3) We are analyzing all of the patients who met the requirements of the protocol.

4) Even when seen from the viewpoint of the completed study, the protocol was well chosen and relevant.

5) Only one end point was contemplated in advance, and this is the one we are using.

6) We have only looked seriously at the data a few times, each of which was fixed well in advance.

What then?

Notice first that there may well be excellent reasons, often involving knowledge gained during the study, which can make any one or more of these desiderata either unwise to attempt or impossible to have. Real studies often have real problems, which we must meet as best as we can.

If, however, we have a focused clinical trial with the characteristics just described, we have a study of the best sort anyone knows how to conduct, and our statements of significance are much more likely to mean what they say, especially if we make some allowance for the number of looks, than most of those routinely found in the literature of any field, medico-surgical or not. As a result, we have, I assert, an ethical obligation to take the results of such a study most seriously.

* * * *

I can hardly claim to have made any of our tasks easier by bringing forward the problems I have discussed. But it would not really have helped us to go ahead in ignorance of the problems that *are* there whether we like it or not.

The pressures of ethics *do* force us to sharpen our interpretation of the uncertainties of the data. The distinction between clinical inquiries and focused clinical trials *is* important. *Both* have important roles to play. There *are* questions we dare not try to answer. Both knowledge *and* opinion are important and must be managed by the same individuals. Historical controls are *not* an easy out. We *cannot*, ethically, either look only once or look very many times. Yet there is hope.

References and Notes

1. D. P. Byar, R. M. Simon, W. T. Friedewald, J. J. Schlesselman, D. L. DeMets, J. N. Ellenberg, M. H. Gail, J. H. Ware, "Randomized clinical trials: Perspectives on some recent ideas," *N. Engl. J. Med.* **295**, 74 (1976).
2. R. Peto, M. C. Pike, P. Armitage, N. E. Breslow, D. R. Cox, S. V. Howard, N. Mantel, K. McPherson, J. Peto, P. G. Smith, "Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design," *Br. J. Cancer* **34**, 585 (1977); for part II see *ibid.* **35**, 1 (1977).
3. J. P. Gilbert, B. McPeek, F. Mosteller, *Science*, **198**, 684 (1977).
4. H. Robbins, "A sequential test for two binomial populations," *Proc. Natl. Acad. Sci. U.S.A.* **71**, 4435 (1974).
5. W. G. Cochran, "Improvement by means of selection," *Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability* (Univ. of California Press, Berkeley, 1951), pp. 449–470.
6. P. Armitage, C. K. McPherson, B. C. Rowe, "Repeated significance tests on accumulating data," *J. R. Stat. Soc. A* **132**, 235 (1969).
7. C. K. McPherson and P. Armitage, "Repeated significance tests on accumulating data when the null hypothesis is not true," *J. R. Stat. Soc. A* **134**, 15 (1971).
8. MSC, paper on "Group sequential designs for clinical trials," by S. J. Pocock, *RSS News & Notes* (Royal Statistical Society) **3** (No. 9), 5 (1977).

9. The text of this article was prepared in part in connection with research at Princeton University, sponsored by the U.S. Energy Research and Development Administration, and was presented at the Birnbaum Memorial Symposium, 27 May 1977, at the Memorial Sloan-Kettering Cancer Center, New York.

Statistics and Ethics in Surgery and Anesthesia

John P. Gilbert, Bucknam McPeek, Frederick Mosteller

Ethical issues raised by human experimentation, especially in medicine, have been of increasing concern in the last half of the 20th century. Except for issues of consent and capacity to consent, ethical concerns raised by controlled trials center about the fact that individuals are being subjected, randomly, to different treatments. Two arguments are raised, and in each the patients are seen to be the losers. The first argument is an expression of the fear that the trial, by withholding a favorable new therapy, imposes a sacrifice on the part of some of the patients (the control group). The second argument raises the opposite concern that, by getting an untested new therapy, some patients (those in the experimental group) are exposed to additional risk. To a large extent, both arguments imply that investigators know in advance which is the favorable treatment.

Some empirical evidence on these issues can be obtained by examining how potential new therapies are evaluated and what the findings are. How often do new therapies turn out to be superior when they are tested, and how much better or worse is a new therapy likely to be than the standard treatment? We have investigated such questions for surgery and anesthesia.

The Sample of Papers

For an objective sample we turned to the National Library of Medicine's MEDLARS (Medical Literature Analysis and Retrieval System). For almost 15 years, this computerized bibliographic service has provided exhaustive coverage of the world's biomedical literature. Articles are classified under about 12,000 headings, and computer-assisted bibliographies are prepared by cross-tabulating all references appearing under one or more index subjects. For example, all articles indexed under prostatic neoplasms, prostatectomy, and postoperative complications might be sought.

We obtained our sample from the MEDLARS system by searching for prospective studies and a variety of surgical operations and anesthetic agents, such as cholecystectomy, hysterectomy, appendectomy, and halothane (*1*). The papers appeared from 1964 through 1973.

We found 46 papers that satisfied our four criteria: The study must include (i) a randomized trial with human subjects, (ii) with at least ten people in each group, (iii) it must compare surgical or anesthetic treatments, and (iv) the paper had to

J. P. Gilbert is staff statistician at the Office of Information Technology, Harvard University, Cambridge, Massachusetts 02138 and assistant in biostatistics in the Department of Anesthesia, Massachusetts General Hospital, Boston 02114. B. McPeek is anesthetist to the Massachusetts General Hospital and assistant professor of anesthesia, Harvard University. F. Mosteller is professor of mathematical statistics in the Department of Statistics and chairman of the Department of Biostatistics, School of Public Health, Harvard University, 7th Floor, 677 Huntington Avenue, Boston, Massachusetts 02115.

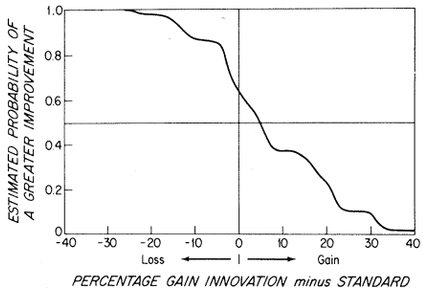

Fig. 1. Secondary therapies: estimated cumulative distribution of true gains (reduction in percentage with a particular complication).

be written in English because of our own language limitations. All the papers we found, by the MEDLARS search, that met these criteria are included in the sample. Although this sample is neither a strictly random sample nor a complete census of the literature of the period covered, the method does largely exclude personal biases in selection.

These papers evaluated two types of therapy. One type is designed to cure the patient's primary disease. An example is the trial of radiation therapy in addition to surgery for the treatment of cancer of the lung (*2*). The second type of therapy is used to prevent or decrease the rate of an undesirable side effect of the primary therapy. Examples are the various trials of anticoagulants to decrease the incidence of thromboembolism after operations on the hip. Because we felt that these two types of therapies might differ in the distributions of improvements we wished to study, as indeed they seemed to, we have recorded them separately using the terms primary and secondary therapies, respectively. While each of our sample papers has provided important information concerning the treatment of a specific disease or condition, prognosis, complications, the natural history of disease, and the like, we have concerned ourselves only with the comparison of effectiveness between competing therapies.

We have classified each therapy either as an innovation or as the standard treatment, to which the innovation was being compared. Although this distinction is usually clear, in a few instances some readers might disagree with our decisions. We took the position of the investigators, who usually indicated which therapies they regarded as the standards for comparison. Some papers report trials where several innovations were tested against one standard, or one innovation was sometimes tested against several standards, or the comparison was made for several distinct types of patients. To prevent one paper from having an undue effect on the total picture, no more than two comparisons were taken from any one paper, the choice being based on the importance of the comparisons for the surgery. When two comparisons were used, each was weighted one-half. When several papers reported the same investigation, we used the most recent one.

Comparisons of Innovations and Standards

To give a rough qualitative idea of how the innovations (I) compared with the standards (S), we have classified the outcomes by "highly preferred to" (>>), "preferred to" (>), and "about the same as" (=) in Table 1. In the first set designated =, the innovation was regarded as a success because it did as well as the standard and did not have other disadvantages, such as high cost, dangerous side effects, or the requirement of extra skill or training in its administration. Thus, it offers the surgeon an extra therapy when the standard may have drawbacks.

In the second set designated =, the investigators seemed indifferent to the equality; in the third set, the innovations were regarded as a disappointment because of undesirable features. The preferences reported reflect closely the views of the original investigators.

About 49 percent of the innovations were successful when compared to their matched standards, and 13 percent were highly preferred. Among pairs of primary therapies, the innovation was highly preferred in 5 percent, and among pairs of secondary therapies the innovation was highly preferred in 18 percent of the comparisons. Indeed, the totals of the two extreme categories were smaller in the primary comparisons than in the secondary—10.5 percent as compared to 27 percent.

The overall impact of the data in Table 1 is to suggest that, when assessed by randomized clinical trials, innovations in surgery and anesthesia are successful about half the time. Since innovations brought to the stage of randomized trials are usually expected by the innovators to be sure winners, we see that in the surgery and anesthesia area the evidence is strong that the value of the innovation needs empirical checking.

Quantitative Comparisons

In addition to the qualitative comparisons of Table 1, we want to compare the performance of the innovation more quantitatively with the standard. For those primary therapies where survival gives a suitable measure of performance, we examine the distribution of the difference in survival percentages (I minus S). For the secondary therapies, we compare the percentages of patients not getting a specific complication such as abdominal infection or thrombosis. (Where we have used two complications in one study, each has been weighted one-half, as in Table 1.) If we merely take the observed differences, they are subject to variation over and above the true differences because of sampling error due to the finite samples used in the experiments. To adjust for these sampling errors, we use an empirical Bayes procedure, as described in the appendix. Efron and Morris (*3*) describe the general idea through an instructive sports example:

If we observed the batting averages for their first 50 times at bat for 200 major league batters, we might find them ranging from 0.080 to 0.450, yet we know that major league averages for a season ordinarily run from about 0.200 to 0.350 these days. The excess spread comes from the sampling error based on only 50 times at bat rather than the season's total experience. To adjust this, we can shrink the results toward the center of the distribution (roughly 0.275). How this is done is explained by Efron and Morris (*3, 4*) and more simply by them in (*5*); the explanation is given in detail for the present situation in (*6*).

After the shrinking is carried out, we can estimate the distribution of the true gains or losses associated with the innovation by methods discussed in the appendix and in (*6*). In Fig. 1, we give the estimated cumulative distribution for the

Table 1. Qualitative comparisons between innovations (I) and standards (S) stratified by primary and secondary therapies. Where a paper had two comparisons, each was weighted one-half.

Preference	Primary	Secondary	Total	Percent
I > > S	1	5	6	13
I > S	4	4½	8½	18
I = S (success)	2½	6	8½	18
I = S (indifferent)	1½	1	2½	5
I = S (disappointment)	6	5	11	23
S > I	3	4	7	15
S > > I	1	2½	3½	7
Total	19	28	47*	(99)

*One paper contributed to both the primary and the secondary column.

true gains of secondary innovations. The graph suggests that about 80 percent of the innovations offer gains between −10 percent and +30 percent. In about 24 percent of the studies, gains of at least 20 percent occur. In about 10 percent of the studies, gains of more than 30 percent occur. About 12 percent of the time, losses of more than 10 percent occur. The sharp dip just to the right of zero improvement in Fig. 1 could, in a replication, move a few percent to the left or right of its present position. We have to emphasize that the cumulative is based essentially on a sample of 24 papers (not all secondary papers in Table 1 could be used here); but each paper is worth rather less than one whole observation of the difference because of the sample sizes in the investigations. If the sample sizes were infinite, we would not have the shrinking problem, and each paper would provide a full observation.

Gains or losses of modest size, such as 10 percent, while extremely valuable, are hard to detect on the basis of casual observation. We need careful experimentation and good records to identify such gains and losses. To get an idea of how hard it is to detect a difference of 10 percent, say that between 55 percent and 45 percent, it may help to know that two samples of size about 545 are required to be 95 percent sure of detecting the difference, by a one-sided test of significance at the 5 percent level. To be 50 percent sure requires samples of 136. Such large trials were rare in our samples.

Nonrandomized Controlled Trials

In addition to the randomized clinical trials, 11 less well-controlled trials seemed appropriate for reporting. Results are shown in Table 2 in a manner similar to that used in the randomized trials. By and large, the distribution leans more favorably toward innovations than that seen in Table 1. A tendency for nonrandomized trials to favor innovations is frequently noted. Although speculation is easy, the reasons for this are unclear. While in general a randomized trial provides stronger evidence than a corresponding nonrandomized trial, there are occasions where a nonrandomizing trial may be convincing. A nonrandomized study of abdominal stab wounds seems especially instructive because it provides strong evidence favoring a new policy. The hospital's standard policy had been to perform a laparotomy (surgical exploration of the abdominal cavity) on all patients with abdominal stab wounds. In 1967, the hospital instituted a change in policy, the results of which Nance and Cohn (*7*) report. The new policy demanded exploration only when the attending surgeon judged it necessary. (A patient might be observed for a period and then explored.)

Table 2. Summary for controlled nonrandomized trials.

Preference	Primary	Secondary	Total
I > > S	2	3	5
I > S	1	1	2
I = S (disappointment)		2	2
S > *I*		1	1
S > > I		1	1
Total	3	8	11

The investigators give a record of (i) the substantial number of complications (25 percent) emerging from routine laparotomy when, in retrospect, the patient had not required surgical repair for the stab wound; (ii) the recovery without complications in the approximately 8 percent of patients who declined or otherwise passed by the former administrative rule of always performing a laparotomy; and (iii) evidence that delay before exploration under the old policy was not associated with an increase in the complication rate. These observations suggest that omitting the laparotomy for selected patients might be good practice.

Some might have said, on the basis of the data presented in (i), (ii), and (iii), that the proposed new policy of judgmental surgical decisions would be clearly preferable to routine laparotomy. Nevertheless, such inductive leaps have often failed in other attractive circumstances, sometimes because the new policy loses some advantages that the old one had, or falls prey to the fresh problems that may arise when any policy is totally changed. Changing from set policy to the regular use of judgmental surgical decisions plus keeping records provided an inexpensive type of quasi-experiment. The method has a grave weakness because the time period is not common to the differently treated groups; and, therefore, causes other than the change in treatment may produce at least part of the observed differences.

For the stab wounds, the need for a randomized clinical trial is not now compelling for the hospital partly because, in addition to the logic and data of (i), (ii), and (iii) above, the final quasi-experiment produced a large improvement. Although the percent requiring repair of the stab wound was about the same under the old and new policies (30 percent as compared to 28 percent), the overall complication rate dropped substantially from 27 to 12 percent. One fear would be that the unexplored group would produce a proportion of very severe complications. The evidence goes the other way. Among those not explored, the number without complications remained at zero even though the number not explored rose from 38 to 72 patients, and the percent explored fell from 92 to 40 percent. The average length of hospitalization over all patients dropped from 7.9 to 5.4 days. Had the effect been small, one might still be concerned whether possible biases and other changes could have given misleading results. All told, the evidence favoring the new policy seems persuasive for this hospital.

Comparisons of Degrees of Control

Although randomized clinical trials are not the only strong form of evidence about therapies in humans, weakly controlled investigations may not give the same results as better controlled ones. Chalmers and his colleagues have compared (*8, 9*) views of many investigators who had make studies of a single therapy, with respect to the degree of control used in each investigation. We give the results of one example of such collections of investigations (*8*).

Table 3. Degree of control versus degree of investigator enthusiasm for portacaval shunt operation in 53 studies with at least ten patients. The table is revised from Grace, Muench, and Chalmers (*8*), table 2, p. 685 (©1966, Williams and Wilkins, Baltimore). Chalmers advised us of two additional studies to add to the well-controlled to moderate cell, raising the count from 1 to 3.

Degree of control	Degree of enthusiasm			
	Marked	Moderate	None	Total
Well controlled	0	3	3	6
Poorly controlled	10	3	2	15
Uncontrolled	24	7	1	32
Total	34	13	6	53

Table 3 shows the association between degree of enthusiasm and degree of control for the operation of portacaval shunt [slightly revised, by adding two cases, from Grace, Muench, and Chalmers (*8*)]. The counts in Table 3 are not of patients but of investigations. Table 3 shows that, among the 53 investigations, only six were classified as "well-controlled." Among the 34 associated with "marked enthusiasm," none were rated by the in-

vestigators as "well-controlled." The "poorly-controlled" and the "uncontrolled" investigations generated approximately the same distribution of enthusiasm: about 72 percent "marked," 21 percent "moderate," and 6 percent "none." The six "well-controlled" investigations split 50-50 between enthusiasm levels "moderate" and "none." Muench, who participated in collecting these data, has a set of statistical laws (*10*), one of which says essentially that nothing improves the performance of an innovation as much as the lack of controls. Because tables for other therapies have given similar results, one must be cautious in accepting results of weakly controlled investigations.

In Table 3, the rows for "poorly controlled" and "uncontrolled" studies suggest that repeated, weakly controlled trials are likely to agree and build up an illusion of strong evidence because of the large count of favorable studies.

Not only may this mislead us into adopting and maintaining an unproven therapy, but it may make proper studies more difficult to mount, as physicians become less and less inclined, for ethical reasons, to subject the issue to a carefully controlled trial lest the "benefits" of a seemingly proven useful therapy be withheld from some patients in the study.

Strengths of Belief

A controlled trial of innovative therapy may sometimes impose a sacrifice on the part of some patients by withholding the more favorable of a pair of treatments. However, prior to the trial we do not know which is the favorable therapy. Only after the trial can the winner be identified. Some will say that the physician must have an initial guess, however ill-founded. It is unlikely that his view of the two competing treatments is exactly 50-50. The question then arises: If the physician fails to act on such a preference, is the patient getting responsible care? To help consider this question, let us review information obtained from experiments on incidental information.

Alpert and Raiffa (*11*) have performed a number of experiments on guessing behavior. Individuals were asked to estimate quantities about which they might have been expected to have some incidental information, such as the fraction of students in their class having a particular characteristic. Subjects were graduate students in the Faculty of Arts and Sciences and in the Graduate School of Business Administration at Harvard University. In addition to the basic estimate, the graduate students were asked to provide numbers below which various subjective probabilities would lie. If we think of the upper and lower 1 percent intervals as ones where a responder would be seriously surprised to find the true answer (that is to say, the responder felt 98 percent sure that the answer would lie between the chosen 1 percent and the 99 percent levels), then these responders were seriously surprised in 42.6 percent of the guesses or about 21 times as often as they should have been if the subjective estimates matched the true frequencies. Alpert and Raiffa's work (*11*) shows that experienced adults are likely to overrate the preciseness of their estimates. These people were too sure of their information. Although these people were not physicians in a patient relation, they were well educated and engaged in thoughtful work. Until we get contrary information from more relevant studies, such data suggest that strong initial preferences for therapies yet to be tested by controlled trials should be viewed with reserve. And, of course, the distribution shown in Fig. 1 and the results of Table 1 also show that, for therapies tested in trials, holding a view not far from 50-50 has some empirical foundation for surgery.

Shapiro (*12*) gives examples of wide variation among different physicians' estimates of probabilities in therapeutic situations, data pertinent to this discussion, but not the same as the Alpert-Raiffa point. Shapiro shows that physicians differ a great deal in their estimates; Alpert and Raiffa show that people are very frequently much further off than they expect to be.

Do We Owe the Past or Future?

Let us consider the question of whether a present patient should give up something for future patients. We, or our insurance carriers, pay the monetary cost of our care. What we do not pay for is the contribution to the medical system by past patients. These patients, through their suffering and participation in studies, have contributed through their illness and treatments to the present state of evidence for all patients. Such contributions cannot be purchased by money but can be repaid in part by making, when appropriate, a contribution to the same system. One good way is through participation in well-designed clinical trials when the patient falls into the limbo of medical knowledge. Other nonmonetary ways are donating blood and organs. So one may feel an obligation to the system of medicine that has reached its present state without his or her assistance, and in addition each person has an interest in its general improvement as we next explain. [For a recent treatment of this point see Almy (*13*).]

In some circumstances, participation in the trial may turn out to be of help to the patient. Aside from the luck of getting the best therapy of several that are offered, this occurs, for example, when the patient has a disease for which treatments can be readily changed after the trial. Nevertheless, there are circumstances when the treatment is not reversible and when the chances are that the specific trial will be of little individual benefit—that is, when it has but slight chance of being a benefit to the patient, his family, or friends.

Under these circumstances, the patient may still be willing to participate in a trial. If the trial is recognized as part of a general system of trials in which patients participate only on such occasions as they qualify and when a trial seems necessary, then the patient may well benefit in the future not so much from the results of the particular trial he or she participates in but from the system that gives rise to it. Findings will come forward on many other diseases and the patient, or someone dear to him, will be likely to suffer from some of those diseases whose trials will have produced useful findings. It is not so much, then, the direct payoff of this present trial that we should have our eye on, but pooled benefits of the whole system. The longer the patient lives, the more likely it is that he or she will suffer from some other of the diseases being studied by careful trials. And insofar as they are not studied by careful trials, the appropriate conclusions may be slow in coming. By putting off the day when strong evidence is obtained, we reduce the patient's chances of benefiting most fully from modern medicine. Thus the patient has an interest not only in the trial he or she has the opportunity to engage in, but also a stake in a whole system that produces improved results that may well offer benefits in the future, if the patient survives the present difficulty. Thus, the social system will likely offer benefits through the larger system even when a particular component of the system may fail to pay off directly for a patient, his family, friends, or some other social group he belongs to.

A further statistical point that may not be much appreciated by potential participants in randomized trials is that the inferences apply primarily to the popu-

lation sampled in the study. To the extent that individuals or groups decline to participate in studies, and to the extent that their responses may differ from those of the rest of the population (an interaction between participation and response to therapy), the treatments selected may not apply to them as well as to participants and people "like" them. For example, if those in the lower economic status were less likely to participate and if economic status related to the differential effectiveness of therapies, say, through additional lack of compliance, the study will not properly appreciate the value of the therapy for the nonparticipating group.

The lone individual may seem to have little incentive to participate because one seems so few among many. But the stake is not in any one person appearing in this study; it is in having people from segments of the population that represent that individual being properly represented in this and other studies so that the results of the whole system may be more assuredly applied to this patient when disease strikes. The idea is similar to that of being told not to complain of the system when one does not vote. But the extra feature here is that one gets to vote on certain special occasions, and then only a few are admitted to the booth, and so each opportunity to vote weighs much more heavily than usual.

If certain groups tend not to participate in the evaluative system, then they will not find medical evaluations of therapies as well pointed to their needs as if they did participate. Thus, each individual has a stake in wanting people like themselves represented. Since it is hard to say what "people like themselves" means, the good solution is to have the whole appropriate population volunteering in all the therapies tested. Participating presumably encourages others like me to participate too, and vice versa.

The main point of this discussion is that if participation seems to the patient to be a sacrifice, it should be noted that others are making similar sacrifices in aid of the patient's future illnesses. So even if the particular trial may not help the patient much, the whole system is being upgraded for his or her benefit. We have a special sort of statistical morality and exchange that needs appreciation.

Responsibility for Research

Much of current popular discussion of the ethnical issue takes the position that physicians should use their best judgment in prescribing for a patient. To what extent the physician is responsible for the quality of the judgment is not much discussed, except to say that he must keep abreast of the times. Some physicians will feel an obligation to find out that goes beyond the mere holding of an opinion. Such physicians will feel a responsibility to contribute to research. In similar fashion, some current patients may feel a responsibility to contribute to the better care of future patients. The current model of the passive patient and the active ongoing physician is not the most effective one for a society that not only wants cures rather than sympathy, but insists on them—a society that has been willing to pay both in patient cooperation and material resources for the necessary research.

Quality of Life

In addition to a society willing to support medical research through responsible experimentation on human beings, in addition to physicians dedicated to acquiring knowledge on behalf of the sick, we must be certain that controlled trials are designed to seek answers to the appropriate questions. In our survey, we found most concern with near-term outcomes, both mortality and morbidity.

We need additional data about the quality of life of patients. Among our initial sample of 107 papers drawn through the MEDLARS search, quality of life seemed often to be a major consideration, although rarely did papers address more than a few features of that quality (*14*). Because much of medicine and surgery is intended to improve quality rather than to save life, measuring the improvement is important. As we have indicated above, different therapies frequently produce about the same mortality and morbidity, and so the ultimate quality of life achieved would bear heavily on the choice. Thus, for proper evaluation of alternatives, we need to assess the patient's residual symptoms, state of restored health, feeling of well-being, limitations, new or restored capabilities, and responses to these advantages or disadvantages.

For surgery, we need long-term follow-up and both objective and subjective appraisals of the patient's quality of life. Frequently, the long-term follow-up is carried out, but overall quality of life is rarely measured. For example, among 16 cancer papers in the initial sample of 107, follow-ups ranged from 2 months to 2 decades. With few exceptions, survival and recurrence data were the principal information given, and because different treatments usually had similar rates, it would be fruitful to report contrasts among the treatments in the quality of life or death experienced by patients with the same disease but having different treatments. This might be especially appropriate because the therapies involved such features as castration, hormones, irradiation, chemotherapy, and various amounts of surgery. Developing and collecting suitable measures for quality of life after surgery requires leadership from surgeons and the cooperation of social scientists. We hope these developments will soon take place.

Summary

Approximately half the surgical innovations tested by randomized clinical trials provide improvements. For those where reduction in percent of complications was a useful measure, we estimate that about 24 percent of the innovations gave at least a 20 percent reduction in complications. Unfortunately, about 12 percent of the innovations gave at least a 10 percent increase in complications.

Therefore, keeping gains and discarding losses requires careful trials. Gains of these magnitudes are important but are hard to recognize on the basis of incidental observations. When well-controlled trials have not been used, sometimes data have piled up in a direction contrary to that later found by well-controlled trials. This not only impedes progress but may make carefully controlled trials harder to organize. Most of the trials we studied did not have large sample groups. To dependably identify gains of the magnitude we found in the discussion on surgery and anesthesia, trials must be designed carefully with sufficient statistical power (large enough sample sizes) and appropriate controls, such as may be provided by randomization and blindness. As Rutstein (*15*) suggests:

> It may be accepted as a maxim that a poorly or improperly designed study involving human subjects . . . is by definition unethical. Moreover, when a study is in itself scientifically invalid, all other ethical considerations become irrelevant. There is no point in obtaining "informed consent" to perform a useless study.

When we think of the costs of randomized trials, we may mistakenly compare these costs with those of basic research. A more relevant comparison is with the losses that will be sustained by a process that is more likely to choose a less desir-

able therapy and continue to administer it for years. The cost of trials is part of the development cost of therapy. Sometimes costs of trials are inflated by large factors by including the costs of the therapies that would in any case have been delivered rather than the marginal cost of the management of the trial. This mistake is especially likely to be made when a trial is embedded in a large national program, and this is also the place where trials are highly valuable because their findings can be extended to a whole program.

Surgical treatment frequently trades short-term risk and discomfort for an improved longer term quality of life. While long-term follow-up is frequently reported, a vigorous effort is needed to develop suitable measures of quality of life.

Table 1 gives empirical evidence that, when surgical trials are carried out, the preferable treatments are not known in advance. Although a common situation in a trial would be that the innovation was expected to be a clear winner, the outcome is in grave doubt. Empirical evidence from nonmedical fields suggests that educated "guesses" even by experienced, intelligent adults are way off about half the time. For these reasons we discount the pretrial expectations or hunches of physicians and other investigators.

Most innovations in surgery and anesthesia, when subjected to careful trial, show gains or losses close to zero when compared to standards, and the occasional marked gains are almost offset by clear losses. The experimental group is neither much better nor much worse off than the control group in most trials, and we have little basis for selecting between them prior to the trial.

The one sure loser in this system 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. Let us recall that our whole financial industry is based on a continuing return of a few percentage points.

All in all, the record in surgery and anesthesia is encouraging. We regard a finding of 50 percent or more successes for innovations in surgical and anesthetic experiments as a substantial gain and a clear opportunity for additional future gains. Well-conducted randomized clinical trials are being done. All of us, as potential patients, can be grateful for a system in which new therapeutic ideas are subjected to careful systematic evaluation.

Appendix

Estimating the distribution of gains. The model of the process is that of two-stage sampling. We regard the innovation and its paired standard as drawn from a population of pairs of competing therapies. Let Z be the random variable corresponding to the improvement offered by the innovation (innovation minus standard), with mean M and variance A. For the ith innovation with true gain Z_i, the experiment assesses the gain as W_i, and W_i has mean Z_i and variance D_i.

If we assume as an approximation that the distributions of Z_i and W_i are normal, then the posterior distribution of Z_i has mean

$$Z_i^* = M^* + e_i(W_i - M^*)$$

where

$$e_i = A^*/(A^* + D_i)$$

A^* is an estimate of A, and M^* is an estimate of M. The posterior distribution of Z_i is approximately normal with mean Z_i^* and variance $(1 - B_i)W_i$, where

$$B_i = D_i/(A^* + D_i)$$

In the current problem the D's are estimated from binomial theory because the W's are the difference between two independent observed proportions. Details of obtaining A^* and M^* are given in (*6*).

To estimate the cumulative distribution of Z, we compute for each observation W_i

$$c_i = \frac{z - Z_i^*}{(1 - B_i)D_i}$$

then using normal theory we compute

$$\Phi(c_i) = P(X < c_i)$$

where X is a standard normal random variable. Thus

$$\Phi(c_i) = [1/\sqrt{2\pi}\,] \int_{-\infty}^{c_i} \exp(-\tfrac{1}{2}x^2)\,dx$$

Finally

$$\sum_{i=1}^{k} \Phi(c_i)/k$$

estimates $P(Z < z)$ for each value of z. We thus release ourselves from the original normal approximation for Z and get a new distribution that is not normal but should be an improved approximation of the true distribution. When weights were used because one study gave two comparisons they modified both the estimation of A and W and the estimation of $P(Z < z)$.

References and Notes

1. For a discussion of MEDLARS, see M. Day, *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **33**, 1717 (1974). Indexing is done by specially trained abstractors at the National Library of Medicine. The MEDLARS contents vary over time as additions are made to correct omissions. Our initial search turned up 36 randomized clinical trials. These are listed in (*6*), appendix 9-1, pp. 145–154. A repeat search, approximately 18 months later, done according to the same search instructions, revealed 13 additional randomized clinical trials as follows: R. B. Noone, P. Randall, S. E. Stool, R. Hamilton, R. A. Winchester, *Cleft Palate J.* **10**, 23 (1973); R. Smith, *Trans. Ophthalmol. Soc. Aust.* **27**, 17 (1968); B. Brehmer and P. O. Madsen, *J. Urol.* **108**, 719 (1972); J. E. Rothermel, J. B. Wessinger, F. E. Stinchfield, *Arch. Surg.* **106**, 135 (1973); R. K. Laros, G. I. Zatuchni, G. J. Andros, *Obstet. Gynecol.* **41**, 397 (1973); W. H. Harris, E. W. Salzman, R. W. DeSanctis, R. D. Coutts, *J. Am. Med. Assoc.* **220**, 1319 (1972); J. W. Roddick, Jr., and R. H. Greenelaw, *Am. J. Obstet. Gynecol.* **109**, 754 (1971); I. L. Rosenberg, N. G. Graham, F. T. DeDombal, J. C. Goligher, *Br. J. Surg.* **58**, 266 (1971); R. Brisman, L. C. Parks, J. A. Haller, Jr., *Ann. Surg.* **174**, 137 (1971); J. M. Lambie, D. C. Barber, D. P. Dhall, N. A. Matheson, *Br. Med. J.* **2**, 144 (1970); D. B. Haverstadt and G. W. Leadbetter, Jr., *J. Urol.* **100**, 297 (1968); J. A. Haller, Jr., *et al.*, *Ann Surg.* **177**, 595 (1973); D. J. Pinto, *East Afr. Med. J.* **49**, 643 (1972).
2. A. B. Miller, W. Fox, R. Tall, *Lancet* **1969-II**, 501 (1969).
3. B. Efron and C. Morris, *J. Am. Stat. Assoc.* **68**, 117 (1973).
4. ———, *J. R. Stat. Soc. B* **35**, 379 (1973).
5. ———, *Sci. Am.* **236**, (No. 5) 119 (1977).
6. J. P. Gilbert, B. McPeek, F. Mosteller, in *Costs, Risks, and Benefits of Surgery*, J. P. Bunker, B. A. Barnes, F. Mosteller, Eds. (Oxford Univ. Press, New York, 1977), chap. 9, pp. 124–169. For formulas see pp. 156–161.
7. F. C. Nance and I. Cohn, Jr., *Ann. Surg.* **170**, 569 (1969).
8. N. D. Grace, H. Muench, T. C. Chalmers, *Gastroenterology* **50**, 684 (1966).
9. T. C. Chalmers, J. B. Block, S. Lee, *N. Engl. J. Med.* **287**, 75 (1972).
10. J. E. Bearman, R. B. Loewenson, W. H. Gullen, "Muench's postulates, laws, and corollaries," *Biometrics Note 4* (National Eye Institute, Bethesda, Md. 1974).
11. M. Alpert and H. Raiffa, "A progress report on the training of probability assessors," unpublished paper, Harvard University (28 August 1969).
12. A. R. Shapiro, *N. Engl. J. Med.* **296**, 1509 (1977).
13. T. P. Almy, *ibid.* **297**, 165 (1977).
14. 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), chap. 10, pp. 170–175. The results of the initial sample of 107 are reported in (*6*). Of these, 36 were randomized and 34 could be used, 11 were nonrandomized controlled trials, and 59 were series (study of one therapy). Our additional sample added 13 randomized trials for use in Table 1. Chapter 10 discusses quality of life.
15. D. Rutstein, *Daedalus* **98**, 523 (1969).

16. This work was facilitated by NIH grant Gm 15904 to Harvard University, by the Miller Institute for Basic Research in Science, University of California, Berkeley, and by NSF grant SOC75-15702 A01. We appreciate the advice and assistance of A. Bigelow, M. Ettling, M. Gasko-Green, D. Hoaglin, V. Miké, A. Perunak, K. Soper, J. W. Tukey, and G. Wong.