributed to the increase in health care costs (23,45). When used widely, they may not only raise costs directly but also indirectly, by increasing the use of other services (34). Attempting to control the costs of new and expensive technologies, the federal government has implemented policies such as requiring drug manufacturers to of--fer discounts to state Medicaid programs (75). Also, although no formal mechanism exists, Medicare officials have suggested that they may establish cost-effectiveness as a criterion for coverage (42), as is done in Australia and other countries (29). The importance of economic evaluation was recently reconfined in Congress's mandate to the Agency for Health Care Policy and Research (AHCPR) when it was reauthorized in 1992 (Public Law 102-410).

Still, many technologies continue to make their way into routine medical practice without being accompanied by economic information to promote cost-effective use. In part, this reflects the fact that economic information based on experiment and observation has not been widely available to those who determine the use of health care technology. The analyses that have been performed have primarily used data on efficacy derived from clinical trials, but have resorted to the use of economic data that are not derived empirically and that are sometimes derived outside of the context of the clinical use of the technology (60). Most of these economic evaluations of technologies have been performed using decision modeling, claims data analysis, and other secondary data analytic techniques.

To date, relatively few economic analyses have been based on the economic data collected in clinical trials. In part, this reflects concerns about how adequately the economic consequences of treatment can be appraised at the same time that clinical benefits are evaluated. Nonetheless, integrating the collection and analysis of economic data in clinical trials is a growing practice with several potential benefits:

- Clinical trials, particularly randomized trials, provide a scientifically rigorous method of assessing the clinical benefits (e.g., efficacy and safety) of new technologies (38).
- Clinical trials conducted for the FDA-approval process provide opportunities to collect economic data at the time when they may be most needed for planning and guiding the appropriate use of a treatment by health care providers.
- In contrast to clinical trials, studies using secondary data often incorporate data from disparate and sometimes incompatible sources, which makes the results difficult to interpret or apply. Thus, relevant economic data collected early and rigorously could be especially useful when coupled with strong experimental designs.

Role in Evaluating Health Technologies

Clinical-economic trials provide helpful data for organizations and individuals who must decide how to develop, pay for, or use medical technologies. The decisionmakers include insurers, providers, manufacturers, and panels that formulate national clinical practice guidelines.

Each type of decisionmaker evaluates the economic aspects of a particular technology from a different perspective, which affects what kinds of economic data are collected and how they are analyzed and interpreted. An insurer may want information about the technology's effect on claims for health services in order to promote the appropriate use of the technology and to adjust premiums; a provider may be concerned about how the technology would affect the cost of providing care; a manufacturer may use economic data to facilitate the development and application of new technology, the availability of health insurance coverage, and the strategies for marketing; a national guideline panel may use economic data to arrive at recommendations that meet the public interest in conserving national resources. The consumer's perspective influences the decision about which data to collect in a clinical-economic trial and about how to analyze and interpret the data.

Insurers

Private insurance companies are concerned about how expensive medical technologies will affect their ability to set competitive insurance premiums and about whether new technologies will result in unexpected payments that exceed the revenue from premiums. Early and better information about the cost-effectiveness of treatments is thus increasingly valuable to private insurers who are attempting to predict and control their costs. The need will be even greater if managed care organizations such as health maintenance organizations (HMOS) continue to spread.

In the future, it is even possible that denials of payments for particular health care services might be defended on the ground that the benefits are small in relation to the costs or that other services could result in equal benefits at lower costs. Data on cost-effectiveness would be crucial in such a debate.

Public health insurance programs, such as Medicaid and Medicare, are under pressure to control the rising costs that have resulted from increased enrollment and the expansion of federally mandated benefits. Like their private counterparts, public insurers are concerned about the costs arising from the widespread use of expensive technology. Data on the economic consequences of such technology are needed for informed decisions about coverage and payment policies. Rules requiring public insurers to provide or withhold coverage based on cost-effectiveness (e.g., the proposed addition of cost-effectiveness to the current "reasonable and necessary" criteria for Medicare coverage of medical services [54 FR 4302]) must be based on credible economic and clinical data and must be promulgated in a timely fashion.

Published economic analyses of treatments received by the Medicare population (e.g., references 51, 64) suggest that the medical community is increasingly aware of the importance of evaluating the cost implications of expensive new technology from the perspective of third-party payers. Many of these studies, however, are published after the technologies and policies for paying for them have become implemented in routine clinical practice—when both practice styles and payment policies are much more difficult to alter than they would have been in the initial stages of the technology's dissemination. For example, if early clinical trials revealed that using recombinant erythropoietin to treat the anemia caused by chronic kidney disease could reduce the costs of hospitalization and transfusion-related illnesses in dialysis patients, the Medicare program—as the principal payer for the care provided to patients with end-stage renal disease—might be willing to pay more for the use of the drug. Thus, early information about the economic effects of treatments under study in clinical trials could promote the spread of cost-effective technology.

Providers

An important result of the pressure for cost containment has been the establishment of the Medicare prospective payment system, under which providers receive predetermined payments for Medicare beneficiaries' hospital stays (Public Law 98-21). Many state Medicaid programs have also adopted per-case prospective payment systems. This payment method may sometimes discourage hospitals from using expensive new technology because the increased cost it entails does not bring commensurate revenue in the short term (3,70). Consequently, hospital administrators want to know whether purchasing and using new technology would not only improve patients' health but also pay for itself. Assessing the economic implications prior to purchasing a new technology has become more common as changes in reimbursement levels have increased the pressure for limiting costs.

Hospitals' pharmacy and therapeutics committees, which are responsible for determining which drugs are placed on the hospitals' formularies, frequently rely on economic analyses in making their decisions. For example, such committees determine whether patients with acute myocardial infarction will usually be given recombinant tissue plasminogen activator or streptokinase, whether low-osmolality or high-osmolality contrast agents will be used in radiologic-imaging examinations (such as computed tomography and cardiac catheterizations), and whether the expensive new anti-emetic drug odansetron will be used instead of traditional anti-emetic drugs.

Manufacturers

Because new medical treatments may be less likely to be used if they are too costly, manufacturers are increasingly concerned about producing technology that is not only safe and efficacious but also cost-effective. In Australia, for instance, pharmaceutical manufacturers must submit evidence that products are cost-effective before they can be included on the government's list of reimbursed products (12). As more countries adopt such requirements in the future, manufacturers will want to be able to use data from clinical-economic trials that address the issues of different international markets. The growing market pressures have led a growing number of manufacturers to evaluate the cost implications of new technologies at earlier stages of development in order to:

avoid making substantial investments in products that are unlikely to be covered by insurance or accepted by providers,

- ensure that data on the economic implications of the technologies are available for marketing purposes, and

facilitate the establishment of prices that will provide adequate returns on the manufacturers' investments, while maintaining the technologies' economic viability.

In some cases, manufacturers might also use economic information to help make other internal decisions as well. For example, a pharmaceutical manufacturer developing a drug with possible applications for a number of different diseases might find economic data valuable for deciding which of the possible indications for the drug it should seek Food and Drug Administration (FDA) approval for.

Guideline Panels

Expert panels convened by federal agencies (and other organizations) routinely develop clinical practice guidelines based on information about the safety and effectiveness of medical technologies. The use of information about the costs and cost-effectiveness of technologies by such panels is less common but not unknown, and it may increase.

Since 1977, for instance, the National Institutes of Health (NIH) has convened conferences to develop statements of consensus about important management issues in medical care. Although the primary purpose of these statements is to comment on the efficacy and safety of treatments, 16 of these statements have used the word cost- effec*tiveness,* and three conferences have addressed the question of cost-effectiveness (19). Cost issues were discussed at 53 of the 93 consensus development conferences held between 1977 and 1992 (19).

One of the functions of AHCPR, which was established in 1989, is to develop clinical practice guidelines (Public Law 101-239). Although the original mandate emphasized the reduction of variations in medical practice and outcomes as a goal, rather than cost containment, legislation reauthorizing the agency in 1992 directed it to incorporate cost-effectiveness information into its technology assessments, where feasible, and to consider health care costs when developing practice guidelines (Public Law 102-410).

A recent AHCPR guideline on cataract management in adults contains a section on the cost of care, but the panel that developed the guideline found no published data regarding the cost of preoperative, intraoperative, or postoperative care (72). Some panels clearly desire the economic data that could be generated from clinical trials, and the demand for such information may increase in the future.

Current Applications

A variety of medical technologies—such as pharmaceuticals, devices, procedures, and other services—have been assessed in economic components of observational or experimental clinical trials. The fact that these trials have addressed diverse populations (e.g., children and elderly people or inpatients and outpatients) and various illnesses suggests that economic analysis is broadly applicable in clinical studies. The diversity reflects the needs of those who use economic data (often the sponsors) and the capacity and interests of the different types of organizations that actually conduct the evaluations (box 5-1).

Several recent clinical-economic trials sponsored by industry and conducted at academic institutions have assessed both the financial and medical effects of new pharmaceuticals:

In a recent study sponsored jointly by Schering-Plough and Sandoz, for instance, researchers at the Memorial Sloan-Kettering Cancer Center examined the costs and benefits of granulocyte microphage colony stimulating factor (GM-CSF) as an adjuvant therapy in relapsed Hodgkins disease (28).

- A study sponsored by Hoechst-Roussel Pharmaceuticals, Inc., and conducted at the University of Southern California School of Medicine, was designed to ascertain the costs and medical outcomes of treating spontaneous bacterial peritonitis with short courses of antibiotics as compared with long courses of antibiotics (59).
- Researchers at the Johns Hopkins Medical Institutions evaluated the relative cost-effectiveness of low-osmolality and high-osmolality radiographic contrast media in patients undergoing cardiac angiography (50) in a study sponsored by Sanofi Winthrop, a manufacturer of radiographic contrast media.

BOX 5-1: Who Performs Economic Analyses in Clinical Trials?

Those who use economic analyses sometimes sponsor and perform their own clinical-economic trials, as do manufacturers and other sponsors. Academic and government researchers also perform the analyses in clinical-economic trials funded by outside sponsors.

Manufacturers. Manufacturers face tremendous incentives to prospectively evaluate the economic implications of new technologies in order to ensure that economic data are available at the time the products are launched. The trials may be initiated by any of several departments with the firms, including the clinical research and marketing departments, Because most manufacturers currently lack the extensive expertise necessary for conducting such studies, academic institutions or other private entities are usually given grants or contracts to conduct them, but many manufacturers are recruiting experts (e.g., doctoral- or master's-level health economists) to improve their in-house capabilities.

Academia. The economic analysis of medical technology has evolved into a discipline in some universities in response both to concern for health policy and financing and to demands from industry (27,51 ,55,62). Many of the analytic techniques applied in economic analyses, whether performed in the context of clinical trials or not, have been developed by academicians, which makes universities a source of expertise. The demand for economic information has led to an initial collaboration between academia and health care providers, especially within academic medical centers. Inasmuch as academic medical centers are often the loci for clinical trials of the efficacy and safety of emerging technologies, economic evaluations in conjunction with these trials are natural extensions.

Private consulting firms. Other private sector organizations, such as consulting firms and think tanks, are often called upon to perform economic analyses of medical technologies. The funding for this work has come, in large part, from manufacturing firms. A firm conducting a clinical trial-either internally or through a grant or contract with an academic organization or health care provider—might turn to consulting firms with expertise in health policy and economics. These firms might be asked to identify the reimbursement and marketing issues associated with a new product and then to collaborate with the investigators who are designing the clinical trial to collect economic information that will be useful for launching the product.

Government. The government is an important sponsor for biomedical research in general, but aside from a few studies funded by the Agency for Health Care Policy and Research, government sponsorship of economic analysis in clinical trials has been limited. Some institutes of the National Institutes of Health have occasionally permitted the collection and analysis of economic data within their clinical trials, although the funding for the economic components has come from elsewhere (such as foundations and AHCPR). The potential exists for more such trials, inasmuch as the National Cancer Institute, the National Heart, Lung, and Blood Institute, and the National Institute on Aging occasionally consult with extramural scientists on economic issues and implications and also have intramural scientists (including economists) engaged in economic studies other than clinical trials. These scientists often perform post-hoc analyses of data using economic modeling. Although the Health Care Financing Administration and other public payers are becoming important consumers of economic information from clinical trials, there IS little evidence that they are conducting or sponsoring such studies.

SOURCE: Neil R Powe and Robert I Griffiths, 1995.

Other types of manufacturers have also provided funding, technical support, and equipment to researchers collecting economic data in clinical trials, for example:

- Support Systems International, which makes air-fluidized beds, provided equipment, consultations, and technical services to researchers who compared the cost-effectiveness of home air-fluidized therapy with that of conventional home therapy for pressure sores (65).
- Burron Medical, Inc. sponsored a study comparing the time and cost of filling syringes with automated versus manual methods (l).
- Researchers at the Nuffield Department of Obstetrics and Gynecology at Oxford received a loan of equipment to compare the costs and outcomes of videopelviscopy with those of laparotomy for treating ectopic pregnancies (4).

Several economic studies have also been performed by health care providers to justify their own costs or to improve efficiency:

- A study conducted by the First Hill Orthopedic Clinic in Seattle, for example, demonstrated that despite requiring relatively long hospital stays, total hip arthroplasty for patients older than 80 was a cost-effective alternative to placing the patients in nursing homes (5).
- A Department of Veterans Affairs (VA) study demonstrated that the costs of VA-hospital-based home care for the terminally ill were comparable to those of community home care or hospice care, and that patients and caregivers expressed the greatest levels of satisfaction with hospital-based home care.
- A cost-effectiveness study comparing erythromycin with mupirocin as treatments for impetigo in children, conducted by researchers in the Department of Pediatrics at the Johns Hopkins Medical Institutions, evaluated not only the costs of the medical treatments but also the nonmedical costs incurred by the families as a result of the illness (58).

Public agencies and private philanthropic organizations have also played important roles in conducting or sponsoring clinical trials with economic components.

- The World Health Organization, for example, was one of the sponsors of a study in which the use of biobrane was compared with the use of l-percent silver sulfidiazine in the outpatient management of partial-thickness bums (22).
- The National Center for Health Services Research[1] sponsored several clinical studies with economic components. The studies investigated the costs and benefits of cyclosporine relative to prednisone and azathioprine in improving the results of renal transplantation (61).

These assessments reflect the diversity of approaches to economic analysis of medical technology such as study design (e.g., perspective of the analysis and types of costs considered). The scope of these trials demonstrates that providers, payers, and patients are concerned with economic issues in all types of health technology applications. Despite the variety of health technologies studied, however, a recent study indicates that few clinical trials (0.2 percent) include economic analyses (2) and that no relationship has been established between the methodology for economic analysis and the quality of the research. Therefore, clinical-economic analyses have so far produced few sound data to which health care policy makers can turn for guidance.

METHODOLOGICAL CONSIDERATIONS

Conducting clinical-economic trials to assess the cost-effectiveness of emerging technologies entails a number of methodological considerations that can challenge researchers and affect the usefulness of the information generated by the trials.

[1] NCHSR was eliminated with the establishment of the Agency for Health Care Policy and Research In 1989.

| Analytic Framework

Traditional Clinical vs. Clinical-Economic Trials

Because a clinical-economic trial is a particular type of clinical trial, many of the methodological and practical issues that arise in traditional clinical trials also pertain to clinical-economic trials. The nonrandom allocation of treatments to groups of patients can bias both economic and clinical findings, because important characteristics of the patients in the experimental and control groups may differ. Also, clinical trials, particularly those conducted in the early stages of a technology's development, require designs that may diverge from normal clinical practice. The early clinical trials of the drug recombinant human erythropoietin (18), for example, included only a small number of relatively healthy dialysis patients (those without systemic illnesses) and were performed in institutions where patients were likely to receive superior care. Although the early trials yielded very encouraging results, a subsequent study of more than 50,000 patients suggested that the efficacy demonstrated in the early trials might not be as high for the general population of dialysis patients, in part because of differences in the patient populations, the physicians' practices, the regulatory influences, and the quality of care (53).

What distinguishes clinical-economic trials from traditional clinical trials is the incorporation of resource usage and costs as outcome measures and their subsequent availability for further analysis. These economic measures and the rationale for collecting them pose distinct issues for researchers. Hypotheses to be tested in a clinical-economic trial include a technology's effects on both the patients' health and the costs of treatment. The clinical trial's protocol and setting may place their own special constraints on the collection of relevant data about costs.

Economic data can be collected prospectively in longitudinal studies ranging from observational studies to experimental studies (e.g., randomized controlled trials). Although the traditional definition of clinical trials excludes studies with historical control groups, some of the considerations that apply to clinical-economic studies probably extend to studies without control groups or to studies with historical controls. The purpose of collecting economic data in observational studies in which no direct comparisons are made between technologies is usually to identify or enumerate the costs of applying specific technologies or the costs associated with specific illnesses. An economic comparative trial, whether experimental or observational, compares the costs or cost-effectiveness of two or more alternative strategies for managing a condition or disease. These distinctions affect the types of conclusions researchers can draw about comparative efficiency and outcomes, because comparative studies yield information on relative outcomes.

Clinical-economic trials may also be viewed in the same way as other economic analyses. Such trials are most commonly performed as part of cost-effectiveness analyses, which assess the comparative costs and effectiveness of alternative technologies (see box 5-2). Within this framework, the trials can be thought of as providing a way to incorporate economic measures into prospective clinical studies. The economic measures include: 1) resources consumed as a result of the application of medical technologies, and 2) the costs of those resources from different perspectives.

Types of Resource Consumption and Costs

Any resources consumed in providing health care, or as a result of illness, cannot be used for other purposes. Resources are typically valued by economists in terms of the next best alternative uses, known as the *opportunity costs.* Because the opportunity costs are reflected in the price one is willing to pay for using resources, the resources are usually valued in dollars, but economists may also speak of the value of the resources in terms of utility. Dollars and utilities are simply different ways of valuing the resources that are consumed.

It is important to distinguish between accounting and economic costs (20). *Accounting costs are* the monetary outlays associated with the con-

BOX 5-2: Types of Economic Analyses

The demand for evaluating the costs and benefits of medical technology has led to three basic types of comparative health economic analyses: cost-identification, cost-benefit, and cost-effectiveness (10,15,73), Each method requires economic data that may be collected during a clinical trial.

Cost-identification analysis enumerates all the costs of applying a technology to a specified population under a particular set of conditions (such as inpatient care). The analysis is usually performed in conjunction with a longitudinal and observational clinical study that does not compare the benefits of one technology with those of alternative technologies. Researchers examine the natural history, in an economic sense, of the group of patients in the study. The resources expended by the providers and the patients for the technology and all associated interventions are measured, and the overall costs are calculated. Analysts can also enumerate the contributions of different types of resources (labor, supplies, and capital), as well as the contributions of particular subgroups of patients, to the overall costs. Researchers can determine, for example, whether labor costs are greater than capital costs or whether the costs of therapy are similar for young and old patients, male and female patients, black and white patients, or high- and low-risk patients. Cost-identification studies are also performed to obtain pilot data for use in planning experimental trials or comparative studies (54). Cost-identification data can also be integrated with other clinical data (from outside the trial) using modeling or simulation approaches to compare technologies.

Cost-benefit analysis enumerates and compares both the costs of applying the technology and the net savings resulting from its therapeutic benefits. One strength of this type of analysis is that it provides a rule for deciding whether to adopt or reject a technology from a strictly economic perspective. Health care providers may want to know not only that a particular drug prevents a certain number of heart attacks per year at a specific cost, but also that the drug saves money for the provider or the insurer. If the sum of the benefits is greater than the sum of the costs of using the technology, the net benefit is positive and the technology should be adopted. One limitation of this approach, however, is that the therapeutic benefits must be expressed in monetary terms. Placing dollar values on decreased mortality or morbidity is highly controversial, and existing techniques may systematically undervalue or overvalue the lives of individuals in certain groups (such as very young, elderly, or impoverished people).

Cost-effectiveness analysis also entails the explicit valuation of the costs and therapeutic benefits of applying medical technology and compares net costs to net benefits. In contrast to cost-benefit analysis, however, cost-effectiveness analysis expresses therapeutic benefits in such reduced-mortality or -morbidity measures as years-of-life-saved or quality-ad justed-years-of-life-saved, The strength of this approach is that it obviates the need to assign dollar values to life-years saved or to reduced morbidity. At the same time, however, it produces no explicit decision rule for adopting or rejecting the technology. Whether a technology whose cost-effectiveness ratio is $100,000 per life-year saved is adopted depends on whether the decisionmaker considers a year of life to be worth at least $100,000.

Cost-identification analyses may be most appropriately incorporated into observational trials, while cost-benefit and cost-effectiveness analyses may be more appropriately incorporated into experimental trials whose objectives include either comparing the costs and benefits of alternative technologies or comparing the costs and benefits of a technology with those that would occur without intervention.

SOURCE Neil R. Powe and Robert I. Griffiths, 1995

sumption of resources, whereas *economic costs* include not only the monetary outlays but also the opportunity costs. For example, the accounting cost of an illness includes only the cost of the treatment, while the economic cost includes both the cost of the treatment and the loss of earnings that results from the patient's morbidity or mortality. A distinction can also be drawn between fixed costs and variable costs (67). *Variable costs* change (at least in the short term) in accordance with the extent to which health services are provided, whereas fixed *costs are* independent of the quantity of health services provided. Variable costs typically include labor and supplies, while fixed costs often include equipment. Economic analysts also distinguish average costs from marginal or incremental costs. *Average costs* include both the fixed costs and the variable costs apportioned across all units of a particular resource, whereas *marginal costs are* the additional variable costs of providing additional services.

The total value of resources consumed for health care can be categorized as direct medical costs, direct nonmedical costs, and indirect economic costs (15,73).

- *Direct medical costs* result from the consumption of medical resources in applying a technology to produce health care services. For instance, magnetic resonance imaging entails the use of capital equipment (the imaging machine), staff time to operate the equipment, and professional time to interpret the results. Medical complications arising from the use of some technologies may also result in the consumption of additional resources, which are counted as direct medical costs of applying the technology.
- *Direct nonmedical costs are* associated with the application of a technology but do not result from the consumption of medical resources. Such costs may include expenditures for travel or parking, food, lodging, or child care in conjunction with medical treatment (43). For several types of chronic debilitating illnesses, the direct nonmedical costs can be substantial.
- *Indirect economic costs* result from the excess morbidity or mortality associated with the application of a technology or with its side effects. Excess mortality or morbidity frequently entails an individual's loss of the opportunity to produce valued resources, goods, or services. Often referred to as the loss of human capital, such costs include lost wages resulting from decreased life expectancy or earning ability reductions resulting from disability.

The *cost savings* associated with a technology are measured in terms of the expenditures that are obviated by the technology's therapeutic benefits. If the application of one technology eliminates the need for an inferior alternative or for related technologies, the medical-resource costs of applying the inferior technology may be counted among the savings of the superior technology. If the technology's therapeutic benefits decrease the use of medical resources, the costs of the unused resources are additional savings. If the application of a technology eliminates the need for a second technology that is associated with side effects, the medical resource costs of treating those side effects may be counted among the savings associated with the first technology. In radiologic procedures, for example, fewer adverse reactions result from the use of low-osmolality contrast media than from the use of conventional high-osmolality contrast media, which means that the cost of managing complications may be lower with the former than with the latter.

Similar principles apply to valuing the indirect cost savings associated with applying one technology in lieu of another. The increased earnings associated with incremental gains in life expectancy or reduced morbidity may be counted among the savings. For instance, if applying a technology increases a patient's life expectancy by two years, during which the individual is expected to earn $60,000, the indirect savings would be $60,000.

Depending on the purpose of a clinical-economic trial, the availability of data, or the availability of resources for conducting the study, researchers may assess a limited range of resources or costs. The resulting picture of the technology's economic implications may therefore be incomplete, and decisionmakers who try to use the data may not understand what has been included in the analysis.

| Selection of an Economic Perspective

An important initial step in any clinical trial is to determine the perspective of the analysis. Analyses can be performed from the point of view of society, of the health care provider (clinic, hospital, or physician), of the payer (Medicare, a private insurance company, or an HMO), or of the patient. The exact methods of collecting economic data and the types of economic data collected may vary for different perspectives.

Choosing a perspective is critical in the design of a clinical-economic trial, because the perspective dictates what resources will be examined, what types of cost data will be collected, how the analysis will be structured, and ultimately what kinds of conclusions and recommendations will emerge. A common problem with economic analyses is the failure to establish or clarify the analytic perspective (24). This failure can result in the collection and synthesis of economic data that are not pertinent to some decisionmakers. For example, using data on a provider's billed charges to estimate the costs of some resources (from the payer's perspective) and using a provider's own cost data to estimate the costs of other resources (from the provider's perspective) yields a hybrid result that represents the view of neither the payer nor the provider.

In addition, there is a common misconception that only one correct perspective exists for measuring costs. Some analysts believe that the societal perspective is best because it addresses the public good. Although the societal perspective may be favored for national allocative decisions and may address the public interest, it does not always address the needs of specific decisionmakers, such as third-party payers, employers who pay for care through health premiums or self-insurance, providers who must decide whether a technology is worth using, or patients who face out-of-pocket costs. The correct perspective is the one that will yield information of relevance to the decisionmaker, whoever that might be.

Another misconception is that the conclusions drawn from a clinical-economic trial will be the same from all perspectives. In fact, however, conclusions can vary substantially. For example, one cost-benefit analysis performed in a clinical-economic trial of the use of low- versus high-osmolality contrast media in cardiac catheterizations suggests that the higher material cost of using low-osmolality contrast media is partially offset by the reduction in the costs of managing adverse reactions, but that the offset is lower from the hospitals' perspective than from society's perspective and that it may not be realized by third-party payers (50).

| Resource Measurement Methods

In addition to clinical information (such as risk factors and outcomes), two types of data are required for a clinical-economic trial: the medical and nonmedical resources that are consumed and the costs of those resources (from the chosen perspective). Quantitative measures of patient preferences for clinical benefits *(patient utilities)* provide another important measure of economic outcomes, but the appropriateness of performing utility measurements on persons enrolled in clinical trials is still the subject of debate. A clinical trial may not always be the appropriate setting for eliciting patients' preferences for outcomes or treatment, inasmuch as the process of the trial may itself influence patients' responses.

Of the several ways to collect data on the resources consumed in the context of a clinical trial, some are better suited than others to particular purposes. Which approach is most useful depends on the characteristics of the technology, the patient population, the clinical setting, and the perspective of the analysis.

Reviewing Charts

The medical records of the patients enrolled in a trial can be reviewed to abstract data about a variety of resources (8,25), including admissions to the hospital, the use of laboratory or imaging tests, other ancillary services (e.g., electrocardiograms, foley catheters and respiratory or physical therapy), any consultations by specialists, and the days spent in a special care unit (e.g., an intensive-care unit or a laminar-air-flow room). Regardless of whether the review is performed to document outpatient or inpatient services, the medical records must be complete, and those persons reviewing charts must be trained in medical record abstraction to avoid problems such as overcounting resources because of imprecise documentation. (For example, if the time of day is recorded inaccurately, abstracters may have difficulty determining whether separate documentation refers to the same test.)

Chart reviews can be problematic where therecordkeeping is below standard or where the records are kept in different places. Problems also arise when test results are not recorded in the medical record until weeks after the tests have been performed and the orders do not document the submission of samples for testing (e.g., when a physician sends blood for thyroid-function tests directly to the laboratory without entering the information in the order sheet). Abstracters face other difficulties if they must use records at more that one place (e.g.. in a multicenter study) and the organization of the records varies from site to site. An assessment of the interrater and intrarater reliability of the record abstracts is important for any study that uses data from charts.

One limitation of medical records is that they generally do not provide data on direct non-medical resources or on the resources used to estimate indirect costs. The records may also fail to document such direct medical costs as nonprescription drugs taken at patients' homes.

Examining Patients' Bills

Patients' bills are often the source of documentation on the use of medical resources (61,65). In some cases, bills can back up poor recordkeeping, but they only capture the resources that are charged to the patient or to a third-party payer. Many of the resources that are consumed never appear on patients' bills, either because the providers do not receive additional reimbursement for the resources or because the provider sometimes neglects to bill for reimbursable services. At the same time, some of the resources that appear on patients' bills may never have been used for those patients. The extent to which these errors occur may vary from one institution to the next.

Using patients' bills to ascertain the consumption of medical resources maybe most appropriate for assessing costs from the perspective of a third-party payer, because the bills reflect the resources for which the payer will be asked to pay. Patients' bills may not be so useful for identifying costs from the provider's viewpoint, inasmuch as some of the resources consumed may not appear on the bills. Another limitation is that bills from institutional providers do not document the provision of physicians' care, which is usually billed separately. In addition, physicians' bills maybe generated from multiple sources, which makes it difficult to collect data on all the care provided by physicians.

Interviewlng Providers or Patients

Another technique for obtaining data on resources is to ask providers about the services they performed or ordered for patients, or to ask patients about the services they received (65). Although interviews are straightforward when used to identify a few obvious and highly visible resources, they can be complex if they include many economic aspects of treatment, such as complications of treatment, disability, and work loss. Detailed interview protocols with branching logic may have to be developed. In addition, providers who participated in only a limited aspect of the patients' care may have no information about the broader and longer-term consequences of the treatment.

In such circumstances, questionnaires regarding the consumption of medical resources should be administered soon after the services have been provided in order to avoid problems with recall. In addition, patients' questionnaires should be simplified to capture only general categories of resource consumption, because most patients cannot provide detailed information either because they were too sick to comprehend their physicians' explanations or because they were never informed in a detailed way (e.g., they knew that they had blood tests or x-rays, but not whether they had complete blood counts or magnetic resonance imaging).

Interviews with patients are particularly useful when patients are served by several health care providers or have multiple sources of payment that are not captured by a single data collection system. To minimize difficulties with recall, patients can also keep diaries or logs of the health care resources they use. Finally, interviews may be the only way to identify the use of resources that comprise direct nonmedical costs or indirect economic costs (such as days lost from work).

Conducting Time-and-Motion Studies

In some studies, researchers must measure the process of producing health services on a more detailed chronological basis (e.g., minute-by-minute) and note all of the labor and nonlabor resources that are expended (17,26,47). To measure the effect of a device for patient-controlled analgesia versus nurse-delivered analgesia for patients who have had recent surgery, for instance, researchers might examine how many minutes nurses use for each strategy to relieve postoperative pain. Typically, in a time-and-motion study, the nurses would be directly observed and the tasks timed with a stopwatch (1).

Although time-and-motion studies yield very accurate results, they are expensive to perform, because they require intensive observation by the researchers. Another concern is that those being observed may alter their behavior in response to the observation.

| Cost Assignment Methods

Once the researchers have measured the resources that are consumed by participants in a clinical-economic trial, they must assign the appropriate costs to those resources. It is widely recognized that the actual costs of providing health care services are likely to differ substantially from the charges that providers submit to patients or third-party payers (20). Charges often reflect what the market will bear, rather than the true cost of there-sources consumed in providing health care services. Therefore, charges are often set arbitrarily and may vary substantially among facilities whose costs for producing health services are similar. Furthermore, submitted charges are not always fully paid. The amount paid can vary by payer, delivery system (e.g., negotiated discounts by a managed care insurer), and geography (e.g., state-mandated inpatient rate-setting in Maryland). These factors demand that researchers do more than simply collect information about what charges were submitted.

In estimating the total costs associated with a medical technology or clinical management strategy, researchers often take one or both of two approaches.

1. They may build up the costs from the level of the individual resources (such as a dose of penicillin or an hour of a nurse's time), an approach often referred to as *microcosting.*
2. *They* may assign costs to resources at an aggregated (bundled) level of resources (such as a hospitalization or a clinic visit). Under this approach, investigators often use cost-to-charge ratios specific to the institutions or cost centers to estimate the actual costs from the charges.

One practical concern is the availability of systems that can yield data on the various types of costs incurred by institutions or providers. Although the systems of some hospitals and clinical practices are sophisticated (40), those of others are archaic. This may limit researchers' ability to perform studies in some settings or limit the type of data that can be collected (e.g., charges versus costs), which may explain why so many studies have reported on charges rather than costs (5,22,28,46,48,49,65,74).

Microcosting

A very time-consuming process of collecting cost data, microcosting usually requires investigators to work with the **staff** of a hospital or clinic to identify the expenses for various resource inputs, such as capital, labor, and supplies (17,32,33,54). In an institution where purchasing and hiring decisions are decentralized, this process may entail contact and discussions with a large number of organizational units (e.g., the pharmacy, nursing, cardiology, laboratory medicine, physical therapy, and professional-fee billing departments).

Some institutions or departments may have sophisticated methods of evaluating the true costs of producing services. For example, the pharmacy may have developed standards for valuing the pharmacists' time, the ancillary supplies, the material used in acquiring a dose of antibiotic from the supplier and getting it to the bedside and into the patient. A detailed and well-documented, centralized cost accounting system can help a great deal. Researchers sometimes turn to published estimates of costs—using, for example, the *Drug Topics Redbook* to arrive at the cost of materials for pharmaceuticals. Although the publication provides useful approximations, in that it lists wholesale costs, the discounts often given to health care institutions by suppliers are not reflected. To assess the use of outpatient resources, researchers might conduct a local survey of providers and calculate the average cost (or charge) for a service (58).

Additional methodological issues in micro-costing concern the allocation of overhead costs (also referred to as indirect accounting costs). Average costs include overhead, but analysts may differ on the extent to which certain categories of overhead—such as departmental overhead (e.g., the department manager's salary) and hospital-wide overhead (e.g., the chief executive officer's salary) should be included (39,54). Different approaches to the inclusion of various types of overhead may yield vastly different results. Standardization is lacking.

Assigning the Costs of Aggregated *Resources*

Whether assigning costs on an aggregate level is appropriate depends on the study's perspective and the availability of data. For instance, aggregation at the hospital discharge level may be appropriate in assigning Medicare's cost for a hospital admission using payments based on diagnosis-related groups (DRGs). Cost-to-charge ratios provide a convenient way to estimate the actual costs for medical services from the charges to payers. A ratio of .80 would imply that the true cost of a service is 80 percent of its charge.

Although cost-to-charge ratios are commonly used to estimate the cost of hospital services (8), there is some debate about how the methods are used. Many researchers have used the ratios from Medicare cost reports, which are widely available, but some investigators advocate the use of department-specific ratios, while others believe that the less complex institutional level ratios are adequate. The distinction is that department ratios purport to value like services the same (e.g., the cost-to-charge ratio for a chest x-ray is the same as that for a knee x-ray, but differs from the ratio for a blood glucose test), whereas institutional level ratios value all services the same (e.g., the ratio is the same for a CT scan and an antibiotic). Thus, institutional-level cost-to-charge ratios may not account for the fact that expensive devices or drugs often have lower ratios than other resources. In addition to obtaining ratios from Medicare cost reports, researchers have used ratios that were generated for purposes other than cost reporting to the Medicare program and that relied less on the grouping of dissimilar services.

Assigning the Costs of Professional Services

Data on the costs of professional services can also be obtained in different ways. The actual costs of physicians' services are difficult to obtain. The new resource-based relative value scale (implemented by Medicare as a basis for physician payment) (35) may be helpful in this regard. It is possible to estimate the actual costs by using the average reimbursements for a geographic area (e.g., by looking at payments by Medicare intermediaries) or from fee schedules maintained by third-party payers.

At the institutional level, it may be possible to identify the actual fees reimbursed by different third-party payers or an average for various payers over a specified period. Some studies have calculated average reimbursements from billed amounts by using a ratio of collected-to-billed charges (40). In examining physicians' costs, knowledge of physicians' billing practices (such as bundling of services) and coding is important.

Assigning Indirect Economic Costs

Researchers estimating indirect economic costs may find it difficult to obtain information about incomes and benefits directly from patients. Alternative sources used by investigators include standard industry profiles merged with primary data (such as the numbers of days lost from work and the nature of patients' occupations).

VALIDITY AND RELIABILITY

The validity of conducting economic analysis during a clinical trial reflects the degree to which the data used for estimation reflect the actual resources or costs of providing services. Its reliability reflects the degree to which investigators would obtain the same results if the study were repeated on the same population of patients and providers. Because of the relative newness of economic data collection in clinical trials of medical treatments, the literature contains few methodologic studies that compare approaches for obtaining economic data. Furthermore, because costs—unlike some clinical variables-change over time, both reliability and validity can be difficult to assess.

Several of the detailed methodological issues and choices described above can potentially affect the validity and reliability of data based on clinical-economic trials. In addition, there are other issues that relate to the basic characteristics of economic data collected during the course of a clinical trial of a technology's efficacy. Some of these broader issues are described below.

| Statistical Distributions of Costs

One particularly problematic issue in measuring and analyzing the use of medical resources is the way in which resource utilization and costs are distributed. The distributions typically are skewed, either with a few persons using a few services (or not using any services) or with a few persons using large amounts of resources. Unfortunately, excluding the patients who use considerable resources may be undesirable because they are important to decisionmakers.

In view of the large variance in costs, large numbers of study participants are usually needed for adequate statistical power. A potential pitfall of incorporating an economic component into a clinical trial is that the sample size needed to test economic hypotheses may exceed that needed to test clinical hypotheses, because of differences in the clinical and economic variables. Several biostatistical techniques, such as transformations of data to achieve normality (e.g., through logarithmic calculations) or the use of hierarchical models (14), are useful both in sample-size calculations and in the analysis of highly skewed data. Some researchers have suggested that in fact costs in trials often need not, and should not, be measured to the same level of statistical precision as health effects in clinical trials (13a,47a).

The problem of disparate sample sizes can be exacerbated by interim findings on the efficacy side of the trial. Clinical trials often have rules for terminating trials when clinically and statistically significant differences in clinical outcomes are observed. For instance, a trial might be stopped when the new treatment is shown to be efficacious at a predefine level of statistical significance after only half the anticipated trial participants have been enrolled. Although stopping a trial because of the clinical results may conserve resources or satisfy ethical considerations, the early termination of a clinical-economic trial could prevent researchers from drawing conclusions about cost-effectiveness if the sample patients had not reached the level needed for examining important economic outcomes.

| The Influence of Clinical Protocols on the Use of Resources

Another important consideration in estimating the use of resources in clinical trials is the extent to which clinical protocols might influence the use and costs of resources. For example, early studies of a new technology often include the performance of laboratory or radiologic tests to monitor patients for serious or unknown side effects. This monitoring is often driven by the clinical research protocol, which can be influenced by the need for data to assist in the FDA approval process. The monitoring can significantly alter the validity of an economic analysis, however, because the resources consumed in the monitoring process do not always become a necessary component of routine clinical practice. This is true for both the treatment group and the control group in a randomized trial.

In the Women's Health Study conducted by NIH to assess hormonal therapy, for instance, women undergo frequent office followups, electrocardiograms, endometrial biopsies, and mammograms to monitor the safety of hormonal therapy. The exclusion of these resources from the data collection or from the accounting of costs is often proposed as a way to solve the problem, but the monitoring can have more profound effects when abnormal tests lead to further testing or treatment.

Studies that examine the use of aggregated resources without examining the components and their relationship to clinical events are more likely to encounter this flaw than studies that take care to attribute the use of resources to clinical events (51). One solution in a multicenter study might be to modify the clinical protocol at some of the institutions and examine how the resource consumption varies depending on whether a center uses the standard or modified protocol. Although excluding some costs may seem reasonable, however, recommendations for dosing and safety monitoring of complications after FDA approval are often based on the protocol established in the clinical trials. Therefore, investigators collecting and analyzing economic data may wish to perform sensitivity analyses-that is, to perform their analyses more than one way, based on the inclusion or exclusion of certain categories of resources.

I Standardization in Multicenter Trials

Increasingly, large clinical trials are conducted at multiple sites. Multicenter studies raise important issues of standardization, and the best centers for collecting clinical data are not always the best centers for collecting economic data. Although investigators may easily develop standardized criteria for the collection and determination of clinical events (e.g., electrocardiographic and cardiac enzyme evidence for acute myocardial infarction), the standardization of costs is difficult because of differences in accounting systems across sites. Some centers may have sophisticated methods of ascertaining their costs or specific billing information, for instance, while other centers do not. Researchers may also find it difficult to standardize the measurement of costs for different types of providers (e.g., HMOs versus fee-for-service practices, hospital outpatient departments versus physicians' offices, or VA hospitals versus private hospitals) and for different geographic areas (e.g., states with inpatient rate regulation versus states with market competition, or Canadian facilities versus U.S. facilities).

| The Effect of Masking on the Use of Resources

Because it minimizes bias related to treatment, the double-blind trial—in which both patients and physicians are unaware of who is receiving which treatment alternative-is considered an important component of the evaluation of the efficacy and safety of a therapy. In assessing the economic implications of a treatment in normal practice, however, an equally important consideration is the fact that awareness of the treatment can influence providers to use resources in a different fashion.

Suppose, for instance, that radiologists in clinical practice were more likely to initiate aggressive and expensive treatments for contrast-induced complications if they knew that the patients had received high-osmolality rather than low-osmolality contrast media. In a masked clinical trial, the radiologists might show less restraint in their use of medical resources to manage complications, because they could not be certain which contrast media had been used. Uncertainty as to treatment (as well as more intensive observation) in the clinical trial might influence radiologists to provide more treatment than they would in normal practice. As a result, the masking might increase the costs.

I The Timing of Clinical and Economic Outcomes

The economic consequences of treatment choices may extend far beyond the time horizon of a clinical trial. For example, thrombolytic therapy (e.g., recombinant tissue plasminogen activator) administered for an acute myocardial infarction can cause a stroke, a clinical endpoint, and the patient could require long-term nursing care, the cost of which could extend for many years. If the clinical protocol stipulated that followup on a patient would end in the event of a stroke, the full economic consequences would not be obtained.

Clinical benefits and costs may accrue at different times in the course of an illness, and clinical benefits or costs may accrue at different times for each treatment strategy being compared in a clinical-economic trial. An analysis of the benefits and costs accruing at different times must take this into account by adjusting the observed costs for the time value of money. Benefits and costs that accrue now are worth more than they would be if they accrued in the future. The procedure for adjusting for the time value of the resources or costs is referred to as *discounting,* in which benefits and costs incurred in the future are valued in current dollars (73).

Discounting is unnecessary if the time covered by the analysis is short (e.g., less than one year). When discounting is necessary, the rate at which to discount is often controversial because the choice can greatly influence the conclusion about a technology's cost-effectiveness. Different consumers of economic data (including those only affected in the future) might advocate different rates. Therefore, analyses might examine the sensitivity of conclusions drawn from clinical-economic trials to the rate of discounting.

| Generalizability

The external validity (i.e., generalizability) of the economic data collected in clinical-economic trials is of great concern. Studies are often performed at individual institutions that are part of, or affiliated with, academic medical centers, where two possible problems influence generalizability.

First, the medical practice may not be similar to that in many other institutions. For example, physicians at teaching institutions may order more tests and consume more resources as a result of their teaching or research activities, which may result in an overestimation or underestimation of costs. More discretionary testing raises cost estimates, although more careful testing may prevent complications of treatment and, therefore, result in the trading of an upfront outlay for a potential reduction in the long-run use of resources. Institutions that adopt technologies early may have the most experience in their application. This experience could lead physicians to select patients more ideally suited for treatment or to be better at identifying or managing side effects. This might translate not only into better outcomes than are realized in general medical practice (51) but also into the more efficient use of resources.

Second, in addition to differences in physicians' practices, the costs of resources vary across institutions of different sizes (economies of scale or scope), location (geographic variation in resource inputs), and organizational characteristics (for-profit versus not-for-profit institutions). Manufacturers that perform economic studies in different countries must be aware of the variability in medical practices, medical costs, and the medical infrastructures required to support use of new technologies.

Other factors to consider are that, for the purposes of clinical trials, investigators may be able to obtain related medical services (such as monitoring tests) at discounted prices or costs. The discounts may not reflect the true economic cost in everyday practice.

It is worth noting that cost-effectiveness analyses based on synthetic or modeled analyses that use the best data available from various published and unpublished sources, including opinions (55), are not immune from the problem of limited generalizability. They, too, may be limited because they project how resources would be used under optimal circumstances. The effects of the assumptions that are made in using such data and the validity of the estimates of the cost-effectiveness or cost-benefit that are generated are often unknown.

FUTURE PROSPECTS

I Research Needs

The analysis of economic data in clinical trials is afield still in its infancy. In view of how many entities are interested in performing studies and what kinds of techniques can be used in the process, there is a need for studies that compare the results generated by different methodologies and techniques. Few studies have addressed the reliability and validity of cost assignment methods by comparing different methods of obtaining, calculating, or modeling costs (32). Such studies are badly needed in order to improve our understanding of how alternative methods affect the results of economic analyses in clinical trials.

Such studies would compare alternative methods of collecting both resource-utilization data and cost data (including modeling) for the same technologies. The studies would also explore the degree to which summary measures, such as cost-benefit or cost-effectiveness ratios, are affected by the data collection methods. This could be done by analyzing the benefits (or effectiveness) and costs that were measured within the same trial using different techniques.

The extent to which the characteristics of a technology dictate the best approach to collecting data on resources and costs is unclear. The approaches required by diagnostic technologies may differ from those required by therapeutic technologies. Inpatient and outpatient treatments may also require different approaches to resource measurement and the assignment of costs. The best approach for one ailment (such as cardiovascular disease) may differ from that for another (such as arthritis), and chronic diseases that have longer durations may require different approaches from those required by acute diseases.

Another gap in the literature is the lack of studies examining the relative gain from careful attention to the precision of assessments. This is important because the costs of collecting data usually rise with more detailed assessments. We also need to understand how much generalizability increases when economic data are collected from more than one institution, inasmuch as it costs more to collect data from multiple sources. Future studies could assess whether methodological shortcuts are possible and yield valid results.

It is not clear that the private sector, particularly industry, can or will support the necessary methodological development or the particular applications to research in this field. Although it may be able to support a specific evaluation related to a particular need, the private sector has little incentive to take on the tasks of developing methods or examining the economic issues from more than one perspective.

No government agency has currently embraced the responsibility for supporting the development of methods for collecting economic data in clinical trials and for integrating them into clinical trials. In part, this reflects the fact that specific funding has been limited or that it competes with other programmatic areas. AHCPR has a mandate for examining the cost of health care services, and NIH has the authority and are supporting large-scale clinical trials of new therapies, but neither agency has undertaken primary responsibility for research that intersects these areas.

| Standardization

The quality of economic analyses is of considerable concern as the methodology evolves (30) with no established guidelines on appropriate techniques and no consistency in technique across studies. If information about cost-effectiveness is to be useful as a criterion for decisions such as whether a drug is to be included in a hospital formulary, or whether a procedure should be covered by insurance, some standards for the types of data and the methods of obtaining them must be developed (41). Questions to be addressed include whether more than one perspective should be considered in economic analyses, what types of costs should be considered, and (in order to make allocative decisions) what constitutes the appropriate patient or provider population for an economic study?

Because there are few standards for the proper conduct of clinical-economic studies, studies are susceptible to accusations of bias, particularly if the study results favor the sponsor product or interests. Much of the concern relates to the fact that invalid or unreliable approaches (such as the incomplete enumeration of resources or costs) may be used selectively to obtain particular desired results. Some of this concern could be alleviated with greater methodological standardization.

There may also be pressures to refrain from publishing results that are unfavorable to the sponsors' interests. The degree of publication bias—the tendency for over-representation in the published literature of studies with statistically significant results, or studies whose results favor currently accepted theories—in the cost-effectiveness field generally is unknown, but some observers believe that studies are less likely to be published if they fail to show that a medical treatment saves money or is significantly cost-effective. Publication bias limits the number of studies and results that can be compared by decisionmakers, and it may lead users to draw incorrect conclusions about a technology's overall cost-effectiveness. In addition, it may lead researchers to take methodologically inferior approaches that are more likely to yield positive results.

Conversely, where results favor the sponsor, they sometimes may be disseminated (e.g., used in marketing efforts) without having been adequately peer-reviewed. Both of these factors make it difficult for decisionmakers to trust cost-effectiveness findings. To address these problems, some researchers have advocated the development of rules of conduct for the dissemination as well as the performance of clinical-economic research (30).

| Cost-Effectiveness of Clinical-Economic Trials

Not all clinical trials are good candidates for economic data collection and analysis. Adding an economic component to a clinical trial adds to the cost of the trial (see box 5-3). In view of the considerable expense of clinical-economic analyses and the limitations in generalizing from them, the collection and analysis of economic data in clinical trials may not always be the best way to reach conclusions regarding anew technology cost-effectiveness.

Nonetheless, economic information is badly needed by patients, providers, and payers alike as the nation grapples with the question of how to provide good care at the lowest cost. The clinical-economic trial generates explicit information about which alternative treatment options are the most cost-effective, and it can provide this information early in the life of a new technology, before its use becomes widespread.

Equivalence trials may be particularly appropriate contexts in which to conduct clinical-economic studies. In contrast to a difference clinical trial, which attempts to demonstrate a difference between two therapies, an equivalence trial is an attempt to discover whether one treatment strategy is equivalent to another, perhaps more expensive, strategy. The combined BARIE/SEQOL study, described in box 5-3, is an example of an equivalence trial in that investigators are seeking to establish whether angioplasty—an alternative with potentially lower initial costs—is clinically and economically equivalent to coronary artery bypass surgery.

| Conclusions

The demand for early information on the costs and effectiveness of new technologies is driven by health care policy makers who hope to improve

BOX 5-3: Resources Required for Collecting Economic Data

Economic evaluations within clinical trials add to the cost of the clinical trials, although the additional (or incremental) costs of collecting economic data are less than they are for studies designed strictly to answer economic questions. A clinical trial funded by the National Institutes of Health illustrates this point. The Bypass and Angioplasty Revascularization Investigation (BARIE) is a $35-million, 1,800-patient, 14-center clinical trial that is randomizing patients to receive either angioplasty or coronary artery bypass surgery for symptomatic multivessel coronary artery disease. The trial's major endpoints are death and other morbid cardiovascular events. The study began in 1988, and researchers finished recruiting patients in 1991; The five-year followup will be analyzed in 1996. An evaluation of the economics and the patients' quality of life (SEQOL) (31) was added to the study and funded by the Robert Wood Johnson Foundation at $4.25 million, which was roughly 12 percent of the cost of the clinical trial.

There are five major determinants of the costs of an economic evaluation in a clinical trial:

1. The number of additional research personnel needed for collecting and analyzing data. Few investigators with backgrounds in clinical medicine or epidemiology have also had formal training in such disciplines as health economics, accounting, or health care finance, but the research staff must include personnel with the training to design and help carry out the economic component of the trial.
2. The number of study participants and the duration of patient followup necessary for observing the care. As is true for any clinical study, a clinical-economic evaluation's cost usually varies positively with the length of the observation period and the number of patients studied.
3. The type and comprehensiveness of the economic data elements collected (such as the number of perspectives chosen and the types of costs included). If investigators want greater detail about the use of resources (e.g., ambulatory as well as inpatient services), the costs of data collection rise when the efforts of the current clinical research is fully expended.
4. The extent to which the use of resources can be measured from automated databases rather than by hand. Comprehensive data systems are extremely efficient, which makes the per-patient cost of collecting economic data decrease as the number of patients rises. Most systems, however, are insufficient for the valid identification of resources and costs (e.g., they may include only data on charges or average costs rather than data on marginal costs). This means that the investigators may have to abstract data on resources from patients' records, patients' bills, or surveys of patients; and to collect data on costs from cost (or expense) reports or from manufacturers' invoices.
5. The extent to which modeling and data collection outside the trials are necessary to answer economic or cost-effectiveness hypotheses. Often, the amount of data needed to perform a cost-effectiveness or other economic analysis cannot be generated solely from the patients who are enrolled in a trial. For example, if the researchers need data on patients' preferences for each of the possible outcomes associated with a technology, some of the data may need to be obtained from patients who are not participating in the trial.1 The substantial modeling of data from the clinical trials to simulate or project economic implications for a collection of providers, a region, or the nation can be a labor-intensive task that is possible only after the results from the primary data collection are available.

[1] The preferences of trial participants may differ from those of patients who were not eligible for the trial. Patients' preferences could also be affected by the trial participationitself.

SOURCE: Neil R. Powe and Robert 1Griffiths, 1995.

medical care without increasing its costs; by providers who want to remain competitive in a cost-conscious environment; by insurers who must make decisions about coverage and reimbursement; and by manufacturers who adapt their research and marketing strategies in response to these concerns. In view of these demands, clinical-economic trials are likely to become increasingly common.

There are tradeoffs between the limitations inherent in clinical-economic trials and the need to anal yze a treatment cost-effectiveness before the technology becomes widely (and perhaps irreversibly) adopted by the medical community. This suggests that there is not one optimal time in the life cycle of a technology to perform a clinical-economic trial, but that researchers and users must understand the limitations in the data (that is, the conditions under which data are generated) and be willing to adjust the estimates in accordance with new medical knowledge or practice patterns.

Despite the demand for sound economic information about medical technologies, the field may not develop in tandem with the needs of policymakers. The need for more methodological research and standardization in particular are potential barriers to the development and wider use of economic evaluation methods in clinical trials.

REFERENCES

1. Achusim, L. E., Weller, T. W., Somani, S. M., et al., "Comparison of Automated and Manual Methods of Syringe Filling, ''*American Journal of Hospital Pharmacy 47:2492 -2495, 1990.*
2. Adams, M. E., McCall, N. T., Cray, D. T., et al., "Economic Analysis in Randomized Control Trials," *Medical Care, 30(3):231-243, 1992.*
3. Anderson, G. F., and Steinberg, E. P., "To Buy or Not To Buy: Technology Acquisition Under Prospective Payment," *New England Journal of Medicine 3* 11: 182-185, 1984.
4. Baumann, R., Magos, A. L., and Tumbull, S. A., "Prospective Comparison of Videopelviscopy with Laparotomy for Ectopic Pregnancy," *British Journal of Obstetrics and Gynecology 98:765-771, 1991.*
5. Boettcher, W. G., "Total Hip Arthroplasties in the Elderly: Morbidity, Mortality, and Cost Effectiveness," *Clinical Orthopedics and Related Research 274:30-34, 1992.*
6. Clinton, B., "The Clinton Health Care Plan," New *England Journal of Medicine 327(11): 804-807, 1993.*
7. Cummings, J. E., Hughes, S. L., Weaver, F. M., et al., "Cost-Effectiveness of Veterans Administration Hospital-Based Home Care: A Randomized Trial," *Archives of Internal Medicine 150:1274-1280, 1990.*
8. de Buitleir, M., Sousa, J., Boiling, S.F., et al., "Reduction in Medical Care Cost Associated with Radiofrequency Catheter Ablation of Accessory Pathway s," *American Journal of Cardiology 68(17):1656-1661, 1991.*
9. Detsky, A. S., "Are Clinical Trials a Cost-Effective Investment?" *Journal of the American Medical Association 262: 1795-1800,* 1989.
10. Detsky, A. S., and Naglie, I. G., "A Clinician's Guide to Cost-Effectiveness Analysis," *Annals of Internal Medicine 113(2): 147-154, 1990.*
11. Dickersin, K., "The Existence of Publication Bias and Risk Factors for Its Occurrence," *Journal of the American Medical Association 263: 1385-1389, 1990.*
12. Drummond, M. F., "Basing Prescription Drug Payment on Economic Analysis: The Case of Australia," *Health Affairs,* 11(4): 197-201, winter 1992.
13. Drummond, M. F., and Davies, L., "Economic Analysis Alongside Clinical Trials: Revisiting the Methodological Issues," *International Journal of Technology Assessment in Health Care 7(4):561-573, 1991.*

13a. Drummond M., and O'Brien, B., "Clinical Importance, Statistical Significance and the Assessment of Economic and Quality-of-Life Outcomes," *Health Economics 2:205-212, 1993.*

14. Duan, N., Manning, Jr., W.G., Morris, C. N., et al., "A Comparison of Alternative Models for the Demand for Medical Care," *Journal of Business Economic Statistics 1(2): 115-126, 1983.*
15. Eisenberg, J. M., "Clinical Economics: A Guide to the Economic Analysis of Clinical *Practices, ''Journal of the American Medical Association 262(20):2879-2886, 1989.*
16. Eisenberg, J. M., Glick, H., and Koffer, H., "Pharmacoeconomics: Economic Evaluation of Pharmaceuticals," *Pharmacoepidemiology,* B.L. Strom (cd.) (New York, NY: Churchill Livingstone, 1989).
17. Eisenberg, J. M., Koffer, H., and Finkler, S.A., "Economic Analysis of a New Drug: Potential Savings in Hospital Operating Costs from the Use of a Once-Daily Regimen of a Parenteral Cephalosporin," *Reviews of Infectious Diseases 6(Suppl. 4): S909-S923, 1984.*
18. Eschbach, J. W., Abdulhadi, M. D., Browne, T.J., et al., "Recombinant Human Erythropoietin in Anemic Patients with End-Stage Renal Disease: Results of a Phase III Multicenter Clinic Trial, ''Annals *of lnternal Medicine* 11(1):992-1000, 1989.
19. Ferguson, J. H., "NIH Consensus Conferences: Dissemination and Impact, ''Annals *of the New York Academy of Sciences 703:180-199, 1993.*
20. Finkler, S. A., "The Distinction Between Cost and Charges," *Annals of lnternal Medicine 96: 102-109, 1982.*
21. Friedman, E., "The Uninsured: From Dilemma to Crisis," *Journal of the American Medical Association 265(19):2491 -2495, 1990.*
22. Gerding, R. L., Emerman, D. L., Effron, D., et al., "Outpatient Management of Partial-Thickness Bums: Biobraneo Versus 1% Silver Sulfadiazine," *Annals of Emergency Medicine 19: 121/29-124/32, 1990.*
23. Ginzberg, E., "High-Tech Medicine and Rising Health Care Costs," *Journal of the American Medical Association 263(13):* 1820-1822, 1990.
24. Goddard, M., and Drummond, M., "The Economic Evaluation of Cancer Treatments and Programmes," *European Journal of Cancer 27(10):1 191-1196, 1991.*
25. Gray, D., MacAdain, D., and Boldy, D., "A Comparative Cost Analysis of Terminal Cancer Care in Home Hospice Patients and Controls," *Journal of Chronic Disease 40(8):801-810, 1986.*
26. Greendyke, R. M., "Cost Analysis: Bedside Blood Glucose Testing," *American Journal of Clinical Pathology 97: 106-107, 1992.*
27. Griffiths, R.I., Bass, E. B., Powe, N. R., et al., "Factors Influencing Third Party Payer Costs for Allogeneic BMT," *Bone Marrow Transplantation 12(1):43-48, 1993.*
28. Gulati, S. C., and Bennett, C. L., "Granulocyte-Microphage Colony-Stimulating Factor (GM-CSF) as Adjunct Therapy in Relapsed Hodgkin Disease," *Annals of Internal Medicine* 116(3):177-182, 1992.
29. Henry, D., "Economic Analysis as an Aid to Subsidisation Decisions: The Development of Australian Guidelines for Pharmaceuticals," *PharmacoEconomics 1(1):54-67, 1992.*
30. Hillman, A.L., Eisenberg, J. M., Pauly, M. V., et al., "Avoiding Bias in the Conduct and Reporting of Cost-Effectiveness Research Sponsored by Pharmaceutical Companies," New *England Journal of Medicine 324(19): 1362-1365, 1991.*
31. Hlatky, M.A., Charles, E. D., Nobrega, F., et al., "Initial Functional and Economic Status of Patients with Multivessel Coronary Disease," working paper, 1993.
32. Hlatky, M. A., Lipscomb, J., Nelson, C., et al., "Resource Use and Cost of Initial Coronary Revascularization: Coronary Angioplasty Versus Coronary Bypass Surgery," *Circulation 82(Suppl. 1V):208-213, 1990.*
33. Hlatky, M. A., Morris, K. G., Pieper, K. S., et al., "Randomized Comparison of the Cost and Effectiveness of Iopainidol and Diatrizoate as Contrast Agents for Cardiac Angiography," *Journal of the American College of Cardiology 16:871-877, 1990.*

34. Holahan, J., Dor, A., and Zuckerman, S., "Understanding the Recent Growth in Medicare Physician Expenditures," *Journal of the American Medical Association 263(12): 1658-1661, 1990.*

35. Hsiao, W. C., Braun, P., Yntema, D., et al., "Estimating Physician's Work for a Resource-Based Relative Value Scale," New *England Journal of Medicine 260:835-841, 1988.*

36. Hughes, S. L., Cummings, J., Weaver, F., et al., "A Randomized Trial of the Cost Effectiveness of VA Hospital-Based Home Care for the Terminally Ill," *Health Services Research 26(6):801-817, 1992.*

37. Institute of Medicine, *Clinical Practice Guidelines: Directions for a New Program,* M.J. Field and K.N. Lohr (eds.) (Washington, DC: National Academy Press, 1990).

38. Institute of Medicine, Committee for Evaluating Medical Technologies in Clinical Use, *Assessing Medical Technologies, C.* Goodman (cd.) (Washington, DC: National Academy Press, 1985).

39. Krueger, H., Goncalves, J. L., Caruth, F. M., et al., "Coronary Artery Bypass Grafting: How Much Does It Cost?" *Canadian Medical Association Journal 146(2): 163-168, 1992.*

40. Larsen, G.C., Manolis, A. S., Sonnenberg, F. A., et al., "Cost-Effectiveness of the Implantable Cardioverter-Defibrillator: Effect of Improved Battery Life and Comparison with Amiodarone Therapy," *Journal of the American College of Cardiology 19:1323-1334, 1992.*

41. Laupacis, A., Feeny, D., Detsky, A. S., et al., "How Attractive Does a New Technology Have To Be To Warrant Adoption and Utilization? Tentative Guidelines for Using Clinical and Economic Evaluations," *Canadian Medical Association Journal 146(4): 473-481, 1992.*

42. Leaf, A., "Cost Effectiveness as a Criterion for Medicare Coverage," New *England Journal of Medicine 321 (13):898-900.* 1989.

43. Leonard, B., Brust, J. D., and Sapienza, J.J., "Financial and Time Costs to Parents of Severely Disabled Children," *Public Health Reports 107(3):302-312, 1992.*

44. Meinert, C. L., *Clinical Trials: Design, Conduct and Analysis (New* York, NY: Oxford University Press, 1986).

45. Meyer, B.R., "Biotechnology and Therapeutics: Expensive Treatments and Limited Resources: A View from the Hospital," *Clinical Pharmacology and Therapeutics 51(4):359-365, 1992.*

46. Miller, M.J., Swartz, W. M., Miller, R. H., et al., "Cost Analysis of Microsurgical Reconstruction in the Head and Neck," *Journal of Surgical Oncology 46:230-234, 1991.*

47. Nicholl, J.P., Brazier, J. E., and Milner, P. C., "Randomized Controlled Trial of Cost-Effectiveness of Lithotripsy and Open Cholecystectomy as Treatments for Gallbladder Stones," *Luncet* 340(8823):801-807, 1992.

47a. O'Brien, B.J., Drummond, M. F., Labelle, R.J., et al., "In Search of Power and Significance: Issues in the Design and Analysis of Stochastic Cost-Effectiveness Studies in Health Care," *Medical Care 32(2):* 150-163, 1994.

48. O'Donoghue, S., Platia, E. V., Brooks-Robinson, S., et al., "Automatic Inflatable Cardioverter-Defibrillator: Is Early Implantation Cost-Effective?" *Journal of the American College of Cardiology 16:1258-1263, 1990.*

49. Peters, J. H., Ellison, E. C., Innes, J.T., et al., "Safety and Efficacy of Laparoscopic Cholecystectomy: A Prospective Analysis of 100 Initial Patients," *Annals of Surgery 213:3-12, 1991.*

50. Powe, N.R., Davidoff, A.J., Moore, R. D., et al., "Net Cost from Three Perspectives of Using Low versus High Osmolality Contrast Media in Diagnostic Angiocardiography," *Journal of the American College of Cardiology 21: 1701-1709, 1993.*

51. Powe, N.R., Grifflths, R. I., and Bass, E. B., "Cost Implications to Medicare of Recombinant Erythropoietin Therapy for the Anemia of End Stage Renal Disease," *Journal of the American Society of Nephrology 3:1660-1671, 1993.*

52. Powe, N. R., Griffiths, R. I., de Lissovoy, G., et al., "Access to Recombinant Erythropoietin by Medicare-Entitled Dialysis Patients in the First Year After FDA Approval, ''Journal *of the American Medical Association 268: 1434-14401992.*

53. Powe, N.R., Grifflths, R. I., Greer, J.W., et al., "Early Dosing Practices and Effectiveness of Recombinant Erythropoietin," *Kidney International 43:* 1125-1133, 1993.

54. Powe, N. R., Steinberg, E. P., Erickson, J. E., et al., "Contrast Medium-Induced Adverse Reactions: Economic Outcome," *Radiology 169(1): 163-168, 1988.*

55. Praecon Inc., "Adverse Reactions Associated with Contrast Media: The Economic Consequences" (proceedings of a conference), Winthrop-Breon Laboratories, 1986.

56. Physician Payment Review Commission, *Annual Report to Congress* (Washington, DC: U.S. Government Printing Office, 1991).

57. Reinhardt, U. E., "Commentary: Politics and the Health Care System," New *England Journal of Medicine 327(11):807-81* 1,1992.

58. Rice, T.D., Duggan, A. K., and DeAngelis, CO, "Cost-Effectiveness of Erythromycin Versus Mupirocin for the Treatment of Impetigo in Children," *Pediatrics 89(2):210-214, 1992.*

59. Runyon, B. A., McHutchison, J.G., Antillon, M. R., et al., "Short-Course Versus Long-Course Antibiotic Treatment of Spontaneous Bacterial Peritonitis: A Randomized Controlled Study of 100 Patients," *Gastroenterology* 100: 1737-1742, 1991.

60. Schulman, K. A., Glick, H. A., Rubin, H., et al., "Cost-Effectiveness of HA-IA Monoclonal Antibody for Gram-Negative Sepsis: Economic Assessment of a New Therapeutic Agent," *Journal of the American Medical Association 266(24):3466-3471, 1991.*

61. Showstack, J., Katz, P., Amend, W., et al., "The Association of Cyclosporine with the l-Year Costs of Cadaver-Donor Kidney Transplants,' 'Journal *of the American Medical Association 264(14): 1818-1823, 1990.*

62. Steinberg, E. P., Moore, R. D., Powe, N. R., et al., "Safety and Cost Effectiveness of High Osmolality as Compared with Low-Osmolality Contrast Material in Patients Undergoing Cardiac Angiography," New *England Journal of Medicine 326(7):425-430, 1992.*

63. Steinberg, E. P., Sisk, J. E., and Locke, K. E., "The Diffusion of Magnetic Resonance Imagers in the United States and Worldwide," *International Journal of Technology Assessment in Health Care 1(3):499-514, 1985.*

64. Steinberg, E. P., Topol, E.J., Sakin, J. W., et al., "Cost and Procedure Implications of Thrombolytic Therapy for Acute Myocardial Infarction,''Journal *of the American College of Cardiology 12:58 A-68A, 1988.*

65. Strauss, M.J., Gong, J., Gary, B. D., et al., "The Cost of Home Air-Fluidized Therapy for Pressure Sores: A Randomized Controlled Trial," *Journal of Family Practice 33(1):52-59, 1991.*

66. Sullivan, L. W., "The Bush Administration's Health Care Plan," New *England Journal of Medicine 327(1 1):801-804, 1992.*

67. Trisolini, M.G., McNeil, B. J., and Komaroff, A. L., "The Chemistry Laboratory, Development of Average, Fixed, and Variable Costs for Incorporation into a Management Control System," *Medical Care 25(4):286-299, 1987.*

68. Udvarhelyi, 1. S., Colditz, G. A., Rai, A., et al., "Cost-Effectiveness and Cost-Benefit Analyses in the Medical Literature: Are the Methods Being Used Correctly?" *Annals of Internal Medicine 116(3):238-244, 1992.*

69. U.S. Congress, Congressional Budget Office, *Projections of National Health Expenditures* (Washington, DC: October 1992).

70. U.S. Congress, Office of Technology Assessment, *Medicare's Prospective Payment*

System: Strategies for Evaluating Cost, Quality and Medical Technology, #1%-86-184 296/AS (Washington, DC: National Technical Information Service, 1985).

71. U.S. Congress, Office of Technology Assessment, *Evaluation of the Oregon Medicaid Proposal* (Washington, DC: U.S. Government Printing Office, May 1992).
72. U.S. Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research, "Cataract in Adults: Management of Functional Impairment," *Clinical Practice Guideline,* Number 4, prepared by D.M. O'Day, A.J. Adams, E.H. Cassem, et al., AHCPR Pub. *No. 93-0542 (Rockville, MD:* February 1993).
73. Weinstein, M.D., and Stason, W.B., "Foundations of Cost-Effectiveness Analysis for Health and Medical Practices," New *England Journal of Medicine 296(13): 716-721, 1977.*
74. Welch, H.G., Walsh, J. S., and Larson, E. B., "The Cost of Institutional Care in Alzheimer's Disease: Nursing Home and Hospital Use in a Prospective Cohort," *Journal of the American Geriatric Society 40:221-224, 1992.*
75. Wilensky, G. R., "Achieving Prescription Drug Savings," (letter), *Health Affairs 9(4):222, 1992.*