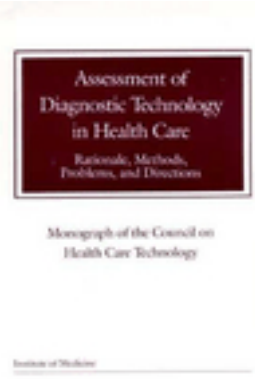

Assessment of Diagnostic Technology in Health Care: Rationale, Methods, Problems, and Directions

Council on Health Care Technology, Institute of Medicine

ISBN: 0-309-53588-3, 152 pages, 6 x 9, (1989)

Monograph of the Council on Health Care Technology

Assessment of Diagnostic Technology in Health Care

Rationale, Methods, Problems, and Directions

Harold Sox, Susan Stern, Douglas Owens, and Herbert L. Abrams

Institute of Medicine

NATIONAL ACADEMY PRESS
WASHINGTON, D.C. 1989

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

THE COUNCIL ON HEALTH CARE TECHNOLOGY was established in 1986 by the Institute of Medicine of the National Academy of Sciences as a public-private entity to address issues of health care technology and technology assessment. The council is committed to the well-being of patients as the fundamental purpose of technology assessment. In pursuing that goal, the council draws on the services of the nation's experts in medicine, health policy, science, engineering, and industry.

This monograph was supported in part by a grant to the Council on Health Care Technology of the Institute of Medicine from the National Center for Health Services Research of the U.S. Department of Health and Human Services (grant 5 R09 HS055 26 02). The opinions and conclusions expressed here are those of the authors and do not necessarily represent the views of the Department of Health and Human Services, the National Academy of Sciences, or any of their constituent parts.

Library of Congress Catalog Card Number 89-62666
International Standard Book Number 0-309-04099-X
Additional copies of this report are available from: National Academy Press 2101 Constitution Avenue, NW Washington, DC 20418

Printed in the United States of America

S033
First Printing, November 1989
Second Printing, September 1990

Council on Health Care Technology

Preface

In the recent past the interests of different groups concerned with health care have focused on the use of medical technologies—their impacts on safety, efficacy, and effectiveness; cost-effectiveness and cost-benefit; quality; and their social, legal, and ethical implications. The sum of these varied interests is the field of health care technology assessment.

The Council on Health Care Technology was created to promote the development and application of technology assessment in health care and the review of health care technologies for their appropriate use. The council was established as a public-private enterprise at the Institute of Medicine, a component of the National Academy of Sciences, through the Health Promotion and Disease Prevention Amendments of 1984 (P.L. 98-551, later amended by P.L. 99-117). In 1987 the U.S. Congress extended support for the council as a public-private venture for an additional three years (by P.L. 100-177).

The goals and objectives of the council, as stated in the report of its first two years of operations, are "to promote the development and application of technology assessment in medicine and to review medical technologies for their appropriate use. The council is guided in its efforts by the belief that the fundamental purpose of technology assessment is to improve well-being and the quality of care." In pursuing these goals, the council seeks to improve the use of medical technology by developing and evaluating the measurement criteria and the methods used for assessment, to promote education and training in assessment methods, and to provide technical assistance in the use of data from published assessments.

The council conducts its activities through several working and liaison panels. Members of these panels reflect a broad set of interested constituencies—physicians and other health professionals, patients and their families, payers for care, biomedical and health services researchers, manufacturers of health-related products, managers and administrators throughout the health care system, and public policymakers. In addition, it carries out councilwide activities that utilize the specific assignments of more than one panel.

This monograph contributes to the series of occasional publications produced by the council in carrying out its several missions. It examines two issues of special concern to the council—collection of primary data

and the assessment of diagnostic technologies—and explores innovative mechanisms, particularly reliance on multi-institutional approaches to assessment, to improve both the evaluation and the use of medical technology in ways that coincide with patient well-being.

William N. Hubbard, Jr., Chairman
Jeremiah A. Barondess, Co-Chairman

Contents

Introduction

In an era in which the pressure for containing health care costs is steadily mounting, the case for improving our methods of technology assessment is compelling. The rate of increase in health care expenditures during the past five years has been higher than at any time in our history (Califano 1987). A major element in cost is hospital care (Waldman 1972, Gibson and Mueller 1977). New and improved technologic services, procedures, and techniques (for example, drugs, diagnostic tests, cardiac surgery) are responsible for a large fraction of the rise in hospital costs (Waldman 1972). Innovative approaches have at times diffused rapidly, without carefully defined prospective studies to evaluate their precise role in diagnosis or therapy (Abrams and McNeil 1978b,c; Cooper et al. 1985). The dangers of *over*utilization have been appropriately emphasized (Abrams, 1979), but the biologic and monetary costs of *under*utilizing new and successful methods have not been adequately clarified (Doubilet and Abrams 1984). Proper utilization is impossible without timely, sophisticated, and precise evaluation of new methods (Abrams and McNeil 1978a,b,c).

While safety, cost, and patterns of diffusion are matters of profound concern, the most important and difficult aspect of technology assessment is the determination of efficacy (Abrams and McNeil 1978a, McNeil 1979, Adelstein 1982, Eddy 1982, Greer 1981, Yeaton and Wortman 1984, Petitti 1986). This requires consideration of a number of critical problems: timing, bias, primary data collection, new methods of secondary data analysis, the nature of the "laboratory" in which technology as

sessment is best accomplished, the "exploitative" character of assessment research, ethical issues, and diffusion.

In 1975, Congress asked the Office of Technology Assessment (OTA) to look into the problem of medical technology assessment. The OTA report (OTA 1976) led to the establishment of the National Center for Health Care Technology in 1978 (Perry and Eliastam, 1981). During its three-year history, the center acted as an important catalyst for more scientific evaluations of new and existing technologies and as a vehicle for disseminating information about them. The center ceased to exist in December 1981, after the Administration refused to request funds for its continuation out of deference to the two main sources of opposition, the American Medical Association and the Health Industry Manufacturers Association. The AMA opposed it because "the relevant clinical policy analysis and judgments are better made … within the medical profession." The Health Industry Manufacturers Association feared that the center "might constrain industry's freedom in the market place" (Perry 1982). The American College of Physicians, insurance carriers, the Association of American Medical Colleges, the Association for Advancement of Medical Instrumentation, and many other groups supported its continuance.

But the problem did not disappear. In a changing economic climate, all of the constituencies concerned with health care became more fully aware of the need for objective data on medical technologies.

By way of response, in 1983 the Institute of Medicine of the National Academy of Sciences issued a planning study report on assessing medical technology (Institute of Medicine 1983). From the report, there ultimately emerged a congressional bill—Public Law 98-551—which required that the already established National Center of Health Services Research add to its name and mission "Health Care Technology Assessment" and that a new Council on Health Care Technology be organized as an oversight group.

This time, the Health Industry Manufacturers Association reversed itself and endorsed the technology assessment provisions of the law, realizing that, in an era of restricted resources, the ever more cautious hospitals and buyers would require better data on efficacy, safety, and cost (Perry 1986).

The legislation was approved by the Congress, and both the center and the council were funded in 1986. The council is nongovernmental, an arm of the Institute of Medicine of the National Academy of Sciences. It represents a number of different constituencies: the physician providers;

the organized consumers; those who pay, including insurance companies and corporation health groups; the hospital community; the manufacturers and developers; the health maintenance organizations (HMOs); the nursing profession; and groups concerned with health policy, management, and economics.

Congress charged the council to serve as a clearinghouse for information on health care technologies and health care technology assessment, to collect and analyze data, to identify needs in the assessment of specific technologies, to develop criteria and methods for assessment, and to stimulate assessments of technologies that were potentially important to the health care of the nation. At the outset, the council formed a Methods Panel, to look at techniques and to identify the methodologic problems in assessing new technologies, and an Information Panel, which is responsible for developing a usable and accessible clearinghouse of assessment data. Subsequently, an Evaluation Panel was organized to identify the technologies most urgently requiring exploration (Abrams and Hessel 1987).

As one of the 16 members of the council, and as co-chairman of the Methods Panel, I have been impressed by the breadth and quality of the work that has been produced by the council members, the panel participants, and the staff alike. The Methods Panel has dealt, in particular, with a number of secondary approaches to technology assessment, including meta-analysis and developments and improvements in consensus methods.

During my career in diagnostic radiology, both at Stanford and at Harvard universities, I have observed and participated in the introduction of a wide array of innovations, including visceral angiography, angiocardiography, coronary arteriography, lymphangiography, ultrasound, radionuclide methods, computed tomography, digital radiography, magnetic resonance imaging, among others. Many of these are based on complex technologies. In each case, before they could be properly utilized, their appropriate role had to be defined, and their efficacy and safety had to be addressed.

Furthermore, my background and experience have exposed me—through direct participation—to the array of impediments to first-rate primary data acquisition (Abrams and McNeil 1978a,b,c; Abrams 1979; Hillman et al. 1979; McNeil et al. 1981; Abrams et al. 1982; Adelstein 1982; Hessel et al. 1982; Alderson et al. 1983; Doubilet et al. 1985; Doubilet and Abrams 1984; Abrams and Hessel 1987). As a consequence, I have emphasized at meetings of the council and of the panel that

secondary data analysis based on inadequate primary information will necessarily be flawed. In the course of our exchanges, a request emerged that a more systematic depiction of the barriers to scientific primary assessments of diagnostic technologies be developed.

This brief tract is a direct response to that request. Initially, I conceived of it simply as a summary of the mechanical problems inherent in completing well-designed, prospective efficacy studies. Out of a series of discussions with my colleague, Dr. Harold Sox, we became persuaded that a somewhat broader approach would be of value.

The resulting monograph contains seven chapters. Chapter 1 presents the case for evaluating diagnostic tests and procedures. We adopt three perspectives: that of the patient, the physician, and the public. Chapter 2 introduces a central theme: the evaluation of a diagnostic test should provide the information required to decide if an individual patient needs the test. The chapter is a primer on the concepts of decision analysis. Chapter 3 represents a critique of the present state of assessment of diagnostic technologies. The focus is on the design of studies that avoid the problem of systematic bias in selecting patients to participate.

Chapter 4 explores the practical problems encountered in performing studies of diagnostic technology. These problems are at least as formidable as the issues of study design dealt with in Chapter 3. Chapter 5 identifies the costs associated with diagnosis technology assessment. The evaluation of diagnostic tests is expensive, although not as expensive as the failure to evaluate. Chapter 6 describes a national multicenter program for technology evaluation, designed to avoid some of the problems discussed in preceding chapters. Chapter 7 incorporates a review of the principles and problems of multi-institutional studies.

There is, of course, what some might consider an important gap: we have deliberately chosen to omit a consideration of the ethical issues in technology assessment. We are fully aware of the questions that have been raised about controlled clinical trials and of the need to buttress the assessment of health technologies in humans with the underlying principles germane to any clinical research: respect for persons, beneficence, and justice. These issues are sufficiently wide-ranging, however, to warrant a full exposition of the arguments on both sides, and they could not possibly be dealt with within the limited agenda that we had set for ourselves.

Dr. Sox has played the most important role in fulfilling our objectives and has contributed much of the creative thinking in a number of chapters. Susan Stern has done an outstanding job in gathering and organizing

much of the material on rationale, barriers to implementation, and multi-institutional studies. Dr. Douglas Owens has developed the material on cost and funding. All of us have reviewed, criticized, and modified each section in a joint effort to render them clear, pertinent, and useful.

The support of the Methods Panel in reviewing each section has been invaluable, and the final editing by Mrs. Jeffery Stoia has helped eliminate both duplication and lack of clarity where they existed.

Meant more as a primer than as an exhaustive treatment of the subject, it is our hope that the product will indicate both problems and their potential solutions, while at the same time highlighting the inherent complexity of the area and the need for adequate resource allocation.

The contents of this brief monograph reflect the experience, observations, and opinions of the authors and should not be construed as a statement by the Council on Health Care Technology as a whole, or by its Methods Panel. Although we have been helped by many comments and suggestions from council members, there has been no formal review or endorsement by the council, nor would we have considered that appropriate.

A responsive and well-developed system of technology assessment can provide a strong impetus to rapid application of essential technologies and prevent the wide diffusion of marginally useful methods. In both of these ways, it can increase quality and decrease the cost of health care. The requisite investment of funds, intellect, creativity, and supporting personnel is a small price to pay for the potential good that may be achieved.

Herbert L. Abrams

REFERENCES

Abrams, H.L. The "overutilization" of x-rays. New England Journal of Medicine 300:1213-1216, 1979.

Abrams, H.L., and Hessel, S. Health technology assessment: Problems and challenges. American Journal of Roentgenology 149:1127-1132, 1987.

Abrams, H.L., and McNeil, B.J. Computed tomography: Cost and efficacy implications. American Journal of Roentgenology 131:81-87, 1978a.

Abrams, H.L., and McNeil, B.J. Medical implications of computed tomography (CAT scanning). New England Journal of Medicine 298:253-261, 1978b.

Abrams, H.L., and McNeil, B.J. Medical implications of computed tomography (CAT scanning). II. New England Journal of Medicine 298:310-318, 1978c.

Abrams, H.L., Siegelman, S.S., Adams, D.F., et al. Computed tomography versus ultrasound of the adrenal gland: A prospective study. Radiology 143:121-128, 1982.

Adelstein, S.J. Pitfalls and biases in evaluating diagnostic technologies. In McNeil, B.J., and Cravalho, E.G., eds., Critical Issues in Medical Technology, pp. 67-79. Boston, Auburn House Publishing Company, 1982.

Alderson, P.O., Adams, D.F., McNeil, B.J., et al. Computed tomography, ultrasound, and scintigraphy of the liver in patients with colon or breast carcinoma: A prospective comparison. Radiology 149:225-230, 1983.

Califano, J. Quoted in David, M. U.S. health care faulted in senate. New York Times, January 13, 1987, p. 10.

Cooper, L.S., Chalmers, T.C., and McCally, M. Magnetic resonance imaging (MRI) (NMR). Poor quality of published evaluations of diagnostic precision. Abstract. Clinical Research 33:597A, 1985.

Doubilet, P.M., and Abrams, H.L. The cost of underutilization: Percutaneous transluminal angioplasty. New England Journal of Medicine 310:95-102, 1984.

Doubilet, P., McNeil, B.J., Van Houten, F.X., et al. Excretory urography in current practice: Evidence against overutilization. Radiology 154:607-611, 1985.

Eddy, D.M. Pitfalls and biases in evaluating screening technologies. In McNeil, B.J., and Cravalho, E.G., eds., Critical Issues in Medical Technology, pp. 53-65. Boston, Auburn House Publishing Company, 1982.

Gibson, R.M., and Mueller, M.S. National health expenditure, fiscal year 1976. Social Security Bulletin 40:3-22, 1977.

Greer, A.L. Medical technology: Assessment, adoption, and utilization. Journal of Medical Systems 5:129-145, 1981.

Hessel, S.J., Siegelman, S.S., McNeil, B., et al. A prospective evaluation of computed tomography and ultrasound of the pancreas. Radiology 143:129-133, 1982.

Hillman, B., Abrams, H.L., Hessel, S.J., et al. Simplifying radiological examinations. The urogram as a model. Lancet (May 19, 1979) 1:1069-1071.

Institute of Medicine. A Consortium for Assessing Medical Technology. Washington, D.C., National Academy Press, 1983.

McNeil, B.J. Pitfalls in and requirements for evaluations in diagnostic technologies. In Wagner J., ed., Proceedings of the Conference on

Medical Technologies. DHEW Publication (PHS) 79-3254, 1979:33-39.

McNeil, B.J., Sanders, R., Alderson, P.O., et al. A prospective study of computed tomography, ultrasound, and gallium imaging in patients with fever. Radiology 139:647-653, 1981.

Office of Technology Assessment. Development of Medical Technology: Opportunities for Assessment. Washington, D.C., U.S. Government Printing Office (OTA-H-34), 1976.

Perry, S. The brief life of the National Center for Health Care Technology. New England Journal of Medicine 307:1095-1100, 1982.

Perry, S. Technology assessment: Continuing uncertainty. New England Journal of Medicine 314:240-243, 1986.

Perry, S., and Eliastam, M. The National Center for Health Care Technology. Journal of the American Medical Association 245:2510-2511, 1981.

Petitti, D.B. Competing technologies. Implications for the costs and complexity of medical care. New England Journal of Medicine 315:1480-1483, 1986.

Waldman, S. The effect of changing technology on hospital costs. Research and statistics note. DHEW Publication 72-11701, 1972:1-6.

Yeaton, W.H., and Wortman, P.M. Evaluation issues in medical research synthesis. In Yeaton, W.H., and Wortman, P.M., eds., New Directions for Program Evaluation, vol. 24, Issues in Data Synthesis, pp. 43-56. San Francisco, Jossey-Bass, December 1984.

1

Rationale for Assessment of Diagnostic Technology

Over the past 15 years, there has been rapid growth in the use of innovative diagnostic technologies, such as digital radiography, ultrasound, computed tomography (CT), and magnetic resonance imaging (MRI) (Hillman and Schwartz 1985). The use of diagnostic tests in general is increasing even more rapidly than the cost of medical care (Schroeder 1981). Furthermore, there is evidence that diagnostic technologies consume a significant portion of our health care resources. For example, Scitovsky has estimated that of all expenditures for outpatient care in 1975, 29 percent were for laboratory tests and X rays (Scitovsky 1979).

In medicine, the primary goal of new technology is to improve the quality of care. Nevertheless, the recent past has been marked by early diffusion of new technologies without adequate measurement of their effects on the quality of care (Abrams and McNeil 1978a,b; Guyatt et al. 1986). A number of authors have pointed to the lack of prospective, controlled studies (Harris 1981, Sheps and Schechter 1984, Schwartz 1986, Cooper et al. 1988, Kent and Larson 1988). Although individual diagnostic tests have been studied at length for accuracy, direct comparisons of a new test with an older test have been all too infrequent (McNeil et al. 1981, Abrams et al. 1982, Hessel et al. 1982, Alderson et al. 1983, Inouye and Sox 1986). What are the consequences of the medical profession's failure to evaluate diagnostic tests in a timely and rigorous manner? Why is such an evaluation important?

These questions may be approached from many perspectives, but three are especially relevant: that of the patient, that of society, and that of the physician. We will consider each of these perspectives in terms of safety, efficacy, and economic efficiency—all critical standards for a diagnostic test, but standards whose relative importance depends upon the perspective under consideration.

THE PATIENT'S PERSPECTIVE

The impact of a diagnostic test is felt most directly by the patient. The patient's concern is improving his or her health status. From the patient's perspective, the most important characteristics of a diagnostic test are its safety and efficacy. The economic aspect of testing may also assume importance, but it is often a secondary consideration.

Safety

A *safe* test does not cause an unacceptable degree of direct harm to the patient. It is important to the patient that the diagnosis be made with a minimum of inconvenience and discomfort. If a test is hazardous, he or she will be directly at risk. Obviously, some "safe" tests will have side effects, and the acceptability of such effects will be determined by weighing their severity against the need for the information that the test provides.

Efficacy

After safety, the starting point for any assessment of health care technology must be efficacy or effectiveness. *Efficacy* refers to the potential benefit to patients in a defined population when a test is applied to a specified problem under ideal conditions of use (Willems et al. 1977). *Effectiveness* is measured under the usual conditions of medical practice. This distinction is important when designing a technology assessment and interpreting its result; in the present discussion, however, we will refer only to efficacy. The efficacy of a diagnostic test should be measured in terms of the test's safety, its technical quality, its accuracy, its therapeutic impact, and its impact on the health of the patient (Fineberg et al. 1977).

The first step in an efficacy assessment is to determine the test's *technical capability*. Does the test meet the standards attributed to it? For

a diagnostic imaging method, the first stage of assessment might involve using cadaver specimens to see how well the technique is able to demonstrate the anatomy of various regions in the body. The second stage of an efficacy assessment is to define the test's *diagnostic accuracy*. In this regard, three commonly used expressions are true-positive rate, true-negative rate, and accuracy. The *true-positive rate*, or *sensitivity*, is a measure of the test's ability to detect disease correctly when it is present. The *true-negative rate*, or *specificity*, measures the test's ability to exclude disease in those patients who do not have it. *Accuracy* is the proportion of test results that are correct (*true-positive* results plus *true-negative* results divided by the total number of test results) when the test is used in a specified population. Thus, it is a reflection of both the sensitivity and the specificity of the test.

To call a test result a true positive or a true negative, one must determine the true state of the patient. This is usually accomplished by doing another test, called the "gold standard," which is considered sufficiently reliable to reveal the true state of the patient, and either confirm or refute the study test result. For example, coronary angiography has been used to verify the presence of coronary artery disease in patients participating in an efficacy study of the stress electrocardiogram. For an ideal test, there should be little disagreement between its result and the result of the "gold standard": the test should have both high sensitivity and high specificity.

Although quantitative measures of test performance are important, a study of efficacy should not focus solely on its technical aspects (that is, on the machine). Rather, an assessment should include data on *diagnostic impact* and on *therapeutic impact*, including outcomes that are relevant to the patient. These are the third and fourth levels of an efficacy assessment. The following questions might be asked in the third stage of an assessment. Does the result of the technique change the diagnosis? Does the technique add clinically significant information? At the fourth level of an efficacy assessment, the question would be: Is the diagnostic impact one that changes the management of the patient?

These two levels are often, although not always, interrelated. For example, MRI has an unparalleled capacity to detect demyelinating lesions in the brains of patients with multiple sclerosis. MRI is therefore an extremely important tool for assessing prognosis and also for evaluating new therapeutic approaches as they are developed. Nevertheless, there is currently no effective therapy for this disease. The results from a study that focuses solely on the efficacy of MRI in assessing the *progression* of multiple sclerosis would have limited clinical use, because knowledge of the *extent* of the lesions does not currently play a part in deciding which

therapy the patient should receive (Feeny et al. 1986). *Many studies of diagnostic tests fail to consider the impact of a test result on patient management and clinical outcome*. In this sense, the patient's perspective is not always fully reflected in contemporary technology assessment.

Economic Issues

Although the primary concern of the patient may not be financial, the cost of testing and of receiving high-quality medical care is clearly important to them. Unnecessary tests, inappropriate treatment, and disability may all prove very expensive. Although most patients are somewhat insulated from the costs of their care by third-party payment mechanisms, they are still usually responsible for paying a *deductible* portion of their care, that is, a fixed amount that must be paid before insurance coverage begins. This may be quite large in some types of insurance. In addition, some individuals have limited insurance coverage, and 35 million people in the United States have no health insurance (Annas 1986). The RAND health insurance experiment demonstrated that paying patients are sensitive to the costs of health care and that individuals responsible for almost all of their health care costs incurred expenses about 50 percent less than individuals receiving free care (Newhouse et al. 1981).

The Benefits of High-Quality Studies

There are many potential benefits for the patient of a test that has been shown to be both safe and effective. By improving the quality of diagnostic information, the test result may lead to timely, correct therapy. Earlier diagnosis may result in reduced or postponed morbidity and mortality. Tests may resolve uncertainty (Abrams and McNeil 1978b), reassure both the patient (Sox et al. 1981) and the physician, and increase the patient's confidence in the physician (Marton et al. 1982). Comprehensive assessment provides the patient with information about the accuracy of the test, its risks, and the monetary costs associated with its use. With this information, the patient is better able to make an informed decision about accepting a recommendation to undergo a diagnostic procedure.

The Risks of Poor or Absent Assessment Data

Tests are often used when there is no proof of efficacy. Patients may suffer in several ways when their diagnosis depends on a test lacking established validity and known false-positive and false-negative rates.

Tests can lead to unnecessary confirmatory tests or to incorrect therapy. Consider the outcome of using a test that has a high false-positive rate (low specificity): a large proportion of "normal" individuals may mistakenly receive treatment for the condition in question or may have to undergo further tests to define their true state. They may suffer direct harm from the unnecessary diagnostic interventions or treatment and may experience considerable anxiety. In the initial assessments of radionuclide ventriculography as a test for coronary artery disease, bias in the selection of study patients led to an overly optimistic conclusion about the test's performance. The test was later found to be far less specific (49 percent) than early studies had shown (93 percent) (Rozanski et al. 1983). This high false-positive rate may have caused many patients without coronary artery disease to be referred for an unnecessary invasive angiographic procedure.

Tests can also cause harm through false reassurance. A methodologically flawed assessment may lead to the conclusion that a test has a much lower false-negative rate than is really true. Many patients who have a disease will have a negative result and be told that the disease is absent. They will then have a false sense of confidence, treatment may not be started, and the disease may advance beyond the point where it can be cured. Patients with suspected colorectal cancer could have suffered if their physicians had used a normal carcinoembryonic antigen (CEA) level to conclude that there was no cancer. This test had been used widely in the initial diagnosis of colorectal cancer, but later studies showed that CEA levels are often normal in the early stages of the disease. The test may detect as few as 30 percent of patients with early colorectal cancer (Fletcher 1986).

Even if a test provides accurate information, the test result may have no impact on the therapeutic plan or on patient outcome. For example, randomized trials of emergency endoscopy for patients with upper gastrointestinal bleeding have shown that endoscopy provides additional diagnostic information, but that this information does not alter surgical rates, length of hospital stay, or patient mortality (Peterson et al. 1981, Dronfield et al. 1982). On balance, the examination cannot be considered beneficial.

Summary: The Patient's Perspective

For the patient, the most important attributes of a diagnostic test are its safety and clinical efficacy. A safe, efficacious test should reveal the true

state of the patient, with a minimum of inconvenience and adverse effects. The benefits of a true-positive finding include timely diagnosis and treatment. The benefits of a true-negative include reassurance and protection from unnecessary treatment and further procedures. A correct test result should increase patients' confidence in their care. Patients may also be very concerned about the costs of testing, particularly if they do not have health insurance. The role of good technology assessment is to identify tests that are reliable, that provide useful incremental information, and that may have a positive effect on patient outcome, as well as to single out those tests that fail to measure up to these standards.

SOCIETY'S PERSPECTIVE

Safety and Efficacy

Society's concern about the safety and efficacy of technology has two components. First, our government's role as "guardian of the public safety" (Foote 1987)—traditionally quite limited in the past—has been expanded over recent years. The power of the Food and Drug Administration to regulate the safety and efficacy of drugs and, more recently, the safety and efficacy of medical devices is evidence of this expanded obligation (Foote 1986). Thus, society's interest in safety and efficacy stems partly from a perceived ethical duty.

A second component of this interest is economic. Society must be concerned about test efficacy because tests that are not efficacious are not likely to be economically efficient. Tests with high false-positive rates (low specificity) expose patients to the risk of a needless workup and increase the direct costs for their care. Missed diagnoses, which result from the use of tests with a high false-negative rate, may increase total health care costs if treatment that is begun at a later stage of the illness consumes more resources. Society's concern about the safety and efficacy of technology is matched by an equally strong interest in economic efficiency.

Economic Issues

Many analysts agree that new medical technology has been an important factor in the rise of health care costs, although they disagree about the magnitude of its contribution (Waldman 1972, Feeny et al. 1986). The Office of Technology Assessment (OTA) has estimated that the technol

ogy components of care are responsible for nearly 30 percent of the rise in Medicare costs for the period between 1977 and 1982 (OTA 1984). Because resources are limited, and because government at all levels has assumed a larger role in paying for health care, society has an interest in assuring that the available resources are utilized efficiently.

A cost-effectiveness analysis is the method used to measure the efficiency with which dollars are translated into health outcomes. The OTA defines cost-effectiveness analysis as a comparison of the positive and negative consequences of using alternative technologies (OTA 1980a,b). The key to cost-effectiveness analysis is that it is *comparative*. It compares the cost and outcome of using one test for a diagnostic problem with the cost and outcome of using another test. If there is no existing diagnostic technology, the new technology can be compared to doing nothing. The results of a cost-effectiveness analysis are usually expressed as the cost per unit of outcome (average cost-effectiveness) or the change in costs per change in unit of outcome (marginal cost-effectiveness). Compared with existing technology, an efficient new technology would achieve the same outcome at a lower cost, a better outcome at the same cost, or a better outcome at a lower cost.

From the societal perspective, cost-effectiveness analyses help policymakers decide which technologies should be encouraged by reimbursement policy. In principle, given a fixed budget, the use of this approach to allocate resources to programs results in the greatest impact on clinical outcome. A complete cost-effectiveness analysis will measure monetary and other costs, the effectiveness of the technology in achieving its intended objectives, and the positive and negative effects from both intended and unintended consequences (Arnstein 1977). Without complete assessments and physician education, certain technologies will be overutilized, while others may be underutilized (Abrams 1979, Doubilet and Abrams 1984). The net result is wasted resources and lost opportunities.

Summary: Society's Perspective

Society necessarily has a deep interest in the costs of diagnostic technology, although safety and efficacy are also important. As government has assumed a much larger role in financing health care, the importance of efficient use of diagnostic procedures has also grown. Policymakers need accurate information about which technologies consume the least resources for a given outcome so that they can allocate limited health care

resources. Nevertheless, the first considerations in deciding whether a diagnostic technology is a good societal investment must always be safety and efficacy.

THE PHYSICIAN'S PERSPECTIVE

The importance of diagnostic tests and technology assessment to the physician must be examined in the context of the physician's dual role as the agent of both the patient and society.

Safety and Efficacy

The physician's role with respect to a patient is traditionally that of a fiduciary: the principal (the patient) entrusts the fiduciary (the physician) with the power to act on his behalf. For example, a patient usually undergoes a diagnostic procedure at the request of a physician, and therefore both have the same interest in knowing that a test is both safe and efficacious. Tests are done when the patient's history is consistent with a particular illness but the true disease state remains uncertain.

The purpose of a diagnostic test in this clinical setting is twofold. First, it should provide reliable information about the patient's condition. Second, the result of the test should influence the physician's plan for managing the patient. A test can serve these functions only if the physician knows how to interpret its result.

Adequate assessment of diagnostic technology is important to clinicians because it provides the data needed to interpret test results. The result of a test whose sensitivity and specificity have been measured reliably can be used as the basis for sound clinical decisions. For example, in cases where there is an effective treatment for a disease, a positive test result may raise the probability of disease sufficiently to convince the clinician to start treatment.

Tests that have not been adequately assessed are not as useful to the physician because the meaning of their results is ambiguous. For example, *if a test result is negative, should the physician trust that result and assume that the disease is not present?* When the sensitivity of a test is unknown, the physician has no way of knowing the proportion of patients who have the disease despite a negative result. Frequently, the false-negative rate and false-positive rate of a test are stated but are inaccurate. The physician may believe that a second test is needed when it is not, or may think that it is unnecessary when in fact it should be done.

Without *comparative* assessment data, a physician cannot be sure whether a new test should replace an older test, should be used in conjunction with the older test, or should not be adopted at all. The routine use of intrapartum electronic fetal monitoring (EFM) in place of periodic auscultation for all deliveries provides an example. Because EFM is a very accurate diagnostic tool, it met with early acceptance. Its use, however, may lead to a higher rate of operative deliveries. For low-risk pregnancies, periodic auscultation provides adequate information with fewer adverse side effects (Thacker 1987). Similarly, a recent comparative trial has demonstrated that even in high-risk pregnancies, auscultation yields equivalent results with EFM in terms of both the fetus and the mother (Luthy et al. 1987). Comparative assessments, such as that conducted by Luthy et al., along with physician education, are needed to prevent irrational and inefficient use of diagnostic technology.

Ideally, the interests of the physician and the patient can be equated, because, in principle, the physician is acting solely for the benefit of the individual patient. In addition to patient benefit, however, physicians may also be concerned about the financial and legal repercussions of diagnostic or therapeutic errors. A good deal of "defensive" medicine is practiced—with consequent overutilization—because of the fear that the omission of a diagnostic test may be construed as malpractice. Errors might be caused by using a test whose efficacy is uncertain. A test that yields many false-negative results may lead to missed diagnoses. One that yields many false-positive results may lead to excessively complex workups with untoward effects.

Thus, a new technology that has made its way into clinical practice without adequate assessment may adversely affect the health care provider as well as the patient. Conversely, when a thorough assessment indicates that a test is both highly sensitive and specific, both the physician and the patient benefit.

Economic Issues

Physicians as well as hospitals must be concerned about economic efficiency. In an increasingly competitive environment, economic success will depend on efficiency as well as on quality of care. Adopting a new technology may attract more patients and increase a physician's competitive advantage. Some new technologies may reduce the costs of providing health care, but many new technologies are more expensive than those they are designed to replace. The cost of diagnostic technology

is clearly of concern to physicians practicing within an organization where income is based on a fixed payment per patient rather than fee-for-service, because they may be at direct financial risk for expenditures over this fixed amount. The ultimate net effect of a new technology cannot be assumed to be beneficial; once clinical efficacy has been established, the cost-effectiveness of the technology should be evaluated. These comparative assessments should then influence hospitals' decisions to purchase new technology, as well as physicians' decisions to use it.

Economic efficiency is important to the physician at yet another level. After the primary responsibility to the patient, the physician also has a societal obligation to help to contain the costs of health care. Physicians play the principal role in controlling the services patients receive, and they have a large influence on aggregate health care expenditures. Because diagnostic technologies are an important component of these costs, clinicians can exert their influence by using diagnostic technologies efficiently.

Without high-quality technology assessments, practice habits may change inappropriately. For example, a new, expensive diagnostic method may replace an older, less expensive—but equally efficacious—technology. The case of intrapartum EFM in place of periodic auscultation of the fetal heart is again illustrative. Initial studies of EFM documented its high level of technical and diagnostic accuracy and suggested that its use was associated with a reduction in perinatal morbidity and mortality. Widespread diffusion of this costly technology followed. Nevertheless, two recent critical examinations of the literature on EFM concluded that there is little rigorous evidence that *routine* use of the method leads to a beneficial impact on patient outcome. The conclusion is that EFM should have been thoroughly evaluated by comparative studies at an early stage of its diffusion into routine practice *before* it replaced the less costly alternative of traditional auscultation (Shy et al. 1987, Thacker 1987).

Summary: The Physician's Perspective

This perspective reflects the concerns of both the patient and the society, because the physician serves as a crucial link between the two. For example, physicians play a large role in controlling the flow of society's health care resources. Our society expects physicians to make responsible decisions on how often and under what circumstances expensive diagnostic technologies will be used. To this end, they must have reliable, accurate, and comparative information on test performance. In

addition, accurate tests help to avoid errors that could lead to legal and financial difficulties. Economic efficiency is an important personal concern for physicians as well, and those in private practice or in large hospitals that use diagnostic technology inefficiently may be unable to compete effectively in the health care market. Most important, the physician needs information about diagnostic technology in order to provide high-quality health care to every patient. Technology assessment is one mechanism for obtaining this information.

SUMMARY

High-quality, timely assessment is the prerequisite for the safe, efficacious, and economically efficient use of diagnostic technology. The data to be obtained will depend on the perspective adopted for the purpose of the assessment. Nevertheless, the three parties most directly affected by the use of diagnostic technology share many of the same concerns.

Although there is a well-defined methodology for assessing diagnostic technology, few studies have satisfied all of the methodological criteria. Even well-designed studies have encountered problems in the course of collecting primary data. The following examples should serve to illustrate this point.

- In a cooperative study of computed tomography (CT) and radionuclide (RN) studies on the brain, data were collected by five hospitals over a five-year period. Of 3,000 patients who entered into the study, only 136 patients had technically adequate and available CT and RN studies that could then be used in the final data analysis (McNeil 1979).
- One of the few prospective, comparative studies of diagnostic imaging techniques to date ran into similar difficulties. The authors of the cooperative study on computed tomography, ultrasound, and gallium imaging in patients with fever stated: "We spent 17 months collecting data from two major teaching institutions [the Peter Bent Brigham and Johns Hopkins hospitals]. Full-time research assistants at each institution tried to obtain all cases in random order. Yet even this concerted effort produced only 156 cases, and then only 50 percent of them included objective proof of disease" (McNeil et al. 1981).
- Another class of problems is highlighted in Philbrick's analysis of studies of exercise testing in the diagnosis of coronary artery disease. He found a wide range of both sensitivity (35 percent to 88 percent) and specificity (41 percent to 100 percent). The results of this review of 33

studies of patients undergoing both a stress ECG and coronary angiogram "suggest that a principle source of variation may be methodological defects in research design" (Philbrick et al. 1980). These defects included bias in patient selection, referral for the coronary angiogram, and test and gold-standard interpretation, as well as inadequate reporting.

There are many obstacles to good studies of diagnostic technologies. The examples above touch on only a few. The problems of conducting clinical trials of *therapeutic* technologies have been well documented, and considerable research has been done with the aim of improving therapeutic trials (for example, see Meinert 1986). Until recently, however, there has been less interest in trials of diagnostic technology. Although the science of assessment of diagnostic technology has made considerable progress over the last decade, the art of conducting this type of study remains underdeveloped. This monograph focuses, therefore, on the problems that arise in the attempt to collect high-quality *primary data* for diagnostic technology assessment.

REFERENCES

Abrams, H.L. The "overutilization" of x-rays. New England Journal of Medicine 300:1213-1216, 1979.

Abrams, H.L., and McNeil, B.J. Medical implications of computed tomography ("CAT" scanning). New England Journal of Medicine 298:255-261, 1978a.

Abrams, H.L., and McNeil, B.J. Medical implications of computed tomography ("CAT" scanning). New England Journal of Medicine 298:310-318, 1978b.

Abrams, H.L., Siegelman, S.S., Adams, D.F., et al. Computed tomography versus ultrasound of the adrenal gland: A prospective study. Radiology 143:121-128, 1982.

Alderson, P.O., Adams, D.F., McNeil, B.J., et al. Computed tomography, ultrasound, and scintigraphy of the liver in patients with colon or breast carcinoma: A prospective comparison. Radiology 149:225-230, 1983.

Annas, G.J. Your money or your life: "Dumping" uninsured patients from hospital emergency wards. American Journal of Public Health 76:74-77, 1986.

Arnstein, S.R. Technology assessment: Opportunities and obstacles. IEEE Transactions on Systems, Man and Cybernetics Health SM-7:571-582, 1977.

Cooper, L.S., Chalmers, T.C., McCally, M., et al. The poor quality of early evaluations of magnetic resonance imaging. Journal of the American Medical Association 259:3277-3280, 1988.

Doubilet, P., and Abrams, H.L. The cost of underutilization: Percutaneous transluminal angioplasty for peripheral vascular disease. New England Journal of Medicine 310:95-102, 1984.

Dronfield, M.W., Langman, M.J.S., Atkinson, M., et al. Outcome of endoscopy and barium radiography for acute upper gastrointestinal bleeding: Controlled trial of 1,037 patients. British Medical Journal 284:545-548, 1982.

Feeny, D. New health technologies: Their effect on health and the cost of health care. In Feeny, D., Guyatt, G., and Tugwell, P., eds., Health Care Technology: Effectiveness, Efficacy and Public Policy. Montreal, The Institute for Research on Public Policy, 1986.

Fineberg, H.V., Bauman, R., and Sosman, M. Computerized cranial tomography: Effect on diagnostic and therapeutic plans. Journal of the American Medical Association 238:224-230, 1977.

Fletcher, R.H. Carcinoembryonic antigen. Annals of Internal Medicine 104:66-73, 1986.

Foote, S.B. From crutches to CT scans: Business-government relations and medical practice innovation. In Post, J.E., ed., Research in Corporate Social Policy and Performance. Greenwich, Conn., JAI Press, 1986.

Foote, S.B. Assessing medical technology: Past, present and future. The Millbank Quarterly 65:59-80, 1987.

Guyatt, G., Drummond, M., Feeny, D., et al. Guidelines for the clinical and economic evaluation of health care technologies. Social Science and Medicine 22:393-408, 1986.

Harris, J.M. The hazards of bedside Bayes. Journal of the American Medical Association 246:2602-2605, 1981.

Hessel, S.J., Siegelman, S.S., McNeil, B.J., et al. A prospective evaluation of computed tomography and ultrasound of the pancreas. Radiology 143:129-133, 1982.

Hillman, A.L., and Schwartz, J.S. The adoption and diffusion of CT and MRI in the United States: A comparative analysis. Medical Care 23:1283-1294, 1985.

Inouye, S.K., and Sox, H.C., Jr. Standard and computed tomography in the evaluation of neoplasms of the chest. Annals of Internal Medicine 105:906-924, 1986.

Kent, D.L., and Larson, E.B. Diagnostic technology assessment: Problems and prospects. Annals of Internal Medicine 108:759-761, 1988.

Luthy, D.A., Shy, K.K., van Belle, G., et al. A randomized trial of electronic fetal monitoring in preterm labor. Obstetrics and Gynecology 69:687-695, 1987.

Marton, K.I., Sox, H.C., Jr., Alexander, J., and Duisenberg, C.E. Attitudes of patients toward diagnostic tests: The case of the upper gastrointestinal series roentgenogram. Medical Decision Making 2:439-448, 1982.

McNeil, B.J. Pitfalls in and requirements for evaluations of diagnostic technologies. In Wagner, J., ed., Proceedings of a Conference on Medical Technologies. DHEW Pub. No (PHS) 79-3254, pp. 33-39. Washington, D.C., U.S. Government Printing Office, 1979.

McNeil, B.J., Sanders, R., Alderson, P.O., et al. A prospective study of computed tomography, ultrasound, and gallium imaging in patients with fever. Radiology 139:647-653, 1981.

Meinert, C.L. Clinical Trials: Design, Conduct and Analysis. New York, Oxford University Press, 1986.

Newhouse, J.P., Manning, W.G., Morris, C.N., et al. Some interim results from a controlled trial of cost sharing in health insurance. New England Journal of Medicine 305:1501-1507, 1981.

Office of Technology Assessment, U.S. Congress. The Implications of Cost-Effectiveness Analysis of Medical Technology. Stock No. 051-003-00765-7. Washington, D.C., U.S. Government Printing Office, 1980a.

Office of Technology Assessment, U.S. Congress. The Implications of Cost-Effectiveness Analysis of Medical Technology. Background paper #1: Methodological issues and literature review. Washington, D.C., U.S. Government Printing Office, 1980b.

Office of Technology Assessment, U.S. Congress. Medical Technology and the Costs of the Medicare Program. OTA-H-227. Washington, D.C., U.S. Government Printing Office, 1984.

Peterson, W.L., Barnett, B.S., Smith, H.J., et al. Routine early endoscopy in upper gastrointestinal tract bleeding. A randomized trial. New England Journal of Medicine 304:925-929, 1981.

Philbrick, J.T., Horwitz, R.I., and Feinstein, A.R. Methodologic problems of exercise testing for coronary artery disease: Groups, analysis and bias. American Journal of Cardiology 46:807-812, 1980.

Rozanski, A., Diamond, G.A., Berman, D., et al. The declining specificity of exercise radionuclide ventriculography. New England Journal of Medicine 309:518-522, 1983.

Schroeder, S.A. Medical technology and academic medicine: The doctorproducers' dilemma. Journal of Medical Education 56:634-639, 1981.

Schwartz, J.S. Evaluating diagnostic tests: What is done—What needs to be done. Journal of General Internal Medicine 1:266-267, 1986.

Scitovsky, A.A. Changes in the use of ancillary services for "common" illness. In Altman, S.H., and Blendon, R., eds., Medical Technology: The Culprit Behind Health Care Costs? Proceedings of the 1977 Sun Valley Forum on National Health, pp. 39-56. DHEW Pub. No.

(PHS) 79-3216. Washington, D.C., U.S. Government Printing Office, 1979.

Sheps, S.B., and Schechter, M.T. The assessment of diagnostic tests: A survey of current medical research. Journal of the American Medical Association 252:2418-2422, 1984.

Shy, K.K., Larson, E.B., and Luthy, D.A. Evaluating a new technology: The effectiveness of electronic fetal heart rate monitoring. Annual Review of Public Health 8:165-190, 1987.

Sox, H.C., Jr., Margulies, I., and Sox, C.H. Psychologically mediated effects of diagnostic tests. Annals of Internal Medicine 95:680-685, 1981.

Thacker, S.B. The efficacy of intrapartum electronic fetal monitoring. American Journal of Obstetrics and Gynecology 156:24-30, 1987.

Waldman, S. The effect of changing technology in hospital costs. U.S. Department of Health, Education and Welfare, Social Security Administration, Office of Research and Statistics. Research and Statistics Note No. 4-1972 (February 28).

Willems, J.S., Banta, H.D., Lukas, T.A., and Taylor, C.A. The computed tomography scanner. In Altman, S.H., and Blendon, R., eds., Medical Technology: The Culprit Behind Health care Costs? Proceedings of the 1977 Sun Valley Forum on National Health, p. 131. DHEW Pub. No. (PHS) 79-3216. U.S. Government Printing Office, Washington, D.C., 1977.

2

The Use of Diagnostic Tests: A Probabilistic Approach

Diagnostic tests and the information that they convey are too often taken for granted by both physicians and patients. The most important error is to assume that the test result is a true representation of what is really going on. Most diagnostic information is imperfect; although it changes the physician's perception of the patient, he or she remains uncertain about the patient's true state.

As an example, consider a hypothetical test. With this test, 10 percent of patients who have the disease will have a negative result (a *false-negative* result), and 10 percent of the patients who do not have the disease will have an abnormal result (a *false-positive* result). Thus, when the result is abnormal, the clinician cannot be certain that the patient has the disease: abnormal results occur in patients who have the disease *and* in patients who do not. There is similar uncertainty if the test result is negative. As long as tests are imperfect, this uncertainty is intrinsic to the practice of medicine.

The physician who acknowledges the imperfections of a diagnostic test will ask, "In view of this test result, how uncertain should I be about this patient?" Fortunately, there is a method for answering this question: the theory of probability. This chapter is a primer for applying probability theory to the interpretation of test results and deciding when to do a test rather than treat or do nothing.[1] It is divided into five parts: (1) first

[1] This chapter is adapted from an article written by one of the authors (Sox 1986). The material is covered in greater depth in standard textbooks (Sox et al. 1988, Weinstein 1980).

principles; (2) interpreting test results: the posttest probability; (3) estimating the pretest probability; (4) measuring test performance; (5) expected-value decisionmaking; and (6) the choice among testing, starting treatment, or doing nothing.

FIRST PRINCIPLES

The way in which one decides to do a diagnostic test is based on two principles.

PRINCIPLE I: Probability is a Useful Representation of Diagnostic Uncertainty.

Uncertainty is unavoidable. How can we best respond to it? A starting point is to adopt a common language. Some express their uncertainty as the *probability* that the patient has a specified disease. By using probability rather than ambiguous terms such as "probably" or "possibly," the clinician expresses uncertainty quantitatively. More important, probability theory allows one to take new information and use Bayes' theorem to calculate its effect on the probability of disease. These advantages are compelling, and our approach to test evaluation is based on providing the information required to use probability theory to interpret and select diagnostic tests.

> *Example:* In a patient with chest pain, past history is very useful when trying to decide whether he or she has coronary artery disease. Patients whose pain is typical of angina pectoris and is also closely linked to overexertion are said to have "typical angina pectoris." Over 90 percent of men with this history have coronary artery disease. When anginal pain is less predictably caused by exertion, the patient is said to have "atypical angina." About two-thirds of men with this history have coronary artery disease.
>
> Physicians who are uncertain about the meaning of a patient's chest pain often ask the patient to undergo an exercise test. The probability of coronary artery disease after a positive exercise test may be calculated with Bayes' theorem. If the history is typical angina, the probability after a positive test is nearly 1.0. If the history is atypical angina, the probability after a positive test is about 0.90.
>
> *Comment:* Estimating the probability of coronary artery disease helps to identify the situations in which the probability of disease will be altered dramatically by an abnormal test.

PRINCIPLE II: A Diagnostic Test Should Be Obtained Only When Its Outcome Could Alter the Management of the Patient.

A test should be ordered only when forethought shows that it could lead to a change in patient management. How does one decide if a test will alter the management of a patient? There are several considerations.

The effect of a test result on the probability of disease. If the probability of disease after the test will be very similar to the probability of disease before the test, the test is unlikely to affect management. The posttest probability of disease can be calculated by using Bayes' theorem, as discussed later in this section.

> *Example:* The probability of coronary artery disease in a person with typical angina pectoris is 0.90. If an exercise test result is abnormal, the probability of disease is 0.98. If the result is normal, the probability of disease is 0.76. Many physicians would conclude that the effect of the results is too small to make the test worthwhile for diagnostic purposes.

The threshold model of decisionmaking. This approach is based on the concept that a test is judged by its effect on the probability of disease (Pauker and Kassirer 1975, 1980). The model postulates a *treatment threshold probability*, below which treatment is withheld and above which it is offered. In this situation, a test is only useful if, after it is performed, the probability of disease has changed so much that it has crossed from one side of the treatment threshold probability to the other. If the posttest probability were on the same side of the threshold as the pretest probability, the decision of whether or not to treat would be unaffected by the test results, and the test should not be ordered. One must estimate the benefits and the harmful effects of treatment in order to set the treatment threshold probability.

> *Example:* Some patients with suspected pulmonary embolism are allergic to the contrast agents that are used to perform a pulmonary arteriogram, the definitive test for a pulmonary embolism. Many physicians say that if faced with this situation, they would start anticoagulation if they thought that the patient had as little as a 5 to 10 percent chance of having a pulmonary embolism. Thus, their treatment threshold probability is 0.05 to 0.10.

Effect of test results on clinical outcomes. Even if a test result leads to a change in management, if the patient will not benefit, the test should not

have been done. Thus, one is concerned not only with the test itself but also the efficacy of the actions that are taken when its result is abnormal.

> *Example:* Investigators have calculated the average improvement in life expectancy that results from the management changes following coronary arteriography in patients with stable angina pectoris. The analysis shows that middle-aged men will gain, on average, approximately one year from undergoing coronary arteriography and coronary bypass surgery if severe disease is present (Stason and Fineberg 1982). This test does have an effect on clinical outcomes.

Marginal cost-effectiveness of the test. This measure of test performance is a way to characterize the efficiency with which additional resources (dollars) are translated into outcomes (longevity). It takes into account the increased costs from doing a test and the incremental benefit to the patient. A test result may lead to a good outcome, such as improved longevity, but the increase in cost for each unit of increase in longevity may be so high that there is a consensus that the test should not be done.

INTERPRETING TEST RESULTS: THE POSTTEST PROBABILITY

The interpretation of a test result is an important part of technology assessment. A test with many false-negative and false-positive results will be interpreted with far more caution than a test with few such misleading results. Therefore, measuring the performance characteristics of a test is important, because the clinician must know them in order to interpret the result.

Important Definitions[2]

The probability of disease after learning the results of a test is called the posttest probability of disease. It is the answer to the question, "What does this test result mean?" One calculates the posttest probability with Bayes' theorem, which is derived from the first principles of probability and requires both the pretest probability of disease and two measures of the accuracy of the test. One measure is called the *sensitivity* of the test

[2] See also the Glossary of Terms at the end of this chapter.

(true-positive rate, or TPR). It represents the likelihood of a positive test in a diseased person, as is shown in the following equation:

$$\text{Sensitivity} = \frac{\text{number of diseased patients with positive test}}{\text{number of diseased patients}}.$$

Example: There have been many studies of the exercise electrocardiogram. In these studies, a patient with chest pain undergoes both the exercise electrocardiogram and a definitive test for coronary artery disease, the coronary arteriogram. About 70 percent of patients who had a positive arteriogram also had a positive exercise electrocardiogram (as defined by the presence of at least 1 mm of horizontal or downsloping ST segment depression). Thus, according to this result, the sensitivity of an exercise electrocardiogram for coronary artery disease is 0.70.

The second measure of test accuracy is its *false-positive rate*, the likelihood of a positive result in a patient without disease. *Specificity*, the true-negative rate (TNR), is 1 minus the false-positive rate.

$$\text{False-positive rate} = \frac{\text{number of nondiseased patients with positive test}}{\text{number of nondiseased patients}}.$$

Example: The studies of the exercise electrocardiogram have shown that about 15 percent of patients who did not have coronary artery disease nonetheless did have an abnormal exercise electrocardiogram. Thus, the false-positive rate of the exercise electrocardiogram for coronary artery disease is 0.15.

Likelihood ratio. The likelihood ratio is a measure of how much the result alters the probability of disease.

$$\text{Likelihood ratio} = \frac{\text{probability of result in diseased patients}}{\text{probability of result in nondiseased patients}}.$$

We can use this definition to define a positive test result and a negative test result. A positive test result raises the probability of disease, and its

likelihood ratio is >1.0. The likelihood ratio for a positive test result is abbreviated as LR(+). A negative test result lowers the probability of disease, and its likelihood ratio is between 0.0 and 1.0. The likelihood ratio for a negative test result is abbreviated LR(-).

> *Example:* If an exercise test is positive, the likelihood ratio is 0.70 divided by 0.15, or 4.666. The odds of having coronary artery disease increase by a factor of 4.666 if an exercise test is abnormal. (Odds are defined in the Glossary of Terms.) If an exercise test is negative, the likelihood ratio is 0.30 divided by 0.85, or 0.35. When an exercise test is negative, the odds of having coronary artery disease are 0.35 times the pretest odds.

Bayes' theorem uses data on test performance in the following way. In these formulas, TPR (true-positive rate) is used in place of sensitivity, and FPR is used to denote false-positive rate. TNR denotes the true-negative rate, and FNR denotes the false-negative rate (these terms are defined in the Glossary). The pretest probability of disease is represented by p(D).

$$\text{Probability of disease if test is positive} = \frac{p(D) \times TPR}{p(D) \times TPR + [1 - p(D)] \times FPR}.$$

$$\text{Probability of disease if test is negative} = \frac{p(D) \times FNR}{p(D) \times FNR + [1 - p(D)] \times TNR}.$$

The probability of a positive test result equals the probability of a true-positive result plus the probability of a false-positive result.

$$\text{Probability of positive test result} = p(D) \times TPR + ([1 - p(D)] \times FPR.$$

Bayes' theorem can be written in a simplified way that facilitates calculation. This form is called the odds-ratio form of Bayes' theorem.

$$\text{Posttest odds} = \text{pretest odds} \times \text{likelihood ratio}.$$

Example: A clinician is planning to use an exercise test with a sensitivity (TPR) of 0.7 and a false-positive rate (FPR) of 0.15. Suppose the pretest probability of disease, p(D), is 0.30:

$$\text{Probability of disease if test positive} = \frac{p(D) \times TPR}{p(D) \times TPR + [1 - P(D)] \times FPR}$$

$$= \frac{.30 \times .70}{.30 \times .70 + .70 \times .15} = \frac{.21}{.21 + .105} = .667.$$

The pretest odds are .30/.70 = 0.43 to 1.0. The likelihood ratio for the test is .70/.15 = 4.667.

$$\text{Posttest odds} = \text{pretest odds} \times \text{likelihood ratio}$$
$$= 0.43 \times 4.667 = 2.0 \text{ to } 1.0.$$

Odds of 2.0 to 1.0 are equivalent to a probability of 0.66.

The importance of Bayes' theorem in interpreting a test is that it defines the relationship between pretest probability and posttest probability, which is shown in Figure 2.1. The relationship between these two entities has several implications.

The interpretation of a test result depends on the pretest probability of disease. If a result is positive, the posttest probability increases as the pretest probability increases (Figure 2.1a). If the result is negative, the posttest probability decreases as the pretest probability decreases (Figure 2.1b). The consequence of this relationship is that *one cannot properly interpret the meaning of a test result without taking into account what was known about the patient before doing the test*. This statement is inescapably true, because it is based on first principles of probability theory.

The effect of a test result depends on the pretest probability. The vertical distance between the 45-degree line in Figure 2.1 and the curve is the difference between the pretest and the posttest probability.

When the clinician is already quite certain of the patient's true state, the probability of a disease is either very high or very low. When the pretest probability is very *low*, a negative test has little effect, and a positive test has a large effect. When the probability is very *high*, a

negative test has a considerable effect, and a positive test has little effect. This example shows that a test result that confirms one's prior judgment has little effect on the probability of disease. Tests have large effects when the probability of disease is intermediate, which corresponds to clinical situations in which the physician is quite uncertain. Tests can also be useful when their result does not confirm the prior clinical impression—for example, a negative result in a patient who is thought very likely to have a disease.

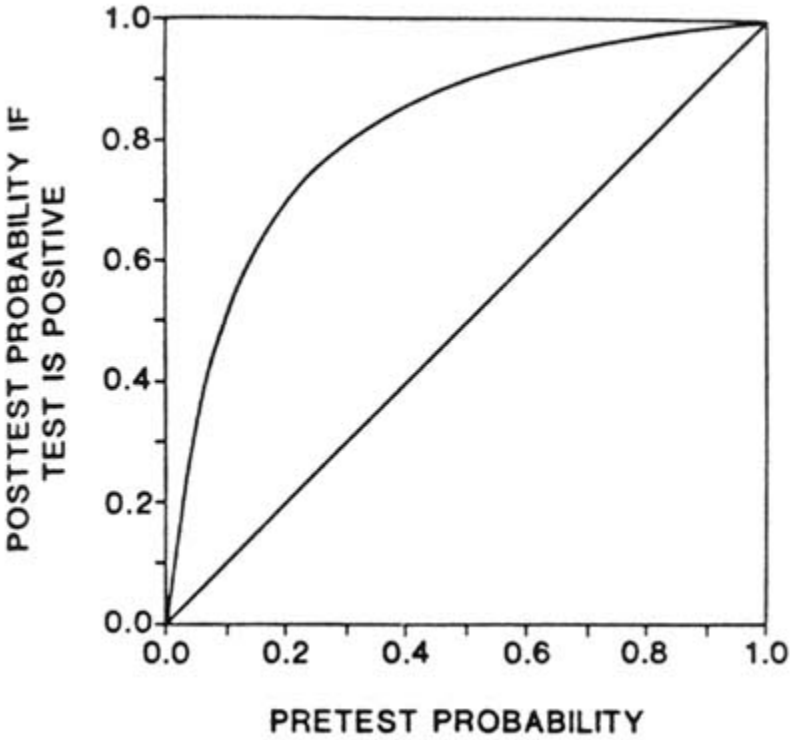

FIGURE 2.1 Relationship between pretest probability and posttest probability of disease.

Figure 2.1a
The posttest probability of disease corresponding to a positive test result was calculated with Bayes' theorem for all values of the pretest probability. The sensitivity and specificity of the hypothetical test were both assumed to be 0.90.

The pretest probability affects the probability that a positive or negative test result will occur. The higher the pretest probability, the more likely one is to experience a positive test. Conversely, a negative test is less likely as the pretest probability increases.

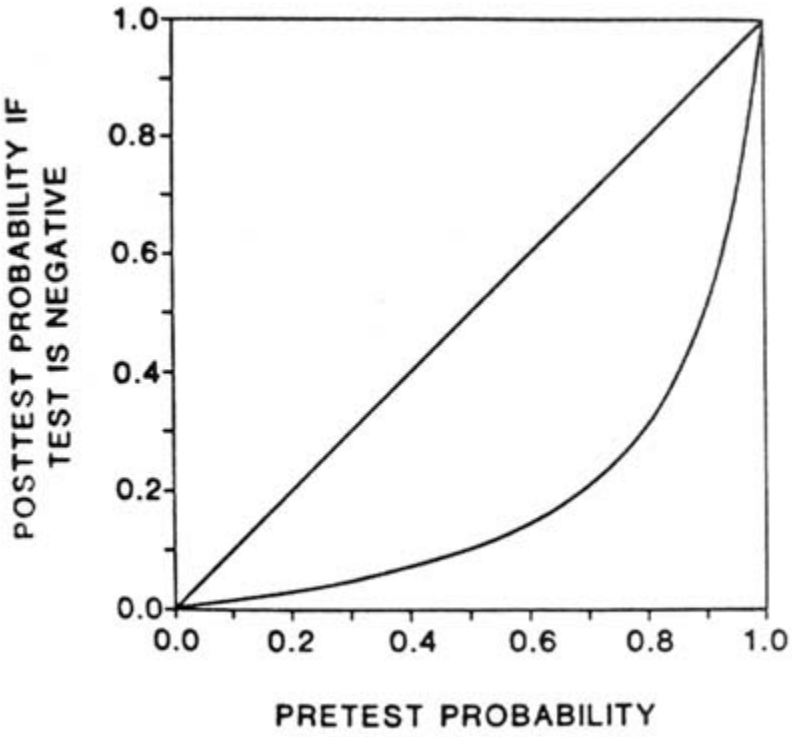

Figure 2.1b
The posttest probability of disease corresponding to a negative test result was calculated with Bayes' theorem for all values of the pretest probability. The sensitivity and specificity of the test were both assumed to be 0.90.

The posttest probability depends on the sensitivity and the false-positive rate of the diagnostic test. This relationship is one reason to be concerned about measuring test performance accurately.

The Assumptions of Bayes' Theorem

Bayes' theorem is derived from first principles of probability theory. Therefore, when it is used correctly, the result is reliable. Errors in using Bayes' theorem can occur when people ignore several assumptions.

One assumption of Bayes' theorem is that sensitivity and specificity are constant, regardless of the pretest probability of disease. This assumption can be false. A test may be less sensitive in detecting a disease in an early stage, when the pretest probability is low, than it would be in an

advanced stage, when there are many signs and symptoms and the pretest probability is high. This error may be avoided by dividing the study population into subgroups that differ in the extent of clinical evidence for disease (Weiner et al. 1979).

A second assumption is that the sensitivity and false-positive rate of a test are independent of the results of other tests. This *conditional independence assumption* is important when Bayes' theorem is used to calculate the probability of disease after a sequence of tests. The posttest probability after the first test in a sequence is used as the pretest probability for the second test. In an ideal study of two tests, both tests in the sequence and a definitive diagnostic procedure have been performed on many patients. The sensitivity and specificity of the second test in the sequence are calculated twice: in patients with a positive result on the first test and in patients with a negative result on the first test. If the sensitivity and specificity of the second test are the same, they are said to be *conditionally independent* of the results of the first test, and the conditional independence assumption is valid. In practice, the conditional independence assumption is seldom tested, and the clinician should be cautious about using recommendations for sequences of tests.

THE PRETEST PROBABILITY OF DISEASE

Why is the pretest probability of disease an important concept in understanding the assessment of diagnostic technology? The pretest probability is required to calculate the posttest probability of disease, and thus to interpret a diagnostic test; it is also the cornerstone of the decision whether to treat, to test, or to do nothing. A patient's pretest probability of disease encodes the individual's own clinical findings and is one of the ways in which a decision can be tailored to the patient. Knowing how to estimate the pretest probability is an essential clinical skill and is described in the section that follows.

When is testing particularly useful? Patients are particularly *unlikely* to benefit from testing when the pretest probability is very high or very low. *If the pretest probability is very high*, the physician is likely to treat the patient unless a negative result raises doubts about the diagnosis. The posttest probability of disease after a negative test may be so high that treatment is still indicated.

Example: In a patient with typical angina pectoris, the posttest probability after a negative exercise test is 0.76. Most physicians

would begin medical treatment for coronary artery disease even if the probability of disease were considerably less than 0.76. For these physicians, the decision to treat would not be affected by the normal exercise test result.

If the pretest probability is very low, as occurs in screening asymptomatic individuals, the clinician is likely to do nothing unless a positive test result raises concern. If, for example, the pretest probability is less than 0.001, the posttest probability may be less than 0.01. In this situation, a change in management is not indicated.

Figure 2.1 shows that the greatest benefit from testing is likely to occur when the pretest probability of disease is *intermediate*. This corresponds to a clinical situation in which there is uncertainty about the patient's true state. Patients are also likely to benefit from testing when the pretest probability is close to a treatment threshold probability. At this point, it requires only a small change in the probability of disease to cross the threshold and alter management.

Physicians customarily use their intuition to estimate the probability of disease. The two principal influences on probability estimates are personal experience and the published literature.

Using personal experience to estimate probability. To estimate probability, the physician should recall patients with characteristics similar to the patient in question, and then try to recall what proportion of these patients had disease. This cognitive task is forbiddingly difficult. In practice, the assignment of a probability to a clinical situation is largely guesswork.

There are several cognitive principles for estimating probability (Tversky and Kahneman 1974). These principles are called *heuristics*.

A clinician is using *the representativeness heuristic* when he or she operates on the principle that "If the patient looks like a typical case, he probably has the disease." Thus, if a patient has all the findings of Cushing's disease, he is thought very likely to have the disease itself. The representativeness heuristic can be misleading, because it leads the physician into ignoring the overall prevalence of a disease. It can also lead to error if the patient's findings are poor predictors of disease or if the physician overestimates probability when there are many redundant predictors. Additionally, the clinician's internal representation of the disease may be incorrect because it is based on a small, atypical personal experience.

Clinicians are using the *availability heuristic* when they judge the

probability of an event by the ease with which similar events are remembered. This heuristic is usually misleading.

Individuals often adjust from an initial probability estimate (the anchor) to take account of unusual features of a patient. The *anchoring and adjustment heuristic* is an important principle. It is equivalent to someone planning a trip by public transportation; the first step is to identify the subway station that is closest to the destination. The person then walks through the neighborhood of the station to the final destination. Bayes' theorem is the best way to make adjustments from the initial anchor point.

Published experience. The reported prevalence of a disease in a clinical population is a useful starting point for estimating the probability of the disease. The physician can then modify this initial estimate to take into account the patient's clinical findings. Most published studies have two important shortcomings.

The first drawback is that these studies usually lack the data required to estimate probability. A typical description will report the prevalence of a clinical finding in patients with a disease, rather than the prevalence of various diseases in patients with a clinical symptom. The anchor point for estimating probability is the prevalence of a disease in patients with a particular clinical finding or diagnostic problem. Thus, a typical study will report the prevalence of weakness in patients with Cushing's disease, when what is needed is the prevalence of Cushing's disease in people complaining of weakness.

Published studies fall short in another way. They report the prevalence of a finding in patients with a disease, which is the *sensitivity* of the finding; they do not report its prevalence in patients who do not have the disease, which is the *false-positive rate* of the finding. The most useful type of study also reports the prevalence of a finding in patients who were initially suspected of having the disease but were proven not to have it.

One reason for these shortcomings is that studies are often done by specialists who report on patients referred to them with a disease. Studies should be done by primary care physicians who keep track of everyone in their practice with a particular clinical complaint, eventually identifying all patients as either having or not having a particular disease.

Clinical prediction rules. These are derived from systematic study of patients with a diagnostic problem, and they define how combinations of clinical findings may be used to estimate probability (Wasson et al. 1985). One well-known rule is designed to help a preoperative consultant estimate the probability that a person scheduled for surgery will have a cardiac complication during surgery (Goldman et al. 1977). The rule

designates the clinical findings that are the best predictors of a complication and assigns a numerical weight to each. The clinician measures the "preoperative score" by taking the sum of the numerical weights of each finding. He or she then estimates the probability of a complication by noting the frequency of complications in prior studies of patients with similar scores.

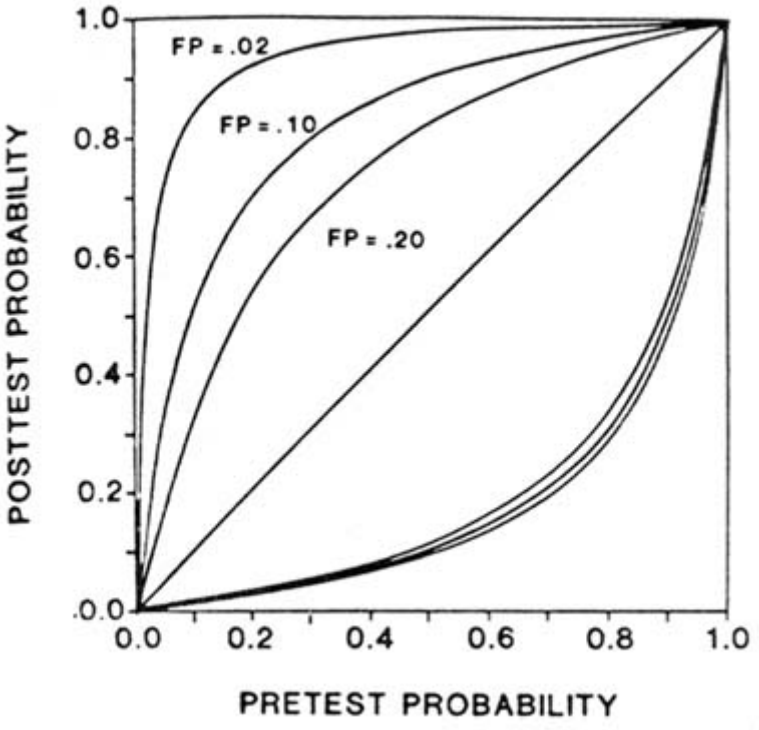

FIGURE 2.2 Effect of test sensitivity and specificity on posttest probability.

Figure 2.2a
As seen in the upper family of curves, the *falsepositive rate* (denoted by FP) of a test is an important factor in determining the posttest probability after a *positive* test. The false-positive rate, however, has a very small effect on the posttest probability after a *negative* test result, as seen in the lower family of curves.

MEASURING THE PERFORMANCE CHARACTERISTICS OF DIAGNOSTIC TESTS

This section describes what many would regard as the central issue in the assessment of diagnostic tests: how to measure their performance characteristics. As shown in Figure 2.2, the sensitivity and specificity of

a test determine its effect on the probability of disease and, therefore, how the test should be interpreted.

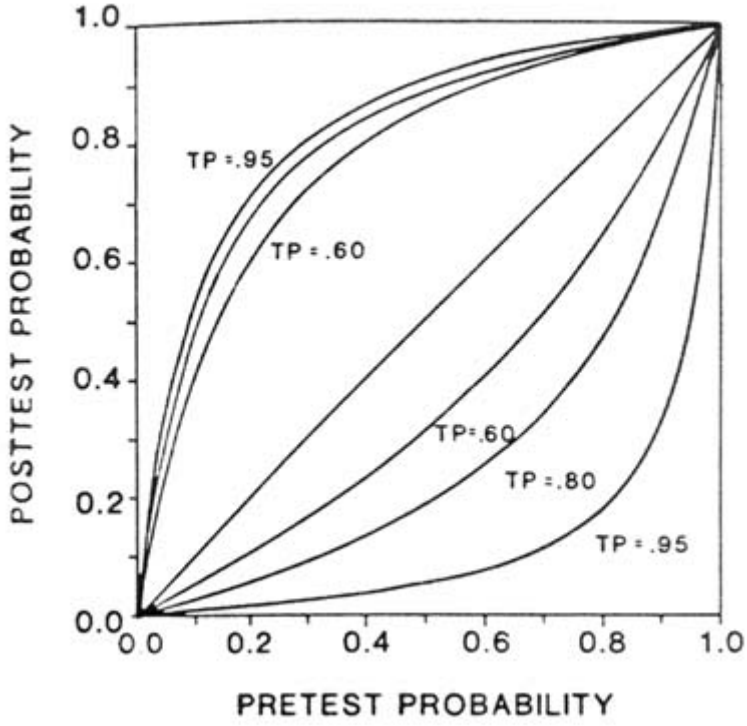

Figure 2.2b
Test *sensitivity* (denoted by TP) has relatively little effect on the posttest probability after a *positive* test, as seen in the upper family of curves. It does affect the posttest probability after a *negative* test, however (lower family of curves), particularly when the pretest probability is high.

Studies that measure the sensitivity and false-positive rate of a test are important, but they are difficult to perform. Many apply only to a narrow spectrum of patients, and studies of the same test in different institutions may lead to discrepant results.

Example: Computed tomography (CT) is often used to determine the extent of a newly discovered lung cancer and thus whether removing the cancer has any chance of curing the patient. As shown in Table 2.1, a survey of studies of CT in lung cancer patient shows wide variation in the results.

TABLE 2.1 True-Positive Rate and False-Positive Rate of Computed Tomography for Detecting Mediastinal Metastases from Lung Cancer

Study	True-Positive Rate	False-Positive Rate	Likelihood Ratio (+)	Likelihood Ratio (-)
1	.29	.54	.5	1.5
2	.51	.14	3.6	.6
3	.54	.32	1.7	.7
4	.57	.15	3.8	5.0
5	.61	.19	3.2	.5
6	.74	.02	37.0	.3
7	.80	.24	3.3	.3
8	.85	.11	7.7	.2
9	.88	.06	14.7	.1
10	.94	.37	2.5	.1
11	.95	.36	2.6	.08
12	.95	.41	2.3	.08
13	.95	.32	3.0	.07

SOURCE: Inouye and Sox 1986.

Figure 2.3 shows the consequences of this wide variation in measured test performance characteristics: the probability of mediastinal metastases if the CT scan is abnormal and if it is normal. The data used to calculate the posttest probability, for a pretest probability of 0.50, were taken from two of the studies in Table 2.1. Depending on which study is used, the interpretation of the test varies widely. In one case, one may interpret a test result as indicating that disease is present if the test is positive and absent if the test is negative. Using data from another study, one cannot conclude anything from a test result, because the probability of disease is changed very little by the test results. This example shows forcefully how much clinical decisions can depend on high-quality studies of test performance.

The discussion of the measurement of test performance characteristics starts with a description of some of the terms used in describing and interpreting studies of test performance. The design of a typical study is as follows. A series of patients undergo the test under study and a second test that is assumed to be a perfect indicator of the patient's true state. The results are displayed in Table 2.2.

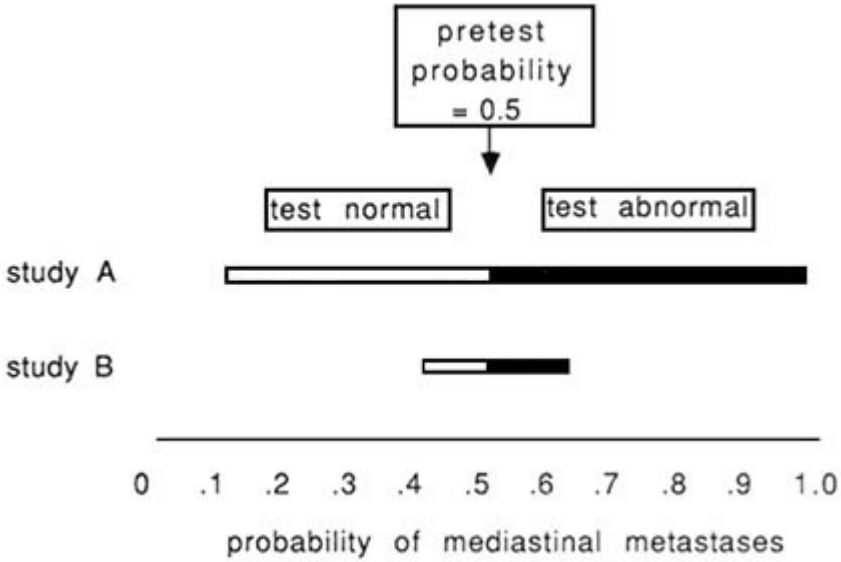

FIGURE 2.3 Posttest probability of mediastinal metastases. The pretest probability is 0.5. Study A and Study B denote two studies of the performance of the CT scan in detecting mediastinal metastases. The posttest probability was calculated with Bayes' theorem, using the true-positive rate and false-positive rate from each of the two studies.

TABLE 2.2 Test Performance Measurement

Test Result	Disease Present	Disease Absent
Positive	A	B
Negative	C	D
Total	A + C	B + D

NOTE: True-positive rate (sensitivity) = A/(A + C); false-negative rate = C/(A + C); false-positive rate = B/(B + D); true-negative rate (specificity) = D/(B + D).

Designing Studies of Test Performance

The principal problem with most studies of the operating characteristics of tests is that the clinically relevant population differs from the study population (for definitions of unfamiliar terms, see Glossary of Terms). Selective referral may result in as few as 3 percent of the clinically relevant population who received the index test being referred for the gold-standard test (Philbrick et al. 1982). Those who design studies of test performance need to ask the following questions, which are based on the work of Philbrick and Feinstein (Philbrick et al. 1980).

Do the Patients in the Study Population Closely Resemble the Patients in the Clinically Relevant Population?

Early in the history of a test, the discrepancy between these two groups may be particularly striking. Often the nondiseased subjects are normal volunteers, for whom the false-positive rate of the test will be lower than the expected in the clinically relevant population. Often the diseased patients are very sick indeed, because an early goal of study is to be sure the test can detect disease. If only the sickest patients are included, the true-positive rate will be higher than in the clinically relevant population.

Was an Abnormal Result on the Index Test a Criterion for Referring the Patient for the Gold-Standard Test?

Ideally, the answer is no, and the index test is obtained routinely on patients who have been referred to have the gold-standard test for other reasons. Referring physicians are much more apt to refer patients with an abnormal index test result and are unlikely to refer patients with a negative index test, because the latter is seen as presumptive evidence against the disease. When the index test is a referral criterion (*workup bias*), the true-positive rate and the false-positive rate will both be higher than would be expected in the clinically relevant population.

If the Index Test or the Gold-Standard Test Required Visual Interpretation, Was the Observer Blinded to All Other Information About the Patient?

When the observer's interpretation of one test is influenced by knowledge of the results of the other test, the concordance between the two

results is likely to increase. *Test-review bias* refers to the situation in which the index test is interpreted by someone who knows the results of the gold-standard test. *Diagnosis-review bias* refers to the opposite situation, in which the gold-standard test is interpreted by someone who knows the results of the index test. Both of these biases increase the true-positive rate and reduce the false-positive rate.

Were the True-Positive Rate and False-Positive Rate of the Test Measured in Clinically Relevant Subgroups of Patients?

Most study populations contain a spectrum of patients, whose disease state varies in clinical severity and in anatomic extent. An average figure for true-positive rate and false-positive rate may conceal clinically important differences among subgroups. The true-positive rate may be higher, for example, in patients with extensive disease than in those with early or mild disease. The ideal study provides the true-positive rate and false-positive rate in clinically defined subgroups and in subgroups defined by anatomic extent of disease.

Was Interobserver Disagreement Measured?

Experts often disagree on the interpretation of images or tracings. Two clinicians can provide different answers to the same question. Which interpretation is to be believed? The study protocol should provide for independent interpretation of study data by two or more observers. Interobserver disagreement should be calculated.

Is the Gold-Standard Procedure an Accurate Measure of the True State of the Patient?

The sensitivity and false-positive rate should be measures of a test's ability to predict the patient's true state. In fact, they are measures of the index test's ability to predict the results of the gold-standard test. If the gold standard does not reflect the patient's true state perfectly, one will be unable to interpret the results of a test as a measure of disease.

Is the Study Population Described Carefully Enough to Allow Comparison to the Clinically Relevant Population?

The demographic and clinical characteristics of the study population must be presented in enough detail to permit a determination of the applicability of the findings to the patients in a particular clinical setting.

Choosing a definition of an abnormal result. Most studies of test performance define sensitivity and specificity in relation to a single cutoff value of a continuous variable. Much information may be lost when test results are defined as dichotomous variables, such as "positive" and "negative." Many test results are expressed as a continuous variable, such as the serum concentration of creatine phosphokinase. A very high serum concentration of creatine phosphokinase is much more indicative of myocardial infarction than a serum concentration that is just above the upper limit of normal. When sensitivity and specificity are known for each point on a continuous scale, the posttest probability can be calculated for any test result.

The relationship between the true-positive rate and the false-positive rate of series of cutoff points may be expressed graphically. The graph is called a receiver operating characteristic (ROC) curve. The ROC curve was first used to express the performance of radar systems in distinguishing warplanes from other objects on the radar screen. Figure 2.4 shows a ROC curve for the exercise electrocardiogram. The ROC curve expresses graphically a basic rule: as you adjust the cutoff point to detect more diseased patients, you inevitably label more nondiseased patients as having disease.

> *Example:* The ROC curve in Figure 2.4 shows that there are very few false-positive results when one chooses 2.5-mm ST-segment depression as the definition of an abnormal result. Nevertheless, few patients with coronary artery disease have such an extreme result on the exercise electrocardiogram, and there would be many false-negative results if this cutoff point were chosen. By instead choosing 1-mm ST-segment depression to define an abnormal result, one detects many more patients with disease, but there are far more false-positive results than when 2.5-mm ST-segment depression was chosen.

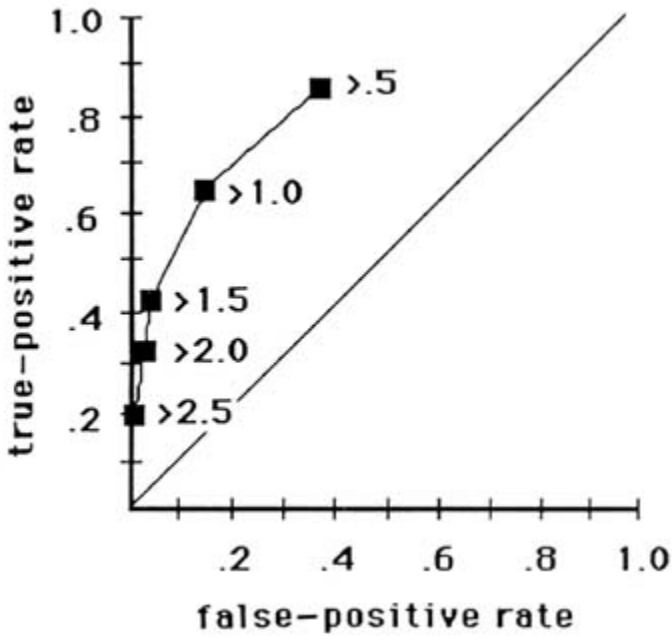

FIGURE 2.4 A ROC curve for the exercise electrocardiogram as a predictor of significant coronary artery disease. The numbers represent the amount of ST-segment depression (measured in millimeters) that is used to define an abnormal exercise test.

How does one choose the optimum cutoff point on the ROC curve? The optimum point is determined by the pretest probability of disease (p[D]) and by the ratio of the costs of treating nondiseased patients as if they had disease (C) to the benefits of treating diseased patients (B) (Metz 1978).

$$\text{Slope of ROC curve at the optimum operating point} = \frac{(1 - p[D])}{p[D]} \times \frac{C}{B}.$$

The slope of the ROC curve is relatively steep for points that are close to the origin, where both the true-positive rate and the false-positive rate are low. The clinician should choose a cutoff point near the origin when the disease is rare or the treatment is dangerous; this choice will serve to minimize both the number of false-positive results and the danger to nondiseased patients. The slope of the ROC curve is flat near the upper

right hand corner, where the true-positive rate is very high and false-negative results are uncommon. The clinician should choose a cutoff point in this area when the patient is very likely to have disease or when the treatment is safe and effective. This choice will minimize false-negative results in a situation where they would be very harmful.

EXPECTED-VALUE DECISIONMAKING

Expected-value decisionmaking is the central idea behind quantitative approaches to decisionmaking when the outcomes are uncertain. Physicians cannot be right all the time. Given our limited understanding of the biologic factors that underlie the response to treatment, some patients will always have done better if they had been treated differently. Since the physician cannot always make the right recommendation for an individual patient, he or she should choose a decisionmaking strategy that will maximize the number of good outcomes that are seen during a lifetime of making decisions. This strategy is called *expected-value decision making*.[3] The decisionmaker chooses the option that has the largest benefit when averaged over all patients.

Expected-value decisionmaking is a straightforward concept. The problem lies in applying it to patient care. How does one calculate an average value for a management alternative? How does one place numerical values on the outcomes of an illness? These questions are best answered by example.

Example: Consider a treatment decision for a patient with chronic pancreatitis. The patient himself and the internists caring for him favored operating on the patient's pancreas. The surgeons were not enthusiastic, citing the high mortality of the operation. In making their case, the internists decided to calculate the expected value of medical therapy and surgery. They represented choice between surgery and continued medical management by the decision tree shown in Figure 2.5. The first node (represented by a square) represents the decision to operate or not, and there are two branches, one for the surgery option and one for the medical treatment option.

[3] We use "value" as a general term for that which one tries to maximize in decisionmaking. Strictly speaking, one might speak of *expected-outcome* decisionmaking, in which the outcome could be life expectancy or a measure of preference for the outcome states.

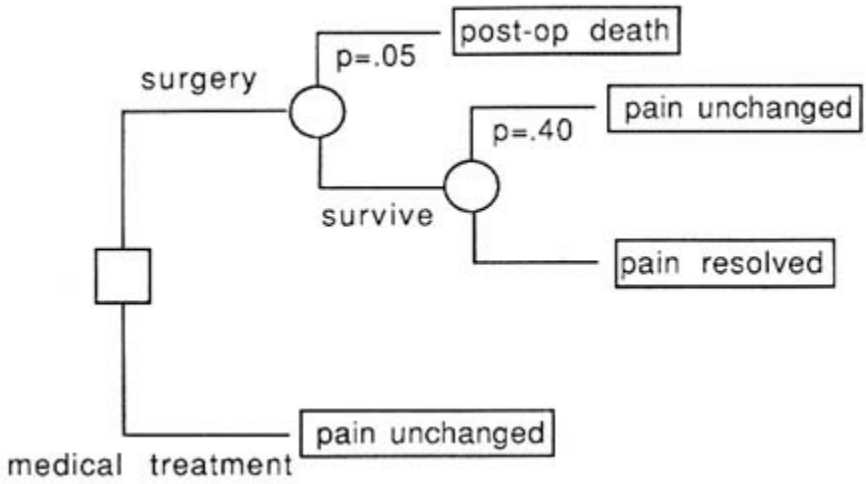

FIGURE 2.5 A decision tree for deciding between surgery and medical management of chronic pancreatitis. The square represents a decision node. The circles represent chance nodes. The square represents a decision node. The rectangles represent terminal nodes for the various outcome states, and the numbers enclosed within the rectangles are the products of the length of life and the quality of life in the outcome state.

Setting Up the Decision Tree

Surgery: The first node on the surgery branch is a chance node (represented by a circle), which represents uncertainty about whether the patient would survive the operation. The patient may survive or may die, but the true outcome of the operation is unknown and can only be represented by a probability. On average, the mortality rate of the operation is 5 percent, which seemed a reasonable representation for this patient, who was otherwise well. The next uncertainty was the outcome of treatment. Only about 60 percent of patients obtain relief of pain after surgery. The possible outcomes are represented by terminal nodes (shown as a rectangle). Each outcome is assigned a quantitative measure, such as the life expectancy in that outcome state. This patient's life expectancy was 20 years.

Medical treatment: Because management associated with the medical option does not change, there are no chance nodes, and the patient's life expectancy is 20 years.

Weighing the Outcomes for Quality of Life

The patient's life expectancy was 20 years if he survived the operation, and it was thought to be the same regardless of whether he experienced chronic pain or was pain-free. The patient pointed out that 20 years of life with chronic pain was equivalent to 12 years of being pain-free. In other words, to be free of pain he was willing to give up eight years of life with chronic pain. This method for weighing the length of life in a certain state of health by a factor that represents the quality of life in that state is called the "time trade-off" method. It is described in standard textbooks (Weinstein et al. 1980, Sox et al. 1988).

Calculating the Expected Value of the Treatment Options

The average (or expected) outcome is calculated by taking the product of all the probabilities along a path to a terminal node and multiplying it by the value assigned to the terminal node. The management alternative with the highest expected value is usually the preferred choice. In this case, the expected length of life, measured in healthy years, was 16.8 years for surgery and 12 years for medical management. The surgeons were convinced by this analysis and scheduled the patient for surgery.

Note that expected-value decisionmaking allows one to balance the risks and benefits of treatment. These factors are usually considered intuitively. By assigning a value to each outcome and weighing it by the chance that it will occur, expected-value decisionmaking allows one to integrate risks and benefits.

THE CHOICE AMONG DOING NOTHING, TESTING, OR STARTING TREATMENT

The art of medicine is making good decisions with inadequate data. Physicians often must start treatment when still uncertain about whether the patient has the disease for which the treatment is intended. If treatment is started, there is a risk of causing harm to a person who does not have the disease, as well as the prospect of benefiting the person who does. If treatment is withheld, a person who is diseased will be denied a chance at a rapid, effective cure. This situation is often unavoidable, and the physician has three choices.

- *Do nothing:* the chance of disease is low, treatment is either harmful or ineffective (or both), and a false-positive result might occur and lead to harmful treatment for someone who does not have disease.
- *Get more information:* do a test or observe the patient's course in the hope that the correct choice will become apparent.
- *Start treatment now:* the chance of disease is relatively high, treatment is safe and effective, and a false-negative result might lead to withholding useful treatment from someone with disease.

The method for solving this problem analytically is called the *threshold model of medical decisionmaking* (Pauker and Kassirer 1975, 1980; Doubilet 1983). The threshold model is an example of expected-value decisionmaking that is applied to a particular type of decision.

The key idea is the *treatment threshold probability*, which is the probability of disease at which one is indifferent between treating and not treating. The basic principle of the threshold model is the following dictum: *Do a test only if the probability of disease could change enough to cross the treatment threshold probability*. Three steps are required to translate this idea into action.

Step 1: Estimate the pretest probability of disease.

Step 2: Set the treatment threshold probability. This step is difficult because it requires the clinician to express the balance of the risks and benefits of treatment in a single number. One can use clinical intuition to set the treatment threshold probability. This task is made easier by the following relationship:

$$\text{Treatment threshold} = \frac{C}{C + B},$$

where C is the cost of treating nondiseased patients, and B is the benefit of treating diseased patients (Pauker and Kassirer 1975). The cost and the benefit must be expressed in the same units, which can be dollars, life expectancy, or a measure of the patient's attitudes toward treatment and the disease.

Note that when the costs of treating nondiseased patients equal the benefits of treating diseased patients, the treatment threshold is 0.5. Thus, for many treatments, the treatment threshold probability will be less than

0.50. For a safe, beneficial treatment, such as antibiotics for community-acquired pneumonia, the treatment threshold probability may be less than 0.10. If there is good reason to suspect disease, the pretest probability will be above the treatment threshold. In deciding whether to perform a test, the clinician must ask whether the posttest probability after a negative test result would be below the treatment threshold probability. This question is answered by taking Step 3, described below.

One can also use analytic methods to set the treatment threshold probability (Sox et al. 1988). Consider the decision tree in Figure 2.6, which shows a hypothetical problem in which treatment must be chosen despite uncertainty about whether the patient has the disease for which the treatment is intended.

To use the decision tree to estimate the treatment threshold, recall that this threshold is the probability of disease at which one is indifferent between treating and not treating. First, one assigns values to each of the probabilities and outcome states except for the probability of disease. Second, one calculates the expected value of the two options, leaving the probability of disease as an unknown. Third, one sets the expression for the expected value of the treatment option equal to that of the nontreatment option. Fourth, one solves for the probability of disease.

To use a decision tree, one must assign a probability to each chance node and a numerical value to each outcome state. The latter value could be *life expectancy*. Alternatively, as shown in Figure 2.6, one could assign each outcome state a *utility*, which is a quantative measure of relative preference. A utility of 1.0 is assigned to the best outcome, and a utility of 0.0 to the worst. The utility of each intermediate outcome state is then assessed on this scale of 0.0 to 1.0. When utility is used as the measure of outcome, the alternative with the highest expected utility should be the preferred alternative.

> *Step 3: Use Bayes' theorem to calculate the posttest probability of disease*. If the pretest probability is above the treatment threshold, one must calculate the probability of disease if the test is negative. If the pretest probability is below the treatment threshold, one must calculate the probability of disease if the test is positive.

If the pretest probability is far enough above or below the treatment threshold, a test result will not affect management because the posttest probability will be on the same side of the treatment threshold as the pretest probability. There is a pretest probability for which the posttest probability is exactly the point at which one is indifferent between not

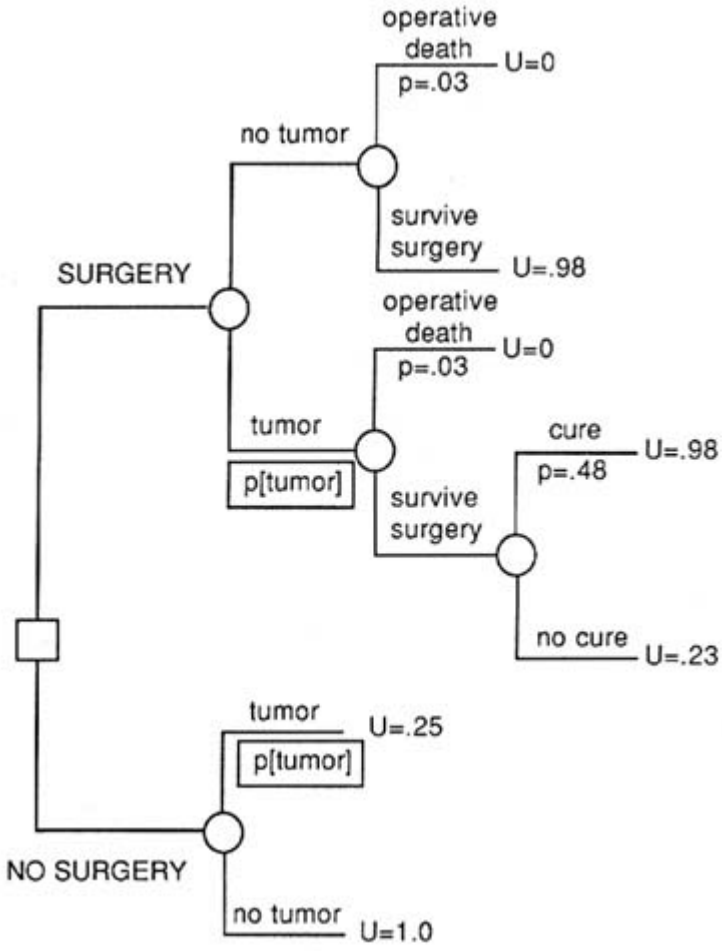

FIGURE 2.6 A decision tree for choosing between treatment and no treatment when the clinician does not know whether the patient has the disease for which treatment is indicated.

treating and testing (the treatment threshold probability). Below this pretest probability, called the no treat–test threshold (Pauker and Kassirer 1980), a positive test result could not increase the probability of disease enough to cross the treatment threshold, and both testing and treatment should be withheld. Above this threshold, the posttest probability will exceed the treatment threshold, and testing is indicated. These concepts are illustrated in Figure 2.7.

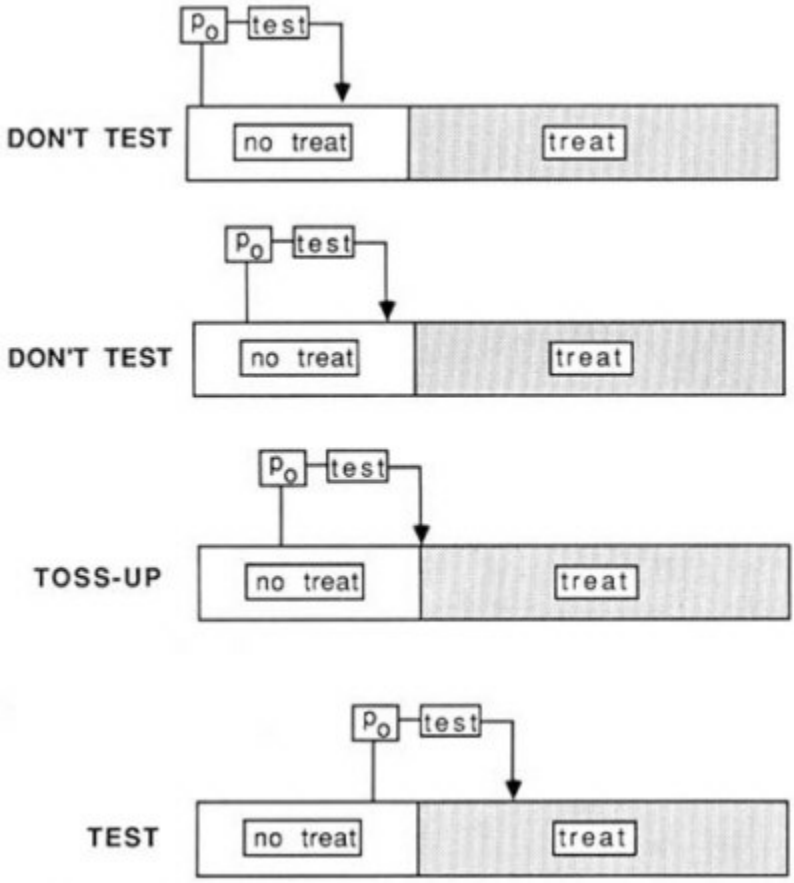

FIGURE 2.7 Illus ration of how to set the no treat–test threshold. As p_0, the pretest probability, is gradually increased, the posttest probability is first below the treatment threshold, then equal to it, and finally above it. At the point where p_0 equals the treatment threshold, one should be indifferent between not treating and testing. This probability is the no treat–test threshold.

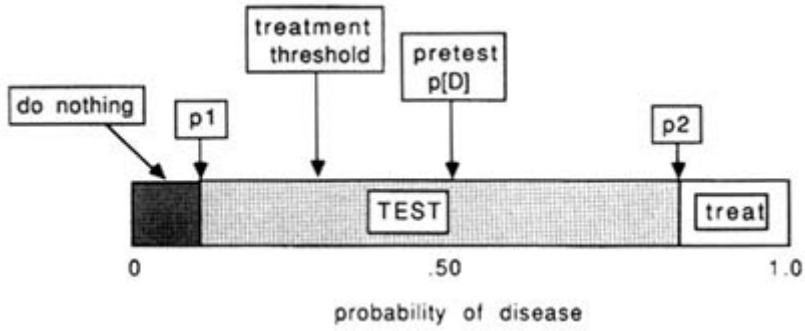

FIGURE 2.8 Using the treatment threshold probability to help decide whether to do a test.

One can use the same approach to calculate the point at which one should be indifferent between testing and treating (the test–treatment threshold). Both of these thresholds are a function of the true-positive rate of the test, the false-positive rate of the test, the treatment threshold, and a measure of what experiencing the test means to the patient (the cost of the test). Figure 2.8 shows the three zones of the probability scale. Using Figure 2.8, one needs only to estimate the pretest probability to know whether testing is the preferred action or whether one should treat or do nothing.

SUMMARY

The purpose of this chapter is to provide a working knowledge of how probability theory and expected-value decisionmaking are used to help make decisions about diagnostic testing. Past studies of diagnostic tests have measured only test performance. A complete evaluation should provide information about the treatment threshold and how to estimate the pretest probability of disease. With this information, the clinician can decide when a test will alter management and can use the results to choose the action that will most benefit the patient.

Glossary Of Terms

Bayes' theorem:	an algebraic expression for calculating the posttest probability of disease if the pretest probability of disease [p(D)] and the sensitivity and specificity of a test are known.
Clinically relevant population:	the patients on whom a test is normally used.
Cost-effectiveness analysis:	comparison of clinical policies in terms of their cost for a unit of outcome. *Marginal cost-effectiveness:* the increase in cost of a policy for a unit increase in outcome.
False-negative rate:	the likelihood of a negative test result in a diseased patient (abbreviated FNR).
False-negative result:	a negative result in a patient with a disease.
False-positive rate:	the likelihood of a positive test result in a patient without a disease (abbreviated FPR).
False-positive result:	a positive result in a person who does not have the disease.
Gold-standard test:	the test or procedure that is used to define the true state of the patient.
Index test:	the test for which performance is being measured.
Likelihood ratio:	a measure of discrimination by a test result. A test result with a likelihood ratio >1.0 raises the probability of disease and is often referred to as a "positive" test result. A test result with a likelihood ratio <1.0 lowers the probability of disease and is often called a "negative" test result.

$$\text{Likelihood ratio} = \frac{\text{probability of result in diseased persons}}{\text{probability of result in nondiseased persons}}.$$

Negative test result: a test result that occurs more frequently in patients who do not have a disease than in patients who do have the disease.

Odds: the probability.

$$\text{Odds} = \frac{\text{probability of event}}{1 - \text{probability of event}}.$$

Positive test result: a test result that occurs more frequently in patients with a disease than in patients who do not have the disease.

Posttest probability: the probability of disease after the results of a test have been learned (synonyms: posterior probability, posttest risk).

Predictive value negative: probability of the absence of the disease if a test is negative.

Predictive value positive: probability of a disease if a test is positive.

Pretest probability: the probability of disease before doing a test (synonyms: prior probability, pretest risk).

Probability: an expression of opinion, on a scale of 0.0 to 1.0, about the likelihood that an event will occur.

Sensitivity: the likelihood of a positive test result in a diseased person (synonym: true-positive rate, abbreviated TPR).

$$\text{Sensitivity} = \frac{\text{number of diseased patients with positive test}}{\text{number of diseased patients}}.$$

Specificity: the likelihood of a negative test result in a patient without disease (synonym: true-negative rate; abbreviated TNR).

$$\text{Specificity} = \frac{\text{number of nondiseased patients with negative test}}{\text{number of nondiseased patients}}.$$

Study population:	the patients for whom test performance is measured (usually a subject of the clinically relevant population).
Treatment threshold probability:	the probability of disease at which the clinician is indifferent between withholding treatment and giving treatment. Below the threshold probability, treatment is withheld; above the threshold, treatment is given.
True-negative result:	a negative test result in a person with a disease.
True-positive result:	a positive test result in a person with a disease.

REFERENCES

Doubilet, P. A mathematical approach to interpretation and selection of diagnostic tests. Medical Decision Making 3:177-195, 1983.

Goldman, L., Caldera, D.L., Nussbaum, S., et al. Multifactorial index of cardiac risk in non-cardiac surgical procedures. New England Journal of Medicine 297:845-850, 1977.

Griner, P.F., Mayewski, R.J., Mushlin, A.I., and Greenland, P. Selection and interpretation of diagnostic tests and procedures: Principles and applications. Annals of Internal Medicine 94(part 2):553-560, 1981.

Inouye, S.K., and Sox, H.C. Standard and computed tomography in the evaluation of neoplasms of the chest. Annals of Internal Medicine 105:906-924, 1986.

Metz, C.E. Basic principles of ROC analysis. Seminars in Nuclear Medicine 8:283-298, 1978.

Pauker, S.G., and Kassirer, J.P. The threshold approach to clinical decision making. New England Journal of Medicine 302:1109-1117, 1980.

Pauker, S.G., and Kassirer, J.P. Therapeutic decision making: A cost-benefit analysis. New England Journal of Medicine 293:229-234, 1975.

Philbrick, J.T., Horwitz, R.I., Feinstein, A.R., Langou, R.A., and Chandler, J.P. The limited spectrum of patients studied in exercise test research: Analyzing the tip of the iceberg. Journal of the American Medical Association 248:2467-2470, 1982.

Philbrick, J.T., Horwitz, R.I., and Feinstein, A.R. Methodologic problems of exercise testing for coronary artery disease: Groups, analysis, and bias. American Journal of Cardiology 46:807-812, 1980.

Ransohoff, D.F., and Feinstein, A.R. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. New England Journal of Medicine 299:926-930, 1978.

Sox, H.C., Blatt, M.A., Higgins, M.C., and Marton, K.I. Medical Decision Making. Boston, Butterworth, 1988.

Sox, H.C. Probability theory in the use of diagnostic tests: An introduction to critical study of the literature. Annals of Internal Medicine 104:60-66, 1986.

Stason, W.B., and Fineberg, H.V. Implications of alternative strategies to diagnose coronary artery disease . Circulation 66(Suppl. III):80-86, 1982.

Tversky, A., and Kahneman D. Judgment under uncertainty: Heuristics and biases. Science 185:1124-1131, 1974.

Wasson, J.H., Sox, H.C., Neff, R.K., and Goldman, L. Clinical prediction rules: Applications and methodologic standards. New England Journal of Medicine 313:793-799, 1985.

Weiner, D.A., Ryan, T.J., McCabe, C.H., et al. Exercise stress testing: Correlation among history of angina, ST-segment response and prevalence of coronary-artery disease in the Coronary Artery Surgery Study (CASS). New England Journal of Medicine 301:230-235, 1979.

Weinstein, M.C., Fineberg, H.V., Elstein, A.S., Frazier, H.S., Neuhauser, D., Neutra, R.R., and McNeil, B.J. Clinical Decision Analysis. Philadelphia, W.B. Saunders, 1980.

Weinstein, M.C., and Stason, W.B. Foundations of cost-effectiveness analysis for health and medical practices. New England Journal of Medicine 296:716-721, 1977.

3

Assessment: Problems and Proposed Solutions

The purpose of this chapter is to describe some of the problems one encounters in evaluating diagnostic technology and to propose an approach that avoids many of them. Our underlying premise is that the public is ill-served by current approaches to assessment of diagnostic technology.[1]

Medical tests are more difficult to evaluate than medical treatments. A treatment is typically evaluated through a clinical trial in which patients are randomly assigned to the treatment group or to a control group, which may receive a placebo or conventional therapy. The endpoint of the trial may be physiologic (for example, blood pressure) or functional (such as the ability to walk without developing chest pain), but most often the endpoint is the development of a disease (for example, acute myocardial infarction) or death. In the best trials, therapy subsequent to randomization is controlled, and the only variable that differentiates the intervention and control groups is the intervention. Under these circumstances, the investigators are often able to attribute differences in outcome to the intervention. By contrast, evaluation of a diagnostic test can occur at several levels, as outlined by Fineberg (Fineberg et al. 1977).

[1] Parts of this chapter are adapted from a technical report (Sox 1987).

MEASURES OF CLINICAL EFFICACY

Technical Capability

This measure answers the question, "Does the test do what the manufacturer says it does?" For example, an MRI scanner meets this criterion if it produces a crisp image of the brain, regardless of whether that image faithfully reflects the true state of the brain. The Food and Drug Administration currently requires this level of assessment for diagnostic technologies before it will issue premarketing approval.

Sensitivity and Specificity

These two measures of test performance are the most widely used indicators of efficacy. They may help one to decide which of several diagnostic tests is superior, but the verdict is sometimes a split decision: one technology has a lower false-negative rate while the other has a lower false-positive rate. Furthermore, these measures are not sufficient to indicate whether the test should be done. In many cases, the test is so inaccurate and the treatment so safe and effective that the patient should be treated without testing in order to avoid the possibility of being misled by a false-negative result.

Diagnostic Impact

Do the test results alter the pattern of diagnostic testing? Does the test replace other tests, including some that are more hazardous or costly? This outcome is relatively easy to measure, and, because it occurs in the near term, one can often attribute an effect on patterns of testing to a new technology. Noninvasive methods of imaging internal organs have had a major impact on medical care because the information they provide has reduced the number of invasive diagnostic studies performed. CT scanning of the head has reduced the number of craniotomies for head trauma (Ambrose et al. 1976). This measure of efficacy, however, is not sufficient to answer the important question, "Should I do this test on this patient?"

Resolution of diagnostic uncertainty is one measure of diagnostic impact. There is ample evidence that patients seek relief from uncertainty and that diagnostic tests play a role in satisfying them (Sox et al. 1978, Marton et al. 1980). The physician must use reassurance when it is

indicated, as when a test result reduces the probability of disease to the point where no further intervention is needed. Reassurance following a negative result on a test that has a high false-negative rate may not be appropriate in some cases, particularly if the physician strongly suspected that disease was present before doing the test.

Therapeutic Impact

If a test alters the choice of the treatment for the patient, it meets this criterion for efficacy. The threshold model is built around the assumption that an effect on therapy is the *sine qua non* for doing a test. But, as indicated in Chapter 2, a test may alter therapy in one patient but not in another, depending on the pretest probability and the treatment threshold.

Impact on Clinical Outcomes

The ultimate measure of a test is its ability to alter the patient's outlook by leading to changes in management that reduce symptoms or prolong life. The determinants of long-term outcome are many. The accuracy, cost, and morbidity of a test may be much less important than when it is done in the natural history of the illness. The most important determinant of clinical outcome is therapy rather than diagnosis (Abrams and McNeil 1978). Improved imaging of metastases to the liver from a colon cancer does not improve the patient's outcome because there are no highly effective treatments for metastatic colon cancer. Imaging of metastases may, however, spare a patient from abdominal exploratory surgery that cannot alter the long-term prognosis. An improved short-term outcome does not necessarily imply an improved long-term outcome.

This summary of the measures of clinical efficacy indicates the futility of basing a decision about a technology on a single dimension. The way through this dilemma is to focus on the patient's needs. The right question about a technology is "Will this maximize this patient's chances for the best achievable outcome?" In some cases, the answer to this question is the same for a large class of patients, and one can formulate a general recommendation. In others, the answer depends on the value that the individual patient places on the outcomes that the illness and its treatment may entail. In this case, a general recommendation may not be possible. In this chapter, we show how a technology assessment can provide the data that allow a physician to identify which management alternative will maximize the patient's chances for a good outcome.

THE SYSTEM OF TECHNOLOGY ASSESSMENT IN THE UNITED STATES

The Pyramid

The system of technology assessment in the United States is a pyramid with several layers. The broad base of the system consists of clinical studies in which physicians subject patients to a technology and observe them for its effects. These studies include a research question, which usually calls for the comparison of several technologies; they also include a rigorous study design and meticulous implementation of the study protocol. In studies of diagnostic technology, the index test, one or more competing tests, and a gold-standard test are performed in a series of patients. The published reports of these studies seldom put their findings into a form that helps the clinician to decide which patients will benefit most from the test or procedure.

A second layer is made up of individuals who review the literature and try to distill the evidence into recommendations that are true to the facts. These individuals are typically clinicians who have training in the disciplines of meta-analysis, clinical epidemiology, decision analysis, and cost-effectiveness analysis. This method has been used to identify outmoded, overused, and ineffective technologies. More recently, it has been used to make recommendations for using a diagnostic test or choosing among tests.

On the next layer are organizations that do technology assessment. They differ in their approach, but the starting point is frequently a technical background paper written by an individual who reads the literature and proposes guidelines for using the technology. The conclusions of this paper are reviewed by others, and clinical policy recommendations are forged by some consensus process. The American College of Physicians' Clinical Efficacy Assessment Program (CEAP) is a prototype of this approach.

Policymakers sit atop the pyramid and are the ultimate consumers of technology assessment. What they consume is the product of analysis and consensus, and it generally takes the form of recommendations about the usefulness of a technology. The individual physician, who bases decisions about using technology on published reports of assessments, is a policymaker. Other policymakers work for third-party payers, who exert control over medical practice by their coverage policy.

This description of the technology assessment system in the United States shows that many individuals and organizations depend on good

studies of technology. We use the term "primary technology assessment" to denote studies in which clinical data are obtained systematically on patients who have been subjected to a health intervention, such as a diagnostic test or treatment. In the next section, we discuss some of the methodological problems that are encountered in doing primary assessments of diagnostic tests.

PROBLEMS WITH THE CURRENT SYSTEM

Standards of evidence are incomplete. The standard of evidence for the efficacy of therapeutic technologies, such as surgical operations or drugs, has become the randomized clinical trial. This standard may be insufficient for clinical decisionmaking. One drug is considered superior to another if there is a statistically significant difference in a measure of outcome, such as survival. Achieving this criterion does not mean that the drug should be used in all patients. This decision may depend on the characteristics of the individual patient, including the value he or she places on the benefits, adverse effects, and costs of the drug. One can use expected-value decisionmaking to identify the best alternative for an individual patient.

According to the decision model described in Chapter 2, the usefulness of a test depends on the clinical circumstances. Among these is the pretest probability of disease. One of a pair of competing tests may be preferred in patients with a low pretest probability of disease, while the other test should be preferred in patients with a high pretest probability. In summary, we suggest that the efficacy of a test is context-dependent.

Studies do not gather the data needed for decisions in individual patients. Large-scale, randomized trials sometimes lead to the conclusion that a given therapy is preferred only in a subgroup of patients. They do so by gathering the clinical data necessary to subclassify patients. Studies of diagnostic tests could be carried out in the same way, but they seldom are. For example, published studies of diagnostic tests infrequently report clinical prediction rules for estimating pretest probability.

Studies of technology often apply only to a narrow spectrum of patients. As discussed in Chapter 2, the patients who are enrolled in a study of a diagnostic test are often a small minority of those who actually receive it (Philbrick et al. 1980). Similarly, randomized clinical trials may exclude many patients, such as those with more than one disorder, in order to maximize the chances of obtaining an unequivocal answer. The results of these studies may not apply to many patients of concern to clinicians.

Studies of diagnostic tests often do not compare a new test with an established test. Randomized clinical trials of treatments usually compare a new therapy to an established therapy or a placebo. Studies of diagnostic tests often do not compare one test with a competing test. When competing tests are compared, the design of the study usually precludes a complete answer to such questions as, "Should I do Test A but not Test B? Both Test A and Test B? Test A followed by Test B only if Test A is negative?"

Studies are seldom timely. The earliest studies of a new technology tend to be misleadingly optimistic about its performance, often because the study populations are not clinically relevant (Ransohoff and Feinstein 1978). Practice patterns are often established on the basis of early studies. Similarly, when hospital managers decide to invest in a new technology, they must often base their decision on early studies. Therefore, the quality of early studies must be improved.

Technology is constantly changing. By the time a study is completed, the test or imaging device has changed, and no one believes that the results apply to the new, improved technology. Technical changes may improve the image provided by a scanner, but they do not necessarily lead to a lower false-negative or false-positive rate, nor do they guarantee improvement in clinical outcomes. Technology assessment should be done quickly. For example, a multi-institutional study could take but a few months. Also, there should be a system for monitoring, and perhaps reevaluating, the technology as it matures.

The results of a study may apply to a narrow spectrum of the users of the technology. Published assessments of a diagnostic technology are usually done in academic medical centers. The use of the technology in such centers may differ greatly from its use in a community hospital. The indications for using the test, the spectrum of patients, the technique for using the equipment, and the skill of the clinician who interprets the results are but a few of the areas in which an academic medical center may differ from a community hospital.

Two recent case reports illustrate some of the difficulties that are caused by inadequate primary technology assessment.

> *Case Report 1. Premature obsolescence: standard chest X-ray tomography*. Computed tomography (CT) was widely adopted before it had been compared with what was then the standard method for imaging the chest, standard X-ray tomography. Relatively few studies had compared the tests in the same patients. A review that compared their accuracy brought out some unexpected findings

(Inouye and Sox 1986). CT was superior to standard tomography for some indications. When 16 studies of chest tomography for mediastinal metastases were reviewed, however, the frequency of false-negative results was lower for CT, but the frequency of false-positive results was lower for standard tomography. Furthermore, the differences in accuracy were too small to be important for decisionmaking. By now, however, most radiologists consider standard tomography to be obsolete in the study of most intrathoracic disorders.
Comment: Large-scale, multi-institutional, prospective studies comparing CT and standard tomography should have been done very early in the history of the new technology. These might have shown that the two procedures were equivalent in most patients and might have defined patient subgroups in which one test was clearly superior.
Case Report 2. Premature adoption of a new technology: magnetic resonance imaging. Magnetic resonance imaging (MRI) is being adopted by hospitals throughout the United States and may eventually replace computed tomography (CT) in studies of the central nervous system (Steinberg et al. 1985). MRI provides a remarkable definition of central nervous system structures. The images are striking in their detail, but those who purchase MRI scanners or use them should ask several pertinent questions: Does the improved image lead to lower false-negative rates without increasing false-positive rates? Does MRI lead to useful changes in diagnostic certainty, choice of therapy, or even clinical outcome? The answers to these questions were not available when many MRI scanners were purchased, because most early studies of MRI were relatively unsatisfactory (Kent et al. 1988, NIH Consensus Conference 1988).

We now turn to a discussion of how diagnostic tests should be evaluated.

RANDOMIZED TRIALS OF DIAGNOSTIC TESTS

A well-designed and well-executed randomized clinical trial is widely regarded as the most powerful method for comparing technologies. Sources of ambiguity in data interpretation are, in principle, removed by randomization, because this process assures that all potentially influential vari

ables, known and unknown, are distributed equitably among the study groups. Blinding of the investigator and the patient to the assigned intervention reduces bias in obtaining data from patients. A well-conducted trial has internal safeguards to assure strict adherence to the study protocol.

Limitations of Randomized Trials

The cost may be high. Randomized trials can be very costly if standardization of the intervention requires special care for patients. By focusing on effectiveness (measuring effects under usual patient care conditions) rather than efficacy (measuring effects under ideal circumstances), the costs of a randomized trial can be kept to a minimum.

The study population may be too small. Many chronic diseases progress slowly, and outcome events accumulate slowly unless the study population is very large. Evaluating an intervention in subgroups of patients may require an unrealistically large number of patients. A large study population is also required if the intervention is expected to have a small effect. These problems can often be avoided by self-discipline when formulating the study hypotheses. Sometimes the requirement for a large sample size is unavoidable, and many medical centers may be required to assemble a sufficient sample of patients.

The technology may become obsolete before the study is complete. Studies that continue for many years run the risk that the results will be irrelevant because of technological advances that have occurred during the years of the study.

The results may apply to a narrow spectrum of patients. Most randomized trials exclude many patients. For example, only 12.7 percent of the patients in the Coronary Artery Surgery Study were randomized to receive surgery or medical therapy (CASS Principal Investigators, 1983). The remainder were not enrolled because they met one of many exclusion criteria. A study performed in a single institution may have a limited spectrum of study patients. Because of these problems, the results of a study may not necessarily apply to patients who are important both to clinicians and to policymakers. The exclusion of patients older than age 65 from the Coronary Artery Surgery Study is an example (CASS Principal Investigators 1983). Ideally, a randomized trial should include a wide spectrum of care facilities and should enroll patients who might be excluded from other studies.

The trial may not measure outcomes of clinical interest. By focusing on the principal clinical hypothesis, past randomized control trials have

often failed to study other measures of the effect of the intervention. Return to work, psychological status, and social function all measure the impact of successful treatment. Many observers feel that these "secondary endpoints" are as important as the primary endpoint of the study, which is usually mortality from the disease. For example, cost-effectiveness is becoming a study endpoint in many trials.

Trial Design

A randomized trial of a diagnostic test is a powerful method for evaluating its effects on patient care. The test can be compared against another test or against no test. In general, a comparison with no test is not ethically sound. Most would agree that a patient could be randomized to no test only if the clinical history could be relied upon to be sure that the patient does not have the disease in question. Under these circumstances, little can be learned about the effect of the test, other than its psychologically mediated effects (Sox et al. 1981). In general, a randomized trial will compare two putatively similar diagnostic tests, such as MRI and CT. There have been few such studies, and this approach deserves greater use.

One of the advantages of a randomized trial is that the principal study endpoint is a clinical outcome (for example, length of hospitalizations, use of other tests, morbidity, or mortality). In contrast to studies that measure test accuracy, there is no need to perform a potentially dangerous gold-standard test on all patients. This advantage suggests two types of randomized trials of diagnostic tests.

"OFF THE GOLD STANDARD"

If the goal of the study is to measure clinical outcomes rather than test accuracy, one can ethically enroll anyone who needs the index test. Study patients are randomly assigned to have the index test or the alternative test and are then monitored for the occurrence of the endpoint of the study, which could be length of hospitalization, total cost of care, or functional status one month after enrolling. Being able to enroll all patients means that there is little problem with bias in selecting patients, and the findings will apply to primary care populations. A randomized study can show which diagnostic test is superior, and subgroup analysis can identify patients who benefit particularly from a given test. Nevertheless, this type of randomized trial cannot measure the true-positive and false-positive rates of the index test, because the gold-standard test will be performed irregularly, or perhaps not at all. Therefore, this study does not provide all

the information that is required to interpret a test or to decide if it is necessary to perform the test.

"ON THE GOLD STANDARD"

Studies that directly compare two tests are important. One way to do these studies is to perform both tests on each of a series of patients. This approach is costly, time-consuming, and potentially risky for the patient, and many patients may refuse to enroll. The alternative is to allocate patients at random to one of two putatively equivalent tests, perform the gold-standard test on all patients, and then measure clinical outcomes. This approach allows one to compare accuracy and effect on short-term clinical outcomes, such as short-term morbidity and mortality, reduction in diagnostic uncertainty, and altered choice of therapy and other technologies. The shortcoming of this approach is that many patients in the clinically relevant population will not be enrolled because their physicians do not refer them for the gold-standard test. The result will be biased measures of test performance and a relatively select study population, which compromises the generalizability of the outcome studies.

A randomized trial that compares the effect of two diagnostic tests on clinical outcomes poses another potential problem. In a trial, the test is done on all patients who are assigned to have it, rather than on those selected because the test was indicated. If there is a narrow range of pretest probabilities for which a test is likely to be useful, few patients who are randomly assigned to get the test will benefit from it. As a result, the number of patients needed to detect a clinically significant effect on outcomes may be very large, and there is a particularly high probability that a negative result will fail to detect a clinically significant true difference.

The randomized trial of the effect of a test on clinical outcomes has been underutilized and deserves greater attention from investigators. Much of this attention should be directed at the potential problems of study design and interpretation.

A PROPOSAL FOR MODEL-DRIVEN TECHNOLOGY ASSESSMENT

Most studies of diagnostic tests have measured little more than the false-negative rate and false-positive rate of a given test. This section describes an approach that we call "model-driven." In model-driven technology assessment, the data to be obtained are specified by a method

for making decisions (Sox 1987, Phelps and Mushlin 1988). We have used the threshold model for test-treatment selection to illustrate this discussion, but the particulars of the model are less important than the principle—that is, that one should obtain the data that will enable the clinician to identify the decision alternative that will be most useful to the patient.

A Technology Should Be Compared with a Competing Technology

The decision to adopt a new technology often means abandoning an older technology. In evaluating a technology of any kind, one should ask in what ways it is better than another (its marginal effectiveness). Many studies of diagnostic technology have not been comparative. There have been very few randomized trials comparing the effects of tests on outcomes. Too few studies have compared the accuracy of two tests by doing both of them on a series of patients.

The ideal study. The marginal effectiveness of a technology may be measured and its true value discovered only by comparison with another clinical method. A new technology may be compared with an old one, or two established technologies may be compared.

There are two types of studies of diagnostic tests. Ideally, a diagnostic test will always be compared to some other method for obtaining information, such as the patient's history and physical examination or another diagnostic test.

First the effects of two or more tests on clinical outcomes can be compared. The marginal effect of a new test may be discerned by a randomized trial in which the effect of the test on patient care outcomes is measured directly, rather than inferred from probabilistic and decision-analytic models. The potential limitations of this approach are discussed in the preceding section.

The performance characteristics of the tests can be compared. Comparing the frequency of false-negative and false-positive results in two or more tests provides the necessary data for a decision model that will help to indicate which test is preferred. Patients can be randomized to have one test or the other, or both tests can be done for each patient.

STUDIES SHOULD BE PLANNED BEFORE ENROLLING THE FIRST PATIENT

Most studies of diagnostic tests have retrospectively analyzed data that had been obtained for another purpose. Thus, they describe clinical experience rather than planned research. Typically, the index test has

been performed on many patients, but only a few have had the gold-standard test. The characteristics of the index test group are seldom compared with the characteristics of those who also undergo the gold-standard test. Other clinical data have been obtained irregularly. Most of the defects of past studies are attributable to their retrospective character.

The ideal study. A study should be planned in advance to assure adherence to a uniform data collection protocol. Bias in selecting patients and interpreting data can be reduced by planning. In a multicenter study, all participants follow the same data collection protocol.

ALL THE DATA THAT ARE NEEDED FOR CLINICAL DECISION MAKING SHOULD BE COLLECTED

Past studies have measured the accuracy of a test, but they have not collected all the data required to help physicians make decisions concerning individual patients. For instance, sequences of tests are not reported, although physicians must often choose between doing such a sequence or doing one test. Using Bayes' theorem to interpret the second test in a sequence usually requires assuming that the performance of the second test is conditionally independent of the results of the first. In some studies, two tests have been performed on a series of patients, and the operating characteristics of each test have been reported; seldom reported, however, is the frequency of each combination of results (both positive, Test A positive and Test B negative, Test A negative and Test B positive, both negative) in diseased and nondiseased patients.

Test results are not reported as continuous variables. The interpretation of a test usually depends on the extent of the abnormality. Thus, an orange-size lung mass is more likely to be malignant than a pea-size mass. To make use of this information in decisionmaking, the false-negative rate and false-positive rate for an orange-size mass should be reported separately from these rates for a pea-size mass. In most published studies, the results have been reported simply as "normal" or "abnormal." As discussed in Chapter 2, the operating characteristics of a test really reflect the criterion for calling a particular result "abnormal." Thus, optimal decisionmaking requires reporting the true-positive rate and false-positive rate of each of a series of definitions of an abnormal result.

The ideal study. The ideal study of technology is model-driven: the data to be obtained are those required by a model of the decisionmaking process. According to the principles of expected-value decisionmaking,

the clinician must know the pretest probability of disease and must be able to calculate the patient's expected utility for each of the decision alternatives: treat without testing, do Test A, do Test B, or do nothing. To provide the data needed for decision making, studies of diagnostic tests should:

Develop clinical prediction rules for estimating pretest probability. Clinical prediction rules estimate the probability of disease from the history and physical examination and other data (see Chapter 2). To develop a clinical prediction rule, one must obtain a complete problem-related data set on each patient and do a gold-standard test to define his or her true state. These data are easily obtained at little additional cost in a prospective study to measure the false-negative rate and false-positive rate of a test.

Measure the false-negative rate and false-positive rate of sequences of tests. In a study that compares several diagnostic tests, each test should be performed on each patient in the study, and the results of one test will be reported separately under each set of results for the other tests. For example, suppose that Test A and Test B are performed on all study patients. The false-negative rate and false-positive rate of Test B will be reported in patients who had a positive result on Test A and in patients who had a negative result on Test A.

Report the operating characteristics for several different results of the test. A study should enroll enough patients to report the accuracy of the test in subgroups of patients who show increasingly abnormal results. Results should be reported as receiver operating characteristic (ROC) curves. Tests can be compared by calculating the area under their ROC curves, although a more clinically useful comparison is the range of disease probability over which the test is preferred.

Provide a decision model for identifying the preferred option. The clinician can use the principles of expected-value decisionmaking to identify the decision that will maximize the patient's chances for a favorable outcome. The principles of expected-value decisionmaking, and of designing a decision model or decision tree, are described in Chapter 2. The decision tree requires one to estimate the probabilities at the chance nodes, which are usually obtained from published studies but could be obtained from analysis of insurance claims data (Barry et al. 1988). The

tree also requires a quantitative measure for each outcome. This measure could be life expectancy or a measure of patient preference, such as utility. Each study patient's utility for each outcome state will be measured using standard utility assessment techniques.

Consider treatment threshold probabilities. For many problems, the threshold model of decisionmaking can help the physician decide whether to treat, withhold treatment, or do a test or sequence of tests. The physician can use intuition to estimate an individual patient's treatment threshold or can use the analytic methods that were described in Chapter 2. The treatment threshold will vary from patient to patient because of their different outcome preferences and their different clinical characteristics. The distribution of treatment thresholds will provide an essential background for physicians as they estimate the individual patient's threshold. If the range of treatment thresholds is relatively narrow, one can make general recommendations for using diagnostic tests.

BIAS IN PATIENT SELECTION SHOULD BE AVOIDED

In past studies of diagnostic tests, the study population has differed significantly from the patients who undergo the test in the usual course of medical care. This defect of past studies is the most important and the most difficult to solve. Chapter 2 contains a description of the selective forces that lead to a biased spectrum of study patients. This defect leads to test measurements that lack external validity and could seriously mislead the clinician.

The ideal study. All patients who receive an established test in customary and usual practice should be included in the study population if possible. Exclusion and inclusion criteria, if they are needed, should be stated in the study protocol. The most troublesome selective factor is "workup bias." There are several ways to avoid this problem.

The best way is to avoid using a gold-standard test that is unpleasant, costly, and risky. For example, in evaluating the accuracy of rectal ultrasound for evaluating prostate nodules, one can use needle biopsy of the prostate as the gold standard. This procedure can be performed so easily that there is no barrier to referring patients.

Another way to avoid workup bias is to be sure that a positive index test is not used as a criterion for obtaining the gold-standard test. One way to assure compliance is to obtain the index test only in patients who have had the gold-standard test.

A third way to avoid workup bias is to use long-term follow-up as an ultimate measure of whether or not the patient had the disease. Thus, all patients who do not get an invasive gold-standard test for cancer would be evaluated periodically for the appearance of a cancer that was initially missed by the index test.

PATIENTS SHOULD BE OBSERVED FOR ADVERSE EFFECTS OF THE INDEX TEST

Most studies of diagnostic tests have not included any clinical outcome measures other than diagnosis. Ill effects have seldom been assessed, other than to note direct complications (death and disability from the procedure itself). Other ill effects—such as psychological dependence on test results (Sox et al. 1978), expensive workup of false-positive results, and mistakenly labelling the patient as diseased—have seldom been investigated.

The ideal study. All patients should be monitored to detect any delayed effects of the test. A prospective study can incorporate these important study endpoints at a small additional cost. A research assistant can perform clinical follow-up of each patient by administering a questionnaire and by reviewing the patient's medical record.

INTERPRETATION OF DATA SHOULD BE FREE OF BIAS

The index test and the gold-standard test should be interpreted independently to avoid having the results of one influence the interpretation of the other. In some published reports, each test has been interpreted independently, but the protocol for interpreting the index test and the gold-standard test is not usually described. One way to avoid biased interpretation is to have standardized, written criteria for classifying test results.

The ideal study. The gold-standard test and each test being evaluated are interpreted independently, according to standardized criteria. To achieve this goal will require the active cooperation of the clinicians who perform and interpret the test.

INTEROBSERVER DISAGREEMENT SHOULD BE MEASURED

Studies have often shown considerable disagreement among observers in labelling an image or tracing as abnormal (Koran 1975). Very few studies of diagnostic tests have included measures of interobserver disagreement.

The ideal study. At least two people should examine images or tracings and categorize the result according to prospectively defined criteria. These test result categories could be limited to normal and abnormal or could include several degrees of abnormality. The level of agreement should be characterized quantitatively.

THERE SHOULD BE ENOUGH PATIENTS TO REPORT THE RESULTS IN CLINICALLY USEFUL SUBGROUPS OF PATIENTS

Typical studies of diagnostic tests enroll fewer than 100 patients, far too few to evaluate the performance of a test in clinically important subsets of patients. One large clinical study has shown that the accuracy of a diagnostic test varies among clinically defined patient subgroups (Weiner et al. 1979). Patients who appear very sick often have extensive disease that a test can detect easily. Disease is often less extensive, and therefore less easily detected, in patients who do not appear ill. Applying results obtained in very sick patients may lead to incorrect interpretation of test results in other patients.

The ideal study. The study should enroll enough patients to measure test performance in subgroups of patients, and it should prospectively establish criteria for different categories of disease severity. The operating characteristic of the index test should be measured in these subgroups, as well as in the entire patient population.

SUMMARY

The chief importance of this chapter is that it sets out expectations for future studies of diagnostic tests. There are a few basic principles. *Do comparative studies:* a test can be compared with a competing test, either by randomly allocating patients to one test or the other or by performing both tests on all patients. *Do clinically relevant studies:* the investigators should gather all the data that are required to implement a model for making clinical decisions. *Avoid bias*: the study population should be all those who get the index test in the course of usual care.

REFERENCES

Abrams, H.L., and McNeil, B.J. Medical implications of computed tomography ("CAT" scanning). New England Journal of Medicine 298:261, 310-318, 1978.

Ambrose, J., Gooding, M.R., and Uttley, D. E.M.I. scan in the management of head injuries. Lancet 1:847-848, 1976.

Barry, M.J., Mulley, A.G., Fowler, F.J., and Wennberg, J.W. Watchful waiting vs. immediate transurethral resection for symptomatic prostatism. Journal of the American Medical Association 259:3010-3017, 1988.

CASS Principal Investigators. Coronary Artery Surgery Study (CASS): A randomized trial of coronary artery bypass surgery: Survival data. Circulation 68:939-950, 1983.

Fineberg, H.V., Bauman, R., and Sosman, M. Computerized cranial tomography: Effect on diagnostic and therapeutic plans. Journal of the American Medical Association 238:224-230, 1977.

Haughton, V.M. MR imaging of the spine. Radiology 166:297-301, 1988.

Inouye, S.K., and Sox, H.C. A comparison of computed tomography and standard tomography in neoplasms of the chest. Annals of Internal Medicine 105:906-924, 1986.

Kent, D.L., and Larson, E.B. Magnetic resonance imaging of the brain and the spine. Annals of Internal Medicine 108:402-423, 1988.

Koran, L.M. The reliability of clinical methods, data, and judgment. New England Journal of Medicine 293:642-646, 695-700, 1975.

Marton, K.I., Sox, H.C., Wasson, J.H., and Duisenberg, C.E. The clinical value of the upper gastrointestinal series . Archives of Internal Medicine 140:191-195, 1980.

Modic, M.T., Steinberg, P.M., Ross, J.S., Masaryk, T.J., and Carter, J.R. Degenerative disk disease: Assessment of changes in vertebral body marrow with MR imaging. Radiology 166(part I):193-199, 1988.

NIH Consensus Conference. Magnetic resonance imaging. Journal of the American Medical Association 259:2132-2138, 1988.

Phelps, C.E., and Mushlin, A.I. Focusing medical technology assessment using medical decision theory. Medical Decision Making 8:279-289, 1988.

Philbrick, J.T., Horwitz, R.I., and Feinstein, A.R. Methodologic problems of exercise testing for coronary artery disease: Groups, analysis, and bias. American Journal of Cardiology 46:807-812, 1980.

Philbrick, J.T., Horwitz, R.I., Feinstein, A.R., et al. The limited spectrum of patients studied in exercise test research: Analyzing the tip of the iceberg. Journal of the American Medical Association 248:2467-2470, 1982.

Ransohoff, D.F., and Feinstein, A.R. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. New England Journal of Medicine 299:926-930, 1978.

Sox, H.C. Centers for Excellence in Technology Assessment: A proposal for the national program for the study of health care technology. In Roe, W., Anderson M., Gong, J., and Strauss, M., eds., A Forward Plan for Medicare Coverage and Technology Assessment. Washington, D.C., Department of Health and Human Services, 1987.

Sox, H.C., Margulies, I., and Sox, C.H. Psychologically mediated effects of diagnostic tests. Annals of Internal Medicine 95:680-685, 1981.

Steinberg, E.P., Sisk, J.E., and Locke, K.E. X-ray CT and magnetic resonance imagers: Diffusion patterns and policy issues. New England Journal of Medicine 313:859-864, 1985.

Weiner, D.A., Ryan, T.J., McCabe, C.H., et al. Exercise stress testing: Correlations among history of angina, ST-segment response, and prevalence of coronary-artery disease in the Coronary Artery Surgery Study (CASS). New England Journal of Medicine 302:230-235, 1979.

4

Primary Assessment of Diagnostic Tests: Barriers to Implementation

In the two preceding chapters, we have presented a series of guidelines for conducting the ideal study of a diagnostic technology. The goal of this chapter is to examine the practical difficulties that are often encountered when the guidelines are applied to the design and execution of a typical protocol. We will also address briefly a number of methodological issues concerning the interpretation and reporting of the study data. Five specific stages of a primary technology assessment will be addressed: planning and protocol development, recruitment, implementation, interpretation, and reporting.[1]

PLANNING AND PROTOCOL DEVELOPMENT

The key to designing a useful technology assessment is that *studies should be model-driven*; the data that are gathered should be specified by a model of decisionmaking. We pointed out in Chapter 3 that the best way to obtain the information that will make the model usable (that is, the data needed by a physician to make a decision about the care of an individual patient) is to plan the study *before* enrolling the first patient. What are the elements of the planning and protocol development stage? What specific issues must be addressed and how can they be resolved? Weinstein (1985) has discussed many of these issues as they relate to

[1] Parts of this chapter are adapted from a paper published previously by one of the authors (Abrams 1987).

planning a trial of cost-effectiveness of diagnostic technology, and we draw on his work.

First, the planners must clearly delineate the objectives of the study. A number of critical questions must be asked:

- Which clinical condition should be investigated?
- Which patient population should be included in the study?
- Will the endpoint(s) be accuracy, outcome, or both?
- What type of study design will be used?
- Will the assessment also be an economic evaluation?
- Will the study assess efficacy or effectiveness?
- What is the appropriate comparison technology?
- How large a sample will be needed?
- When should the study be conducted?
- Is there institutional support for the study?

The answers to these questions will greatly influence the design of the protocol and the nature of the data to be gathered. We will therefore consider each of them in more detail.

Choosing a Clinical Condition

A diagnostic imaging technique has numerous potential applications. For example, it was estimated in 1984 that MRI examinations might be used in up to 250 diagnosis-related groups (DRGs) (Steinberg and Cohen 1984, Weinstein 1985). Defining the role of MRI for *each* of these categories would require many studies and a tremendous investment of time and resources. Recognizing that society may not be able to afford to assess every application of a diagnostic technology, we must establish priorities for technology assessment.

In choosing the clinical problem to be evaluated in a diagnostic technology assessment, policy-oriented investigators would use criteria such as the frequency of a condition, the cost of the technology, and the potential impact of the study result on clinical practice. Other factors that might influence the choice include the potential effect of the test on patient management and outcome and deficiencies in existing diagnostic methods (Figure 4.1) (Guyatt and Drummond 1985). Planners may use policy considerations to select a study problem that will have a significant societal impact; but they must also ask if the study is feasible.

The feasibility of the study depends on a number of variables, such as cost and the availability of a gold standard. (The costs of studies of

diagnostic technology are discussed in detail in Chapter 5.) Open-ended questions and poorly defined goals may limit feasibility. Assessing the efficacy of CT or MRI of "the liver" ignores the sharp distinctions among biliary obstruction, mass lesions, and diffuse hepatocellular disease. Each topic requires separate consideration. One prospective study of CT, ultrasound (US), and scintigraphy focused on the tests' ability to detect metastatic liver disease from several types of primary carcinoma. No difference was observed in the diagnostic capabilities of these technologies (Smith et al. 1982). Nevertheless, the results of a more recent study, restricted to patients with carcinoma of the breast or colon, suggest that differences do exist in the diagnostic yield of the three modalities when pathologically distinct lesions are analyzed separately. These differences may have been obscured in the first study because the clinical problem was too broadly defined (Alderson et al. 1983).

FIGURE 4.1 Factors that influence the choice of the clinical condition to be studied.

One possible way to set priorities for technology assessment is to use decision-analytic techniques for determining the value of perfect information. Suppose we are considering an assessment of the accuracy of a new test for patients with condition X. Let us assume that the new test provides perfect information, thereby resolving all uncertainty about the true state of the patient, and that we can determine the value of the

information in dollars. According to the model, if we find that the cost of *performing the test* is greater than we would be willing to pay for perfect information, using the new test to diagnose patients with condition X would not be worthwhile (Phelps and Mushlin 1988). The model uses a hypothetical test that is 100 percent accurate to maximize its potential value. If the information from an *ideal* test is not worth the test cost, we can expect, all other things being equal, that the information from a *real, imperfect* test would be worth even less. It follows that we would not want to expend resources to *evaluate* the test's performance in this clinical situation. This methodology provides a powerful tool for determining beforehand whether we should expend the resources necessary to evaluate a particular use of a technology.

Patient Population

The study population must be well defined. When certain subsets of eligible patients are excluded because of other, coexisting disease, a physician may be unable to generalize the study result to the whole spectrum of patients encountered in clinical practice. (See Chapters 2 and 3 for a more thorough discussion of sources of bias in selecting patients and their negative impact on studies of diagnostic technology.) Choosing a specific clinical problem for a study of diagnostic technology defines the diagnostic category of patients who may participate in the study. Within this category, the population should include a representative spectrum of patients.

Inclusion and exclusion criteria are needed to define the boundaries of the study population. They must be explicit and they must be applied consistently. In the University Group Diabetes Study, these criteria were not applied uniformly, leading to the admission of a number of ineligible patients and the exclusion of some patients who were eligible. These errors compromised the generalizability of the study conclusion and wasted resources (Feinstein 1971).

Wide variance in test performance (that is, accuracy) within the study population may obscure differences in the performance of two tests. Investigators may need to specify and analyze the results of a test in subgroups of patients for whom they suspect the test will perform differently. For example, the sensitivity and specificity of the exercise thallium treadmill, used to diagnose coronary artery disease, are different in groups of patients segregated according to the severity of their chest pain (Weiner et al. 1979). Although there may be no significant difference in test

performance when the population is considered as a whole, there may be differences when subgroups within the population are compared.

Endpoints and Study Design

The endpoint of a diagnostic test assessment will determine how the results will be used; it is, therefore, critical. Fineberg has proposed the following hierarchy for the evaluation of diagnostic tests: technical capability, diagnostic accuracy, therapeutic impact, and impact on patient outcome (Fineberg et al. 1977). Early reports of excellent technical capability are often the basis for the later studies of diagnostic accuracy and clinical value (impact on therapy and patient outcome). The critical question in the planning and protocol development stage is: Will the study attempt to measure diagnostic accuracy (that is, sensitivity and specificity), the impact of the test on clinical outcome, or both? (Note that we define the outcome of a diagnostic test as any change in the posttest process. It should not be considered synonymous with the terms morbidity and mortality.)

ACCURACY

Studies of diagnostic accuracy use a "gold standard" to verify the presence or absence of disease. A potential difficulty in a study of accuracy occurs when there is no accepted "gold standard"; it may not be clear which of the available reference standards should be used (Schwartz 1986). All reference standards are imperfect. The coronary angiogram is used as the gold standard in studies of diagnostic tests for coronary artery disease, such as the stress electrocardiogram. Yet, pathologic examination of tissue from patients who have had an angiogram demonstrates that the radiologic procedure underestimates the severity of disease (Abrams 1982). Physicians must interpret the results of studies of accuracy in this context. Perfect or not, in practice the appropriate gold standard will be the test or procedure that physicians use to define the true state of patients with a particular disease.

OUTCOME

Because the purpose of diagnostic technology is to provide information that will improve patient outcome, patient outcome is an important endpoint in technology assessment. Making inferences from data on outcome

may be more difficult than interpreting data from studies of diagnostic accuracy. When long-term measures of outcome are used, the technology may be obsolete before the study is completed. Furthermore, long-term outcome may be an unrealistic criterion, "because the impact of diagnostic technologies generally is subordinated to that of other factors, such as the nature of the disease process itself, patient compliance, the efficacy of treatment, etc." (McNeil 1979, p. 37).

Improvement in long-term outcome may not be the most important effect of a test. If intervening variables act to obscure differences in the long-term effects of two technologies, perhaps the differences are not really important. Investigators must keep in mind that two patients with identical long-term outcomes may have experienced very different posttest processes.

A variety of intermediate variables may be important indicators of the effects of a test. Furthermore, these variables may be more practical to evaluate than long-term effects. For example, a study could measure the ability of a diagnostic technology to obviate the need for further invasive diagnostic procedures. In patients with lung cancer, thoracotomy could be avoided if a test could accurately predict the presence of mediastinal metastases. The test will not improve the five-year survival of such patients, but, avoiding an unnecessary thoracotomy would be a major benefit (McNeil et al. 1978) and would therefore represent an improvement in the posttest process. Outcome studies must track intermediate outcomes and patients' attitudes toward those outcomes.

COMBINED STUDIES: ACCURACY AND OUTCOME

The alternative combinations of study design (randomized or nonrandomized) and endpoint (accuracy and/or outcome) are depicted in Figure 4.2. The design of a technology assessment influences the feasibility of conducting each type of study. In a randomized design, each patient undergoes only one of the study tests; in a nonrandomized design, each patient would undergo all of the study tests, although randomization may be used to assign a patient to a particular sequence of tests. The advantages and disadvantages of a randomized design have already been discussed in Chapter 3.

The following example illustrates that a study design may not be compatible with the endpoint(s) selected for evaluation. In an ideal study to compare the accuracy of two tests, each patient would have both examinations. Guyatt and Drummond have suggested that investigators

use this approach to assess both the accuracy and the impact on outcome of two relatively noninvasive imaging modalities, such as MRI and CT, in a single study. To compare the effects of the tests on outcome for the same patient in whom accuracy is determined, however, the result of one of two tests would have to be withheld from the patient's physician (Guyatt and Drummond 1985).

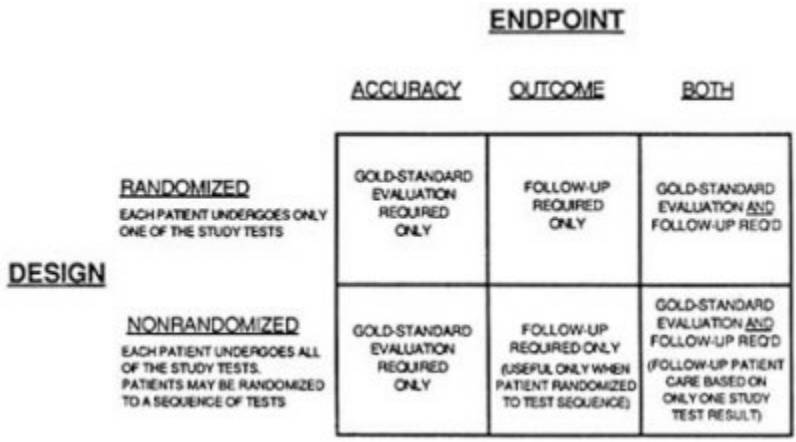

FIGURE 4.2 Alternative combinations of endpoint and study design.

The design of this study poses ethical problems because patients will undergo a diagnostic examination that cannot affect their care (Weinstein 1985). Patients and physicians alike may be reluctant to participate. In Chapter 3, we suggest that a randomized design may be preferred to a nonrandomized design for assessing outcome. Planners could also shift their focus from long-term to short-term outcomes.

SHORT-TERM OUTCOMES: A SYNTHETIC APPROACH

The synthetic approach is a method for assessing short-term outcomes, such as the impact of a diagnostic test on the management of the patient. It involves obtaining detailed information from physicians about their pretest treatment strategies and comparing them to the posttest management of the patient (Guyatt et al. 1986). In the example above, each physician would write down a plan for managing the patient before knowing the CT and MRI results. Using a randomized scheme, the result of one of the two tests would be given to each physician, who would then

formulate and record a treatment plan based on the test result. Next, the result of the other test would be revealed, and the patient's care would ultimately be based on all available information. A test has had an impact if the physician's plans changed because of the test result.

Economic Analysis

In the current era of cost containment and limited resources, a consideration of cost will be an important study endpoint. First, the investigators planning a technology assessment must decide which type of analysis to use (for example, net resource costs, cost-effectiveness analysis, or cost-benefit analysis). Since cost-effectiveness analysis is comparative and does not require that health outcomes be valued in monetary terms, it is the type of analysis used most frequently. Second, investigators must choose an appropriate perspective for the analysis, because this will greatly influence which costs and effects are included. The societal perspective is the broadest, and it is adopted when the results of the cost-effectiveness analysis are needed to guide government decisions about how to allocate resources. Third, the investigators must recognize the degree to which additional time and personnel will be needed when an economic evaluation is included in the study design. (For a complete discussion of these issues see Weinstein and Stason 1977; OTA 1980a,b; or OTA 1981.)

Efficacy Versus Effectiveness

The conditions of the study can imitate real life or they can be idealized. The choice between *efficacy*, the performance of the test under ideal conditions, and *effectiveness*, its performance under ordinary conditions of clinical practice, will determine the type of question the study can answer. Consider a study designed to assess the diagnostic accuracy of barium enema (BE) in detecting colonic polyps (also see Figure 4.3).

A study of *effectiveness* would enroll all patients who are referred for BE in clinical practice. Patients would be given the usual pretest instructions and, although some would be less than optimally prepared, all would undergo the examination. This would be performed under usual conditions, by the individuals who normally perform it—radiology staff or house staff. It would be interpreted by clinicians at varying levels of skill who would be provided with whatever clinical information is generally

available at the time. There would not be a protocol for subsequent patient care.

	EFFICACY	EFFECTIVENESS
PATIENT POPULATION	MORE HOMOGENEOUS; SCREENED FOR COEXISTING ILLNESSES/ COMPLIANCE	HETEROGENEOUS; INCLUDES ALL PTS WHO USUALLY HAVE PROCEDURE
PROCEDURES	STANDARDIZED	MORE FLEXIBLE
TESTING CONDITIONS	IDEAL	CONDITIONS OF EVERYDAY PRACTICE
TEST INTERPRETATION	BLINDED TO CLINICAL DATA	USING OTHER CLINICAL DATA
TYPE OF OUTCOME DATA	OBJECTIVE, "HARD" EVENTS, E.G., DEATH	MORE SUBJECTIVE, "SOFT" EVENTS, E.G., IMPROVED QUALITY OF LIFE

FIGURE 4.3 Differing requirements: studies of efficacy vs. studies of effectiveness.

A study of *efficacy* should assess the potential benefit of the technology when applied to a specific clinical problem in a defined population under ideal conditions. The protocol would be designed to maximize the chance that the true accuracy of the test will be demonstrated by reducing sources of variability. Thus, a study of efficacy would: (1) enroll a more select group of patients; (2) ensure that all patients were adequately and consistently prepared prior to the exam; (3) use only state-of-the-art equipment; (4) employ the most skilled clinicians to perform and interpret the test; and (5) make sure that interpreters were blinded to other clinical information. It would also standardize aftercare.

The individuals who develop the protocol may disagree about which type of assessment is most appropriate, making the choice a difficult one. Feinstein (1983) has suggested a useful way to conceptualize the two approaches—the "fastidious" and the "pragmatic"—to design.

The fastidious approach. Fastidious designers might include the bio

statistician or the scientist who developed the technology. This group would argue that a study of efficacy is the only way to determine the "true" value of the technology. For example, an efficacy design will increase the chances of arriving at an unequivocal answer to the study question by standardizing procedures and removing many of the sources of variability that characterize clinical practice. If such a study concluded that a test was not efficacious, there would be no need to perform further evaluations.

The pragmatic approach. This approach would be adopted by the practicing clinician. The "clean" results of an efficacy assessment may have little value for the physician whose patients will receive their tests under "usual" rather than "ideal" conditions. The pragmatist would argue that only studies of effectiveness, which attempt to mimic clinical reality, provide the information physicians need to make decisions about individual patients.

Resolution of the conflict between the fastidious and pragmatic approaches may involve combining features of both. In any case, the protocol as it is actually carried out may end up as a hybrid, because protocols that have been designed to assess efficacy will often encounter real-world obstacles that make the ideal arrangement impossible. These problems will be covered in detail in the section of this chapter that considers implementation.

Comparative Assessment

In Chapter 3 we emphasized that technology assessments must be *comparative* if they are to provide useful data to the practicing physician. For example, the physician may need answers to either or both of the following questions: (1) When used *instead of* existing tests, does the new test have a greater impact on the outcome of the patient? (2) When used *in combination with* existing methods, does the new test add information that will improve the outcome of the patient?

These questions suggest two comparative designs (see Figure 4.4). In one design, the study would detect any positive impact when the new test is substituted for the old. Patients would be randomized to either the new technology or the existing technology. In the other design, the study would evaluate the impact of a technology when it is used as an addition. Patients would be randomized to undergo either the old test *and* the new test in sequence or the old test alone. Both designs could be used to compare the diagnostic accuracy of the tests (or combination of tests) and their impact on the outcome of disease. (A nonrandomized design, in

which all patients undergo all tests, would also be appropriate for a study of accuracy; refer to the section, "Endpoints and Study Design," pp. 77-80.)

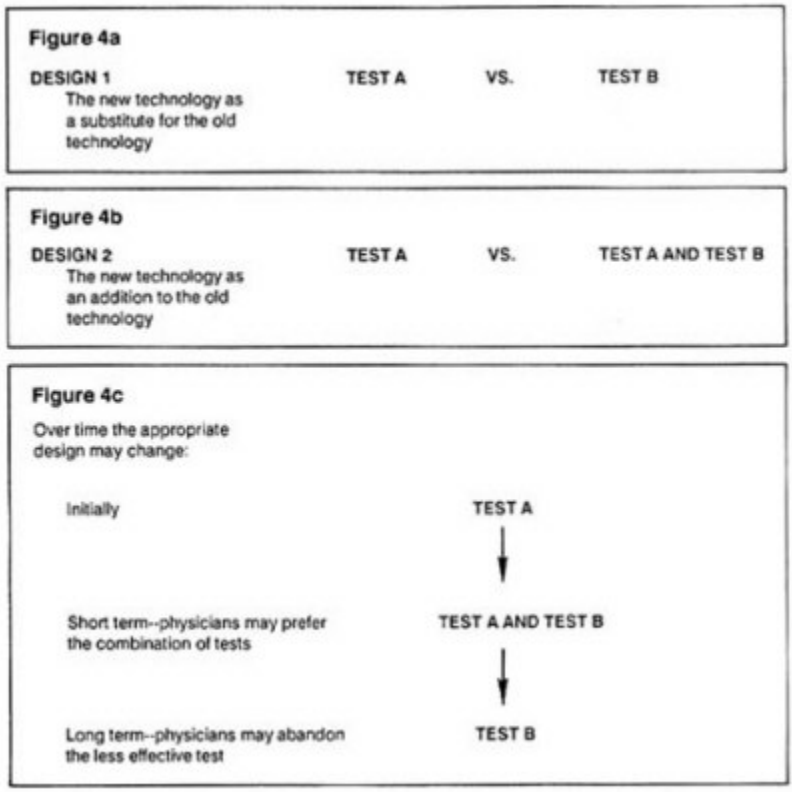

FIGURE 4.4 Designs for comparative assessment.

Which of these designs is more appropriate when comparing an existing technology with a new one? Despite promising reports, physicians may be hesitant, in the short term, to change over completely to a new technology. An additive design would answer their questions about using the new test as part of a sequence. Nevertheless, one goal of technology assessment is to foster appropriate changes in practice habits and to discourage the use of additional tests whenever they will not have an impact. In the long term, if we want to encourage physicians to abandon ineffective tests in favor of more effective ones, we will need to evaluate the technology's substitutive value as well (Weinstein 1985).

Another problem that must be addressed in the design stage is the high level of accuracy of existing diagnostic methods. As diagnostic methods improve, the measurement of small differences in accuracy requires large sample populations. In a study of 279 patients for pancreatic disease, CT had a sensitivity of 0.87 and sonography had a sensitivity of 0.69 in detecting an abnormal pancreas (a difference of 0.18) (Hessel et al. 1982). To show that another technique (such as MRI) is superior to CT, a far larger sample would be required, because the maximal attainable difference in sensitivity is 0.10 or less.

Before embarking on a comparative study, investigators should ask whether such small differences in sensitivity are clinically significant. Here, it may sometimes be helpful to use the threshold model described in Chapter 2. To accomplish this, we must choose a specific clinical problem and estimate the pretest probability of disease. We must also calculate a treatment threshold probability. This step is best accomplished by decision-analytic modeling. When the posttest probability of disease following a negative result from the old test (with sensitivity X) exceeds the threshold, we treat the patient. Let us assume that the new test has a maximum sensitivity of X + 0.10. If the probability of disease after a negative result from the new test remains above the threshold, our treatment strategy will not change. Thus, the difference in sensitivity has no practical implications for patient care (Sox 1986).

A full analysis of the impact of a test on decisionmaking requires knowledge of the distribution of pretest probabilities in the study population and knowledge of patient utilities. These additional data requirements are substantial. Of course, even if the differences in accuracy are not worth assessing, we may want to evaluate other features of a new test, such as increased safety or decreased cost.

Sample Size

As part of the planning stage, investigators must calculate the sample size required to ensure adequate power for the study. A review of 71 "negative" clinical trials found that 50 of the trials had a greater than 10 percent chance of missing a true 50 percent therapeutic improvement because of small sample sizes (Freiman et al. 1978).

A study should avoid two errors of inference. *Type I error* (α-type error) occurs if we reject the null hypothesis, H_0, when it is true. For example, we may conclude that there is a difference in the accuracy of CT and MRI when no difference exists. (The null hypothesis refers to the

basic hypothesis being tested—often one of no difference between the two entities being compared.) *Type II error* (β -type error) occurs if we accept the null hypothesis, H_0, when it is false. For example, we may conclude that there is no difference in the accuracy of CT and MRI when a difference exists.

The *power* of a study is equal to $1 - \beta$; it is the probability that the null hypothesis (that is, MRI is no better than CT) will be rejected when the alternative hypothesis (MRI is better than CT) is true (Brown and Hollander 1977). Viewed from another perspective, as we increase the power of the study, we decrease β, the probability of missing a true difference between the effects of two tests. Increasing the sample size is the primary way to increase the power of the study without increasing α, the risk of concluding that a difference exists when it does not. Investigators must consider how large a population of patients they will need to screen to achieve the appropriate sample size.

Many of the decisions during the planning stage of the trial will affect the size of these two populations. The sample size calculation depends on which outcome variables are selected and the magnitude of the difference in accuracy or outcome that we wish to detect. The factors that directly influence the *size of the sample* include heterogeneity of the study population and the degree of accuracy of existing diagnostic methods. Other factors influence the *size of the screened population*: the frequency of the condition under study, the breadth of the study focus, and the likelihood of patient withdrawal. If these factors are ignored, the result may be a gross underestimation of the number of patients needed and a high probability of type II error. *The problem of a large sample size can be overcome by using a multi-institutional cooperative design rather than attempting to conduct the study at a single center*. We present a proposal for this type of study and discuss the advantages and disadvantages of the multicenter design in Chapter 6.

Timing the Assessment

Selecting a specific clinical problem and determining the appropriate comparisons for a technology assessment are important and challenging aspects of the planning process. Perhaps the greatest challenge in this stage of a trial, however, is to determine *when to conduct the study* (Kent and Larson 1988). For an assessment to have an impact on the use of a new technology, some would argue that the results must be available before clinicians have made subjective judgments about the value of the

technology and widespread diffusion has occurred (Fineberg and Hiatt 1979). Studies must be initiated (and completed) as early in the "lifetime" of the technology as possible. Very early assessment may be difficult, if not undesirable, however, because of the inherently unstable nature of new technology (Alperovitch 1983).

The rapid pace of technological change affects diagnostic techniques as it does other types of technology. For example, a diagnostic method, especially one of the complexity of the MR scanner, is rarely introduced into practice in its most effective form. Rather, the technology continues to develop and improvements are made based on information derived from its early use in practice. Changes may include new configurations of the hardware and improved techniques for using it. As physicians gain experience with the method, their interpretive skills increase (Sheedy et al. 1977). Thus, a study conducted too early in the lifetime of a technology may fail to reflect its true potential. It may also be considered unethical to expose patients to an "unproven" technology, especially when insufficient time has elapsed to allow for the effects of the learning curve.

The decision about which "version" of the technology to assess is important. For example, "MRI is not a homogeneous diagnostic test, but offers a range of related, but different, diagnostic tests" (Weinstein 1985, p. 570). As Weinstein points out, such technical flexibility creates some difficult decisions for trial planners: should a study "freeze" the technology and specify standard hardware and techniques, or should a study systematically compare the efficacy of alternative hardware configurations? In the first case, there is the risk that the chosen configuration will become obsolete while the study is still in progress. In the second case, by the time the study is complete, the diffusion of MRI might not be an issue anymore.

Using technology assessment as a means of controlling the diffusion of new technologies may not be practical. Some have argued that studies should not be conducted until the technology has stabilized, at least with respect to certain clinical conditions. A randomized design would not be appropriate for this type of study, because, as Weinstein points out, stabilization often occurs just as physicians are beginning to consider it unethical to withhold the technology from patients they believe will benefit from it (Weinstein 1985). A nonrandomized design could be used, however, since each patient would undergo the new test. It would then become possible to predict the range of patients in whom the mature

technology would be useful, and the results could be used to influence reimbursement decisions.

Obtaining Institutional Support

Another important step that must be taken in the early stage of a trial is to enlist institutional support. In the National Cooperative Gallstone Study (NCGS), a randomized, controlled trial of chenodiol for the dissolution of gallstones, two study centers had to be deleted because they were unable to meet the required rate of randomization. The source of the problem was a lack of institutional support. When the study coordinators reviewed the applications of institutions seeking to replace the deleted centers, they added an evaluation of the level of administrative and departmental support at each of the centers to the review process (Marks et al. 1984).

The investigator in a study of diagnostic technology depends on the support of each department or group that will be involved in carrying out the study. The most useful approach is to enlist the direct participation of at least one interested, committed, and sophisticated member of the department or group. The time to do this is shortly after the study is conceived, so that these individuals can participate in developing the research design.

Summary: Planning and Protocol Development

Planning and developing a protocol for a primary technology assessment is a time-consuming activity. Protocol development for the Prospective Investigation of Pulmonary Embolic Diagnosis (PIOPED) study of diagnostic tests for pulmonary embolism took 15 months (Vreim 1988). Without institutional support, the project cannot succeed. Investigators will be confronted with a multitude of decisions, and conflicting views on study design may require compromises between the more practical approach (effectiveness) and the ideal approach (efficacy). Technological change will strongly influence the timing of the study.

In general, the most useful results will be achieved with a focused study in a patient population that is as representative of the clinically relevant population as possible. Accuracy is an important endpoint, but the study should also evaluate intermediate outcomes that are important to patients. The study should also include a cost-effectiveness analysis. The

nature of the tests being compared influences the decision between a randomized and a nonrandomized design and the decision between assessing the technology as an addition to existing methods or as a substitute for them. The choice between a randomized or nonrandomized design also depends on the whether the study evaluates accuracy, outcome, or both and whether the technology is mature or just emerging. All these decisions will influence the size of the sample and, therefore, the magnitude of the effort that will be needed to recruit patients for the study.

RECRUITMENT

The experience of many investigators conducting randomized controlled trials (RCTs) of drugs or other therapeutic interventions has borne out Muench's Third Law: "The number of patients promised for a clinical trial must be divided by a factor of at least 10" (Prout 1979, p. 695). For the study of a diagnostic technology, one might expect recruitment to be at least as great a problem as it has been for the trials of therapeutic regimens.

There are three steps leading to patient enrollment. First, a patient must be referred for diagnostic evaluation by the technique under study. Next, eligibility must be established according to the criteria set forth in the protocol. Finally, the patient must give informed consent to participate.

Referring Physicians

The success of the first step, referral, depends on the cooperation of the individuals who usually refer patients for imaging studies: the internist, surgeon, pediatrician, or obstetrician-gynecologist who requires either a solution to a diagnostic problem or the confirmation of a presumptive diagnosis. Unfortunately, these individuals cannot necessarily be counted on to provide the requisite number of cases (Croke 1979, Marks et al. 1984).

In past therapeutic RCTs, there have been a variety of reasons for low rates of referral. Investigators in the Coronary Drug Project found that physicians did not appear to have problems with the concept of the trial but seldom took the initiative to refer. When their patients were identified by records or self-referral, these same physicians were usually very supportive (Schoenberger 1979).

By contrast, in the National Surgical Adjuvant Project for Bowel and Breast Cancer trial to compare segmental mastectomy with total mastec

tomy, many physicians had ethical and other problems with the trial. A survey of surgeons participating in the trial revealed the following concerns:

- There would be negative effects of an RCT on the doctor-patient relationship.
- They were uncomfortable with admitting uncertainty about which treatment was best.
- They felt conflict between the role of clinician (doing what is best for the individual patient) and the role of scientist (adhering to a protocol with randomization).
- They were uncomfortable with the requirement of informed consent.

A significant number of respondents described the process of obtaining informed consent as an "arduous task." In addition, a number of the surgeons already had strong convictions about which treatment was superior (Taylor et al. 1984).

Through the few prospective studies of diagnostic technology that have been performed, the magnitude of the referral problem has become clear. Consider two examples:

- In a comparative study to assess the value of using MRI instead of CT to detect intracranial mass lesions, patients may be randomized to receive one of the two technologies. Because preliminary studies have suggested that MRI has a high level of diagnostic accuracy and the method is now available at many institutions, a physician may feel that it would be unethical to deny the patient an MRI scan.
- Similarly, in the ideal study designed to compare the accuracy of two diagnostic imaging methods in detecting metastatic disease to the liver, each patient would undergo both of the examinations. How do we convince the referring physician to allow the patient to undergo the second test when the physician is perfectly satisfied with the first examination, which shows the presence or absence of metastases?

Physicians may also be reluctant to refer patients for studies that involve the use of an invasive gold standard to verify the presence or absence of disease. They may feel that the study question lacks clinical relevance or that it will be irrelevant by the time the trial is completed. Furthermore, physicians may be hesitant to take on the additional, often involved, paperwork associated with a trial. They may perceive the trial

as an interference in the care of their patients or be concerned that patients referred to a large center for a trial will not return (Ferguson 1988). Clearly, before the investigation begins, the referring physician must feel comfortable with the role of the scientific investigator and agree to participate in the study.

Some physicians have already arrived at a subjective judgment of which of two technologies is better. These physicians may not refer patients to the study for fear that they might be randomized to the other technology. Referral bias can result in the exclusion of important subsets of patients, limiting the external validity of the study and providing unreliable estimates of sensitivity and specificity. A possible mechanism to avoid referral bias might be to assess each physician's preconceptions about the technologies under study through a private questionnaire. These responses could be used to classify patients into subgroups that could be examined later for particular trends.

Patients and Informed Consent

The public approves of research in principle—but not necessarily in practice. When a sample of patients and the general public were surveyed concerning their attitudes about clinical trials, most respondents (71 percent) believed that patients should serve as research subjects and cited the potential benefit to others and the opportunity to increase scientific knowledge as major reasons. When they were queried indirectly about their own willingness to participate in a trial, however, "a more self-concerned, less altruistic standard seemed to prevail" (Cassileth et al. 1982).

The first obstacle to patient participation is randomization. Patients, like physicians, may be uncomfortable with the notion that the method used to diagnose their illness will be chosen randomly. They may not perceive the two arms of the trial as comparable, and randomization does not allow them to express their preferences (Angell 1984). Cassileth's survey also found that many patients believe that "doctors know privately which one of the investigated treatments is best." If this holds true for diagnostic techniques, patients might prefer to have the physician's recommendation rather than risk enrollment in a trial. Many may wish to have the "latest" or "state-of-the-art" procedure, even when its superiority over existing methods has not been rigorously demonstrated.

Patients may also be concerned that, as trial participants, they will be treated by scientific investigators bound by a protocol rather than by a

clinician attentive primarily to their personal needs. A survey of *participants* in the Aspirin Myocardial Infarction Study (AMIS) and the Beta-Blocker Heart Attack Trial (BHAT) cited better quality of care and the availability of second opinions as one of the benefits of participation (Mattson et al. 1985). In the survey of attitudes of *potential participants* discussed above, however, many respondents felt that patients receiving physician-recommended treatment received better care than clinical trial participants (Cassileth et al. 1982). Thus, patients may be worried about the quality of care in a trial.

The requirement for informed consent is another obstacle to enrollment. The *content* of informed consent raises issues that may discourage both referral and enrollment. Moreover, the *process* of informed consent may itself be a barrier (Lidz et al. 1983). Patients may perceive the consent documents as "legalistic, undesirable intrusions into the physician-patient relationship" (Cassileth et al. 1980). They may prefer that the decision they make explicitly when they give consent be made for them implicitly by their physician.

How do we ensure the patient's consent and cooperation with the study? The full involvement of the personal physician is paramount. Without it, truly informed consent is not possible, nor is a clear understanding of the study's potential benefit. Even with the physician's involvement, a research assistant or the investigator must spend time with the patient to explain the protocol and its potential benefits and risks. In particular, they must discuss the randomization of test order; otherwise, the patient will ask why the tests are performed in an apparently strange sequence. Finally, patients may wish to avoid a number of features of a well-designed protocol, such as multiple examinations, additional tests, and return clinic visits for follow-up (Mattson et al. 1985). Patients may feel that the inconvenience, discomfort, and additional time involved in participating is too great.

Summary: Recruitment

A primary technology assessment is a form of human experimentation, with all its associated ethical challenges. Lack of interest, informed consent, randomization, and prior judgments of technological superiority may lead to low rates of referral or biased referral and patients who refuse to participate. Physicians may feel a conflict between their responsibility to their patients and their commitment to the trial. These problems can be

reduced if investigators take the time to explain the study and its goals to the referring physicians and to patients. The study should be planned with a large margin for error in forecasts of referral rates.

IMPLEMENTATION

Implementing a study protocol in the clinical setting presents a number of formidable obstacles. Some occur because the study must be conducted within the context of a health care delivery system that is not specifically set up to accommodate the often artificial circumstances of an experimental protocol. Other problems arise because both patients and physicians may have negative feelings toward this type of research. Without the cooperation of all parties, even the most well-designed protocol is likely to fail. The following section will deal specifically with the logistics of technology assessment.

Logistics of Randomization

Whether the patient is to undergo only one of two examinations or is to have all of several tests under study, some form of randomization is necessary to avoid bias. For example, it would be desirable to randomize the order of two tests in a study in which both tests are performed on each patient. Randomization employs a chance mechanism to assign patients to an arm of the trial—for example, ultrasound or CT. The process of random assignment must be carefully specified in the protocol; it may, for example, involve opening a sequentially numbered, sealed envelope with a code designating the diagnostic procedure to be performed. The process must be followed for each study patient. Some data must be obtained on all patients who withdraw or cannot be randomized so that the population that did not participate can be adequately characterized. Patients can be randomized at the time of enrollment or at the time the examinations are scheduled.

In a busy institution, scheduling a subset of patients according to requirements that differ from the norm may further complicate an already complex task. The cooperation of the hospital staff is needed. For example, in the radiology department of a large teaching hospital, 400 to 700 examinations are performed each day. Someone in the department must take the responsibility for identifying study patients and reviewing all requests for particular examinations. The technology being studied is used in a host of different clinical conditions, and a busy receptionist may

simply schedule the examination rather than take the time to indicate to the physician-investigator that this patient is one for whom the order of exams is to be determined by chance. Two prospective studies that compared multiple imaging techniques encountered problems with scheduling that interfered with the ideal random arrangement of examinations and prevented all of the studies from being performed in all patients (McNeil et al. 1981, Alderson et al. 1983).

Randomizing patients at the time of enrollment is preferable to doing so at the time of scheduling, because fewer people will have this key responsibility. Thus, the research assistant (RA) who responds to referrals could see the patient well before the scheduled tests, obtain consent, administer prerandomization questionnaires, and randomize the patient. The RA would then schedule the patient so that the order of the examinations is as specified by the protocol.

Obstacles to Data Gathering

Gathering data, the most important part of a technology assessment, occurs after the patient has been enrolled and randomized. The quality of a study is greatly influenced by the quality of the data collected. Yet, problems in this area frequently jeopardize the validity of the study (Feinstein 1971). This segment of the study can be divided into three phases: collecting "input" data, performing the study tests, and follow-up studies. Each of the phases involves a number of individuals, including patients, referring physicians, study physicians, technicians, nurses and RAs. Similar factors affect the outcome of each phase: the level of interest of the study personnel, their perception of the relevance of the study, and their comprehension of the protocol requirements. The following section will examine problems that may be encountered in the process of data gathering.

COLLECTING INPUT DATA

"Input" data are obtained before the patient undergoes the study test(s). There are several reasons for gathering such data. First, when the study's aim is to define the marginal increment of information that a new technology adds in a particular clinical situation, input data may consist of the history, physical examination, and laboratory values that constituted the basis for requesting the examination. Second, background information on each patient is extremely important because the generalizability of a study

result depends on a full characterization of the study population. Third, the protocol should include the collection of the clinical data needed to classify patients into subgroups in which test performance might be better or worse than in the total population. Fourth, data may be needed to create clinical prediction rules for estimating the pretest probability of disease. Increasing the *quantity* of data collected, however, also increases the likelihood of decreasing its *quality*.

Past trials, whether of therapeutic or of diagnostic technology, have had two main problems with data collection: missing data and inaccurate data. It is important to collect all pertinent and relevant data initially, because data specified but not collected during a prospective study are usually difficult to acquire retrospectively. The forms for recording the data may be a source of trouble if they require elaborate detail or are complicated to fill out. Inadequate data collection was a problem in the University Group Diabetes Program trial of the effects of oral hypoglycemics on the development of subsequent vascular complication of diabetes mellitus (Feinstein 1971). In the NCGS trial, laboratory reports, radiology forms, and patient history forms were incomplete or incorrect as much as 50 percent of the time; correcting these deficiencies was extremely difficult (Marks et al. 1984). One consequence of missing data is to reduce the number of patients that can be analyzed, which jeopardizes the statistical power of the study. The study population may be inadequately characterized, which compromises the generalizability of the study conclusions.

Often, studies must rely on busy physicians to provide key clinical data. If these individuals have been excluded from the planning and design stages of the trial, they may feel that they are performing additional chores in order to satisfy the curiosity and advance the interests of a third party outside the patient-physician relationship. They may also view the research as "exploitative," and on this foundation resistance is built. The participation of individuals with negative attitudes may be damaging to a study because they will not be concerned about the quality of the data (Hopwood et al. 1980).

Before the start of the study, the investigator should meet with the individuals involved in data collection and monitoring: referring physicians, house staff, nurses, and RAs. The investigator should explain the goals and design of the study, indicate its value and the information required, and answer any questions. Once the physicians and others understand that all parties will gain by the acquisition of more accurate data, chances for effective cooperation are greatly enhanced.

Perhaps the best way to avoid the problem of incomplete or inaccurate data is to bypass physicians altogether by employing an RA to gather all input data. If physical examination data are needed, a nurse practitioner can fill the role of RA. Having one person serve as RA means greater standardization of data collection methods.

PERFORMING THE STUDY TESTS

Success in this phase of data gathering is influenced by many of the same factors that affect the quality of the input data. Protocol compliance and exam quality are the two most important concerns. Complying with the protocol includes not only following the specifications for performance of the diagnostic procedures but also performing all procedures, including the gold standard, in every patient who is supposed to have them.

A detailed, cookbook type of protocol does not assure compliance and may be an obstacle to collecting error-free data. The individuals performing the tests may simply fail to read the complete protocol, may fail to understand the procedures, or may forget the protocol. To ensure that each patient enrolled in the study has a standardized examination, the protocol may specify that a test be performed in a manner that differs from the manner in which it is usually performed. But the individual performing the test may decide not to follow these instructions if the change in procedure requires a great deal of additional time or if the change is perceived as "bad medicine."

These issues can best be illustrated with examples.

- In a multicenter study comparing CT and radionuclide (RN) studies, the protocol carefully specified sodium pertechnetate for the RN studies. One institution, however, "used a mercury isotope and a type of imaging instrument unique to it and virtually unknown to other nuclear radiologists." Another institution participating in the same study obtained fewer than the specified number of images when performing the CT scan (McNeil 1979, p. 34).
- A therapeutic trial used a test of visual acuity to assess outcome. Although the protocol specified a patient-to-chart distance of 20 feet, trial monitors found that the participating clinics used different distances (Ferris and Ederer 1979).

In these two studies, the failure to comply with the protocol could have

reduced the number of cases that could be used in the final analysis, prolonging the recruitment effort and increasing the cost and time required to complete the study. Furthermore, combining data obtained with different procedures may compromise the validity and generalizability of the study conclusions.

Summarized below are some other problems encountered in a prospective study evaluating the diagnostic value of ventilation-perfusion scanning in patients with suspected pulmonary embolism (Hull et al. 1985):

- The index test, ventilation scanning, could not be performed in 20 of the patients because of the lack of availability of ^{137}Xe or for other "technical reasons."
- In 2 other patients, the results of the scan "were inadequate for interpretation."
- Of the potentially eligible patients, 51 were too ill to undergo the gold standard, pulmonary angiography.
- The gold-standard test was not performed in additional patients: 4 patients were allergic to contrast agents; 2 patients were pregnant; 11 patients were too ill; 9 patients refused permission; and 4 patients were excluded for other "technical" reasons.

What are the consequences of these difficulties? Besides reducing the number of patients available for the final analysis, these problems can change the character of the study population. When patients are excluded for ill-defined reasons or do not have the required follow-up with the gold-standard test, the study population, and thus the patients to whom the study conclusions apply, become difficult to define.

FOLLOW-UP STUDIES

An assessment designed to evaluate the impact of a diagnostic test on patient outcome will require clinical follow-up. In addition, when studies of diagnostic accuracy employ a risky gold standard, patients with negative index tests may not be referred for the gold-standard test, and clinical follow-up may be used as a substitute.

There are several ways to conduct follow-up studies.

First, responsibility for collecting the data and filling out the forms can be placed with the referring physician or with physicians and staff at the study center. This approach is useful when physical examinations and testing are part of the follow-up plan. The method is cheap, but risky;

physicians may fail to gather all the data or may use nonstandard methods. Patients may move or may fail to keep follow-up appointments. Such patients are considered "lost to follow-up" and present a challenge to the individuals who must analyze the data.

Second, a research assistant can conduct a structured telephone interview with the patient in order to assess outcome. This method may be more convenient for the patient and may increase the chance of successful follow-up on patients who have moved. It is not useful if tests or a physical examination are needed.

Third, patients can fill out a follow-up questionnaire and return it to the study center by mail. This approach is the least expensive, but compliance is likely to be poor and the cost of contacting noncompliers will be high.

Follow-up can be complicated by a number of factors, particularly if it requires observation or data collection over a period of years or requires the assessment of other than dichotomous variables. Some factors relate to *patients*. Patients may perceive follow-up as a continued intrusion into their lives and simply refuse to cooperate. The patient may experience a change in health status that makes evaluation of outcome more difficult. In a randomized study, the patient may "cross over" and have a diagnostic evaluation for the same indication by the competing technology, making the assessment of the impact of the first test nearly impossible. Furthermore, the patient is not always a reliable source of information. In one study, only 60 percent of patients with heart disease and 70 percent of patients with asthma reported these diagnoses when asked what condition they had (Ludwid and Coletti 1971).

The *environment* in which follow-up is conducted may also present a problem. The technology under study may change, or a newer technology may be developed so that the answer to the study question seems much less important. When interest wanes, follow-up may be inadequate.

The nature of the endpoint chosen for evaluation can also influence the success of follow-up studies. A dichotomous variable such as life or death is easy to assess. Obtaining and coding subjective information about the impact of a test on the patient's functional status or quality of life requires more complex methods. Researchers have recognized the importance of these endpoints and have developed the tools needed to conduct these types of follow-up studies.

Some studies of diagnostic accuracy determine the patient's true state by using the gold-standard test in certain patients and clinical follow-up for those who do not undergo the gold-standard test. Follow-up is very

important in such studies. In McNeil's (1979) evaluation of the CT/RN study, she states that inadequate follow-up made it impossible to determine whether some of the patients entered into the study did or did not have neurological disease. There must be a contingency plan for patients who do not comply with follow-up, and the costs of follow-up must be included in the study budget.

Summary: Implementation

The obstacles encountered in this stage of a technology assessment may be the most difficult to resolve. Randomization, data collection, test performance, and follow-up are all subject to poor compliance and poor performance. To facilitate compliance, the requirements of the protocol should be as explicit and as simple as possible, and they should be written out in detail. The study should be planned to minimize the number of patients who must be randomized at the time examinations are scheduled. The most important element, however, is the motivation of the patients, physicians, and other staff who carry out the protocol. Those individuals involved in carrying out the protocol should receive training before the study begins, and there should be ongoing monitoring of study personnel (Cummings et al. 1988). The best way to avoid implementation problems is expensive: hire a research assistant and assign as many data collection chores as possible to this person.

TEST INTERPRETATION

The choice between efficacy and effectiveness is important in designing the interpretation stage of an assessment. In a study of efficacy, test interpretation must be as accurate, consistent, and objective as possible. The ideal study would include multiple interpretations of both the index test and the gold standard for the purpose of determining interobserver variability. In a study of effectiveness, tests would be interpreted as they are in usual clinical practice. The procedure for interpretation would not necessarily be standardized.

Accuracy

Many factors affect the accuracy of data interpretation. Some, such as physician fatigue (Brogden et al. 1978), are difficult to control. Data from one early study indicated a substantial improvement in radiologists' use

of CT to detect pancreatic carcinoma after the first 1,000 body scans (Sheedy et al. 1977). Improvements in physicians' skills with experience clearly demonstrates the importance of the learning curve. Early estimates of the accuracy of a new test, when physicians' experience is limited, may be a better reflection of their interpretative skills than the potential accuracy of the method.

Consistency and Multiple-Test Interpretations

Consistency is best guaranteed by having the same observer interpret all examinations for a particular technology and by using a standardized definition of an abnormal test result. Ideally, all interpreters using the different methods should be at a similar level of experience. ROC analysis is appropriate for assessing tests with results expressed as continuous variables (Metz 1978). By determining a series of true-positive/false-positive pairs in which different criteria separate the normal from the abnormal, the ROC curve neutralizes observer biases associated with excessively conservative or liberal strategies (Hanley and McNeil 1982).

In a large-scale study, data interpretation might require a full-time commitment from specialists, such as radiologists. It may be difficult to find someone who will devote this amount of time to a study, and equally difficult to recruit the group of specialists who will be needed to reinterpret at least a selected sample of the exams for the purpose of determining interobserver variability. The participation of these individuals should be solicited early, and their time should be a budgeted expense of the project.

Objectivity

How can objectivity of interpretation be obtained? There must be no cross-talk between those who interpret different examinations on the same patient. A physician interpreting the index test should be blinded to the result of the gold standard to avoid *test-review bias*; similarly, a physician interpreting the gold standard should be blinded to the result of the index test to avoid *diagnosis-review bias* (see Chapter 3). Both types of bias can lead to an overestimate of the true-positive and false-positive rates of the index test. Blinded interpretation of the index test and the gold-standard test is absolutely essential. Yet most reports of studies of diagnostic tests do not indicate that this precaution has been taken.

In a study of efficacy, blinded interpretation is the most objective way to determine the accuracy of a test. It may not be ethically sound,

however, to make decisions about patient care based on a test that was interpreted without the benefit of all relevant clinical data. In a study of effectiveness, the interpretation would depend on the combination of clinical information and that derived from the specific imaging examination. This method, although less objective, is the one used in clinical practice.

A study can be designed to accommodate interpretation under both "ideal" and "usual" conditions. There should be two separate interpretations—one (nonblinded) interpretation used for patient care (and thus for effectiveness) and the other (blinded) used for efficacy studies. In general, if we separate study interpretation from interpretation related to patient care, we can blind observers to all other data more ethically.

REPORTING

The clinical utility of an otherwise well-executed diagnostic technology assessment depends on the success with which the results are communicated to physicians who use the tests. In addition, meta-analysis, a form of secondary technology assessment that synthesizes recommendations from published reports, depends on thorough reporting of methods and results (Pillemer and Light 1980, Hunter 1982). A number of authors have proposed standards for assessing and reporting randomized controlled trials; many of these standards can be applied to studies of diagnostic technology. Two groups in particular (Mosteller et al. 1980, Chalmers et al. 1981) have described 16 key features of a good report.

1. a precise statement of the study question, including any prior hypotheses regarding specific subgroups in whom the value of the tests might differ;
2. a complete description of the study population, of inclusion and exclusion criteria (if used), and of patients who were rejected or who may have withdrawn from the study, so that clinicians can determine how their patients compare to the study population, with particular attention to clinical issues that define the spectrum of severity of disease;
3. the dates of the enrollment period, to allow interpretation of the results in light of other developments that may have occurred during that time (for example, technological advances);
4. a detailed description of the study protocol, including the methods for performing tests (or appropriate references for the methodology) and the procedure for randomization (if applicable);

5. a statement of the acceptable level of type I and type II errors, and the size of the sample required to detect the specified difference in study endpoint;
6. presentation of the distribution of pretest variables (for randomized studies) so that clinicians can check for biased assignment of patients to study groups;
7. an indication of the level of compliance with the protocol, with a description of deviations and how they were handled;
8. specification of the reference standard used to define the true state of the patient, taking care to show that there is no use of index test results (or clinical data used for clinical prediction rules) to define the diseased and nondiseased states;
9. the results of the index test(s) and gold-standard test (in a 2-by-2 table, if applicable), with appropriate statistical analyses (for example, ROC for studies of test accuracy where results can be expressed as continuous variables);
10. subgroup analysis: results of tests as in no. 9 in patient subgroups of interest;
11. the results of follow-up (when patient outcome is an endpoint) with confidence limits, life-table analysis, or other statistical analyses as appropriate;
12. a description of the method for handling postintervention withdrawals and patients lost to follow-up;
13. a description of the method used to avoid test-referral bias;
14. a description of the method used to blind those who interpret the index and gold-standard tests;
15. the number of tests that were technically suboptimal or were considered uninterpretable; and
16. the source of funding for the study, to allow identification of possible conflicts of interest.

Two of these items deserve additional attention, because they can be sources of hidden bias in a study of diagnostic technology. Number 8 refers to the pitfall of "circular assessment," which must be avoided when choosing a reference standard. This occurs when the result of one of the index tests in a comparative study is used to define the true state of the patient. To obtain a valid measure of each test's performance, they must be assessed independently of one another, using a different method to verify the presence or absence of disease.

Number 15 in the list above alludes to another potential source of bias:

reports of studies of diagnostic technology seldom include the number of test results that were considered uninterpretable or indeterminate. In one review of ten papers on CT, only five dealt explicitly with the number of unsatisfactory exams. Such information is essential, however, if efficacy is to be judged. For example, if a test detects renal lesions in 70 of 100 patients, misses them in 10, and results in technically suboptimal examinations in 20, the overall sensitivity is 70 over 100 (70 percent). Frequently, the 20 poor-quality exams are excluded, and the sensitivity reported is 70 divided by 80 (88 percent) (Abrams 1981). Thus, if investigators fail to consider the impact of ignoring poor-quality exams, the true-positive and false-positive rates may be artificially inflated (Begg et al. 1986).

CONCLUSION

In this chapter, we have examined the difficulties encountered in each stage of a primary technology assessment, from the planning and design process through the production of the final report. The solutions to some of these problems are relatively straightforward. For example, we have methods to avoid test-review and diagnosis-review bias. We also know that increasing the level of cooperation among participating individuals and institutions will go a long way to improving the outcome of a study. The solutions to other problems, such as when to conduct the assessment or which application to assess, are less obvious. In emphasizing some of the barriers to primary data collection, we have attempted to forestall such difficulties in future assessments. In posing a number of unanswered questions, we would hope to encourage the research necessary to resolve these problems, and thus enhance the value of diagnostic technology assessment.

REFERENCES

Abrams, H.L. Evaluating computed tomography. In Alterman, P.S., Gastel, B., and Eliastam, M., eds., Assessing Computed Tomography, pp. 1-17. National Center for Health Care Technology Monograph Series, Washington D.C., U.S. Department of Health and Human Services, May 1981.

Abrams, H.L. Garland lecture. Coronary Arteriography: pathologic and prognostic implications. American Journal of Roentgenology 139:1-18, 1982.

Abrams, H.L., and Hessel, S. Health technology assessment: problems and challenges. American Journal of Roentgenology 149:1127-1132, 1987.

Alderson, P.O., Adams, D.F., McNeil, B.J., et al. Computed tomography, ultrasound, and scintigraphy of the liver in patients with colon or breast carcinoma: A prospective comparison. Radiology 149:225-230, 1983.

Alperovitch, A. Controlled assessment of diagnostic techniques: Methodological problems. Effective Health Care 1:187-190, 1983.

Angell, M. Patients' preferences in randomized clinical trials. New England Journal of Medicine 310:1385-1387, 1984.

Begg, C.B., Greenes, R.A., and Iglewicz, B. The influence of uninterpretability on the assessment of diagnostic tests. Journal of Chronic Diseases 39:575-584, 1986.

Brogden, B.G., Delsey, C.A., and Moseley, R.D. Effect of fatigue and alcohol on observer perception. American Journal of Roentgenology 130:971-974, 1978.

Brown, B.W., Jr., and Hollander, M. Statistics: A Biomedical Introduction. New York, John Wiley & Sons, 1977.

Cassileth, B.R., Lusk, E.J., Miller, D.S., and Hurwitz, S. Attitudes toward clinical trials among patients and the public. Journal of the American Medical Association 248:968-970, 1982.

Cassileth, B.R., Zupkis, R.V., Sutton-Smith, K., et al. Informed consent: Why are its goals imperfectly realized? New England Journal of Medicine 302:896-900, 1980.

Chalmers, T.C., Smith, H., Jr., Blackburn, B., et al. A method for assessing the quality of randomized control trial. Controlled Clinical Trials 2:31-49, 1981.

Croke, G. Recruitment for the National Cooperative Gallstone Study. In Roth, H.P., and Gordon, R.S., Jr., eds., Proceedings of the National Conference on Clinical Trials Methodology, October 1977. Clinical Pharmacology and Therapeutics 25:691-694, 1979.

Cummings, S.R., Hulley, S.B., and Siegel, D. Implementing the study: Pre-testing, quality control and protocol revisions. In Hulley, S.B., and Cummings, S.R., eds., Designing Clinical Research: An Epidemiological Approach. Baltimore, Williams and Wilkins, 1988.

Drummond, M. Guidelines for health technology assessment: Economic evaluation. In Feeny, D., Guyatt, G., and Tugwell, P., eds., Health Care Technology: Effectiveness, Efficacy and Public Policy. Montreal, The Institute for Research on Public Policy, 1986.

Feinstein, A.R. An additional science for clinical medicine: II. The limitations of randomized trials . Annals of Internal Medicine 99:544-550, 1983.

Feinstein, A.R. Clinical biostatistics-VIII. An analytic appraisal of the University Group Diabetes Program (UGDP) study. Clinical Pharmacology and Therapeutics 12:167-191, 1971.

Ferguson, J.H. Director, Office of Medical Applications Research. Personal communication, 1988.

Ferris, F.L., and Ederer, F. External monitoring in multiclinic trials: Applications from ophthalmologic studies. In Roth, H.P., and Gordon, R.S., Jr., eds., Proceedings of the National Conference on Clinical Trials Methodology, October 1977. Clinical Pharmacology and Therapeutics 25:720-723, 1979.

Fineberg, H.V., Bauman, R., and Sosman, M. Computerized cranial tomography: Effect on diagnostic and therapeutic plans. Journal of the American Medical Association 238:224-230, 1977.

Fineberg, H.V., and Hiatt, H.H. Evaluation of medical practices: The case for technology assessment. New England Journal of Medicine 301:1086-1091, 1979.

Freiman, J.A., Chalmers, T.C., Smith, H., Jr., and Kuebler, R.R. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. New England Journal of Medicine 299:690-694, 1978.

Guyatt, G., and Drummond, M. Guidelines for the clinical and economic assessment of health technologies: The case of magnetic resonance. International Journal of Technology Assessment in Health Care 1:551-566, 1985.

Guyatt, G.H., Tugwell, P.X., Feeny, D.H., et al. The role of before-after studies of therapeutic impact in the evaluation of diagnostic technologies. Journal of Chronic Diseases 39:295-304, 1986.

Hanley, J.A., and McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29-36, 1982.

Hessel, S.J., Siegelman, S.S., McNeil, B.J., et al. A prospective evaluation of computed tomography and ultrasound of the pancreas. Radiology 143:129-133, 1982.

Hopwood, M.D., Mabry, J.C., and Sibley, W.L. A first-order characterization of clinical trials. Prepared for the National Institutes of Health by the Rand Corporation. R-2653-NIH; September 1980: 61-62.

Hull, R.D., Hirsh, J., Carter, C.J., et al. Diagnostic value of ventilation-perfusion lung scanning in patients with suspected pulmonary embolism. Chest 88:819-828, 1985.

Hunter, J.E. Meta-analysis: Cumulating Research Findings Across Studies. Beverly Hills, California, Sage Publications, 1982.

Kent, D.L., and Larson, E.B. Diagnostic technology assessment: Problems and prospects. Annals of Internal Medicine 108:759-761, 1988.

Lidz, C.W., Miesel, A., Osterweis, M., et al. Barriers to informed consent. Annals of Internal Medicine 99:539-543, 1983.

Ludwid, E.G., and Coletti, J.C. Some misuses of health statistics. Journal of the American Medical Association 216:493-499, 1971.

Marks, J.W., Croke, G., Gochman, N., et al. Major issues in the organization and implementation of the National Cooperative Gallstone Study (NCGS). Controlled Clinical Trials 5:1-12, 1984.

Mattson, M.E., Curb, J.D., McArdle, R., et al. Participation in a clinical trial: The patients' point of view. Controlled Clinical Trials 6:156-167, 1985.

McNeil, B.J. Pitfalls in and requirements for evaluations of diagnostic technologies. In Wagner, J., ed., Proceedings of a Conference on Medical Technologies, DHEW Pub. No (PHS) 79-3254, pp. 33-39. Washington, D.C., U.S. Government Printing Office, 1979.

McNeil, B.J., Sanders, R., Alderson, P.O., et al. A prospective study of computed tomography, ultrasound, and gallium imaging in patients with fever. Radiology 139:647-653, 1981.

McNeil, B.J., Weichselbaum, R., and Pauker, S.G. Fallacy of the five-year survival in lung cancer. New England Journal of Medicine 299:1397-1401, 1978.

Metz, C.E. Basic principles of ROC analysis. Seminars in Nuclear Medicine 13:283-298, 1978.

Mosteller, F., Gilbert, J.P., and McPeek, B. Reporting standards and research strategies for controlled trials: Agenda for the editor. Controlled Clinical Trials 1:37-58, 1980.

Office of Technology Assessment, U.S. Congress. The Implications of Cost-Effectiveness Analysis of Medical Technology. Stock No. 051-003-00765-7. Washington, D.C., U.S. Government Printing Office, 1980a.

Office of Technology Assessment, U.S. Congress. The Implications of Cost-Effectiveness Analysis of Medical Technology. Background paper #1: Methodological issues and literature review. Washington, D.C., U.S. Government Printing Office, 1980b.

Office of Technology Assessment, U.S. Congress. The Implications of Cost-Effectiveness Analysis of Medical Technology. Background paper #2: Case studies of medical technologies. Case Study #2: The feasibility of economic evaluation of diagnostic procedures: The case of CT scanning. Washington, D.C., U.S. Government Printing Office, 1981.

Phelps, C.E., and Mushlin, A.J. Focusing technology assessment using medical decision theory. Medical Decision making 8:279-289, 1988.

Pillemer, D.B., and Light, B.J. Synthesizing outcomes: How to use research evidence from many studies. Harvard Education Review 50:176-195, 1980.

Prout, T.E. Other examples of recruitment problems and solutions. In Roth, H.P., and Gordon, R.S., Jr., eds., Proceedings of the National Conference on Clinical Trials Methodology, October 1977. Clinical Pharmacology and Therapeutics 25:695-696, 1979.

Schoenberger, J.A. Recruitment in the Coronary Drug Project and the Aspirin Myocardial Infarction Study. In Roth, H.P., and Gordon, R.S., Jr. , eds., Proceedings of the National Conference on Clinical Trials Methodology, October 1977. Clinical Pharmacology and Therapeutics 25:681-684, 1979.

Schwartz, J.S. Evaluating diagnostic tests: What is done—what needs to be done. Journal of General Internal Medicine 1:266-267, 1986.

Sheedy, P.F., Stephens, D.H., Hattery, R.R, et al. Computed tomography in patients suspected of having carcinoma of the pancreas: Recent experience (abstract). Presented at the scientific assembly and annual meeting of the Radiological Society of North America, Chicago, Ill., November 1977.

Smith, T.J., Kemeny, M.M., Sugarbaker, P.H., et al. A prospective study of hepatic imaging in the detection of metastatic disease. Annals of Surgery 195:486-491, 1982.

Sox, H.C., Jr. Probability theory in the use of diagnostic tests: An introduction to critical study of the literature. Annals of Internal Medicine 104:60-66, 1986.

Steinberg, E.P., and Cohen, A.B. Office of Technology Assessment, U.S. Congress. Nuclear Magnetic Resonance Imaging Technology: A Clinical, Industrial, and Policy Analysis. Technology case study 27. Washington, D.C., U.S. Government Printing Office, 1984.

Taylor, K.M., Margolese, R.G., and Soskoline, 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.

Vreim, C. Project officer, Prospective Investigation of Pulmonary Embolic Diagnosis project (PIOPED). Personal communication, 1988.

Weiner, D.A., Ryan, T.J., McCabe, C.H., et al. Exercise stress testing: Correlation among history of angina, ST-segment response and prevalence of coronary artery in the Coronary Artery Surgery Study (CASS). New England Journal of Medicine 310:230-235, 1979.

Weinstein, M.C. Methodologic considerations in planning clinical trials of cost-effectiveness of magnetic resonance imaging (with a commentary on Guyatt and Drummond). International Journal of Technology Assessment in Health Care 1:567-581, 1985.

Weinstein, M.C., and Stason, W.B. Foundations of cost-effectiveness analysis for health and medical practices. New England Journal of Medicine 296:716-721, 1977.

5

Costs and Sources of Funding

The focus of this chapter will be on the costs associated with studies that prospectively gather primary data on diagnostic technology. The aim of these studies may be to establish safety, efficacy, effectiveness, or cost-effectiveness, and several research designs may be used, including randomized controlled trials, nonrandomized comparative studies, or retrospective studies. Our goals are to understand the factors that contribute to the costs of these studies and to examine briefly the question of who should pay for them. In the process, we will take a closer look at several studies of diagnostic technology.

FACTORS THAT AFFECT STUDY COSTS

Remarkably little information is available on costs of studies of diagnostic technology. Clinical trials have received more attention, in part because of their seemingly large price tags. The Coronary Drug Project, the first major clinical trial sponsored by the National Institutes of Health, cost approximately $50 million (Levy and Sondik 1982).

Several general issues warrant discussion. From the perspective of a funding agency, the important costs to consider are the incremental costs of doing a study. These represent expenses that are beyond the costs of usual patient care. For example, in a study of a drug that is already used in a group of patients, the cost of the drug itself is not a cost of the study; the drug will be prescribed whether or not the study is done. A similar situation occurs with respect to diagnostic tests. Often a study will be

performed on a test that is already in clinical use. The incremental costs of the study will usually involve only the additional tests necessary for the study. There are, of course, exceptions: some projects may require funding for all tests or for activities that could be considered part of the usual patient care. In addition, the incremental costs should include any patient-related expenses brought about by adverse effects of the additional testing. As a practical matter, however, these costs will often be impossible to predict.

Another difficulty in determining the cost of a particular project arises when more than one question is studied simultaneously. A study of CT may simultaneously examine the effectiveness of different contrast agents. It may be difficult to allocate the costs of the study of contrast agents within the context of the entire research project. In large trials in which many questions are addressed, this problem may be intractable.

Many of the costs we discuss are typically included in the budget of a research proposal; some are not budgeted, however, and represent "hidden" costs. For instance, the time and effort involved in protocol development may not be reimbursed, or, in some large projects, the budget may include investigator and consultant time for protocol development (see "Examples from Actual Studies" later in this chapter). Nonetheless, all of the costs should be explicitly anticipated, whether or not they will appear in a submitted budget.

The specific costs of a study depend on the technology under consideration and on the clinical problem for which the technology is being used. Thus, rather than attempt to make detailed estimates, we will raise issues that warrant consideration by investigators and by the policymakers who must allocate resources to support studies of diagnostic technology. We consider costs, including unreimbursed costs to patients, relevant to the funding of these studies. We have arbitrarily divided study costs into three categories: costs associated with planning and protocol development, costs associated with implementation, and costs associated with data interpretation. Meinert (1986) and Piantadosi (1987) discuss the cost of clinical trials; many of the issues that they consider are applicable to the assessment of diagnostic technology, and we draw on their work.

Planning and Protocol Development

PERSONNEL

Investigator time is a substantial cost during all stages of the study. Often it represents a hidden cost, because it may not appear in grant

budgets; in many cases, the institution underwrites all or part of the investigator's salary.

Statisticians, research assistants, data entry personnel, data analysts, and administrative and secretarial staff are necessary for most projects. The costs associated with these personnel accrue at different times during the study but should be anticipated during the planning phase (McNeil 1979).

PROTOCOL DEVELOPMENT

An early step in most studies will be to specify the study protocol(s) and to develop the forms for data collection. It will usually be appropriate to consult with statisticians and data analysts at this stage. The associated costs may be trivial or substantial, depending on the scope of the project. The Veterans Administration Cooperative Studies program estimated the average planning costs of multi-institutional clinical trials to be $20,000 to $26,000, or about 1 to 2 percent of the total cost (Henderson 1980). The Prospective Investigation of Pulmonary Embolic Diagnosis (PIOPED) study of diagnostic tests for pulmonary embolism, a $7.6 million multi-institutional technology assessment sponsored by the NIH, required over a year to develop the necessary protocols (Vreim 1988).

SAMPLE SIZE

A critical activity during the planning phase is estimating the sample size needed to give the study adequate power to detect clinically meaningful differences in test performance. Sample size will strongly affect the costs of the study: the cost of diagnostic tests and the cost of follow-up will directly depend on the number of subjects in the study. Administrative and personnel costs are likely to increase with larger sample sizes as well.

Several factors affect the size of the sample needed for the study. The cost of the study may rise dramatically as the power of the study is increased or when the study is designed to detect smaller differences in outcomes (Detsky 1985). The power of a study is the probability that the study will successfully detect a difference in outcome if the difference actually exists. A study designed to reveal a 20 percent difference in clinical outcome at a given power may be many times more expensive than a study designed to show a 40 percent difference in clinical outcome at the same power. It may be difficult to show a difference in sensitivity and specificity of a new technology when the test performance of the old

technology is already quite good (Abrams and Hessel 1987). Thus, if an existing technology has a sensitivity and specificity that are nearly 1.0, the maximum possible difference in sensitivity and specificity between the new and old technology will be small; to detect this small difference will require a large sample size. In their study of CT, ultrasound, and gallium scans in the evaluation of patients with an undiagnosed cause of fever, McNeil et al. (1981) calculated that it might require as many as 500 patients to show a significant difference in the receiver operating characteristic (ROC) curves of the different tests.

A possible strategy to reduce study costs without loss of power is to use comparison groups of uneven size (Meydrech 1978, Rosenberg 1983). This approach can be useful when the costs associated with one comparison group are less than the costs of the other. The methodology has been analyzed for case-control studies; whether it will be useful for technology assessment is not yet certain.

Decision-Analytic Modeling

We have suggested that technology assessment be model-driven (see Chapter 3) in order to assure that all the relevant data are specified by the study protocol. This approach involves decision-analytic modeling of the relevant decision problems before actual data gathering begins. Thus, an individual with experience in decision-analytic methods must be involved in the study from the outset. Our experience at Stanford University suggests that modeling the clinical problem, including literature review, may take three to six months of full-time work, depending on the scope of the project. This modeling effort would represent a salary expense of approximately $15,000 to $30,000 to support a qualified investigator.

Implementation

PATIENT ACCRUAL

A number of costs, often unanticipated, relate to the process of recruiting and enrolling patients. (For a general model of predicting accrual costs in clinical trials see Piantadosi 1987.) One of the more common mistakes in the design of clinical trials is to underestimate the difficulty of enrolling the required number of patients (Hulley 1988). A notable example has been discussed by McNeil et al. (1981): in a National Cancer Institute study of CT versus radionuclide studies in patients with sus

pected intracranial disease, 3,000 patients were screened but only 156 were suitable for final data analysis. A variety of misadventures contributed to this poor yield, many unrelated to accrual, but the study serves as an example of how unanticipated difficulties may arise. The study cost $2 million, or $16,441 for each patient analyzed.

Several factors influence the difficulty of patient accrual. As patient eligibility criteria become more restrictive, larger numbers of patients will have to be screened. Once the investigators have found an eligible patient, the ease of enrollment will depend on the nature of the study. We can imagine that the enrollment effort in a study where the design requires doing a barium enema, flexible sigmoidoscopy, and colonoscopy in each subject might suffer because of the distasteful nature of the tests. An additional consideration is attrition of patients. If attrition is likely to be a major problem, the investigators should plan to screen still larger numbers of patients.

A research assistant may be vital to the recruitment effort. As we noted in Chapter 4, it will be impossible to enroll patients without the cooperation of referring physicians. The research assistant may play a role in garnering this cooperation. The help that he or she provides in explaining the protocol, gathering data, and providing follow-up information to referring physicians will facilitate enrollment. McNeil et al. (1981) noted the need for a full-time research assistant at each study site in their evaluations of CT and ultrasound.

DATA COLLECTION

Costs will increase as more data are collected because this effort will require more time from study personnel. In addition, procedures to ensure uniform data collection and accurate data entry, and to provide quality control, will become more complex and costly as the size of the study increases. These problems may be particularly important for multi-institutional studies (see "Examples from Actual Studies," p. 113).

PATIENT FOLLOW-UP

Depending on the details of the research question, patient follow-up may be necessary. First, the determination of the true disease state of the patient may depend on clinical follow-up, particularly when there is not an acceptable and reliable gold-standard test. Likewise, if the gold standard test is considered too dangerous to use in patients with a negative

index test, clinical follow-up will be necessary to determine the patient's disease status. Second, in cost-effectiveness or cost-benefit studies, the investigator must determine patient outcomes, and this will usually involve clinical follow-up. Third, occasionally technology assessments will randomize patients to one technology or another, with clinical outcome being the measure of efficacy.

The type of patient follow-up used will clearly influence costs. A questionnaire will be more economical than chart review, which in turn will be more economical than patient interviews.

ADDITIONAL MEDICAL EXPENSES

A study patient may incur costs as an indirect result of the protocol. For example, additional hospital days may be required. Although these costs will usually be difficult to estimate (and they are unlikely to be budgeted), they may be substantial. For example, if a protocol added an average of one-half day to the hospital stay, this could amount to $300 for each patient, or up to $60,000 for a study with 200 patients. In some cases, these costs will be paid by third-party payers. With a DRG-based reimbursement plan, however, the institution will have to absorb these costs.

Data Interpretation

Costs of data storage and interpretation will depend on the scope of the project (see "Examples from Actual Studies"). Microcomputers will often be sufficient for analysis, and their costs should be budgeted. Statistical consultation will be an important element of the costs of data interpretation.

UNREIMBURSED PATIENT COSTS

Studies of diagnostic technology may involve costs to patients that are not reimbursed. Loss of wages due to time spent away from work is an example. Further, a patient may suffer adverse effects from the tests. The more invasive or time-consuming the technology, the greater these unreimbursed costs are likely to be. They will not directly affect the cost of performing a study, but they may affect patients' willingness to participate in the project.

EXAMPLES FROM ACTUAL STUDIES

We will now illustrate some of the costs described earlier in this chapter by examining past and current studies of diagnostic technology. We will look at studies of different scope and see differences in the magnitude of the relative cost components.

Cost of Diagnostic Tests

Reliable information about the true economic cost of diagnostic tests is difficult to obtain. Both direct and indirect costs should be considered (Travers and Krochmal 1988). Direct costs include the cost of equipment and labor that can be directly related to a particular test. Indirect costs include labor that cannot be directly tied to a particular test (for example, supervisory and administrative personnel) and the costs associated with insurance, maintenance, depreciation, power, and the like. In the past, such information has been impossible to obtain, but cost accounting systems are now being developed that will make it possible to develop accurate estimates of test cost (Travers and Krochmal 1988, Travers in press). Until these systems are widely utilized, however, we must estimate the cost of diagnostic tests from the amount charged for the test. Charges are often a poor estimate of the true cost of a test because the charge reflects factors in addition to the cost of producing the service. Nonetheless, charges are usually the only available information about test cost.

In 1978 Alderson and colleagues prospectively compared CT, ultrasound, and technetium scans of the liver in patients with known breast or colon cancer (Alderson et al. 1983). The authors studied 189 patients, 122 of whom had all three studies. The aim of the study was to construct ROC curves for the three tests. The total of charges for the diagnostic tests, including professional fees, was at least $125,000 in 1978 (the year the study was performed). It would cost approximately $163,000 to perform the tests in 1988, given that year's charges for the tests. If the study paid the full professional fees, triplicate readings of films (to assess interobserver variation) would increase the 1978 total by $48,000, and the 1988 total by $91,000.

In this study the cost of the diagnostic tests was borne by third-party payers. Under different circumstances, the funding agency might have to support the expense (see the description of the PIOPED project, below).

Personnel Costs

A current study of MRI provides an example of personnel costs. This study is designed to examine the cost-effectiveness of MRI and to develop new methodologic approaches to technology assessment (Mushlin 1988). The study involves a substantial amount of methodologic research in the first year. Some of the first-year costs would thus not be applicable to other studies, but a significant component of the early effort involves decision-analytic modeling of the clinical problem as a means to guide the design of the study protocol. Thus, the study is an example of the approach to technology assessment that we have advocated.

The total budget for the project is about $1 million, including the indirect costs that cover overhead and other institutional expenses. Of the budgeted direct costs, personnel costs account for 95 percent. The personnel include the investigators, research assistants, and administrative and support staff. Supply and travel costs each amount to 1 percent of the direct costs; data analysis accounts for about 2 percent. No money is budgeted for the tests or patient care, because third-party payers have agreed to fund the costs of the additional tests required for the comparative analysis.

Multi-Institutional Studies

The Prospective Investigation of Pulmonary Embolic Diagnosis project is one of the most ambitious diagnostic technology assessments ever attempted. The study compared ventilation-perfusion scanning and pulmonary angiography in the diagnosis of pulmonary embolism. The total cost to the NIH was $7.6 million, which included direct and indirect institutional costs (Vreim 1988). The study involved six clinical centers and a data analysis center. The intervention group had 951 patients; the "usual care" group contained 568 patients. The data analysis center was budgeted for $1.8 million, or about 24 percent of the total NIH costs (Table 5.1). Personnel costs, including the cost of consultants, accounted for 60 percent of the total budget (76 percent of direct costs) of the data analysis center, and for approximately 50 percent of the total budget (80 percent of direct costs) of the clinical centers (Table 5.2). The second largest item in the budget of the clinical centers was the cost of angiograms in the intervention group (about $600,000 for the six clinical centers). The NIH did not pay for angiograms in the "usual care" group.

TABLE 5.1 Five-Year Budget of the PIOPED Data-Analysis Center

Cost Element	Amount (dollars)
Professional labor	230,000
Programmer labor	321,000
Clerical labor	241,000
Fringe benefits	140,000
Consultants	140,000
Travel	57,000
Office equipment	35,000
Computer services	125,000
Building rent	66,000
Office costs	54,000
Indirect costs	415,000
Total	1,824,000

TABLE 5.2 Five-Year Budget of a Representative PIOPED Clinical Center (one of six centers)

Cost Element	Amount (dollars)
Investigator labor	142,000
Technician labor	174,000
Clerical labor	78,000
Fringe benefits	117,000
Consultants	6,000
Travel	17,000
Office equipment	3,000
Angiograms	98,000
Other direct costs	10,000
Indirect costs	351,000
Total	996,000

Thus, $7.6 million represents an underestimate of the total cost of the study.

The PIOPED study shows that as the scale of investigation increases, the relative size of the cost components may change. Total data analysis costs will be larger at this grand scale. The cost per unit of data analyzed may increase or decrease, however, depending on the efficiency of the data processing. Much of the expense budgeted to the data analysis center was for personnel. In addition, development of the protocols took approximately 15 months; clearly, in studies of this size, the planning and protocol development phase will involve substantial effort and expense.

SOURCES OF FUNDING

Funding for technology assessment and for clinical trials is likely to come from similar sources (Institute of Medicine [IOM] 1985, Meinert 1982): government institutions, private foundations, industry, and third-party payers. Nonetheless, most observers feel that funding for technology assessment has been inadequate (IOM 1985, Fineberg and Hiatt 1979).

Rapid dissemination of technology removes any incentive for manufacturers to fund technology assessment. As we have noted, technology often becomes widely used before it has been adequately assessed. Fetal monitoring and MRI are well-known examples. Under these conditions, manufacturers have no incentive to fund technology assessment, because the results may only serve to decrease the use of the manufacturer's product. Meinert (1986) analyzes this issue in the pharmaceutical industry.

Antitrust laws may impede funding of technology assessment by third-party payers (IOM 1985, Rose and Leibenluft 1986). Generally speaking, antitrust law is applicable to situations in which there is agreement or concerted action between two entities that leads to an unreasonable restraint of trade. Thus a technology assessment financed by third-party payers could, in principle, be subject to challenge under antitrust legislation if the results of the assessment serve to reduce or foreclose the use of a device or procedure. As Rose and Leibenluft (1986) observe, "even conduct that may ultimately be considered legal may nevertheless be the subject of very costly and lengthy litigation" (p. 1492).

Because the results of technology assessment will be public information, there is little incentive for industry or third-party payers to provide funding when it is likely that some other party will. This reasoning has

led to a variety of proposals for involving insurers and industry in the technology assessment effort.

Proposals for Funding

Various suggestions have been made about how technology assessment should be funded (IOM 1985). In 1980, Relman called for a "major new national program of support for the evaluation of medical procedures of all kinds" (p. 154). He suggested that the work be done primarily in the private sector and that large-scale funding, $200 million to $300 million annually, would be necessary. The proposal recommended that an allocation of 0.2 percent of the Health Care Financing Administration budget (about $100 million in 1980) and a proportion of the budgets of private third parties be earmarked for technology assessment.

In an analysis of the effects of reimbursement on biomedical innovation, Bunker et al. (1982) suggest that insurance coverage of new therapies be contingent on their adequate assessment. IOM reports recommend establishing a public–private sector consortium for technology assessment (IOM 1983, IOM 1985). The consortium would begin with start-up funds from Congress and then be supported from an endowment to be raised by pooling the funds of payers, foundations, professional associations, and other users of the assessments. As noted in the Introduction, the Council on Health Care Technology, a nongovernmental arm of the IOM, is an indirect outgrowth of the IOM reports. The council does not, however, play a direct role in the financing of technology assessment.

The proposals share the common theme that private insurers, industry, and government should share the financial burden of technology assessment. A formal mechanism for ensuring such cooperation has not been established.

SUMMARY AND CONCLUSIONS

This chapter has explored the costs of studies of diagnostic technology. The major expenses in these studies will be for personnel and diagnostic tests. The investigator should anticipate expenses associated with the following:

- protocol development
- decision-analytic modeling
- costs associated with patient accrual

- data collection and analysis
- clinical follow-up.

The sample size of the study will strongly influence the total cost; questions of power will therefore warrant careful consideration.

Funding of technology assessment has been insufficient. Most proposals on the subject recommend that the public and the private sectors share fiscal responsibility. The long-term cost of poorly performed technology assessment is likely to outweigh by far the more immediate cost of well-designed studies of diagnostic technology.

REFERENCES

Abrams, H.L., and Hessel, S. Health technology assessment: Problems and challenges. American Journal of Roentgenology 149:1127-1128, 1987.

Alderson, P.O., Adams, D.F., McNeil, B.J., et al. Computed tomography, ultrasound, and scintigraphy of the liver in patients with colon or breast carcinoma: A prospective comparison. Radiology 149:225-230, 1983.

Bunker, J.P., Fowles, J., and Schaffarzick, J. Evaluation of medical-technology strategies: Effects of coverage and reimbursement. Parts I and II. New England Journal of Medicine 306:620-624, 678-692, 1982.

Detsky, A.S. Using economic analysis to determine the resource consequences of choices made in planning clinical trials. Journal of Chronic Diseases 38:753-765, 1985.

Fineberg, H.V., and Hiatt, H.H. Evaluation of medical practices: The case for technology assessment. New England Journal of Medicine 301:1086-1091, 1979.

Henderson, W.G. Some operational aspects of the Veterans Administration cooperative studies program from 1972 to 1979. Controlled Clinical Trials 1:209-226, 1980.

Hulley, S.B., and Cummings, S.R., eds. Designing Clinical Research. Baltimore, Williams and Wilkins, 1988.

Institute of Medicine. Medical Technology Assessment: A Plan for a Public/Private Sector Consortium. Washington, D.C., National Academy Press, 1983.

Institute of Medicine. Assessing Medical Technologies. Washington D.C., National Academy Press, 1985.

Levy, R.I., and Sondik, E.J. Large-scale clinical trials: Are they worth the cost? Annals of the New York Academy of Sciences 382:411-422, 1982.

McNeil, B.J. Pitfalls in and requirements for evaluations in diagnostic technologies. In Wagner, J., ed., Proceedings of a Conference on Medical Technologies. DHEW Pub. No (PHS) 79-3254. Washington D.C., U.S. Government Printing Office, 1979:33-39.

McNeil, B.J., Sanders, R., Alderson, P.O., et al. A prospective study of computed tomography, ultrasound, and gallium imaging in patients with fever. Radiology 139:647-653, 1981.

Meinert, C.L. Funding for clinical trials. Controlled Clinical Trials 3:165-171, 1982.

Meinert, C.L. Clinical Trials: Design, Conduct and Analysis. New York, Oxford University Press, 1986.

Meydrech, E.F., and Kupper, L.L. Cost considerations and sample size requirements in cohort and case-control studies. American Journal of Epidemiology 107:201-205, 1978.

Mushlin, A.I. Personal communication, 1988.

Piantadosi, S., and Patterson, B. A method for predicting accrual, cost, and paper flow in clinical trials. Controlled Clinical Trials 8:202-215, 1987.

Relman, A.S. Assessment of medical practice: A simple proposal. New England Journal of Medicine 303:153-154, 1980.

Rose, M., and Leibenluft, R.F. Antitrust implications of medical technology assessment. New England Journal of Medicine 314:1490-1493, 1986.

Rosenberg, M.J. Cost efficiency in study planning and completion. American Journal of Medicine 75:833-838, 1983.

Travers, E.M., and Krochmal, C.F. A new method for determining test cost per instrument. Medical Laboratory Observer 20:24-29, 1988.

Travers, E.M. Managing costs in clinical laboratories. In Laboratory Microcost Analysis: Developing Instrument and Test Costs. New York, McGraw-Hill, in press.

Vreim, C. Project officer, Prospective Investigation of Pulmonary Embolic Diagnosis project (PIOPED). Personal communication, 1988.

6

A National Program for Assessing Diagnostic Technology

A MULTICENTER CONSORTIUM

The evaluation of diagnostic technology should be national in scope. The purpose of this chapter is to propose a multicenter consortium for conducting cooperative trials of diagnostic technology. This program could solve many of the problems that have been described in the preceding chapters. All studies will adhere to predefined principles and will focus on patients and technologies of interest to policymakers and to clinicians. Competing technologies will be compared, and the data needed for clinical decisionmaking will be obtained. The consortium will be organized according to the principles discussed in Chapter 7.

This proposal is based closely on a plan that appeared in *A Forward Plan for Medicare Coverage and Technology Assessment* (Roe et al. 1987). Two different methods of technology assessment will be used by the proposed centers;

- *Primary technology assessment:* the process of assessing technology by collecting data from patients. This category includes randomized clinical trials and studies that measure the accuracy of diagnostic tests. The centers should emphasize primary technology assessment. In Chapter 3, we described an approach to studying diagnostic technology; the centers should adopt this or a similar approach.
- *Secondary technology assessment:* the process of using previously published studies to evaluate technology. Secondary assessment is widely

used by government agencies, professional organizations, and individual investigators. The centers should develop models and methods for helping physicians to make decisions concerning individual patients and should do cost-effectiveness analyses of competing technologies.

A multicenter approach to evaluating diagnostic technology has several advantages. Although seldom used in past studies of diagnostic technology, the multi-institutional cooperative study addresses two important problems related to the conclusions reached and to study methods. The validity of the conclusions of prior studies of patient care is often questioned, because too few patients have been studied. With a cooperative effort involving many centers, studies can be large enough to meet the requirements of statistical analysis. The conclusions of earlier studies have also often had limited applicability because of bias in patient selection and because the study was carried out at a single institution. Multicenter collaborative studies can involve many types of hospitals, from HMOs to university medical centers.

Another potential benefit is improvement in study methods. The design of the clinical trial with randomized controls has undergone considerable refinement in the past two decades, in part because experienced investigators from different institutions have worked together on issues of study design. There have been few such advances in studies of diagnostic technology. One advantage of a multi-institutional consortium would be to focus the attention of many different experts on study design.

CHARACTERISTICS OF THE CENTERS

The purpose of the centers is to serve as a standing resource that can perform technology evaluation. For new technologies, the centers will help to determine how the test can best be used and how fast it should diffuse into general use. For existing technologies, they will define the role, if any, of a technology in comparison with newer, competing technologies.

Each center should be located at a major teaching hospital, with access to many patient groups and to a strong supporting faculty. To capitalize fully on these advantages, the faculty of the center must convince their colleagues to cooperate. To do so will not be easy, because teaching hospitals are busy, complex organizations that are under considerable stress.

Conceptually, the teaching hospital must be organized as a technology

assessment laboratory (Figure 6.1). The cast of characters in this laboratory is complex; it includes the director of the center, coinvestigators, the core support groups, research assistants, and the medical school and hospital administration. The director of the center must have a resourceful program coordinator. For the technology assessment laboratory to be able to respond rapidly to new technological innovation, all of the major clinical departments must be involved and represented on its staff.

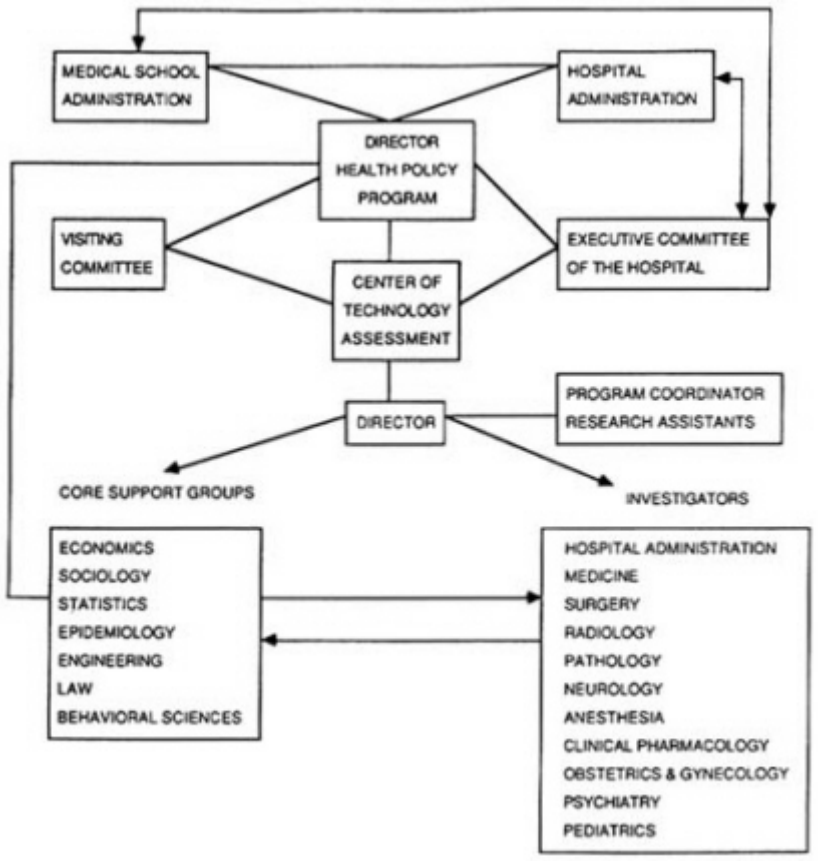

FIGURE 6.1 Components of a center of technology assessment.

Beyond the primary group of investigators, there must be a support structure consisting of faculty in biostatistics, economics, sociology, epidemiology, engineering, law, and the behavioral sciences. The evaluation of technology is neither more nor less difficult than human biomedical

research, with all of the constraints of cost, safeguards for human rights, biases, and legal and ethical problems that characterize any kind of biomedical research in man.

The full support and collaboration of the hospital administration and strong leadership within the laboratory itself are required. Many aspects of the assessment of technology impinge upon the day-to-day functions of the hospital. These must be integrated smoothly into the hospital routine, while assuring that the goals of the technology assessment laboratory can be met.

The medical school has an important role to play in creating and then maintaining an excellent technology assessment laboratory. The medical school determines policy on recruiting and promoting faculty. For too long, medical schools have failed to give adequate recognition to individuals whose research deals with health policy, the quality of medical care, and technology assessment. Recognition is coming slowly, but inevitably, to this important area, at least in many medical schools. Technology assessment laboratories must have the support of the medical school faculty and administration.

Beyond these essential components of a technology assessment laboratory is the obvious need to obtain a large amount of data rapidly, so that the assessments are completed in a timely fashion. To achieve this goal will require four or five national centers working together in a multi-institutional consortium. Depending on the number of centers, the annual cost of maintaining these centers will be several million dollars. What will this support ensure?

Access to patients. Each center should affiliate with several hospitals to increase the diversity of patients and to maximize the rate of enrolling patients. Taken together, the centers will have access to a wide variety of both health care facilities and patients. With four to six centers nationwide, patients can be enrolled rapidly enough to ensure an adequate study of infrequently used technologies. For example, eight patients weekly (400 patients yearly) could be enrolled in a multicenter study of a diagnostic test that is performed once a week in each hospital.

Access to technical expertise. Each center will have a principal investigator, a statistician, and several research assistants. Taken together, the technical expertise of the centers will include clinical prediction rule development, meta-analysis, cost-effectiveness analysis, randomized clinical trials, and clinical epidemiologic methods. Individuals from different centers will work closely in designing studies. The synergy of their joint efforts should be a powerful resource.

Primary technology assessment. Several primary technology assessment studies can be performed simultaneously at all centers. The number of primary technology assessments that can be performed at one time will depend on the rate of enrolling patients and on the level of fiscal support for the centers.

Secondary technology assessment. With modest additional support, each center can perform several secondary technology assessments each year.

The remainder of this section is a description of some of the fiscal and logistical aspects of running the proposed program in technology assessment.

Responding to Requests for a Technology Assessment

The centers will respond to requests from various sources, including the federal government, third-party payers, and professional organizations. These requests might focus on different kinds of problems, such as:

- emerging technologies that have not yet become generally available. Evaluation at this stage of diffusion can prevent indiscriminant adoption of a new technology;
- questions that had not been answered in prior studies (for example, about cost-effectiveness, evaluation in new patient groups, changes in the technology);
- existing technologies that have been found to be effective in past studies but whose usefulness is now being questioned; and
- technologies that have not been studied in a population of patients of special importance for health policy, such as the elderly.

Upon receiving a request, several months may be required to determine the feasibility of doing the study, to identify the need for supplemental funding, and to estimate the time required to accrue enough patients.

Choice of a Study Method

Once a study topic has been decided upon, one individual will be assigned primary responsibility for developing the study protocol. Representatives from each center will meet to refine the protocol.

ESTIMATED TIMETABLE

The time required from receiving a request to completing a patient care

study could be as short as 18 months, depending on the rate at which new patients can be enrolled. The ability to complete an assessment rapidly, before the new technology has diffused into practice, is one of the principal advantages of a multicenter consortium.

REVIEW OF RESULTS AND RECOMMENDATIONS

Expert review before disseminating study results and recommendations will be an integral element in the cooperative studies. The preliminary report of a study will, therefore, be sent to a panel of subspecialty experts for review. The principal investigator will be charged with responding to the comments and incorporating appropriate suggestions and critical analyses into the final report.

STUDY PERSONNEL

Several types of study personnel will be required. The exact mix of these is difficult to establish in advance, but experience with clinical studies suggests the following configuration:

Principal investigator. An investigator from one of the centers will become the principal investigator for a specific study. This person will take primary responsibility for designing, analyzing, and reporting the study, although all investigators will contribute to these phases. The principal investigator will be responsible for making certain that all centers adhere to the study protocol and contribute their expected share of patients. Vesting one investigator with ultimate responsibility for a study is an important principle.

Study co-investigators. Each center that participates in a study will designate an individual to be responsible for ensuring that the center fulfills its obligations to the study. This person will be in charge of the study at the center and will represent the center on the executive committee for the study.

Collaborating subspecialist(s). Subspecialists have the best knowledge of a variety of technologies and the best access to many types of patients. They are essential for studies of diagnostic tests, and their cooperation is necessary to implement protocols for uniform test interpretation. The amount of subspecialist effort will vary from study to study. The budget must include support for these individuals.

Analysts. Each center should have the part-time services of a biostatistician, a decision analyst, and a health economist to advise the principal investigator.

Research assistant. A research assistant will perform follow-up studies on patients, keep records, and enforce uniform patient data collection procedures. The number of research assistants would depend on the number of studies in progress. Each center should have at least two at all times. Additional research assistants can be recruited as needed.

Nurse practitioner. Considerable clinical data must be collected on all patients, including a history of the patient's problem and a directed physical examination. There should be at least one study nurse practitioner at each of the center sites. The nurse practitioner will enroll patients and perform the necessary clinical data collection. The number of nurse practitioners will depend on the number of studies in progress.

Data entry clerk. This individual's assignment will be data entry and data checking. The number of data entry clerks will depend on the number of studies in progress. Each center should have at least one at all times. Additional clerks can be recruited as needed.

Administrative staff. The administrative needs of a center include a full-time secretary to handle correspondence and filing and to assist in data entry. If many studies are being done simultaneously, a full-time study coordinator will be required.

BUDGET

To estimate a budget for each participating center, we assume that each center will provide partial salary support for the principal investigator, several collaborating subspecialists, a statistician, a decision analyst, and a health economist. The center will support full-time positions for two nurse clinicians, two research assistants, one data entry clerk, one secretary, and one study coordinator. We estimate that each center's annual budget would be between $500,000 and $600,000 in 1988 dollars.

MULTICENTER STUDIES—A CAUTION

For all the irreplaceable advantages of multicenter studies, there are many pitfalls to be avoided in running the studies. Chapter 7 contains a

discussion of the logistics of multicenter studies and the potential problems that they present.

ONGOING TECHNOLOGY ASSESSMENT BY MEDICAL INSTITUTIONS

Quite apart from the problem of assessing new diagnostic technologies, the understanding of how methods that have already diffused into practice actually perform is frequently deficient. The diagnostic laboratories and imaging departments of a medical center expend considerable effort on standardizing their product. Accreditation of a hospital depends in part on ongoing quality control of the diagnostic technologies offered within the institution. These activities can be extended in a very important way if an institution makes a commitment to *measure the sensitivity and specificity of diagnostic tests as performed on its patient population*.

The performance of a technology depends partly on the patients on whom it is used and partly on the health professionals who carry out the diagnostic studies. (This point is illustrated dramatically by Table 2.1 in Chapter 2, which describes the results of a series of CT studies in patients with lung cancer.)

The sensitivity and specificity of a diagnostic test are likely to be site-specific. One reason is that different sites have different types of patients. For example, the sensitivity of a test in a primary care population is likely to be lower than in a population of patients that have been referred for treatment of advanced disease. A test generally is able to detect advanced disease more easily than the early disease that is likely to be seen in primary care practice.

Variation in test performance can also arise form variability in both equipment and the competence of the health providers. There have been few interinstitutional comparisons of test performance characteristics in which the institutions have been broadly representative. An exception is a recent study in which samples from a single person were sent to many medical centers for serum cholesterol measurement. The institutions varied widely in the concentration of total cholesterol that they reported (College of American Pathologists 1987). A study of exercise testing came to much the same conclusion (Philbrick et al. 1982).

In these studies, the interinstitutional differences in test performance imply large differences in study populations and technique. Under these circumstances, pooling of several studies to obtain a single number for sensitivity and for specificity may be inappropriate. The data suggest, on the contrary, that each institution is unique. The solution is not to use

pooled data but to measure the performance of the test in the institution. This assignment will become increasingly feasible as hospital computer data bases become more available to clinicians.

In planning a technology assessment, physicians define the gold-standard tests that will be used to determine the true state of the patient. These may vary from surgery, autopsy, or biopsy to the results of another test, or even the results of long-term surveillance to detect disease missed by the index test. Each patient getting a test is entered in a log book. Several months later, their charts are reviewed to see if they had one of the prospectively defined gold-standard tests. If they did, the result is noted, together with the results of the index test. If not, the patient's medical record is flagged, so that he or she can be followed clinically for development of the disease, using clinical criteria that were prospectively defined. The task of identifying medical records would be time-consuming with paper-based medical records, but the increasing use of computers to store coded clinical data will reduce this burden to a straightforward routine.

How are the data from such a program used? The sensitivity and specificity of the test can be calculated on a periodic basis to see if the technology is changing. The sensitivity and specificity of the test can be published in institutional newsletters. The physicians who interpret the test (for example, radiologists or nuclear medicine physicians) can incorporate the institution-specific sensitivity and specificity into probabilistic interpretations of test results that appear in the report to the patient's primary physician.

REFERENCES

College of American Pathologists. Comprehensive Chemistry 1987 Survey. Skokie, Ill., College of American Pathologists, 1987.

Inouye, S.K., and Sox, H.C. A comparison of computed tomography and standard tomographyin neoplasms of the chest. Annals of Internal Medicine 105:906-924, 1986.

Philbrick, J.T., Horwitz, R.I., Feinstein, A.R., et al. The limited spectrum of patients studied in exercise test research: Analyzing the tip of the iceberg. Journal of the American Medical Association 248:2467-2470, 1982.

Roe, W., Anderson, M., Gong, J., et al. A Forward Plan for Medicare Coverage and Technology Assessment. Volume II: Supporting Documentation. Washington, D.C., U.S. Department of Health and Human Services, 1987.

7

Problems of Multi-Institutional Studies

In Chapter 6 we suggested a multi-institutional approach to conducting technology assessments. This strategy has several merits (Sox 1986). These include access to a larger patient population, which could reduce the time needed to obtain the required number of study subjects. Findings from the potentially more diverse population of a multicenter study might be more easily generalized to a wider patient population. In addition, the pooled resources of a number of centers, both in terms of expertise and facilities, will be greater than the resources of a single center (Meinert 1980). To obtain these advantages, however, those planning the assessment must pay careful attention to some requirements and potential problems in four general areas: (1) study structure and organization, (2) study design and protocol development, (3) patient recruitment, and (4) quality control and monitoring.

STUDY STRUCTURE AND ORGANIZATION

Most of the following structural requirements pertain to multi-institutional studies in general; they are not unique to studies of diagnostic technology.[1] All multicenter studies must have a well-defined organizational structure if adequate communication and monitoring are to occur (Meinert 1980). Figure 7.1 depicts an arrangement suggested by a com

[1] This section draws extensively on two reviews of the organization of collaborative clinical trials, Ederer (1975) and Meinert (1981).

mittee of the National Advisory Heart Council (as adapted by Ederer 1975). At the top of the organizational chart is the chairperson, who must be willing to invest a considerable portion of his or her time and to take full responsibility for coordinating the study. This person must be able to provide strong leadership and be sensitive to the politics of the study group.

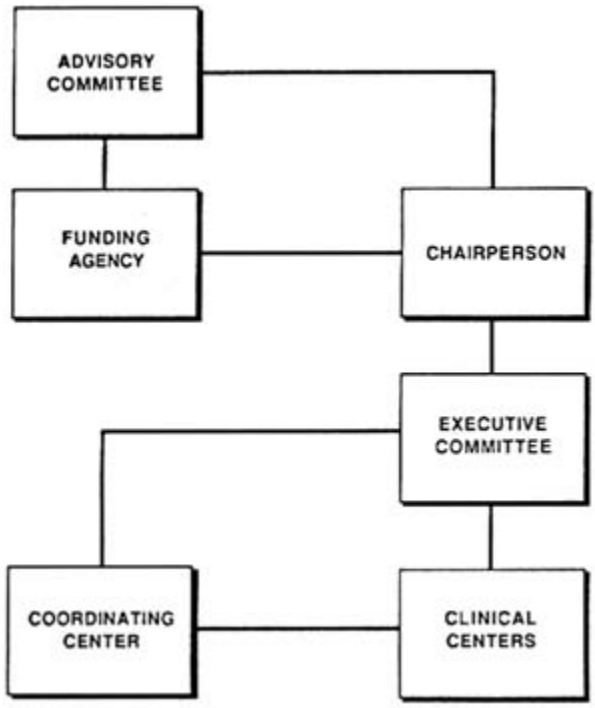

FIGURE 7.1 Organization of a multi-institutional study.

Steering Committee

The steering committee would be composed of principal investigators from the major participating clinical centers and would be responsible for designing the protocol, approving protocol changes, and dealing with operational problems. Approval of the protocol must involve all investigators. Nevertheless, if the number of centers involved in the study is very large, a much smaller subset of the steering committee, an executive committee, may be needed to make timely decisions. The task of control

ling performance would be shared by the study chairperson and the steering committee.

Coordinating Center

This important component of the study would serve a variety of functions, including preparing the manual of operations, developing and pretesting data collection forms, and randomizing the patients to the different arms of the trial. The center would develop the statistical design for the study and would also be responsible for data analysis; its staff would therefore include a full-time biostatistician. In addition, follow-up interviews could be done by telephone from the center to ensure uniformity.

Perhaps the most important function of the coordinating center is its monitoring function. With centralized data management, the data would be monitored for quality, and periodically edited and analyzed.

The coordinating center would be able to detect major drops in the level of participation at any of the clinical units. The center should be in a separate location from the funding agency, which may have a stake in a particular outcome, and from any of the clinical centers, which may try to "dump" extra duties on a center conveniently located within their walls.

Advisory Committee

A group of investigators who are not contributing data to the study would form an advisory committee to review the study and protocol design, recommend changes, adjudicate controversies, and make suggestions about adding or dropping centers. This committee would also advise the sponsoring agency on the design and progress of the trial. These individuals, having no responsibility for the care of patients in the study, would evaluate interim data from the coordinating center for trends that indicate that the study should be terminated early. For example, it would be unethical to continue a study if it became clear that one of the tests was clearly better or had serious unexpected side effects.

Central Observers

Finally, central laboratories or observers may be needed to ensure consistent performance throughout the centers. A patient's entry into the study often depends on the value of a particular laboratory test (for example, the serum glucose) or a specific finding on a diagnostic test

(such as a chest X ray). If such tests are not performed and interpreted in a standardized manner, two patients with the same true state may be evaluated at different centers, but only one is included in the study. In the University Group Diabetes Program (UGDP) study, admission to the study was based in part on the results of a glucose tolerance test, which determined the level of *whole blood glucose*. Four of the clinics, however, substituted *serum glucose* levels for at least a portion of the study (Feinstein 1971). The serum glucose level that defines diabetes mellitus is 20 mg/dl higher than the blood glucose level used to define this disease. Thus, several centers enrolled fewer mild diabetics than did others. Similarly, evaluating the study endpoint may require tests with inherent variability. Eliminating interinstitutional variation is especially important for tests whose results will be used to decide between (for example) inclusion or exclusion and test success or failure (Kahn 1979).

This multitude of requirements indicates the complexity involved in organizing a cooperative study. Recruiting most of the necessary personnel should not be a major obstacle. As we discussed in Chapter 4, however, the requirement for central observers may pose a serious problem. Compensating highly trained subspecialists for devoting a large amount of time to the study could be quite costly. Individuals willing or able to give up time from other commitments may be difficult to find. The ongoing program of technology assessment we have proposed would employ its own staff of subspecialists.

STUDY DESIGN AND PROTOCOL DEVELOPMENT

With the framework of a cooperative study in mind, we can now examine the activities at each level of the organizational chart in more detail. We will begin with the most important tasks of the steering committee: study design and protocol development.

Focus and Compromise

The study design should begin with a carefully identified objective. The protocol must be very detailed and precise. The active participation of many principal investigators with different areas of expertise can produce synergy, but it may produce antagonism as well. The input of the individual investigators, representing a variety of disciplines, may result in a study objective that is too open-ended or overly ambitious (Machin et al. 1979).

Trying to obtain as much data as possible in order to satisfy everyone's requirements may produce a trial that has too many hypotheses. These, in turn, pose a statistical problem, because the larger the number of comparisons, the more likely it is that a difference will appear statistically significant when it is the result of random fluctuations in sampling. Also, the burden of gathering the additional data may be impractically heavy. For example, suppose a patient has agreed to participate and the study has enough funding to monitor his or her clinical condition. The investigators may be tempted to use a diagnostic imaging technique to answer any number of questions about the progress of the patient's disease, rather than—for example—focusing on the technique's ability to detect metastases to the liver. In studies of diagnostic technology, additional procedures may result in both inconvenience and increased costs for the patient. Often the result is poor adherence to the study protocol and patients who withdraw from the study.

Cooperation in a multicenter study will require compromise. When the sterring committee makes a decision about the study objective or the protocol, they must reach unanimity because, as the statistical coordinators of the UGDP study put it, "a majority decision cannot be a substitute where professional ethics and scientific conviction are concerned" (Klimt and Meinert 1966, p. 343). Thus, in a multicenter study, the time needed to agree on a final protocol will be greater than in a single center. A principal investigator who perceives omissions or objectionable provisions in the protocol may refuse to participate from the beginning. Alternatively, he or she may try to find a way "around" the problem or may drop out once the study is in progress. For example, in the methods paper of the extracranial/intracranial (EC/IC) bypass study, the authors state that "some centers have joined the trial with a commitment to exclude patients with no symptoms since their initial carotid occlusions were demonstrated," although such patients were eligible for the study (EC/IC Study Group 1985, p. 399). Compromises aimed at ensuring the participation of particular centers are less problematic if they are made explicit, yet the potential for introducing bias should not be underestimated. In addition, achieving compromise may mean using methods or procedures that "represent a level of consensus which is less than the best scientific basis available" (Klimt and Meinert 1966, p. 343).

Reproducibility Versus Generalizability

In the effort to construct a reproducible protocol, the investigators may choose methods that are agreeable to all but are not widely used in the

day-to-day practice of medicine. There may be a trade-off in the study design between the need for objectivity and precision and the acquisition of clinically useful or relevant data. In the UGDP study, the standard clinical assessment of peripheral neuropathy—which examines touch, pain, and tibial perception of vibration—was replaced by a biothesiometric measurement that was presumably more objective but was not clinically practical. According to Feinstein's critique of the study, the only published results from the biothesiometric procedure were for assessments of vibration in the right index finger, and thus they had "an uncertain pertinence for the problem of peripheral neuropathy in the *legs*" (Feinstein 1971, p. 176).

Techniques providing highly objective, precise results may be more reproducible than traditional clinical methods, and they may be adopted to facilitate standardization between centers. Difficulty arises when clinicians wish to determine if the study results apply to their patients but do not have access to the method used in the study. The data on the characteristics of the study patients obtained with the "foreign" procedure cannot be easily translated into familiar clinical information. If the study does not provide data on these same characteristics, obtained with a commonly used procedure, the physician will be unable to make the necessary comparison and will be uncertain about whether a particular patient is like the study population. Thus, there may be a loss of generalizability of the results of the study.

The need for consistency may result, therefore, in substituting a highly objective para-clinical method for a clinical method that is generally used. It may also lead to using a common procedure in an unusual fashion. For example, the dose of a hypoglycemic agent prescribed for a diabetic patient is usually flexible and is changed according to changes in the patient's status. The UGDP study protocol, however, specified "arbitrarily chosen fixed dosages that were maintained invariantly throughout the project unless the patient dropped out or developed major untoward events" (p. 170). Deviation from usual clinical practice may also be used to facilitate statistical analysis or to maintain blinding (Feinstein 1971). Although standardization among centers is an important goal, departures from ordinary practice may lead to a protocol that is difficult to follow and is frequently misinterpreted.

The protocol for studies of diagnostic technology must specify the procedure for interpreting a test and how the results are to be expressed. These methods must be the same in all centers if data from the different centers are to be combined. Thus, the most objective method for interpre

tation may appear to be the one of choice, although it may not be the one that is commonly used in clinical practice.

Suppose the criterion for including patients in a study is based on the size of a pulmonary nodule on a chest X ray. Patients with nodules of a certain size are then randomized to either of two imaging technologies to determine which is most effective at detecting calcification of the nodule (which indicates a benign condition). According to the study protocol, scanning densitometry is to be used to measure the nodule, because the customary practice of using a ruler is presumed to be too difficult to reproduce. Without access to a densitometer to scan the patient's radiograph, a clinician might find it difficult to determine which of the tests evaluated by the study her patient should receive. The result may be unnecessary adoption of densitometry by physicians who feel compelled to use it in order to apply the study findings to their patients.

Variation in the equipment used by the participating centers is a related problem (McNeil et al. 1981). Which type of equipment should the protocol specify, if any? The problem posed by such variations is that differing performance characteristics of, for example, two CT scanners, one used at center A and the other at center B, may obscure the difference between CT scanning and another imaging technique to which it is being compared.

PATIENT RECRUITMENT

Before recruiting patients, it is important to calculate how many will be needed to answer the study question (Freiman et al. 1978). For example, one might need to determine how large a patient population would be required to demonstrate a difference of a given magnitude between the ability of two tests to detect a lesion in the brain. Because the study population of a multicenter trial is likely to be more heterogeneous than the population from a single center, the sample variance within the intervention groups is likely to be greater. Thus, a larger number of patients may be required in a multicenter study to detect a specified difference between tests. If the condition to be detected by the test is sufficiently rare, it may be difficult to recruit a large enough sample.

The generalizability of the results depends on how study patients compare with all patients with the disease being examined. One advantage of a multi-institutional study is that its combined patient population may more closely approximate the "real world" than the patients from a single institution. As we pointed out earlier, however, a larger sample size

may be needed to generate the statistical power required to answer the study question, and simply adding more patients to the sample may not be adequate. This may mean using a long list of exclusion criteria to try to produce greater homogeneity in the study sample, which in turn would require keeping a log of data on the excluded patients, to verify the representativeness of the population. Imposing highly restrictive exclusion criteria may contribute to increased statistical precision, while detracting considerably from the generalizability of the study results. In addition, stringent exclusion criteria may make it difficult to obtain the required number of patients. Thus, efforts to increase power by reducing heterogeneity may be offset when the study fails to enroll the target number of patients.

Insufficiently explicit inclusion and exclusion criteria can present serious problems as well. If too much room is left for subjective judgment, application of the eligibility criteria may not be uniform at the different centers. The variation may be "random," with each clinic using slightly different interpretations of the criteria. The result may be a diagnostically nonuniform patient population. For example, in the UGDP study, "the clinic physicians used their judgment to screen all patients for absence of life-endangering diseases so as to obtain patients with a minimum life expectancy of five years" (Feinstein 1971, p. 171). There are no quantitative rules for making such prognostic judgments, and specific guidelines were not created for the study. As Feinstein points out, although such criteria would not have guaranteed *correct* predictions, the predictions at the various clinics would at least have been *consistent* (Feinstein 1971).

The result of absent or inconsistent application of admission criteria becomes obvious if the different clinics obtain markedly different results for the same arm of the study. Pooling the data and doing valid statistical analyses may be impossible. In the UGDP study, the widely disparate mortality results obtained with tolbutamide treatment at the different clinics clearly indicated that the tolbutamide treatment groups did not make up a homogeneous population. To use the combined data, a retrospective stratification into groups with similar risk of death was required. Because increased mortality was not an anticipated result of the study, however, baseline information on risk factors for death was never obtained. In some studies, appropriate corrections for differences between clinics may not be possible and a true difference, for example, may fail to achieve statistical significance because of a wide variance.

Besides being uniformly applied, the criteria used to establish the patient's baseline state following enrollment must also be highly specific.

All clinics must use the same specifications for the diagnosis of particular conditions, especially for conditions with a broad clinical spectrum, such as angina pectoris, or if changes in the severity of the conditions are important in evaluating the outcome of the study. Nonuniformity, whether it occurs in the process of enrollment or during the initial workup, can profoundly affect generalizability as well as validity. When patients are excluded for ill-defined reasons, or when they are inadequately characterized, the composition of the study population is unknown, and the study results cannot be applied with confidence to any one patient.

QUALITY CONTROL AND MONITORING

The importance of monitoring in a multi-institutional study cannot be overestimated. Even the most carefully designed study may fail to produce valuable information if the performance of individual centers is not adequately monitored to ensure correctness and thoroughness in the administration of the protocol. Accurate and efficient data entry requires the individuals participating in data collection to cooperate with the staff of the coordinating center. Standardized data forms, designed specifically for the study, must be filled out and returned to the coordinating center without delay (Gaus 1979).

Adherence to the protocol may vary from center to center. With increasing delay between acquiring the data from the patient and completing the data forms, the likelihood of errors or omissions increases. Similarly, there may be substantial delay between the time the forms are completed and when they appear at the coordinating center (Marks et al. 1984). Thus, an error (for example, a patient admitted to the trial by mistake) may not be discovered by the monitoring committee until much later, after considerable time and energy have been invested. In the UGDP study, 69 patients who did not meet the diagnostic criteria for admission to the study were nevertheless included (Feinstein 1971).

Alternatively, some clinics may routinely fail to obtain all the necessary baseline information. In the UGDP study, data specified by the protocol were never obtained for some patients. For example, 311 patients did not have retinal photographs. As Feinstein (1971) points out (p. 177), the absence of this data "adds to the subsequent problems of evaluating risk factors and transitions in groups whose denominator is substantially reduced by the omissions." Another reason for monitoring is to discover attempts to "tamper" with randomization. The test groups from the center with the altered scheme will suffer from selection bias. If such bias is not

detected until late in the study, the biased groups may have to be dropped from the analysis, and the study may fall short of the required number of patients.

Some centers may fail to perform a test or give a treatment as specified by the protocol. In a study comparing the use of computed tomography (CT) with radionuclide (RN) studies in patients with intracranial disease, one institution used a mercury isotope and an unusual type of imaging instrument—although the protocol had carefully specified sodium pertechnetate (McNeil 1979). The cases from this institution, 20 percent of the total, could not be used to analyze the relative lesion detection capacities of CT and RN. Another institution did not follow the provision calling for a minimum of eight cross-sectional views with the CT scanner. Follow-up may be inadequate or may not occur at all. For example, in a study designed to assess the efficacy of a diagnostic technology, some centers may not proceed with the gold-standard test to verify the absence of disease in patients with a negative index test. Inadequate follow-up was a substantial problem in the CT/RN study (McNeil 1979).

For cooperative studies of diagnostic technology, poor technical quality control can be especially disastrous. Large numbers of technically suboptimal exams on eligible patients will reduce the effective number of cases in the study. Hidden bias may enter the study at this point—that is, when patients are removed from the study because of "nondiagnostic or technically unsatisfactory" exams (McNeil 1979). It is important to guard against evaluating only patients on whom the technology "works" (Philbrick et al. 1982). For example, suppose the performance of a new imaging technique is being studied. If some of the "substandard" examinations are really false negatives, and these patients are withdrawn from the study, the sensitivity for the new technique may be artificially inflated (Begg et al. 1986). In a multi-institutional study, the technical problems that limit reproducibility of tests at each of the participating centers must also be considered. Skills of individuals may vary (Harris 1981), and in a multicenter technology assessment, comparable interpretative skills may be just as important as comparable equipment.

Each center's enrollment rate must also be monitored. "Minor" participants, that is, centers that contribute less than a specified number of cases, would have less experience with the protocol and/or the technology involved. Some evidence supports the contention that the quality of the participation of minor participants is lower than that of "major" participants (Sylvester et al. 1981). The cases from minor centers may make a greater contribution to the variance, offsetting any benefit derived from having increased the total number of cases (Gaus 1979).

Poor quality control and inadequate monitoring thus present a number of hazards. In the best case, deviations from the study protocol would be prevented, or at least caught as they occurred. Early detection would reduce the number of patients who were admitted and then dropped at a later date, and it would minimize the costs and resources attendant to such mistakes. Additional patients could be enrolled, and corrections could be made if biases were discovered. If errors are not discovered within a reasonable amount of time, the number of valid cases available for statistical analysis may be reduced. Generalizability may be compromised. In the worst case, poor quality control and inconsistencies would not be detected, and the study conclusion would be erroneous. If the result translates into a general policy that leads the medical community to adopt a less efficient technology or retire one that is still useful, patient care may suffer.

SUMMARY AND CONCLUSION

This chapter has presented some of the difficulties that may be encountered when a multi-institutional framework is used for the assessment of diagnostic technologies. The multi-institutional study requires a more well-defined organizational structure than a single-center study. There must be strong leadership to coordinate the various units, open communication between them, and ensure continuous monitoring. More time will be needed to plan a multicenter trial because the investigators must agree on a focused study question and then approve a protocol designed to answer that question.

Once a multi-institutional study is underway, the greatest challenge will be to obtain a uniform data set, that is, to remove sources of spurious variability between the centers. The validity of statistical analyses on pooled data requires that each center obtain similar results for similar patients. Thus, the protocol in a multi-institutional study must be very precise. It must be easy to follow and reproducible in each of the centers. "Reproducibility," however, must be used with caution as a criterion for protocol design. Data obtained using the most reproducible and objective methods will not be useful if such methods are not those of usual clinical practice. All centers must have access to comparable equipment and technical expertise.

Problems may also arise at the level of patient enrollment. The protocol should include explicit information about which patients are to be enrolled and which are to be excluded. Ideally, all patients who are to have the test under study at a given institution would be enrolled. One of

the chief advantages of the multi-institutional study is access to a wider spectrum of patients, yielding a study population that closely represents the spectrum of patients encountered in clinical practice. If, however, each center *enrolls* a qualitatively different population of patients, pooled analyses may be invalid or impossible. Many of these problems can be avoided or corrected with adequate monitoring. In addition to patient enrollment, protocol adherence and data quality at each center must be monitored continuously. Monitoring is the key to ensuring a uniform data set when many centers are involved.

The multi-institutional model for assessment of diagnostic technology has many advantages (see Chapter 6). Even more than the single-center model, however, it requires careful organization, commitment of those involved, extensive planning, and monitoring if the study is to succeed. When these prerequisites are met, such studies may provide valuable clinical information that would otherwise be impossible to obtain.

REFERENCES

Begg, C.B., Greenes, R.A., and Iglewicz, B. The influence of uninterpretability on the assessment of diagnostic tests. Journal of Chronic Diseases 39:575-584, 1986.

The Coronary Drug Project Research Group. Practical aspects of decision making in clinical trials: The coronary drug project as a case study. Controlled Clinical Trials 1:363-376, 1981.

The EC/IC Study Group. The international cooperative study of extracranial/intracranial arterial anastomosis (EC/IC bypass study): Methodology and entry characteristics. Stroke 16 (3):397-406, 1985.

Ederer, F. Practical problems in collaborative clinical trials. American Journal of Epidemiology 102:11-118, 1975.

Feinstein, A.R. Clinical biostatistics—VIII. An analytic appraisal of the University Group Diabetes Program (UGDP) study. Clinical Pharmacology and Therapeutics 12(2):167-191, 1971.

Freiman, J.A., Chalmers, T.C., Smith, H., Jr., and Kuebler, R.R. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. New England Journal of Medicine 299:690-694, 1978.

Friedewald, W.T., and Levy, R.I. Planning and implementation of large clinical trials. Israel Journal of Medical Sciences 22:191-196, 1986.

Gaus, W. The experience of the EORTC-Gnotobiotic Project Group in planning, organizing, and evaluating cooperative clinical trials. Proceedings

of an EORTC symposium, Brussels, Belgium, April 26-29, 1978. In Tagnon, H.J., and Staquet, M.J., eds., Controversies in Cancer: Design of Trials and Treatment. New York, Mason Publishers, 1979.

Harris, J.M. The hazards of bedside Bayes. Journal of the American Medical Association 246:2602-2605, 1981.

Kahn, H.A. Diagnostic standardization. In Roth, H.P., and Gordon, R.S., Jr., eds., Proceedings of the National Conference on Clinical Trials Methodology, October 1977. Clinical Pharmacology and Therapeutics 25:703-711, 1979.

Klimt, C.R., and Meinert, C.L. The design and methods of cooperative therapeutic trials with examples from a study on diabetes. Chapter 19 (pp. 341-373), in International Encyclopedia of Pharmacology and Therapeutics, Clinical Pharmacology, Vol. 1. New York, Pergamon Press, 1966.

Machin, D., Staquet, M.J., and Sylvester, R.J. Advantages and defects of single-center and multicenter clinical trials. Proceedings of an EORTC symposium, Brussels, Belgium, April 26-29, 1978. In Tagnon, H.J., and Staquet, M.J., eds., Controversies in Cancer: Design of Trials and Treatment. New York, Mason Publishers, 1979.

Marks, J.W., Croke, G., Gochman, N., et al. Major issues in the organization and implementation of the National Cooperative Gallstone Study (NCGS). Controlled Clinical Trials 5:1-12, 1984.

McNeil B.J. Pitfalls in and requirements for evaluations of diagnostic technologies. In Wagner, J., ed., Proceedings of a Conference on Medical Technologies. DHEW Pub. No (PHS) 79-3254. Washington, D.C., U.S. Government Printing Office, 1979:33-39.

McNeil, B.J., Sanders, R., Alderson, P.O., et al. A prospective study of computed tomography, ultrasound, and gallium imaging in patients with fever. Radiology 139:647-653, 1981.

Meinert, C.L. Toward more definitive clinical trials. Controlled Clinical Trials 1:249-261, 1980.

Meinert, C.L. Organization of multicenter clinical trials. Controlled Clinical Trials 1:305-312, 1981.

Philbrick, J.T., Horwitz, R.I., Feinstein, A.R., et al. The limited spectrum of patients studied in exercise test research. Analyzing the tip of the iceberg. Journal of the American Medical Association 248:2467-2470, 1982.

Sox, H.C., Jr. Centers of excellence in technology assessment: A proposal for a national program for the study of health care technology.

In A Forward Plan for Medicare Coverage and Technology Assessment. Report to the Assistant Secretary for Health and Human Services. Washington, D.C, Lewin & Associates, 1986.

Sylvester, R.J., Pinedo, H.M., De Pauw, M., et al. Quality of institutional participation in multicenter clinical trials. New England Journal of Medicine 305:852-855, 1981.

Weiss, D.G., Williford, W.O., Collins, J.F., and Bingham, S.F. Planning multicenter clinical trials: A biostatistician's perspective. Controlled Clinical Trials 4:53-64, 1983.

The Authors

DR. HERBERT L. ABRAMS is professor of radiology at Stanford University School of Medicine and was formerly the Philip H. Cook Professor and chairman of radiology at Harvard Medical School. A member of the Council on Health Care Technology of the Institute of Medicine–National Academy of Sciences, he is co-chairman of the Methods Panel of the Council. He has played an active role in diagnostic technology assessment during the past two decades, and he is the coeditor of a book on the rational utilization of diagnostic imaging procedures.

DR. DOUGLAS OWENS is a clinical assistant professor of medicine at Stanford University School of Medicine, with a major involvement in health policy research. His previous work involves methodologic approaches to technology assessment and test selection.

DR. HAROLD SOX is professor and chairman of the Department of Medicine, Dartmouth Medical School, Hanover, New Hampshire, and formerly professor of medicine at Stanford University School of Medicine. He is the author of numerous papers on health policy and technology assessment, as well as a book on clinical decisionmaking.

SUSAN STERN is a fourth-year medical student at Stanford University School of Medicine, with a major interest in health policy research.