Volume Title: Social Experimentation

Volume Author/Editor: Jerry A. Hausman and David A. Wise, eds.

Volume Publisher: University of Chicago Press

Volume ISBN: 0-226-31940-7

Volume URL: http://www.nber.org/books/haus85-1

Publication Date: 1985

Chapter Title: Toward Evaluating the Cost-Effectiveness of Medical and Social Experiments

Chapter Author: Frederick Mosteller, Milton Weinstein

Chapter URL: http://www.nber.org/chapters/c8377

Chapter pages in book: (p. 221 - 250)

6 Toward Evaluating the Cost-Effectiveness of Medical and Social Experiments

Frederick Mosteller
Milton C. Weinstein

6.1 Introduction

6.1.1 Why Evaluate Medical Experiments?

Although the life expectancy of the U.S. population seems finally to be lengthening, after a prolonged period during which not much improvement was seen (U.S. Department of Health and Human Services 1980), the nation has been increasingly concerned about costs of health care (Fuchs 1974; Hiatt 1975). One possible response to this concern is an accelerated strategy for evaluating the efficacy of medical practices, with the hope that identifying those practices that are not efficacious will lead to their abandonment and, therefore, to substantial savings in health care resources (Cochrane 1972). Undeniably, some medical practices, though costly, may not be efficacious; others may never have had their efficacy evaluated. Often-cited examples are tonsillectomy and adenoidectomy—procedures whose appropriateness has raised doubts for decades, yet only recently has rigorous evaluation of their benefits begun. The Office of Technology Assessment reviewed a substantial number of diagnostic, preventive, and therapeutic practices in several areas of medicine and found that few had been adequately evaluated (U.S. Congress, Office of Technology Assessment 1978).

Frederick Mosteller is professor of mathematical statistics and Milton C. Weinstein is professor of policy and decision sciences, Harvard School of Public Health.

The authors wish to thank John Bailar, Leon Eisenberg, Rashi Fein, Howard Frazier, Alexander Leaf, and Marc Roberts for their comments and suggestions. They are especially indebted to David Freedman, Jay Kadane, and the other participants at the NBER Conference on Social Experimentation for their thoughtful criticism.

This research was supported in part by a grant from the Robert Wood Johnson Foundation to the Center for the Analysis of Health Practices and by National Science Foundation grant SES–75–15702.

An alternative response to the cost problem acknowledges that information on efficacy will not eliminate the need to face trade-offs between increasing incremental costs and diminishing incremental benefits. Given what we know about the benefits offered by available medical technologies, we expect a continuum from more cost-effective to less cost-effective; the more we are willing to spend, the more health we can purchase albeit at increasing incremental costs per unit of benefit. If we want to control costs without sacrificing health benefits, then we must learn how to assess the cost-effectiveness of medical practices and use this information to help us substitute more for less cost-effective practices. Examples of technologies for which proof of efficacy may not be the central issue, but for which cost-effectiveness is, include heart transplants, intensive care for the terminally ill, and possibly artificial implantable organs in the future. Moreover, information on efficacy will not resolve the highly individual and subjective judgments about the value of symptom relief or other aspects of improved quality of life.

These two responses to the health-cost problem are not mutually exclusive, although they lead to different emphases. While we concentrate on evaluation of efficacy as one approach to improving the public health and/or controlling costs, we acknowledge—and, indeed, seek to elucidate—some of the limitations of evaluation of efficacy in the health care system.

Evaluation has its own costs, and so we need to consider how much different kinds of evaluation are worth and what their benefits may be. The long-run goal of the research that we outline here would be to develop and demonstrate a methodology for assessing these benefits and costs.

To oversimplify for a moment, we can identify two possible scenarios that result from evaluating efficacy. In the first, a therapy or diagnostic method that proved ineffective (or, at least, cost-ineffective) would be dropped by the profession, and the money saved would reduce the national medical budget without substantially impairing health. In the second scenario, a procedure is proved effective, leading to more widespread use and the resultant health benefits. We have examples of both scenarios: gastric freezing, for the first, and antihypertensive medications, for the second. We return to these examples below.

Students of policy will recognize both of these scenarios as idealized and unrealistic. Technological changes and changes in practice are ordinarily slow, except in crisis situations. For the first scenario, funds not used for one purpose are quickly and smoothly diverted to other uses, possibly ones that compensate for an abandoned procedure. Advocates of a procedure let go slowly and use ostensibly (and sometimes legitimate) scientific arguments to cast doubt on the validity of the evaluation. For the second scenario, practitioners may be slow to adopt new proce-

dures, even if proven efficacious, unless they perceive the benefits to be immediate and attributable to the intervention (a general obstacle to adopting preventive medical practices).

While we may doubt instant abandonment of procedures, immediate reduction in expenditures, or universal adoption of newly validated practices, we can hope that identifying better procedures will improve the use of medical resources. Improvement may be accomplished by increasing the use of underutilized or innovative practices or programs, by finding more cost-effective or less risky ways to administer care, by weeding out useless or harmful procedures, or merely by speeding up the process that grades procedures as better or worse.

Studies of the effectiveness of evaluations or of the diffusion of medical technology make clear that attempts to evaluate evaluations have a rocky road. For example, the innovation of gastric freezing for treating gastric ulcers was abandoned after substantial evaluations (Miao 1977), but perhaps the procedure was already out of style before the strongest trial had been completed (Fineberg 1976). If the latter was the case, then weaker studies may have had a substantial effect on the total process of introduction, use, and abandonment. At the same time, we know that some techniques such as bleeding, now believed to have no merit, lingered for centuries without evaluation. Consequently, we can afford to approach with modesty a study that aims to develop a basis for evaluating the benefits, risks, and costs of various methods of appraising effectiveness. It is not that we feel the effort unimportant, but that the path has new thickets that replace the old ones as fast as they are cleared.

Although we recognize the difficulty of the task, we are reminded of the need for some rational basis for allocating resources to clinical experiments. Budgets for clinical trials at the National Institutes of Health (NIH) are under constant surveillance, and vigilant congressmen will want to know that the resources have been well spent. Administrators of these agencies, facing contracting budgets, must constantly decide in what medical procedures to invest resources for a clinical trial, recognizing that a trial done in one area means a trial not done in another. Can these administrators not only improve their decision rules for internal-budget allocation, but also determine whether additional resources spent on clinical investigations have a greater expected return than resources spent at the margin elsewhere in the health sector? The economist's test of allocative efficiency (equal shadow prices across and within sectors of the budget) has more than a little conceptual appeal in this domain, but the analytical tasks are formidable.

We realize that the conceptual tools needed for such studies will require repeated refinement. Our first few efforts have no hope of being definitive. From our work thus far, we believe that we cannot get useful handles on this program until we have tried to evaluate a few situations.

We find the candidates extremely varied in their form and background information. Therefore it may be valuable to outline our beginning thoughts and what we foresee as difficulties, in anticipation that criticism will help us streamline and direct a long-term effort or that, informed by the evaluations of our peers, we may even be encouraged to abandon it.

6.1.2 Methods of Evaluation

Initially we defined the problem as that of evaluating the randomized clinical trial (RCT). What is it worth to evaluate a new procedure using RCT? Inevitably the question arises, "Compared with what?" One answer is, "Compared to what would have happened in the absence of an RCT." The possible procedures for comparison are varied: perhaps observational studies of procedures after they are widely practiced, perhaps clinic-based or community-based studies, perhaps systematic efforts using data banks, perhaps NIH consensus-development conferences, perhaps committee appraisals in the Institute of Medicine or the Assembly of the Life Sciences of the National Research Council, or perhaps the review papers in such professional journals as the *British Medical Journal, Journal of the American Medical Association, Lancet,* the *New England Journal of Medicine*, or those devoted to specialties. Whatever the alternatives may be, we do not seem to be able to deal with the RCT, or other methods, in isolation. Obviously this necessity for breadth multiplies our research effort enormously.

Moreover, we need to design the potential study (RCT or otherwise) before evaluating it. (The National Heart, Lung, and Blood Institute uses a planning phase when it prospectively evaluates its clinical trials in which the basic structure of the experimental design is formulated prior to a decision to proceed with full-scale design and implementation [Levy and Sondik 1978].) Since this planning step is also necessary in developing railroads and buildings and weapons systems, we seem to be stuck with it.

Usually we hope that the costs of processing information leading up to a decision, a sort of transaction cost, will be negligible relative to the value of the decision, but if heavy detail is required, such a simplication may be mistaken.

Some general qualitative or operating principles might be developed. For example, we could set up an operating principle that a study involving more than a million people to be followed for twenty years is hopeless. Or, given a choice, that acute-disease studies pay off better than chronic-disease studies, or vice versa. We are not endorsing these as principles, but as illustrations of policies that could emerge from a historical study of medical experiments.

We know that sometimes an RCT is impractical; other times it may not be helpful because other considerations, including value judgments, may overrule it. For example, an RCT helped establish the value of the Salk

vaccine against paralytic polio. Today the Salk vaccine (killed virus) is widely used abroad, while Sabin vaccine (live virus) is largely used in the United States. Both vaccines seem to be highly effective, though it is said that the Sabin leads to a few cases of polio in those exposed to recently vaccinated people (Institute of Medicine 1977). The decision as to which to use seems to depend more on an analysis of the policy of enforcement of administration than on efficacy. A major reason for Sabin use in the United States seems to be our perceived inability to administer booster shots. At another level, some public health officials are considering trying to wipe out the virus totally by administering both Salk and Sabin vaccines to all children ("This Week in Review," *New York Times*, 25 January 1981). To consider and evaluate this idea would require evaluation methods different from the RCT. We will need to consider how to choose methods of evaluation for various purposes, taking into account the value of information produced on the acceptability, risk, cost, and effectiveness of the proposed procedures.

6.1.3 How Are Evaluations Used?

The value of an evaluation depends on how its results are translated into changes in practice. Our approach considers three classes of decision-making models in the presence of information from evaluations: the normative, the descriptive, and the regulatory.

In the normative model—the ideal—physicians act in the best interests of society. They process new information rationally. They allocate resources according to the principles of cost-effectiveness analysis, electing the procedures that yield the maximum health benefits obtainable from the health care budget. Although some future reconfiguration of incentives in our health care system (e.g., explicit resource ceilings, increased competition, increased central management) may move us closer to that state of affairs, the normative model of decision making is best thought of as an unattainable ideal; the value of information under this model is the best we can possibly expect.

In the descriptive model, or models, we would attempt to assess what the response of physicians and other decision makers *would be* to the information from a trial. Here we must rely on past experiences and on what economic, sociologic, and psychologic theories tell us. We need to learn how to predict when the response will be rapid, when slow, when nonexistent, and when paradoxical. Perhaps a model can be developed, based on data from past history, that would identify the characteristics of the procedure, the type of study (e.g., randomized versus observational, large versus small, multi-center versus single institution), the nature of the medical specialty, and other variables that can be combined into a prediction of response.

In the regulatory model, we would allow for the possibility of interven-

tion (by government, by insurers, by professional societies) intended to make medical practice more responsive to information. For example, reimbursement might be preconditioned on evidence of efficacy or otherwise linked to the state of information. FDA-type procedures for practices other than drugs and devices would fall into this category. We recognize many problems inherent in such an approach: establishing criteria for efficacy where outcomes are multi-attributed (including survival and many features of the quality of life), establishing criteria for efficacy to apply to a heterogeneous population when the procedure cannot have been tested in all possible subpopulations. We realize that more decentralized approaches to altering incentives for practice in response to information on efficacy—or even to collecting the information itself—may be possible.

6.1.4 Our Objective

We propose, in section 6.2, a general conceptual model for evaluating the cost-effectiveness of clinical trials. This rather formal, oversimplified model will need more specificity when applied. It likely omits important policy or technological features, either because we have not thought of them or because modeling them presents frustrations.

In section 6.3 we discuss the range of medical problems that might be examined and the range of evaluative options that need to be compared. Our major aim in this section, however, is to describe the kinds of data that may be needed in evaluating the cost-effectiveness of a trial. These data requirements follow from the conceptual model in section 6.2 and from the realities that emerge when some of the simplifying assumptions are relaxed. For example, how do the results of the trial link to medical practice? Who are the decision makers, how will they use the data, and where does the burden of proof lie? We consider also the basis for the required probability assessments, the outcome measures that enter into the definition of "effectiveness," and the costs and risks of the clinical studies themselves. In section 6.4, we turn to some illustrative examples, sketched briefly to make more realistic some of the issues discussed. These sketches should not be confused with what a full study would require. Furthermore, we would presumably need collections of studies to help us base the models on empirical results.

Finally, in section 6.5 we discuss some of the kinds of studies that we believe are ultimately required to make this program a reality.

6.2 A Simplified Decision-Analytic Model for Assessing the Cost-Effectiveness of a Trial

6.2.1 Rationale

Let us clarify our thinking by beginning with a grossly oversimplified model based on admittedly unrealistic assumptions. By studying the

simplified model and then relaxing the assumptions, we can identify the data requirements for actually carrying out a program of evaluating the cost-effectiveness of a clinical trial.

We should point out that ours is not the first attempt at applying decision-analytic concepts to the problem of evaluating evaluations. Thompson, for example, developed a model for evaluating social-program evaluations and applied it to an evaluation of a U.S.-supported health program in Yugoslavia (Thompson 1975). The author admittedly found it difficult to apply the model quantitatively, but did derive qualitative conclusions about the administrative and bureaucratic determinants of effective evaluation. Stafford developed a similar model in relation to evaluations of manpower training programs (Stafford 1979). In the domain of clinical trials, Levy and Sondik (1978) have presented a conceptual framework for allocating resources in the National Heart, Lung, and Blood Institute, but their approach stops short of a formal assessment of the expected value of information. We want to assess the value of information and the costs, risks, and benefits of obtaining it in a practical, but still quantitative manner.

6.2.2 The Cost-Effectiveness Model of Health Care Resource Allocation

Economists turn to cost-effectiveness analysis when resources are limited and when the objective is to maximize some nonmonetary output. This technique is well suited to the assessment of medical procedures, where outcomes do not lend themselves to monetary valuation. The cost-effectiveness of a medical procedure may be evaluated as the ratio of its resource cost (in dollars) to some measure of its health effectiveness (Weinstein and Stason 1977; U.S. Congress, Office of Technology Assessment 1980). The units of effectiveness vary across studies, but years-of-life gained is the most commonly used. The rationale for using such a ratio as a basis for resource allocation is as follows. Let us suppose that the health care budget is B. (In the United States in 1980, B was about $200 billion per year.) Let us further suppose that cost-effectiveness analyses have been performed on each of the N possible uses of health care resources, perhaps defined by procedure and target population. (Of course, N is a very large number.) Suppose the expected net-resource burden of procedure i is C_i, and its expected net effectiveness is E_i. Consider only procedures for which C_i and E_i are both positive, (since the optimal decision rule for procedures with one positive and the other negative is obvious, and because doing a procedure with negative C_i and E_i is equivalent to not doing one with positive, but equal, absolute values). Finally, assume that society's objective is to allocate the budget to achieve the maximum total health effect (setting aside, for later reexamination, equity concerns). In other words, consider total effectiveness to be the sum of individual effectiveness values for each procedure,

regardless of who benefits. Then the problem reduces to the programming problem

$$\max_{\{\delta_i\}} \sum_{i=1}^{N} \delta_i E_i ,$$

subject to the usual constraints

$$\sum_{i=1}^{N} \delta_i C_i \leq B, \; 0 \leq \delta_i \leq 1 ,$$

the solution to which is to select procedures in increasing order of the ratios C_i/E_i until the budget B is exhausted. The C/E ratio for the "last" procedure chosen, λ, is the reciprocal of the shadow price on the budget constraint; that shadow price, in turn, may be interpreted as the incremental health value (in years of life, say, or quality-adjusted years of life) per additional dollar allocated to health care.

Although the cost-effectiveness model is far from being used as a blueprint for health resource allocation in practice, many studies along these lines have helped clarify the relative efficiency with which health care resources are being, or might be, consumed in various areas of medical technology (U.S. Congress, Office of Technology Assessment 1980; Bunker, Barnes, and Mosteller 1977; Weinstein and Stason 1976).

6.2.3 A Cost-Effectiveness Model for Clinical Trials

In the above formulation, the net costs (C_i) and net effectiveness (E_i) are uncertain. For purposes of today's decision making, it may be reasonable to act on their expected values, but we must not obscure the possibility that new information might alter our perceptions of these variables (in the Bayesian sense of prior-to-posterior revision), thus permitting reallocations of the budget in more health-producing ways. In terms of this same objective function, it is reasonable to ask what is the value of information about the effectiveness of a medical procedure. Moreover, since resources for providing such information (e.g., for clinical trials) are limited, it is reasonable to ask what is the cost-effectiveness of a clinical trial, where the "cost" would be the resource cost of the trial and the "effectiveness" would be the expected increase in the health benefits produced, owing to the information. We would also want to take into account the possibility that, if the utilization of a procedure drops as a consequence of the trial (e.g., if the procedure is found not to be effective), the result might be a freeing up of health care resources for other beneficial purposes.

6.2.4 A Simple Model of Two Treatments

We are wary of constructing an elaborate model that is too restrictive in some fundamental way, so we think it best to start with a simple formal model that can be made more realistic as we gain insights from studying

specific examples. For illustrative purposes, this simplified model rests on a strong normative assumption of behavior in response to the information from a trial. If the model were correct, it would yield an upper bound on the value of a trial. More realistic estimates might derive from a model based on predictions of actual decision making in the presence of trial data. Such a model with descriptive assumptions could also be constructed.

Our simple model rests on the following assumptions:

1. An old treatment has been used for some time. It has two possible outcomes: success and failure.
2. A new treatment, about whose efficacy little is known (except from laboratory studies), also may result in either success or failure.
3. Both treatments (a) tend to be used on repeated occasions in individual patients, and (b) make their effects known rather quickly; moreover, (c) we can distinguish "success" from "failure."
4. Let P_O and P_N be the probabilities of success for the old and new treatments, respectively. We start with a joint prior with density $f(P_O, P_N)$. The marginal means are π_O and π_N.
5. The probabilities of success apply to all patients uniformly. This fact is known and unalterable. (This assumption implies that there would be no advantage to stratification.)
6. A controlled experiment with sufficiently large sample sizes can compare the two treatments in such a way that it can be assumed to provide virtually perfect information on P_O and P_N.
7. In the absence of the experiment, the old treatment will continue to be used for T years, in X_t patients in the t_{th} year ($t = 1, \ldots, T$); the experiment lasts T_E years ($T_E < T$). T is known.
8. The unit costs of the treatments are known to be C_O and C_N, for the old and new, respectively.
9. With the experiment, the new treatment will be adopted if and only if its adoption is "cost-effective" in the sense defined in assumption 10 below. Its adoption will be universal, and it will replace the old treatment up to the horizon at year T. (This is the normative assumption of decision making.)
10. The new treatment will be considered cost effective if and only if

$$0 < \frac{C_N - C_O}{P_N - P_O} < \lambda V, \qquad (C_N > C_O),$$

$$P_N > P_O \text{ or } \frac{C_O - C_N}{P_O - P_N} > \lambda V, \qquad (C_N \leq C_O),$$

where V is the health benefit per "success" achieved (in years, or quality-adjusted years). (An important special case arises if $\lambda = \infty$; in

this case health care resources are effectively unlimited, and the shadow price on the budget constraint $[1/\lambda]$ is zero.)

11. The cost of the experiment is C_E; there are no risks.
12. All decision makers are risk neutral.

Consider first the case in which the new treatment is at least as costly as the old ($C_N \geq C_O$). In that case, the trial has value only if it results in a cost-effective improvement in health outcome. This would occur if

$$P_N > P_O$$

and

$$(C_N - C_O)/V(P_N - P_O) < \lambda .$$

Let

$$\Omega = \{(P_O, P_N): (P_N - P_O) > (C_N - C_O)/V\lambda\},$$

and let

$$X = \sum_{t=t_E}^{T} X_t(1+r)^{-t},$$

where r is the discount rate. Then the expected health benefit from the experiment equals

$$VX[\iint_{\Omega} (P_N - P_O) f(P_O, P_N) dP_O dP_N] .$$

Note that we are discounting health benefits at the same rate as costs (Weinstein and Stason 1977). The costs consist of two components: the expected induced treatment cost if the new treatment is adopted, which equals

$$X(C_N - C_O) \iint_{\Omega} f(P_O, P_N) dP_O dP_N ,$$

and the cost of the trial, which equals C_E.

One measure of cost-effectiveness would be given by the ratio of total expected costs to expected benefits:

$$\text{Cost-effectiveness} = \frac{C_E/X + (C_N - C_O) \iint_{\Omega} f(P_O, P_N) dP_O dP_N}{V \iint_{\Omega} (P_N - P_O) f(P_O, P_N) dP_O dP_N} .$$

Now consider the case where the new treatment is less costly than the old ($C_N < C_O$). In that case, the value of experiment might consist of the potential cost savings if the finding is that the new treatment is no less effective than the old.

$$\text{Let } \Psi = \{(P_O, P_N): P_N \geq P_O\} .$$

Then the expected savings consist of

$$S = (C_O - C_N) \iint_{\Psi} f(P_O, P_N) dP_O dP_N ,$$

and the expected health benefits would consist of

$$B = V \iint_{\Psi} (P_N - P_O) f(P_O, P_N) dP_O dP_N .$$

If the expected savings exceed C_E, then the experiment is clearly cost-effective; if not, then a measure of its cost-effectiveness would be given by $(C_E - S)/B$.

If we apply the cost-effectiveness model rigorously in this latter case, the new treatment might be found to be less effective than the old, but not so much less that it would not be cost-effective to adopt it anyway, taking into account its lower cost. This might happen if $P_N < P_O$, but

$$\frac{C_O C_N}{V(P_O - P_N)} > \lambda .$$

Thus, the effect of the experiment might be to make health outcomes a little worse, but in a cost-effective way when compared to other uses of resources. This situation is analogous to one in which an experiment, while proving no benefit, at least gives us reasonable assurance that the procedure in question is cost-ineffective compared to other available health interventions.

We do not want to attach too much importance to the cost-effectiveness ratios themselves. Their meaning depends on a rather stylized and fanciful notion of how decisions get made and how resources are constrained. Rather, we do want to emphasize the approach to estimating, for a trial, the expected change in health outcomes and the expected induced health care costs or cost savings.

6.2.5 Relaxing the Assumptions

Now, let us return to reality and see how, by relaxing the assumptions, we can identify the data required for the kind of evaluation we are proposing.

We will consider the following:

1. How will decisions be made once the information from the trial is in hand? How does this depend on the design and conduct of the trial? How does it depend on the health care institutional structure (e.g., regulation, financing)?
2. How would decisions have been made without the trial? What would have been the course of diffusion and adoption of the procedure?
3. How can the system be improved, by regulation or by imposing more appropriate incentives, so that the results of trials will be used more effectively and efficiently?
4. What do we do if the measure of efficacy is more complicated than "cure"? How do we handle risks, side effects, symptoms? At the very least, we need to estimate these attributes of outcome, but we may

also want to allow for the possibility that the trial will provide information on them. How do we handle effects on morbidity and the quality of life?

5. How do we handle a nonhomogeneous population in which the comparative efficacy (and risks, and perhaps costs) may differ among subsets of the population? Do we need to evaluate alternative experimental designs?
6. How do we assess the information from a trial that does not give perfect information?
7. How do we assess the information we would get from nonexperimental designs?
8. How do we establish the time horizon for the procedures in question, and how do we estimate the numbers of patients who would receive them? How should we decide when in the course of a procedure's dissemination to do a trial?
9. How do we assess prior probabilities of efficacy (i.e., prior to the decision to do the trial)?
10. How do we assess the costs and risks of the experiments themselves?
11. What are some of the other, less direct, benefits of doing trials?

6.3 Problems in Assessing Cost-Effectiveness of Medical Evaluations

In the enterprise we are suggesting, we would hope to have the aid of physicians as well as economists in bring additional realism to the evaluations. The tendency in medical studies, as in judicial review, is to avoid generalizations and focus strongly on specifics. As the following discussion illustrates, the diversity and incomparability of situations forces these constraints. Diagnosis, prevention, therapies, palliation, and health care delivery all fall within the scope of the studies we might try to evaluate. RCT's can be used for any of them or may be a component of evaluation. For example, in considering the dissemination of a new expensive technology, we may require an RCT to help measure the effectiveness of treatment as one component of an evaluation. Another component might relate to utilization patterns, and yet another to costs. We will probably focus on the RCT as a method of providing information on efficacy and take information on other aspects of cost-effectiveness as given. However, we may also want to consider how to assess the value of information on costs or on patterns of use of medical procedures and facilities.

6.3.1 How Decisions Will Be Made with the Experiment

In section 6.2, we offered a stylized model in which resources are allocated "rationally," as if by a benevolent Bayesian dictator. This

model may be helpful to get us started, but it needs to be brought back to reality.

An alternative to the "rational" model is a model that captures the way procedures actually are adopted or abandoned. We do not know very much about how decisions are actually made. When a therapy is evaluated and found useful (or useless or damaging), what can we say of the events that follow? Recent research on the diffusion of medical practices sheds some light on this question (Barnes 1977; Miao 1977; Fineberg 1976), but, as noted earlier, the conclusions regarding the effect of the trial itself are often ambiguous. When the experiment on mammary artery ligation for relief of angina showed the sham operation to be as good as the experimental one (Barsamian 1977), we understand that the experimental operation was dropped. When studies showed that successive diagnoses of the need for tonsillectomy in groups previously diagnosed as not diseased produced the same proportion of "diseased" diagnoses as in the full group, as far as we can see nothing happened (Bakwin 1945).

We need a systematic set of historical studies that tells us the situation before, during, and after the evaluations. (We say evaluations because often more than one is available.) From these, it might be possible to identify the factors that tend to predict the impact of evaluations on practice. For example, how does the effect of an RCT on practice depend on the existence of an inventory of prior observational studies? Does it matter whether the RCT contradicts or confirms the previous studies? Does the second or third RCT make more of a difference than the first? Perhaps, as Cochrane (1972) suggested, we should systematically plan more than one trial, not just for scientific reasons, but because people will pay attention to the results.

Related to the hypothesis about multiple trials is the question of the importance of packaging and public relations for trials. Perhaps trials that show a dramatic effect (or that refute a generally believed large effect) more successfully affect practice than those that deal in small effects. Taking this into account, assuming it is true, should we give priority to trials that are believed ex ante, to be more likely to make a big splash, even if this strategy means sacrificing cost-effectiveness as defined by our hypernormative model?

We may also want to consider the value of making certain that a trial *seems* relevant to a physician's practice, e.g., by conducting it in a community setting, by using a seemingly "typical" cross section of patients. The probability of a successful result may have to be reduced in order to increase the probability of disseminating the findings in practice.

Finally, we observe that with the descriptive, rather than normative, view of decision making, it is very possible that a trial might have negative value. Results get misinterpreted. Expensive procedures found to be

efficacious might be widely adopted even if they are not cost-effective. Procedures often are used in clinical circumstances beyond those for which they were evaluated. Efficacy in the hands of experts may not translate into effectiveness in the hands of a nonexpert, especially complex surgical or diagnostic techniques. Promoting a procedure of questionable efficacy to the status of a trial might give it credibility that it would otherwise lack if left to the "quacks." These and other concerns should be weighed against the benefits of trials, because the medical care system does not always use information, even good information, just as we would want it to.

6.3.2 How Decisions Will Be Made Based on the Literature and Record

Suppose we look at the observational study model when an innovation comes into society, is practiced (or experimented on) for a while, and reports appear about it. How do clinical practitioners respond to these reports and fragments of information? We can draw upon the literature for theoretical insights, but the empirical data base is thin. We see no way to handle this lack of data except to obtain a collection of situations, try to trace them as cases, and then to generalize to some models. For example, by systematically reviewing a surgical journal through the years, Barnes (1977) has provided examples of surgical innovations that later were discarded.

6.3.3 How to Design Institutions to Improve Incentives to Use Information Appropriately

Our third model of the response of health care providers to information from trials (the first two being the normative model and the descriptive model) would allow for intervention, or at least changes in the incentives for decision making owing to changes in the structure of health care insitutions. Regulation is one form of intervention. Weinstein and Sherman (1980) developed a structured framework for considering alternative regulatory and nonregulatory "levers" upon the health care system, taking into account the target of intervention (provider, patient, etc.), the nature of the lever (strict regulation, incentive, persuasion, etc.), and a variety of other dimensions. Our purpose here is not to enumerate all possible forms of leverage, but rather to mention a few as examples.

Various agencies at various levels have some leverage on practice, ranging from regulators such as the FDA to reimbursers such as the Health Care Financing Administration (HCFA) or Blue Cross–Blue Shield. Both HCFA and the Blues have ceased payment for some procedures found to be inefficacious. The National Center for Health Care Technology recommended to HCFA that heart transplants not be reim-

bursed on grounds of cost-effectiveness, although the recommendation was modified to permit certain providers to obtain reimbursement.

Direct linkage of reimbursement to demonstrated efficacy, while appealing in principle, has several limitations. Among these are problems in making the leap from efficacy in a study population to efficacy in individual patients. There would always be the need for some sort of escape clause in exceptional cases. Another problem in such a centralized approach is how to determine to what degree subjectively held concerns for symptoms and the quality of life are legitimate components of efficacy and, if they are, how to weigh them into the standard for reimbursement. Furthermore, fiscal intermediaries seem not to have much of an incentive to engage in such regulatory practices.

Another intervention that seems to work, at least in some settings, involves systematic persuasion within professional societies and peer groups. In one experience, physicians in one hospital reduced their utilization of tonsillectomy when told of their excessive rates relative to other hospitals (Wennberg et al. 1977; Dyck et al. 1977; Lembcke 1959). The Professional Standards Review Organization program was to have had this model as its raison d'être, although it is not clear how successful it has been.

At present we do not have a stable, but a rapidly changing system of control. Thus information, reimbursement principles, and changing regulations may be heavily confounded so that our ability to model a rational or irrational process may be heavily compromised. On the other hand, we may be able to gain insights into the kinds of institutional structures that are well suited to use information provided by clinical trials.

6.3.4 How Shall We Characterize Measures of Efficacy Required for Clinical Decision Making?

Acute and chronic diseases tend to give us different measures of outcome. In acute disease we usually focus on proportion surviving or proportion cured or degree of cure rather than length of survival. Morbidity, measured perhaps by days in the hospital, gives another measure of efficacy. Ideally we would compare costs, risks, and benefits from the new treatment with those from the standard treatment.

In chronic disease, we may be especially concerned with length of survival and with quality of life. Although it is generally agreed that quality of life is important, indeed often the dominant issue, its measurement, evaluation, and integration into cost-benefit studies must still be regarded as experimental (Weinstein 1979). Studies are proceeding in various places. For example, at Beth Israel Hospital, John Hedley-Whyte, M.D., and Allen Lisbon, M.D., are pilot testing a questionnaire on quality of life following surgery. Although the patient reports on the various aspects of life (leisure, family, happiness, ambulatory ability,

etc.), a summary that could be integrated with other attributes of outcome is not in sight. Instead, comparisons of results for different operations is readily available. For another example, in a body of research dealing with "health-status indexes," subjects assign weights to various health states, and a single measure of "weighted life expectancy" or "quality-adjusted life expectancy" is derived (Torrance 1976; Kaplan, Bush, and Berry 1976). Applications of these techniques to specific procedures, using real subjects, are rare.

Another problem arises especially in the evaluation of diagnostic procedures. If a new method of diagnosis successfully detects cases of a disease for which we have no effective treatment, how valuable is the technology? It may be useful for counselling or for research, but the effect on health outcome may be negligible.

6.3.5 Problems with Heterogeneous Populations

Information on homogeneity of response to treatment across patients and providers tells about the uncertainty of improvements a therapy offers. If community hospitals get different results from teaching hospitals, or if various ethnic, age, or sex groups produce differing responses, then efficacy becomes difficult to measure. In these circumstances, we have difficulty nailing down the amount of gain owing to new information.

The problems are especially severe when we must deal with groups that have no theoretical connections among them. Let us mention first a favorable situation. In dose-response-curve work, we often have rough theory and experience to guide the choice of a relation. Since differences in shape of the relation may have modest impact, we can use the information from several groups, and then, say, weight the groups to estimate the effect; we do not lose much information by spreading the information across the groups. But when groups may not be related in their response, the total size of the investigation must be increased. In the extreme case, when we cannot argue from one group to another, each group must be considered separately: pre- versus postmenopausal women, men at various ages and in various stages of diseases. The total sample size for the study would equal the sum of the sample sizes for each group. As groups proliferate, samples become small, too small to determine anything for each group separately. The typical situation lies between these extremes, and we need to learn more about how to model them.

The central point here is that a trial may be valuable in telling us who can benefit from a procedure and who cannot. Such information could save lots of money, even if most procedures are beneficial for some people. But learning how to describe the subpopulations that can benefit may not be easy, especially if we do not have a good predictive model when we allocate patients to treatments and decide how to stratify.

6.3.6 Assessing the Informational Content of Alternative Experimental Designs

The precision of outcome achievable by various designs depends on their size, on their stratification, and on the uniformity of individual and group response. In addition, the measurement technique sets some bounds on precision because simple yes-no responses may not be as sensitive as those achieved by relevant measured variables. When the outcome variables measured are not the relevant ones but proxies for them, we lose both precision and validity.

The RCT, however, is likely to give us values for a rather narrow setting and would need to be buttressed by further information from before and after the investigation.

6.3.7 Assessing the Informational Content of Nonexperimental Designs

Nonexperimental designs run the gamut from anecdotes or case studies of single individuals through observational studies.

Current behaviors toward such studies have great variety; prevailing attitudes include ignoring them, regarding them as stimuli for designing studies with better controls, and regarding them as true, even overriding contradictory results from better-controlled studies. Although it is easy to list reasons often given for these differing behaviors of people and institutions, (reasons such as: physicians like the medical theory; institutions like the implied reimbursement policy; no one has a better therapy, and patients need something; a new generation of physicians is required to understand the new biological theory; patients won't comply), we have difficulty developing a normative basis for judging the information content of the data from the studies.

A Bayesian approach used by Meier (1975) and extended by others in considering the precision of historical controls might be helpful here. In spite of a large sample size, historical control groups may give substantially varying performances depending on the physician or the institution where treatment is given. When we assign reliability to them as if they came from an experiment with sample size n and standard deviation of measurement σ, i.e., using $\sigma/\sqrt{n}$, we overlook the group-to-group variability. By introducing this variability Meier is able to show how much the total variability increases. A difficulty with the approach is agreeing on the size of the group-to-group variability to be introduced. That difficulty arises because (1) we have to define "groups like these," and (2) we have to provide data for the groups, thus establishing a prior distribution for the situation at hand. The first of these may not present much more difficulty than the usual fact that the scientists differ in how they think of their populations. The second requires us to find the data

and implies extensive information gathering and evaluation either in the field or from the literature.

Currently some theoretical work has been going on in statistics oriented toward using Bayes or empirical Bayes methods to evaluate the effects of treatments in studies where the investigator has weak or no control over the allocation of treatments to subjects or patients and where the choice of allocation may itself be related to the probability of a favorable response from the treatment, over and above the value of the treatment itself (Rosenbaum 1980; Lax 1981). For example, more of the slightly ill patients may be allocated to treatment A and more of the seriously ill to treatment B. These investigators are trying to provide ways of untangling such effects. The investigators must, of course, model the various components of the problem.

Since the efforts in this direction are fairly recent, two steps seem appropriate. One is to discover if these approaches can be applied to our problem. The other step is to attempt to find any circumstances where such efforts can be verified. So far, although the methods have been applied, we have no verification.

6.3.8 Predicting the Utilization of Procedures

By assessing numbers of patients with a specific disease and the rates at which the disease occurs and progresses, we can estimate the importance of a procedure and its value. We are, of course, concerned with the value of the information leading to the establishment of the importance or unimportance of a therapy or procedure.

The value of one innovation depends on how soon another at least as good comes along and is adopted. If we have both a new treatment and an old treatment, and the new treatment is better ($P_N > P_O$), then the value of an experiment to establish that fact depends on (1) the rate at which better treatments ($P_B > P_N$) come along to supplant the new, and (2) the rate at which treatments of intermediate efficacy ($P_N > P_I > P_O$) would have come along to supplant the old, prior to the introduction of the better. The situation might develop as shown in figure 6.1, where the shaded area represents the benefit of the trial.

Of course, if physicians are allowed to use the new treatment without a trial, then things get more complicated. We would hope that the more effective the new treatment, the more widely used it will be (as anecdotal evidence spreads and as a sort of "osmotic pressure" builds). Compared with the benefit of the experiment that would pertain under the assumption that the burden of proof falls on the innovation, the benefit will be less if $P_N > P_O$; but it will be positive (and therefore greater) if $P_N < P_O$.

Thinking about the course of diffusion over time raises another important question: At what point in time should a trial be conducted? If we wait too long, the procedure may be established, and practice will be hard

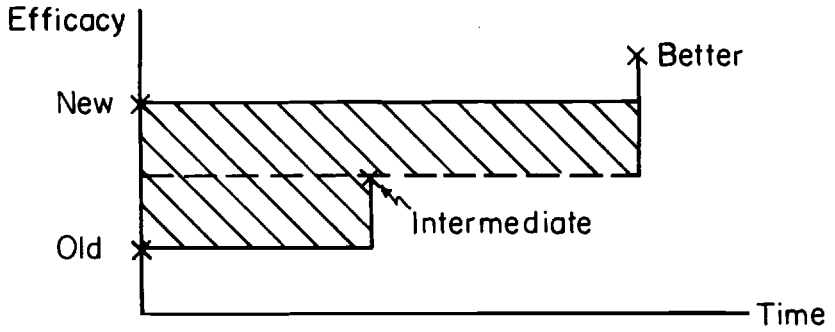

Fig. 6.1 Assessing the potential benefit from a newly validated treatment.

to change. But we also don't want to do the trial too soon, because (*a*) the technology may not be technically mature and may improve over time (in anticipation of which improvement, no one will pay attention to the trial if it shows no benefit), and (*b*) the innovation may turn out to be a poor performer with bad side effects and sink into obscurity. We need to develop strategies that adapt to early signals of a procedure's likely course and to respond promptly (but not prematurely) with a trial when appropriate. In other words, the decision whether to do a trial must be thought of as dynamic, not static.

6.3.9 Assessing Priors

Gilbert, McPeek, and Mosteller (1977) took a small step in the direction of assessing priors by reviewing randomized clinical trials in surgery over a ten-year period. They estimated the distribution of the size of the improvements (or losses), and they separated the experiments into two classes: those innovations intended to improve primary outcomes from surgery and anesthesia, and those intended to prevent or reduce complications following surgery and anesthesia. They found the average gain across studies to be about 0 percent improvement, the standard deviation for the gain in the primaries about 8 percent, and for the secondaries about 21 percent.

Such empirical studies help us assess the prior probabilities of improvements of various sizes brought by innovations. (This is a second-order value of the information from a trial.) Many other research possibilities are available, from observational studies to estimate group variation, from discussions with experts, and sometimes from reasonable considerations of naturally occurring bounds.

6.3.10 Costs and Risks of Studies

If we already have an experimental design, we can probably evaluate its direct costs. Although quarrels can arise about whether the cost of

treatment, for example, should be allocated to the cost of the investigation, we should not have much difficulty evaluating the price of a given trial. On the other hand, in certain cancer trials, the incremental cost may be small because the fixed cost of a multicenter "study group" has already been paid. It is understood that incremental cost is the appropriate measure.

An exception would arise if we had a wholly new therapy or if the insurance system changed suddenly as it did during the swine-flu vaccination program, but we would have to be careful to sort out real costs from transfer payments.

The question of risk is a thorny one that arises when human subjects are given a treatment less effective than the alternative (Weinstein 1974). For treatments with reasonably long horizons, this risk should be considered minor compared to the long-term value of knowing which is the better treatment. However, if horizons are short, these problems may be more important. We may wish to consider alternative experimental designs that reduce the efficiency of the study in order to reduce risks to subjects, e.g., play-the-winner rules and variable allocation formulae, (Zelen 1969; Cornfield, Halperin, and Greenhouse 1969). On balance, we do not want to be diverted too deeply into this thicket, since it is probably not productive. It would be better to concentrate on the primary benefits and costs of studies.

6.3.11 Other Benefits of Trials

One of the great values of combining well-founded facts with good theory resides in the bounds that can be set. For example, with a little theory and a little empirical information, we can reassure ourselves that man need not look forward to running the two-minute mile unless a new method of locomotion (or, should we say, ambulation) is discovered.

Thus a study that gives us solid information about death rates, recovery times, and rates of complications for a variety of treatment groups is likely to provide extra values that go beyond its own problem. The National Halothane Study, for example, not only studied the safety of anesthetics generally, but also provided data used to design other studies, stimulated the further Study of Institutional Differences (in operative death rates), and acted as a proving ground for a variety of statistical methods and encouraged their further development. How shall such information be evaluated? Can we assess a prior distribution for unanticipated benefits, without necessarily being able to imagine what those benefits might be?

Another benefit of clinical trials is that they may reinforce a general professional awareness of the value of scientific evidence of efficacy. The publication of trials in key medical and specialty journals is thus seen as a kind of continuing education, fostering alertness and healthy skepticism

with respect to innovations, and setting high standards for the data base upon which to support clinical decisions. Other benefits that may be attributed to trials include enhancement of the quality of patient care for participants (although perhaps at some added cost), and insights leading to improved efficiency in administration of health care services.

6.4 Examples

Earlier we mentioned the need for a few examples that might be examined with a view towards helping us learn to evaluate the contribution of studies.

6.4.1 Two Fables

We begin with two fables because we need to be more realistic about some problems of evaluation. The examples draw on ideas in the history of gastric freezing and blood pressure control.

Gastric Freezing

Gastric freezing was invented by a famous surgeon, Owen Wangensteen, to replace surgery as a treatment for gastric ulcers. The procedure was supported by biological theory: as a result of cooling, the stomach would have a better chance to heal itself. The patient swallowed a balloon, and coolant from a special machine entered and was withdrawn from the balloon.

A sequence of observational studies reported the performance of treated patients for the period 1962–69, with percentage of relief or definite improvement and sample size as shown in table 6.1. Although the outcomes in tables 6.1 vary a great deal, one notices decreasing performance as follow-up time increases.

Aside from these observational studies, several randomized double-blind trials were carried out comparing improvement following gastric-freezing treatment with that of controls. Table 6.2 shows the outcomes of these investigations. The study labeled "sham" at the bottom of the table was an especially well-controlled study that employed sham freezing as well as regular freezing. The sham and the freezing treatments produced nearly identical results. After this investigation, enthusiasm for the technique waned (Miao 1977). However, some studies of gastric freezing (Fineberg 1976) suggest that the treatment was falling into disfavor already and that the key experiment may have had only a slight effect.

We can probably collect information on the manufacture and sale of gastric-freezing machines; we may not be able to discover actual numbers of treatments. Assuming that we can obtain year-by-year information for treatments, how shall it be used?

The virtues of the findings of a good study are several. If the benefits of

Table 6.1 **Gastric Freezing Observational Studies**

Follow-up Period	Relieved (%)	Number Treated and Observed
Up to 6 weeks	100	19
	majority	86
	100	10
6 weeks–8 months	72	150
	13[a]	13
	78	53
	65	33
8 months–1.5 years	69	185
	14	29
	18	22
	21	60
	31	91
1.5 years–3 years	no studies	
Over 3 years	20	85

[a]A percentage of 13 is not compatible with $n = 13$, but we are not able to recheck this.

Table 6.2 **Gastric Freezing Randomized Trials Together with Sample Sizes (n)**

Time of Follow-up (months)	Gastric Freezing		Control	
	% Improved	n	% Improved	n
6	57	20	30	20
	75	19	29	17
	47	30	21	30
18	76	28	46	24
24	0	8	25	8
24 (sham)	34	82	38	78

treatment are positive, we have strong evidence of a gain. If, on the other hand, the benefits of treatment are zero or negative, we are released from further use of the treatment and can open our minds and resources more wholeheartedly to the use of and search for other treatments.

Hypertension

The association between high blood pressure and cardiovascular mortality and morbidity has been well known for some time. The life insur-

ance industry, for example, published its so-called Build and Blood Pressure Study in 1959. The Framingham Heart Study published its early follow-up results some time thereafter, and they were impressive. Meanwhile, drugs that effectively lower blood pressure were generally available, and their risks seemed small compared to the hypothesized benefits. And yet no randomized trial had proven the benefits of high-blood-pressure control. The practice of antihypertensive medication was limited to cases of malignant hypertension, so-called because of its immediate, dire consequences.

Then in 1967, the first report of the Veterans Administration randomized trial on antihypertensive medication was published (Veterans Administration Cooperative Study Group 1967). The trial established the efficacy of treatment in "severe" hypertensives (those with diastolic blood pressure or DBP above 115 mm Hg). A later report, in 1970, established efficacy in "moderate" hypertensives (DBP between 105 and 114 mm Hg) (Veterans Administration Cooperative Study Group 1970). The findings regarding mild hypertensives (DBP between 90 and 104 mm Hg) were inconclusive. Most of the hypertensives in the United States are mild hypertensives (perhaps 20 million of the 25 to 30 million hypertensives, the remainder being moderate or severe).

Prescriptions for antihypertensive drugs increased following publication of the VA study, but not so rapidly as one might have hoped. By 1973, it was estimated that perhaps 25 percent of hypertensives were receiving medication (although only 15 percent were taking it); however, the diffusion rate in mild hypertension was not markedly less than in moderate hypertension. The Secretary of HEW, Elliot Richardson, launched the National High Blood Pressure Education Program to try to accelerate the practice of treating high blood pressure. This program apparently has been somewhat successful; at least, the proportion of hypertensives who are taking medication has been rising steadily.

About the same time, interest arose in developing a controlled trial that would resolve the uncertainty about the efficacy of treating mild hypertension. This led to the Hypertension Detection and Follow-up Program (HDFP), a community-based, randomized trial in which the controls, instead of receiving a placebo, were allowed to seek whatever treatment they wished.

Many were skeptical whether the results of the HDFP would be useful. Therefore, plans were set for a true double-blind placebo trial in mild hypertension. However, estimates of the study size required to establish a statistically significant ($\alpha = 0.05$) effect with high power ($1 - \beta = 0.90$) ranged from 36,000 to over 200,000, depending on assumptions about compliance, degree of blood pressure reduction, etc. (Laird, Weinstein, and Stason 1979). The prior expectations were based on the Framingham Heart Study.

A calculation of the potential benefits and costs of such a trial was made at that time by one of the authors and his colleagues (Laird, Weinstein, and Stason 1979). First, the cost of the trial was estimated at $135 million, assuming 28,000 subjects followed for five years. Next, the size of the population at risk was estimated to be 20 million, of which 10 percent were already being treated. Now, to simplify considerably, there were three possible results of the trial: not efficacious, efficacious, and inconclusive. If it was found that treatment was not efficacious, and if this finding was translated into practice, then 2 million persons per year would *not* spend an average of $200 on treatment, for a total of $400 million per year. Over ten years, with discounting at 5 percent per annum, the present value is $3 billion.

Now we need to assess some priors. Let us say we assigned a 0.1 probability to the event that treatment is not effective and a 0.2 probability that the study will show conclusively the effect is either zero or small enough to be considered outweighed by risks and costs. (The latter estimate can be made more rigorous by considering study size, a prior distribution of the efficacy parameters, e.g., mortality rates, and the probability that each particular finding would result in reduced utilization.) Under these assumptions, the study has a 0.02 chance of saving $3 billion over ten years, an expected value of $60 million; so this contingency would pay back half the cost of the study. Then we would have to repeat the analysis under the possibility that treatment is efficacious and that the study will so demonstrate. (Now we would have to estimate the health benefits—as Weinstein and Stason [1976] have done—and the additional treatment costs owing to increased utilization.) We would also have to consider the false-negative case (treatment is efficacious, but the study says it is not), and the false-positive case (treatment is not efficacious, but the study says it is). We would then plug all this into the cost-effectiveness model and assess the value of the study.

The epilogue to this fable (although it is by no means over) is that the HDFP reported its results in 1979 (Hypertension Detection and Follow-up Program 1979). There was a significant and important treatment effect, especially in the mildly hypertensive group. Now the controversy continues around whether this community-based study was really measuring the effects of antihypertensive medication or whether other differences between the treatments could have accounted for the difference in mortality. The value of the HDFP—and of the placebo trial that was never conducted—is still not known.

6.4.2 Examples with Other Complications

In the area of large-scale studies that have complications of various sorts we note:

1. *Studies of coronary bypass surgery.* Some coronary bypass surgery studies are experiments, and some are observational studies. Among other difficulties, they present the welcome problems of the improving ability of therapists, the reduction of costs as technology improves, and thus possibly changing findings over time. Although much can be made of these matters, they are a commonplace of the passage of time and improvements in science and technology. Evaluators should have ways of dealing with them. In this sense any therapy is always in an experimental or dynamic state.
2. *Prostate cancer.* A large-scale study of prostate cancer led to dosage and therapeutic recommendations (Byar 1973). As far as we know, the study is not now the subject of controversy, although it was attacked for some time.
3. *Portocaval shunt.* Many portocaval-shunt studies with varying degrees of control show that the weaker the control, the more favorable the outcome of the investigation to the treatment. These studies, and many like them for other diseases collected by Chalmers and his colleagues, go a long way towards undermining the informational content of poorly controlled studies (Grace, Muench, and Chalmers 1966).
4. *University Group Diabetes Project (UGDP).* This diabetes study illustrates many difficulties that arise in practice from optional stopping, from hypotheses generated during the experiments, from data deemed to be important though not collected, and from results that are unpopular with physicians and commercial institutions.
5. *Salk vaccine trial.* This trial went well.
6. *Gamma globulin study.* This gamma globulin study, weakly controlled, was intended to prove the medication effective against polio. At the close of the study little was known.

Although we could lengthen this list, we need to discuss what considerations should go into the choices for detailed study: Can we define a population of studies? Should we study both RCT's and observational studies? Can we measure the follow-up effects of the studies? How?

6.5 Conclusion

The purpose of this paper is to outline a general program of research and the reasons for it. We will benefit from criticism and discussion, recognizing that the total problem is a bit like dipping one's head into a bowl of fruit salad and opening one's eyes.

Are there parts of the research program that can profitably be broken off and studied separately? The approach described has a rather worm's eye view of the problem. We write as if we need to know, or at least make

an estimate of, everything before we can reach conclusions. Are more global, decentralized attacks possible? Meanwhile, three observations seem likely to stand up to further scrutiny:

1. In planning a controlled trial, it would be valuable for those expert in effectiveness, costs, and other data pertinent to the health area of the trial, to perform at least a rough, back-of-the-envelope calculation of potential benefits and cost savings. This sort of meta-analysis cannot hurt, and even if we don't yet know how to implement a full-blown planning model of the type we have outlined, the rough calculations may help.
2. Evidence of efficacy from controlled trials will not solve the health-care-cost problem and will not even eliminate uncertainty from medical decisions. Value judgments related to multi-attributed outcomes (including quality of life) will remain, as will uncertainties at the level of the individual patient. Moreover, the problems of what to do about procedures that offer diminishing (but positive) benefits at increasing costs will always be with us.
3. Clinical trials can help, and we need to learn their value and how to increase it. As a nation, we may try various institutional changes to encourage the use by practitioners of trial information, perhaps by linking reimbursement to demonstrated efficacy, but more likely by providing incentives to be both efficacy-conscious and cost-conscious.

Comment

Joseph B. Kadane

Any governmental activity on which millions of dollars are spent is a worthy subject of analysis. Medical experimentation is a good candidate for such analysis, not only because of the amount of money involved, but also because we are all prospective beneficiaries of the improvements in medical techniques made possible by such experimentation.

In this very interesting paper, Mosteller and Weinstein give us an initial model to guide the choice of which medical experiments to support. The heart of that model is in section 6.2.4 of their paper. Some of their twelve assumptions are heroic, particularly the ninth assumption that all medical decision-makers will instantly adopt a new procedure if it is shown in the experiment to be cost-effective. The authors recognize in section 6.3.1 the desirability of substituting for this normative model a descriptive model of the spread of medical innovation. But they also point out how little we know about the history and sociology of medical innovation. Yet we need good descriptive models of this process to predict what would

Joseph B. Kadane is professor of statistics and social science, Carnegie-Mellon University.

happen in medical decision-making if the experiment were conducted and had various specified outcomes, and what would happen absent the experiment.

Certainly Mosteller and Weinstein are correct to call for rough, back-of-the-envelope calculation of the potential benefits and cost savings of each planned controlled medical trial. I do not understand quite so well, however, what research strategy they would use to make such calculations better informed. What would their research priorities be? Without this information, even a rough calculation of benefits and costs for their own research proposal seems impossible.

References

Bakwin, H. 1945. Pseudodoxia pediatrica. *New England Journal of Medicine* 233: 691–97.

Barnes, Benjamin A. 1977. Discarded operations: Surgical innovations by trial and error. In *Costs, risks, and benefits of surgery*. *See* Bunker, Barnes, and Mosteller 1977.

Barsamian, Ernest M. 1977. The rise and fall of internal mammary ligation in the treatment of angina pectoris and the lessons learned. In *Costs, risks, and benefits of surgery*. *See* Bunker, Barnes, and Mosteller 1977.

Bunker, John P., Benjamin A. Barnes, and Frederick Mosteller. 1977. *Costs, risks, and benefits of surgery*. New York: Oxford University Press.

Byar, D. P. 1973. The Veterans Administration Cooperative Urological Research Group's studies of cancer of the prostate. *Cancer* 32: 1126.

Cochrane, A. L. 1972. *Effectiveness and efficiency: Random reflections on health services*. London: Nuffield Provincial Hospitals Trust.

Cornfield, Jerome, Max Halperin, and Samuel W. Greenhouse. 1969. An adaptive procedure for sequential clinical trials. *Journal of the American Statistical Association* 64: 759–70.

Dyck, Frank J., Fergus A. Murphy, J. Kevin Murphy, David A. Road, Martin S. Boyd, Edna Osborne, Dan deVlieger, Barbara Koschinski, Carl Ripley, Alfred T. Bromley, and Peter Innes. 1977. Effect of surveillance on the number of hysterectomies in the province of Saskatchewan. *New England Journal of Medicine* 296: 1326–28.

Fineberg, Harvey V. 1976. Gastric freezing: A study of the diffusion of a medical innovation. Washington, D.C.: National Academy of Sciences.

Fuchs, Victor R. 1974. *Who shall live*? New York: Basic Books.

Gilbert, John P., Bucknam McPeek, and Frederick Mosteller. 1977. Progress in surgery and anesthesia: Benefits and risks of innovative

therapy. In *Costs, risks, and benefits of surgery*. *See* Bunker, Barnes, and Mosteller 1977.

Grace, N. D., H. Muench, and T. C. Chalmers. 1966. The present status of shunts for portal hypertension in cirrhosis. *Gastroenterology* 50: 684.

Hiatt, Howard H. 1975. Protecting the medical commons: Who is responsible? *New England Journal of Medicine* 293: 235–41.

Hypertension Detection and Follow-up Program Cooperative Group. 1979. Five-year findings of the Hypertension Detection and Follow-up Program. *Journal of the American Medical Association* 242: 2562–77.

Institute of Medicine. 1977. *Evaluation of poliomyelitis vaccines*. Report of the Committee for the Study of Poliomyelitis Vaccines. Washington, D.C.: National Academy of Sciences.

Kaplan, Robert M., James W. Bush, and Charles C. Berry. 1976. Health status: Types of validity and the index of well-being. *Health Services Research* 11: 478–507.

Laird, Nan M., Milton C. Weinstein, and William B. Stason. 1979. Sample-size estimation: A sensitivity analysis in the context of a clinical trial for treatment of mild hypertension. *American Journal of Epidemiology* 109: 408–19.

Lax, David A. 1981. Inference with unobserved variables: Analysis of a non-randomized study. Ph.D. diss., Department of Statistics, Harvard University.

Lembcke, Paul. 1959. A scientific method for medical consulting. *Hospitals*. 33: 65.

Levy, Robert I., and Edward J. Sondik. 1978. Decision-making in planning large-scale cooperative studies. *Annals of the New York Academy of Sciences* 304: 441–57.

Meier, Paul. 1975. Statistics and medical experimentation. *Biometrics* 31: 511–29.

Miao, Lillian L. 1977. Gastric freezing: An example of the evaluation of medical therapy by randomized clinical trials. In *Costs, risks, and benefits of surgery*. *See* Bunker, Barnes, and Mosteller 1977.

Mosteller, Frederick, John P. Gilbert, and Bucknam McPeek, 1980. Reporting standards and research strategies for controlled trials. *Controlled Clinical Trials* 1: 37–58.

Rosenbaum, Paul R. 1980. The analysis of a non-randomized experiment: Balanced stratification and sensitivity analysis. Ph. D. diss., Department of Statistics, Harvard University.

Stafford, Frank P. 1979. A decision theoretic approach to the evaluation of training programs. *Research in Labor Economics*, Supplement 1: 9–35.

Thompson, Mark S. 1975. *Evaluation for decision in social programmes*. Westmead, England: Saxon House.

Torrance, George W. 1976. Social preferences for health states: An empirical evaluation of three measurement techniques. *Socio Economic Planning Sciences* 10:129–36.

U.S. Congress, Office of Technology Assessment. 1980. *The implications of cost-effectiveness analysis of medical technology*. Washington, D.C.: GPO.

———. 1978. *Assessing the safety and efficacy of medical technologies*. Washington, D.C.: GPO.

U.S. Department of Health and Human Services. 1980. *Health United States 1980*. DHHS publication no. (PHS) 81–1232. Hyattsville, Md.: GPO.

Veterans Administration Cooperative Study Group on Antihypertensive Agents. 1970. Effects of treatment on morbidity in hypertension, II: Results in patients with diastolic blood pressure averaging 90 through 114 mm Hg. *Journal of the American Medical Association* 213: 1143–52.

———. 1967. Effects of treatment on morbidity in hypertension, I: Results in patients with diastolic blood pressures averaging 115 through 129 mm Hg. *Journal of the American Medical Association* 202: 1028–34.

Weinstein, Milton C. 1979. Economic evaluation of medical procedures and technologies: Progress, problems, and prospects. In *Medical technology*, U.S. National Center for Health Services Research, DHEW publication no. (PHS) 79-3254. Washington, D.C.: GPO.

———. 1974. Allocation of subjects in medical experiments. *New England Journal of Medicine* 291: 1278–85.

Weinstein, Milton C., and Herbert Sherman. 1980. A structured framework for policy intervention to improve health resource allocation. In *Issues in health care regulation*, ed. Richard S. Gordon. New York: McGraw-Hill.

Weinstein, Milton C., and William B. Stason. 1977. Foundations of cost-effectiveness analysis for health and medical practices. *New England Journal of Medicine* 296: 716–21.

———. 1976. *Hypertension: A policy perspective*. Cambridge: Harvard University Press.

Wennberg, John, Lewis Blowers, Robert Parker and Alan Gittlesohn. 1977. Changes in tonsillectomy rates associated with feedback and review. *Pediatrics* 59: 821–26.

Willems, Jane Sisk, Claudia R. Sanders, Michael A. Riddiough, and John C. Bell. 1980. Cost-effectiveness of vaccination against pneumococcal pneumonia. *New England Journal of Medicine* 303: 553–59.

Zelen, Marvin. 1969. Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association* 64: 131–46.

This Page Intentionally Left Blank