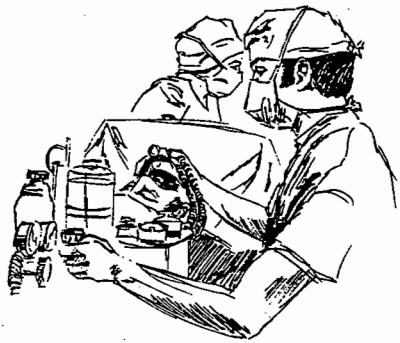


HOW FREQUENTLY DO INNOVATIONS SUCCEED IN SURGERY AND ANESTHESIA?

John P. Gilbert *Harvard University*
Bucknam McPeek *Massachusetts General Hospital, Boston*
Frederick Mosteller *Harvard University*

WHEN THERAPIES are compared for effectiveness, what happens? How often does an innovation appear to be superior to its competitors? When innovations are successful, for example, the Salk vaccine or the development of successful organ transplantation, society gains a major victory. This paper studies the effectiveness of new surgical and anesthetic therapies in their clinical setting.

We reviewed a sample of 107 published papers appraising surgical and anesthetic treatments. Of these therapies sufficiently promising to be tested in human patients, we ask what proportion have proved to be substantial improvements over existing ones? What proportion have been moderately successful? And what proportion have been found to be less effective than had been hoped and expected?

Using these papers, we assess the percentage improvement a new innovation is apt to make, as well as the chance that it will turn out to have been an improvement at all. Thus our aim is to describe the crop of newly tested therapies for effectiveness compared with that of the treatments they are designed to replace.

Except for a major breakthrough like the introduction of antibiotics, we have little reason to suppose that the development of new therapeutic ideas will change drastically. Thus, we assume that a similar distribution of successes and failures will occur in the near future. The results presented here should give realistic expectations at least for the short term.

We drew a sample of papers evaluating different treatments actually given to patients. To get this sample, we turned to the National Library of Medicine's MEDical Literature Analysis and Retrieval System (MEDLARS). Computer-produced bibliographies can be retrieved from this data base which, since January 1964, has provided an exhaustive coverage of the world's medical literature. By searching the system for prospective studies (see the essay by Brown) of specified surgical operations or anesthetic drugs, we were able to gather papers whose authors used human patients to evaluate surgical and anesthetic treatments. The papers appeared between 1964 and 1972.

We considered only papers in English because of our own language disabilities, and papers with ten or more patients in a group because we wanted to study large investigations rather than case studies. Any other bias in the sample selections, then, arose from peculiarities of the MEDLARS indexing system and contents at the time of the search rather than from our prejudices.

The papers included many kinds of studies. To give an idea of the variety, some dealt with ulcers, appendectomy, cirrhosis, cancers, bone operations, colon operations, major vascular operations, stab wounds, antibiotics, clot prevention, drainage, and the impact of anesthetic drugs and techniques.

Our sampled papers reported on three basic types of studies—*randomized controlled trials, non-randomized controlled trials,* and *series.* We use the term randomized controlled trials when the investigator compared two or more treatment groups and assigned patients to the groups by a formal randomization process (such as drawing random numbers to decide which treatment is assigned to each patient). The non-randomized controlled trials did not have such a formal randomization process and varied from comparing groups treated concurrently in the same institution to comparing patients treated previously by one method with patients treated currently with another. The papers reporting on series described sets of patients treated in some specified manner but with no comparison except possibly with other reports in the literature dealing with similar patients. In the rest of this paper we are concerned with only the papers dealing with randomized controlled trials.

If our MEDLARS approach were perfect and produced all the papers, one might think that we have a census rather than a sample of papers. To adopt this attitude would be to misunderstand our purpose. We think of a process producing these research studies through time, and we think of our sample—even if it were a census—as a sample in time from this continuing process. Thus our inference would be to the general process, even if we did have all appropriate papers from a time period.

In appraising the results of comparative investigations, we take several simplifying actions.

First, we classify each therapy as either an *innovation* or as a *standard.* Some diseases have a widely recognized standard therapy against which all others are measured. A good example of this has been (the standard) radical mastectomy for cancer of the breast. In such instances the standard is easy to recognize; all others can be considered as competing innovations regardless of how recent their introduction. We have used the letter "I" to denote the treatment we regarded as an innovation and the letter "S" for the standard.

Ideally one wishes an analysis to produce the maximum amount of information contained in a body of data. It is often impractical to achieve this in practice. Thus in trying to evaluate the difference in performance between standard programs and innovations in our study we were unable to assign a highly accurate and precise value to the observed differences. In many studies we are content with knowing how many differences were positive and how many negative. In our data often the two programs were essentially equal in performance and so it was useful to acknowledge this in the scale. In addition sometimes one program was not only better but was clearly much better, and it was not hard to make a distinction between these two. The five point scale that we have used is a happy solution to this problem because it is relatively simple and easy to apply and it retains most of the relevant information that we need. The five point scale is widely used in both social science and medicine because it allows us to capture much of the information we want in a practical manner.

Second, we speak below of a pair of competitive therapies as having three possible relations: About equal (S=I), the first named preferred to the second named (S>I or I>S), and the first named *highly* preferred to the second named (S>>I or I>>S). We have tried to report on this scale what we think the original investigators would have reported. Usually their words make this clear.

Third, we have divided therapies into two classes: *Primary* therapies intended to cure or ameliorate the patient's primary disease, and *secondary* therapies intended to prevent or treat such complications as infection or thromboembolic disease or to offer improvements in anesthesia or postoperative care. The basic 107 studies included 36 randomized clinical trials. Of these 36 papers, 21 deal with *primary* therapies and 15 deal with *secondary* therapies. For technical reasons* several studies had to be set aside, also

*One study had too many comparisons; another had too small a sample size for its complicated design.

some studies had more than one comparison. In Table I, we deal with comparisons, rather than studies. By coincidence the number of papers equals the number of comparisons in the analysis.

Referring to Table 1 for randomized trials, we see that in five of the 36 comparisons, or about 14%, an innovation was highly preferred to a standard. In 16 comparisons, including the previous five, about 44%, the new therapy was regarded as successful, sometimes because it was no worse than a standard and thus became available as an alternative.

TABLE 1. Summary for Innovations in Randomized Clinical Trials

		PRIMARY	SECONDARY	TOTAL
I > > S:	Innovation highly preferred	1	4	5
I > S:	Innovation preferred	5	2	7
I = S:	About equal, innovation a success	2	2	4
I = S:	About equal, innovation a disappointment	7	3	10
S > I:	Standard preferred	3	3	6
S > > I:	Standard highly preferred	1	3	4
	Comparisons	19	17	36

In 10, or 28%, the equality of an innovation with a standard could be regarded as a disappointment because, although the innovation was more trouble or more costly or more risky, it did not perform better. In 20 comparisons, about 56%, a standard was preferred (counting innovative disappointments) to an innovation.

Overall, Table 1 shows that innovations highly preferred to standard treatments are hard but not impossible to find, and that almost half of the innovations provided some positive gain. It is worth reflecting on what our attitude might be toward extreme findings in either direction. Suppose that nearly all studies, or even the lion's share, found the innovation highly preferred; one would have to conclude that standard therapies were fairly easy to improve on and indeed that the kind of medicine being appraised was in its infancy or else that a sudden breakthrough had been made on all fronts. This is unlikely with as many different diseases and therapies as occur in the sample. At another extreme, if no substantial gains occurred, the suggestion is that the field has topped out, at least during the period of the study, awaiting some new insights.

Figure 1 summarizes 11 primary studies in which survival was an appropriate measure of outcome and plots the percentage of survivors, often after many years, for the standard therapy against that for the innovation. Two papers had two comparisons of a standard against an innovation, making 13 comparisons in all. The seven points below the 45° diagonal line show the

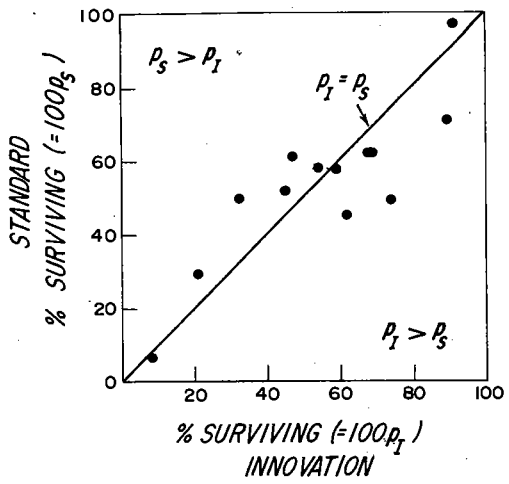


FIGURE 1

Primary Therapies—Survival Percentages

Points falling below the 45° line indicate higher survival rates for the innovation than for the standard. Primary therapies are intended to cure the patient's disease.

innovation performing better; the six points above show the standard performing better. The greatest observed gain (shown by the point farthest below the line with coordinates (74%, 49%)) comes from a study of therapeutic portacaval shunt. Curiously, the point farthest above the line (showing the greatest apparent loss) corresponds to a study of the same operation performed prophylactically, in advance of urgent need (32%, 50%). The overall impression given by the figure is one of points rather closely hugging the diagonal line. The degree of scatter from the line depends in part upon

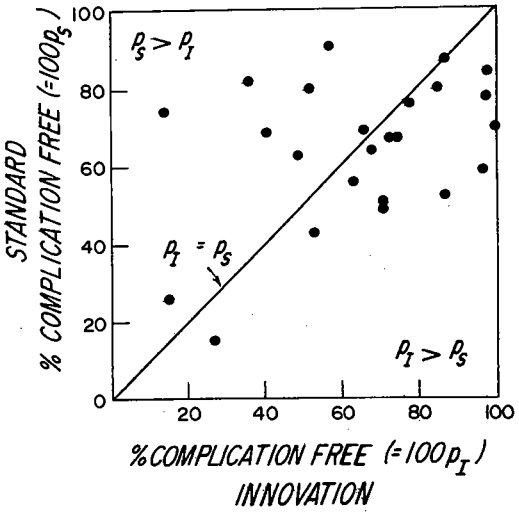


FIGURE 2

Secondary Therapies—Percentage Free of Complication

Percentages avoiding specific postoperative complications. Secondary therapies are intended to reduce the frequency of postoperative complications. Points below the 45° line indicate fewer complications of a specific sort accompanying the innovation than the standard. The innovations have 15 below, 8 above, and 1 on the line.

the size of samples (number of patients used in the studies), and we explore this idea further later.

Figure 2 shows 24 comparisons based on 11 secondary studies (five had 1 comparison, four had 2, one had 3, one had 8). The 15 points below the line indicate the innovation as an improvement over the standard treatment; the 8 above indicate the reverse, and the 1 on the line gives a tie.

The overall scatter about the 45° line in Figure 2 is large, encouraging us to believe that larger percentage differences have been found here than in the studies of Figure 1. By and large, the changes in rate of complications are larger than the changes in survival rate. We make this more quantitative below.

In the work reported so far, some innovations performed better than a standard, others worse. We next regard these outcomes as a sample from the population of all those surgical innovations developed by our medical system and tested by randomized clinical trials. Every study has its uncertainties associated with sampling variability and other sources of unreliability. We want to allow for sampling variability in our description of the gains and losses. The general idea is that if we focus on a particular sort of performance, we may be able to gather strength from several studies even though they deal with disparate operations. For example, among the primary studies we focus on those where the main hope from the operation is the extension of life. Then we might ask about the distribution (variety) of improvements actually achieved by this type of innovation.

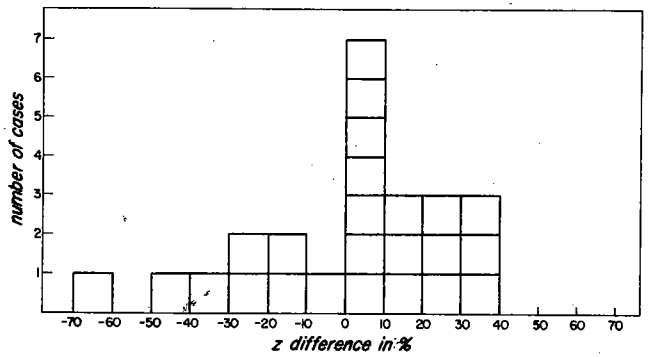
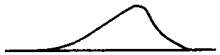


FIGURE 3

Histogram of 24 observed differences in percentages avoiding a complication for several operations. This illustrates one way of representing a frequency distribution function.

We use the idea of distributions, and so we want to illustrate what these are. In our study of post-operative complications we had 24 differences of the form: the percentage of patients who did not develop complication under the innovation MINUS the percentage who did not develop the complication under the standard. We use these 24 differences for illustration. We can ask of the data how many differences were in intervals of length 10 such as between 0 and 9%, 10% and 19%, or between –20% and –29%. This information is presented graphically in Figure 3.

Seven of the differences fell in the interval 0 to 9%, three in each of the intervals 10 to 19%, 20 to 29%, and 30 to 39%, while one was so low that it fell in the interval –60% to –70%. Thus Figure 3 gives us an idea of how these differences are distributed with regard to their values. If we had had many values we could have made much smaller intervals and we could think of it being very like an idealized smooth curve that might look like:



This curve is called the density function of the distribution.

Often it is more relevant or convenient to ask how many data points were larger than a particular value on the scale, rather than asking how many data points were within a particular interval as we did above. If we represent this way of looking at the data graphically we obtain Figure 4.

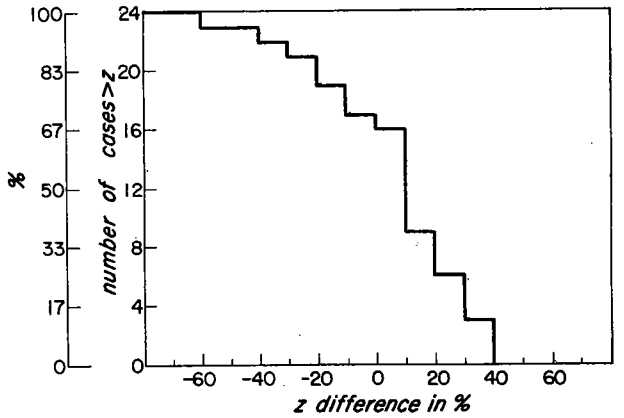
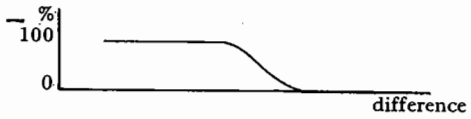


FIGURE 4

Frequency distribution cumulated from the right to show the frequency of getting a given percent difference at least as large as the one on the horizontal axis.

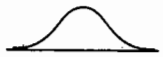
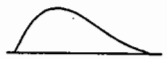
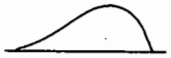
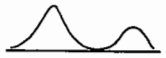
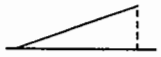
Figure 4 was obtained by adding up the number of squares, i.e. observations, to the right of each point on the horizontal scale of Figure 1. For this reason it is called a cumulative distribution function. If again we think of having many more points and using much smaller intervals we might find the figure to approach a continuous curve that is the cumulative distribution function that corresponds to the continuous density pictured above. This curve looks like:



When the vertical axis is a percent or a proportion, we can read off the estimated probability of a difference at least as large as the one on the horizontal axis.

If every study were based on an enormous sample of patients, so that sampling errors would be very small, the reports of gains and losses would give us the distribution of differences in true performance between innovations and standards in our sample of papers. In turn that sample distribution would estimate the distribution of gains in the population—the process generating these studies and comparisons. But studies are of necessity limited in size, and, in reports of small studies, differences vary more due to sampling error than in large ones. We need to have a way to pool the results of such studies, large and small, that will give an idea of the distribution of *true* gains and losses in the trials.

One such method is to allow for the sampling variability associated with specific randomized trials and come up with a pooled figure. The observed difference may be thought of as having two additive components—the true difference plus the sampling error. A special statistical technique called analysis of variance produces estimates of the average and standard deviation of the sample of true differences. (The standard deviation measures how spread out the distribution is.) We can estimate how often various sizes of gains can be expected to occur by making an assumption about the true differences, namely that the differences approximately follow a normal distribution. (The word normal here applies to a particular shape of distribution—it is not being used in the sense of normal versus abnormal. Heights of adults and scores on achievement tests are examples of distributions that are approximately normal in shape. A normal distribution looks something like


, but other distributions might look like

, or

, or even

, or

for example.)

It is important to understand that this method develops summary statistics for *true* gains and losses *across* studies. The statistics reported are (a) the

estimated average true gain, averaged across comparisons, and (b) the estimated standard deviation of the true gain, averaged across comparisons. If, for example, the average gain were 0% and the standard deviation of the gain 6%, our assumption of a normal distribution allows us to calculate that gains of 10% or more could occur in about one-twentieth of the opportunities. It would still be true that the gain would be positive for about half the innovations and negative (that is, a loss) in half, in agreement with Table 1. It then becomes the goal of clinical research to identify the favorable and unfavorable innovations so that we may use the former and avoid the latter.

The statistics in Table 2 summarize the results.

TABLE 2 Analysis of Variance Estimates of Average and Standard Deviation of True Gains

	ESTIMATED AVERAGE GAIN	ESTIMATED STANDARD DEVIATION OF GAINS
Primary Therapies	1.5%	8%
Secondary Therapies	0.4%	21%

The average gain for the primary therapies is not far from zero, a result that agrees with our more qualitative analysis of Table 1. A zero average gain is consistent with some innovations having substantial improvements balanced by others having substantial losses or with other mixes such as many small gains and a few large losses. The size of the estimated standard deviation of effects of innovations lends added support to such interpretations. (A zero standard deviation would imply that all innovations give essentially the same amount of improvement.) And we know that some of these innovations do produce substantial improvements even when sample size is taken into account.

These figures also yield a rough guess about the proportion of comparisons having true differences favoring the innovation as great as, say, 10%. For the primary therapies, the probability that a new therapy has a positive gain of at least 10%, if the sample represents the future well, is about 0.13, or about 13 chances in 100.

For the secondary therapies, a gain of at least 10% (a 10% reduction in a specific complication) has a probability of 0.32.

The above procedure is rough and ready and leans hard upon an assumption of a normal distribution in its calculation, but the real distribution may not be normal. A new approach called "empirical Bayes" (Efron and Morris 1973) offers an alternative.

If each comparison were based on an infinitely large experiment, we would know the true gain exactly for that comparison. Then to estimate the proportion of gains of more than 10% we would count the number of comparisons with gains larger than 10% and divide by the total number of comparisons. And so if we had 25 comparisons and 5 had gains greater than 10%, we would estimate the probability of a gain of more than 10% as 5/25 or 0.20. This approach does not lean on any assumption about the shape of the distribution of true gains. But we can't use it because we do not have infinitely large experiments.

The new method takes note of the uncertainty associated with each observed gain, primarily using the sample sizes. Instead of regarding an observed gain as greater than 10%, or not greater, it estimates the probability that the true gain is greater than 10%. And so each comparison yields a probability of being greater than 10%, and we average these probabilities from all the comparisons to get our estimate of the overall probability of a gain of more than 10%.

If the observed gain is very large, say 30%, then its probability is nearly 1 (0.99, for example, or 99 chances out of 100) of having a true gain of more than 10% because the variability of the experimental observation is very much less than the 20% difference between 10% and 30%. This 0.99 corresponds to the 1 we would have counted toward the numerator (our 5 of the 5/25) had we known the true gain exactly. If the observed gain is negative, the probability that the true value exceeds 10% will be small, perhaps 0.01. This 0.01 is like the 0 we would have counted for this comparison had we known it to be exactly the true gain. When the observed gain is exactly 10%, the probability is 0.5 that the true gain is larger than 10% and 0.5 that it is smaller.

The same technique applies to finding the chance of gains greater than 5% or 0% or −7%, and so on. The resulting set of probabilities are conveniently graphed and thereby summarized by a cumulative distribution as shown in Figure 3.

Figure 5 shows the estimated cumulative distributions of the true gains in percentages for the primary and for the secondary therapies. By picking a gain in percentage, z, and reading the corresponding vertical axis on the appropriate curve, one can estimate the probability of a new therapy producing a gain as large as or larger than the chosen value of z.

For examples we have:

(a) For primary therapies the chances (i) of a 10% gain or more in survival are about 4 in 100 (we regard this as a better estimate than that given earlier [13 in 100] because we prefer the method rather than because we prefer the answer), (ii) of a 0% gain or more are about 48 in 100, (iii) of a loss of no more than 10% are about 98 in 100 which means that the chances of a loss in excess of 10% are about 2 in 100.

(b) For secondary therapies, the chances (i) of a gain of 10% or more in a specially chosen complication are estimated as 38 in 100 which is close to the earlier 32 in 100, (ii) of a 0% gain or more 57 in 100, (iii) of a loss of no more than 10% as 72 in 100, which means that a loss of 10% or more has chances of about 28 in 100.

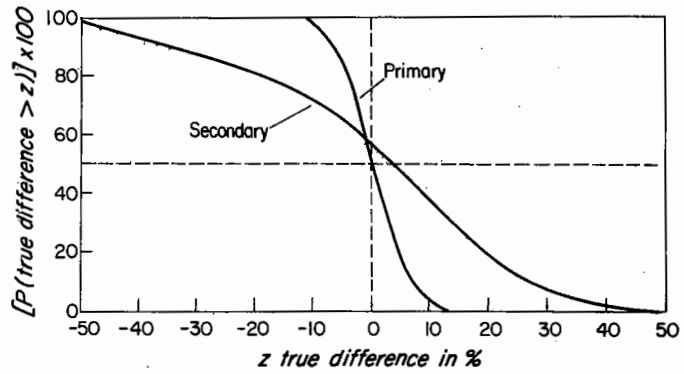


FIGURE 5

The Probability of a New Treatment Producing a Gain as Large as or Larger Than a Chosen Value, z
To find the probability of a difference in percentages greater than a given number z, say z = 10%, erect a perpendicular from 10% on the horizontal axis to the appropriate curve, and read its ordinate off the vertical axis, about 0.04 for the primary and about 0.38 for the secondary.

Is there some reason that secondary therapies are more likely to succeed than primary therapies? Is there something special about a treatment aimed at the disease process itself? We think the difference arises in large measure because in our quantitative analysis we chose to analyze primary therapies in which survival was an appropriate measure of outcome, while for secondary therapies the measure was avoidance of a specific complication. In a way the incidence of a specified complication is a much more discrete measure. One can envision a treatment having a large effect on a specific complication, whereas the difference between life and death may be the sum of the effects of a variety of factors—the primary treatment, the primary disease process, secondary treatments, and a variety of other disease processes and

factors like old age and inter-current disease. Over the last generation the expected length of life has increased only slightly, but great changes have occurred in the variety and extent of postoperative complications as a result of changes in therapy such as the introduction of antibiotics, and of newer anesthetic agents, and techniques.

The sample of published papers is objectively chosen, and we think rather a good one for reflecting the sorts of differences analyzed here. What is less clear is how good a sample it is of therapeutic surgical research on patients generally during this period. First it is likely, and those we have talked with agree, that published papers are reports on better work on the average than that in unpublished research. Second, research that turns out well, our discussants agree, is more likely to be published. This reasoning suggests that the mass of unpublished research, insofar as it might produce measures comparable to those described in this paper, would have a lower average performance for innovations compared with standards than those in our sample. We suppose then that the innovations assessed by randomized clinical trials and reported in the surgical literature, and here, are biased upwards—that is, they present a more promising picture for innovation than if all innovations were subjected to randomized clinical trials. No doubt some innovations are so unsatisfactory that they are quickly abandoned, along with whatever trials were initiated on them. These conjectures suggest that if one were to consider adjusting the distributions shown in Figure 3 to report on all surgical innovations versus standards, the mean of the distribution would be lower and the standard deviation would probably be larger to allow for more frequent large negative differences. We have no grounds but speculation for the amount of such changes.

In a recent review of randomized trials used in evaluating social programs (Gilbert, Light, and Mosteller, 1976), the authors concluded that many new programs do not work and the effects of those that do are usually small. In contrast to these findings in the area of social innovations, this review provides strong evidence for a more optimistic view of the rate of progress in surgery and anesthesia. Almost half of the innovations reported in this series of controlled trials were at least as good as the standard, and a fair number were substantially better. Thus the analyses suggest that four out of ten innovations in secondary therapy produce a reduction in complication rates of 10% or more while two or three out of ten innovations in primary therapy produce a 5% or greater increase in survival. These estimates are for the distribution of the underlying true effects of the innovations. In a sense these results describe the clinical judgment that chooses those innovations as promising enough to test. If innovations were successful in a high proportion of trials, it would suggest that new therapies were being delayed until we were absolutely sure of their success, while if almost none were successful it would suggest a scarcity of new ideas in the field. Thus these distributions also describe research productivity and its effects on the development of better clinical care.

Another view is that this research process tries to reject all the innovations that produce losses and to keep the ones with gains. If this were done, then the median gain (the middle gain) retained for the primary therapies would be about 4%, and that for the secondary therapies would be about 15%. Of course, this would be an idealized state, for we cannot hope to weed out *all* the losses and detect *all* the gains. But it gives an estimated upper limit to what could be accomplished.

We further emphasize that to say that a proportion of innovations are substantial improvements does not serve to identify which they are.

Our findings give us an idea of the sorts of gains that can be made from selecting the better of pairs of therapies that are tested by randomized trials (we do not discuss the others here). The left sides of the curves warn us also that innovations may lose rather than gain, and so evaluation is needed. For example, in secondary therapies losses of as much as 20% could occur about one-fifth of the time. These curves emphasize the size and frequency of the losses as well as the gains. Thus as physicians well know, one cannot assume in advance that a new treatment is an improvement over an old, even when it looks promising enough to warrant a clinical trial. Our distributions show that some innovations provide important gains for the clinical care of patients, such as reducing a death rate by 5%.

To give an idea of the risks represented by a five percent change in death rate, we note that among all the people who died in recent years, 5% were in the ten year age range 40–49. Thus we can think of this rate as corresponding to the natural losses over a ten year period at middle age. Another way to think of 5% is that it is about four times the average surgical death rate from all operations over the country as a whole. Thus its importance is not small.

Reducing a death rate from 35% to 30% may be an important improvement in patient care, but this does not mean that it will be easily identified in the everyday setting of clinical practice. Indeed, statistical theory shows that a well-run randomized controlled trial would need 1,105 patients in each group to be 80% confident of detecting such a difference. Without a large formal trial, the uncontrolled effects of patient selection, of concurrent treatments, and of other factors make the detection of such differences even more difficult.

Since relatively small, even though important, numerical gains or losses are to be expected from most innovations, clinical trials must regularly be designed to detect these small differences accurately and reliably. Our sampled papers, taken as a group, provide an optimistic picture of progress in surgery and anesthesia. This progress depends on a judicious combination of continued development of new therapeutic ideas and their evaluation in good-sized unbiased clinical trials.

PROBLEMS

1. What are the three basic types of studies reported in the sampled papers reviewed by the authors?

2. Describe the scale used in the evaluation of the innovations.

3. What is meant by primary and secondary therapies?

4. Why is it not appropriate to consider a complete enumeration of papers during a certain period of time (say, 1964–1972) as a census rather than as a sample?

5. Using Figure 3, calculate the estimated probability of an innovation being an improvement over the standard. (Hint: In what fraction of the 24 observed differences was I an improvement over S?)

6. Find the estimated probability of an innovation being an improvement over the standard using Figure 4.

7. Does a 0% average gain imply that all the innovations had exactly the same effect as the standards? Why, or why not?

8. Why do the authors prefer the empirical Bayes method to the normal distribution method?

9. Refer to Figure 5. In the primary therapies and in the secondary therapies what are the chances of
   a. a gain of 30% or more?
   b. a loss of no more than 20%?
   c. a gain of more than 0%? Compare with the result for Problems 5 and 6.

10. Why do the authors suppose that the innovations assessed by randomized clinical trials and reported in the surgical literature are "biased upwards"? What is meant by "biased upwards"?

REFERENCES

J. P. Gilbert, R. J. Light, and F. Mosteller, 1976. "Assessing Social Innovations: An Empirical Base for Policy." C. A. Bennett and A. R. Lumsdaine, eds., *Evaluation and Experiment: Some Critical Issues in Assessing Social Programs.* New York: Academic Press.

B. Efron and C. Morris, 1973. "Stein's Estimation Rule and its Competitors: An Empirical Bayes Approach." *Journal of the American Statistical Association,* 68: 117–130.

With permission of the Oxford University Press, this chapter is based on the longer article: J. P. Gilbert, B. McPeek, and F. Mosteller, "Progress in Surgery and Anesthesia: Costs, Risks, and Benefits of Innovative Therapy." J. P. Bunker, B. A. Barnes, and F. Mosteller, eds., *Costs, Risks, and Benefits of Surgery,* Oxford University Press, 1977.