Evaluation of Medical Practices

Howard S. Frazier and Howard H. Hiatt

Summary. Evaluation of the efficacy of a medical intervention requires valid measurements of both its benefits and risks as compared to those of alternative forms of management. The requisite measurements are more difficult to make than this simple description suggests, and the accumulation of information is further inhibited by certain characteristics of our pattern of health care. These features include, for example, discontinuous care by a variety of unrelated providers, inadequate records, the autonomy of physicians as decision-makers, financial disincentives, ambiguities in what we mean by "experimental" and "accepted" forms of therapy, and failure to see continuing evaluation as a necessary component of the cost of providing good medical care. Although no single change will solve all the problems of evaluation, several offer promise of improving our ability to choose from among medical interventions those most likely to be useful.

Dacron grafts have saved the lives of thousands of people with potentially lethal defects in atherosclerotic blood vessels. Antibiotics have spared millions more the catastrophic outcomes of deep-seated bacterial infections. At the other end of the spectrum of medical interventions are those that are carried out exclusively for monetary gain in, for example, a few "Medicaid mills." If medical measures were always directed against acute disease that is lethal in the absence of intervention, the problem of evaluation would be a relatively simple one. Similarly, if the only, or principal, interventions in question were those prescribed by unethical doctors, we could construct corrective measures with reasonable dispatch. Evaluation of medical procedures, however, is a complicated undertaking that defies simplistic or rapid solution. Although the thesis of this article is that we can do better, we believe that our problems do not result from greed, stupidity, or sloth. Rather, they can be ascribed largely to pervasive characteristics of our medical care system and to complexities in the meanings and measurement of risks and benefits.

We shall use the term efficacy to describe the net of benefits and risks of an intervention. Its measurement requires the ordered collection of information about the natural history of the disease for which the intervention is proposed, and the short- and long-term risks and benefits that result from the intervention. Both society and the individual have a stake in the measurement of efficacy. Society is also concerned with the fair representation of the available information on risks and benefits to the individual that result from medical interventions. Further, it has a stake in aggregate risk and benefit, at least along that part of the spectrum that includes death and change in level of disability. Finally, as the largest single source of financial support for health care, society has a growing interest in the cost per unit of net benefit, or the cost-effectiveness of medical interventions.

Empirical information about risk and benefit is probabilistic in form. Because outcomes are value-laden, their measurement often is subject to unconscious bias or other forms of implicit selection. Since these difficulties are common to many other forms of scientific inquiry, we shall refer only briefly to them and consider in more detail some features of our present system of health care which inhibit seriously our ability to develop the kind of information upon which the measurements of efficacy and cost-effectiveness depend. Finally, we shall offer some suggestion designed to palliate, if not to cure.

H. S. Frazier is director of the Center for the Analysis of Health Practices, Harvard School of Public Health, and H. H. Hiatt is dean of the Harvard School of Public Health, Boston, Massachusetts 02115. Both are professors of medicine at the Harvard Medical School.

Examples of the Problem

The evolution of medical practice requires, one might assume, the continual replacement of diagnostic, therapeutic, and preventive measures with ones of greater efficacy. However, the history of medical practice is replete with descriptions of procedures that have been widely employed, only to be discarded when they have been shown to be seriously flawed rather than after the introduction of a more effective measure. For example, Barnes (*1*) describes many "therapeutic" measures that were ultimately abandoned because they were proved completely worthless, including dangerous surgical procedures carried out for such conditions as constipation and dysmenorrhea.

As one example of a diagnostic measure of considerable efficacy but with limitations that were inadequately appreciated for a long period, McDermott (*2*) cites the Wassermann diagnostic test for syphilis. Only after almost 50 years of widespread use was it generally appreciated that half the patients with positive tests did not have syphilis. As a result, many thousands of people were incorrectly diagnosed as having the disease and thereby subjected not only to humiliation, but to the very considerable dangers that then accompanied antisyphilitic treatment.

Many examples can be cited of diagnostic procedures now in widespread use that have not been adequately validated. Fineberg (*3*) has described the proliferating laboratory tests that last year accounted for in excess of $11 billion of our health resource expenditures. He pointed out that there is evidence that much laboratory usage, which is increasing at a rate of 14 percent annually, has little or no beneficial effect on patient care.

Many current therapeutic practices, both medical and surgical, have not been adequately evaluated. For example, over the last 25 years controlled studies have led to a progressive reduction in hospital stay for convalescence from myocardial infarction from 6 weeks to 4 weeks to 3 weeks to 2 weeks and recently to 1 week (*4*). Indeed, some English physicians (*5*) maintain that patients with uncomplicated myocardial infarctions can be cared for at home with no greater mortality than in the hospital. Radical mastectomy remains the most widely practiced operation for breast cancer, but much evidence suggests that it is no more effective than simpler procedures (*6*). Tonsillectomy may be less frequently carried out now than previously, but one can

0036-8075/78/0526-0875$01.00/0

question the justification for most of the almost 1 million operations still performed annually (*7*).

Limitations in Our Knowledge of the Natural History of Disease

Many factors contribute to the complexity of evaluation, but none is more prominent than deficiencies in our understanding of the natural history of most of the chronic illnesses treated in our health care system. McDermott (*2*) cites as a prime example the evolution of our understanding of the late effects of syphilis. Until 20 years ago it was almost universally assumed that a patient with untreated syphilis was doomed to a disastrous outcome as a result of such infection-induced complications as severe mental disease, debilitating neurological problems, serious cardiovascular disease, or a combination of these and others.

Therefore, even though treatment in the pre-antibiotic era carried considerable risk, it was considered essential. Not until 1955 did a prolonged follow-up study of untreated patients by Gjestland (*8*) show that 85 percent of people with untreated syphilis had a normal life-span and that more than 70 percent died without any evidence of the disease. With the advent of antibiotics the risks and benefits of treating syphilis have changed dramatically, and treatment is now appropriately regarded as mandatory for all patients in whom a diagnosis of active disease is made. However, many other illnesses for which risk-laden interventions are now assumed essential probably have a similar natural history that is presently unappreciated. In the absence of baseline data of untreated illness, conclusions are necessarily uncertain concerning the effectiveness of intervention.

Imperfect understanding of the course of both untreated and treated disease, however, is but one of the difficulties confronting those who would attempt evaluation of medical interventions. We shall cite several others.

The Structure of Our Health Care System

Medical care for many Americans is discontinuous. In the absence of a primary care physician, comprehensive medical record keeping is very difficult. We are unable to follow adequately the effects of tonsillectomy on, for example, the frequency of sore throats in the 5 or 10 years following the operation. We cannot even assess the accuracy of the patient's impression of their frequency in the years before the tonsillectomy that is often carried out in response to that impression (*9*). [The problem is made more complex by a limited appreciation that tonsil and ear infections occur less often in unoperated children as they grow older (*7*).] The tendency of many Americans to choose the specialist to whom they take a self-diagnosed problem and to use multiple hospitals accentuates the lack of follow-up. Record systems, even for those patients with primary care physicians, do not begin to have the sophistication of which our technology is capable and that adequate data collection and interpretation demand.

Recent developments in the technology of computers have made possible experiments in the development of specialized data banks for the storage and study of large, standardized data sets concerning patients with a restricted group of medical problems who can be followed over prolonged periods (*10, 11*). Appropriate safeguards of privacy appear feasible, and the aggregated experience may well make important contributions to efficacy questions in the foreseeable future.

Physicians function as largely autonomous decision-makers. Drugs must, of course, be approved by the Food and Drug Administration before they can be adopted for general use. However, the same is not true of other measures employed for diagnosis and treatment. For example, although there is widespread agreement that the coronary artery bypass graft surgical procedure has not been adequately validated, it is estimated that in excess of 80,000 Americans will be subjected to this operation this year. While some people derive major benefit from the operation, many others now operated on have characteristics that most experts agree should exclude them as candidates for the procedure. The independence of what must now be hundreds of surgeons carrying out the procedure makes it impossible to pool their experience in such a way as to maximize the learning experience for them and for society. Thus, a recently published report of a randomized trial of this operation involved about 600 patients studied over a period of up to 3 years (*12*). The conclusions reached from this limited experience have led to great controversy in the medical literature (*13*). We might know more if it were possible to summarize and analyze the experience of the estimated 250,000 patients who have simultaneously been subjected to the procedure under nonexperimental conditions. We would surely know more if all patients had been operated on under experimental conditions.

Reimbursement mechanisms in our present health care system often provide incentives that interfere with proper evaluation. For example, most medical insurance plans cover hospitalization but not home care. Thus, a surgeon wishing to examine experimentally the effects of early discharge of patients subjected to, for example, herniorrhaphy, varicose vein surgery, or gallbladder surgery could do so only by increasing the cost to the patient, if such a measure as home nursing were judged important for adequate follow-up care and were not covered by insurance. Similarly, a group of doctors wishing to pursue the conclusion that home care is equivalent to hospital care for the patient with the uncomplicated myocardial infarction (*5*) could do so in most instances only at great expense to the patient.

In addition, the process by which medical interventions move from "experimental"—and hence not reimbursable—status to that of "accepted practice" depends on criteria that are both insufficiently explicit and variably applied (*14*). The collection of appropriate and credible information on efficacy of new procedures prior to their widespread dissemination could be enhanced by the selective reimbursement of investigators contingent on their use of peer-reviewed protocols during clinical trials of new or controversial interventions. Upon completion of the trials, a well-supported judgment of the relative merit of the therapy could be made, and more general reimbursement begun if appropriate.

The present preoccupation of some in our society with legal action against physicians also militates against adequate evaluation. For example, a large fraction of all patients operated on for breast cancer ultimately die of their disease. Will the surgeon who carries out a simple mastectomy be liable to criticism or even legal suit if the patient unhappily is not cured, and if most surgeons in the community are still recommending radical surgery? Similarly, many physicians order skull x-rays for patients even with minimal head trauma despite studies (*15*) that indicate the efficacy is very low. Today's imperative is a skull film; tomorrow's may be a computerized tomographic scan, at roughly ten times the cost (*16*).

Evaluation Is Not Exclusively a Medical Undertaking

As a result of the interactions of biology and medicine we have seen major progress in recent years in the diagnosis and treatment of disease. However, there has been much less success in combining the activities of physicians and members of other disciplines. Closer interactions with statisticians, for example, might have accelerated the development of clinical trials sophisticated in design and evaluation. Even more difficult is the measurement of quality of life. Most studies describing the results of surgical and medical procedures are couched in terms of survival rates after a certain number of years and recurrence of the problem for which the procedure was applied (*17*). The quality of life that follows the intervention seldom receives attention. In part, this reflects the fact that the physicians are not trained to seek ways of assessing quality of life any more than they are trained in the mathematical techniques required for the development of clinical trial methodology. The inability of medical people to deal optimally with such problems and the diffidence of many nonmedical professionals with the requisite skills to enter the medical arena have had adverse effects not only on the needed assessments, but also on the development of important methodology.

Disappointment often follows the completion of even a well-planned and executed trial because of unrealistic expectations. Many people believe that real advances in medicine take place only in dramatic fashion, such as when a trial with polio vaccine demonstrated virtually complete control of a condition that was previously unmanageable. However, this is clearly the exception. Most innovations in surgery and anesthesia, as described by Gilbert *et al.* (*18*), and probably in other medical spheres as well, occur in much less dramatic fashion. These authors conclude (*18*, p. 143) that "Since relatively small, though important, numerical gains or losses are to be expected from most innovations, clinical trials must regularly be designed to detect these differences accurately and reliably. When a systematic trial of a new therapy is being considered initially, it is frequently subject to great optimism. The advocate for a new therapy may believe, for a number of specific reasons, that it will prove to be greatly superior to the standard and, indeed, preliminary informal experience may seem to provide good evidence for this. As our data show, this initial optimism, though frequently warranted, is often not justified."

Evaluation is not generally perceived as a continuing process. The results may be provider-sensitive; for example, the physicians conducting a trial based in multiple academic centers may be more highly skilled than community-based physicians who have less opportunity to practice these skills. As a possible case in point, in an early series of studies starting in 1964, thermography was first reported to have about 90 percent sensitivity for detection of breast cancer (*19*). By September 1977 the First National Institutes of Health Consensus Development Conference recommended abandoning general use of the procedure because sensitivity as measured in 27 centers involving 258,000 women was only 50 percent, and the false positive rate was high (*20*).

Finally, evaluation must take into account the subjective value systems of physicians and patients. Neutra (*21*) has compared the indications for appendectomy in a variety of countries. In some, surgeons are willing to operate more frequently in an effort to minimize the number of patients with appendicitis whose atypical symptoms might lead to their not being recognized and, therefore, to increased mortality rates. As a result, these surgeons carry out a larger proportion of operations for what ultimately prove to be nonsurgical causes of abdominal pain. Mosteller (*22*) cites this situation as an example of what he terms "the safe surgery dilemma"—in this instance, the incompatible goals shared by both doctor and patient that nobody die of undiagnosed appendicitis, on the one hand, and that nobody with a normal appendix be subjected to appendectomy, on the other. Our skills in recognizing atypical appendicitis are imperfect, and Neutra's data suggest that, at one extreme, saving one additional life may involve so much additional surgery as to result in postsurgical convalescence equivalent to many lifetimes. Such considerations underscore the complexity of the problem of evaluation; a pathology report that a surgically removed appendix was normal does not by itself justify a conclusion that the surgery was unnecessary.

The issue of hysterectomy in postmenopausal women without uterine pathology also heavily involves value systems (*23*). There are those who argue that the removal of a normal uterus is justified in order, for example, to spare the woman the continuing anxiety that the organ may become cancerous. On the other hand, most doctors maintain that such surgery would be in lesser demand if attention were given to periodic examination and if there were more general understanding of both the prevalence and the epidemiology of uterine cancer, on the one hand, and the risks and costs of surgery, on the other.

Ethical Problems

Physicians' primary consideration is the care of the patients for whom they are responsible. Therefore, they can, in good conscience, recommend participation in a randomized trial only if they feel the prospects for benefit are equivalent with the two or more treatments being randomized. When one of the approaches proposed is no treatment, the problem is compounded by the previously described defects in our knowledge of the natural history of the illness being treated.

Premature dissemination of data on unproved intervention may lead subsequently to even greater ethical problems. We have already described the dilemma that may confront surgeons who recommend simple mastectomy for breast cancer, when most of their colleagues use the more radical approach, even though the latter has not been shown to be of greater efficacy. The fact that many other interventions in common practice remain to be fully validated puts into focus the magnitude of the tasks confronting medicine and society.

The Costs of Evaluation

Randomized clinical trials are extremely expensive. At present many are supported by the National Institutes of Health, and there is justifiable concern in the biomedical community that their costs diminish the amount of money available for essential basic biological research and other forms of applied clinical research. It is unfortunate that present arrangements dictate such choices for, as Gilbert *et al.* (*24*) point out, "The cost of trials is part of the development cost of therapy." These authors make two additional points that are also relevant. First, "Sometimes costs of trials are inflated by large factors by including the cost of the therapies that would in any case have been delivered, rather than the marginal cost of the management of the trial." In addition, the costs of trials assume much more realistic proportions

when they are compared "with the losses that will be sustained by a process that is more likely to choose a less desirable therapy and continue to administer it for years. . . . The one sure loser . . . is a society whose patients and physicians fail to submit new therapies to careful unbiased trial and thus fail to exploit the compounding effect over time of the systematic retention of gains and the avoidance of losses."

Conclusions and Recommendations

Uncertainty surrounds the question of safety and effectiveness of a broad range of medical and surgical procedures. The costs in terms of lives, suffering, and dollars are incalculable. The causes are many and complex, and there will be resistance to rapid change. A popular conception of a venal physician willfully prescribing unnecessary interventions is clearly incorrect; such people do exist, but undoubtedly account for only a small fraction of the problem. Much of the difficulty rests in our imperfect understanding of the natural history of the chronic illnesses that lead to the major demands on the resources of the health care system in Western countries. Most interventions for these conditions are designed to improve the quality of life. Since our measures for assessing quality of life are poorly developed, the problem of evaluation is further complicated.

Although no single reform will produce a great improvement in a short period, several reforms could bring about significant progress. Many changes will be difficult to implement, and their benefits will be visible only after several years or decades. However, if we neglect to identify and implement such changes, we doom ourselves and those who follow us to a continuation of the uncertainties, needless risks, and the escalating costs that characterize our present health care system. We shall summarize a few steps that could be constructive.

Medical services should be organized so that primary care and comprehensive health data collection are available to all citizens. Record systems should be kept in uniform or compatible fashion. New procedures should be regarded by physicians and patients as new drugs—not to be used except on an experimental basis until their usefulness has been validated. Methods should be developed so that all physicians and patients involved in a new intervention are enlisted in organized trials. Bunker *et al.* (*25*) have suggested approaches to this problem.

Medical education should be broadened to give much greater emphasis to quantitative analytic methods, including epidemiology and biostatistics. Increased attention also should be given to the social sciences that could help provide approaches to such problems as measurements of quality of life, of costs and benefits, and of the diffusion of medical practices. Simultaneously, physician-investigators should encourage the participation of statisticians, epidemiologists, sociologists, lawyers, and others in research and patient care programs. Courses on health-related topics should be offered to graduate students in other programs besides those in medical school.

Society must understand that a major investment is desirable in the development of methods for the collection, analysis, and storage of medical data. In no other way will we accumulate the required information concerning the natural history of the health problems that afflict our citizens and the effects of intervention. This will be expensive and must be undertaken with an understanding that most payoffs will be deferred for many years. The investment should be made not at the expense of fundamental biological research or applied clinical research. Rather, it must be considered an integral part of medical care.

Many now look to regulation as a means of controlling the use of "unnecessary" or inefficacious procedures. As pointed out elsewhere in this issue (*26*), the protection afforded by the regulatory approach must be balanced against possible social costs resulting from the consequent inhibition of innovation, or the retardation of its spread. More important for the present discussion, credible regulations depend on the existence of adequate evaluations of efficacy. Regulation more commonly intensifies the need for information rather than substitutes for it. The regulatory approach is likely to offer only small advantages to the society that succeeds in making available to its people medical care of demonstrated benefit. To achieve such advantages regulation must be employed only when its benefits outweigh its drawbacks, and when its form is appropriate to the particular need.

References and Notes

1. B. A. Barnes, in *Costs, Risks, and Benefits of Surgery*, J. P. Bunker, B. A. Barnes, F. Mosteller, Eds. (Oxford Univ. Press, New York, 1977), p. 109.
2. W. McDermott, *Daedalus* (winter, 1977), p. 135.
3. H. V. Fineberg, paper presented at the Sun Valley National Forum, Sun Valley, Idaho, 1 to 5 August 1977.
4. J. F. McNeer, G. S. Wagner, P. B. Ginsberg, A. G. Wallace, C. B. McCants, M. J. Conley, R. A. Rosati, *N. Engl. J. Med.* **298**, 229 (1978).
5. H. G. Mather, D. C. Morgan, N. G. Pearson, K. L. Q. Read, D. B. Shaw, G. R. Steed, M. G. Thorne, C. J. Lawrence, I. S. Riley, *Br. Med. J.* **1976-I**, 925 (1976).
6. K. McPherson and M. Fox, in *Costs, Risks, and Benefits of Surgery*, J. P. Bunker, B. A. Barnes, F. Mosteller, Eds. (Oxford Univ. Press, New York, 1977), p. 308.
7. R. J. Haggerty, *N. Engl. J. Med.* **298**, 453 (1978).
8. T. Gjestland, *Acta Derm. Venereol.* **35** (Suppl.) 34 (1955).
9. J. L. Paradise, C. D. Bluestone, R. Z. Bachman, G. Karantonis, I. H. Smith, C. A. Saez, D. K. Colborn, B. S. Bernard, F. H. Taylor, R. H. Schwarzbach, H. Felder, S. E. Stool, A. M. Fitz, K. D. Rogers, *N. Engl. J. Med.* **298**, 409 (1978).
10. R. A. Rosati, J. F. McNeer, C. F. Starmer, B. S. Mittler, J. J. Morris, Jr., A. G. Wallace, *Arch. Int. Med.* **135**, 1017 (1975).
11. S. Weyl, J. F. Fries, G. Weidrhold, *Comput. Biomed. Res.* **8**, 279 (1975).
12. M. L. Murphy, H. N. Hultgren, K. Detre, J. Thompsen, T. Takaro, and participants of the Veterans Administration Cooperative Study, *N. Engl. J. Med.* **297**, 621 (1977).
13. See, for example, "A debate on coronary bypass," *ibid.*, p. 1464.
14. *Computed Tomographic Scanning: A Policy Statement* (Institute of Medicine, National Academy of Sciences, Washington, D.C., 1977).
15. R. S. Bell and J. W. Loop, *N. Engl. J. Med.* **284**, 236 (1971).
16. S. Galbraith, G. Teasdale, C. Blaiklock, *Br. Med. J.* **1976-II**, 1371 (1976).
17. B. McPeek, J. P. Gilbert, F. Mosteller, in *Costs, Risks, and Benefits of Surgery*, J. P. Bunker, B. A. Barnes, F. Mosteller, Eds. (Oxford Univ. Press, New York, 1977), p. 170.
18. J. P. Gilbert, B. McPeek, F. Mosteller, in *ibid.*, p. 124.
19. K. L. Williams, *Ann. N.Y. Acad. Sci.* **121**, 272 (1964); A. G. Swearingen, *Radiology* **85**, 818 (1965); J. Gershon-Cohen, *Ca* **17** (No. 3), 108 (1967); J. D. Haberman, *Proc. Natl. Cancer Conf.* **6**, 157 (1970).
20. H. I. Libshitz, *J. Am. Med. Assoc.* **238**, 1953 (1977); "Mammography screening for breast cancer is the first consensus conference topic," *ibid.* **239**, 486 (1978).
21. R. Neutra, in *Costs, Risks, and Benefits of Surgery*, J. P. Bunker, B. A. Barnes, F. Mosteller, Eds. (Oxford Univ. Press, New York, 1977), p. 277.
22. F. Mosteller, *J. Surg. Res.*, in press.
23. J. P. Bunker, K. McPherson, P. L. Henneman, *Costs, Risks, and Benefits of Surgery*, J. P. Bunker, B. A. Barnes, F. Mosteller, Eds. (Oxford Univ. Press, New York, 1977), p. 262.
24. J. P. Gilbert, B. McPeek, F. Mosteller, *Science* **198**, 684 (1977).
25. J. P. Bunker, D. Hinkley, W. V. McDermott, *ibid.* **200**, 937 (1978).
26. L. Lasagna, *ibid.*, p. 871.

27. We thank H. Fineberg, R. J. Haggerty, F. Mosteller, H. Sherman, and J. Winsten for critiques of the manuscript. The Center for the Analysis of Health Practices is supported by grants from the Robert Wood Johnson Foundation, the Commonwealth Fund, and the Metropolitan Life Insurance Co.