Some Thoughts on Clinical Trials, Especially Problems of Multiplicity

John W. Tukey

Summary. Problems of statistical and conceptual design of experiments are exacerbated by ethical issues in many, if not most, clinical trials. Statutory requirements of demonstrated effectiveness are far from being clearly resolved—either qualitatively or quantitatively. Ethics, bolstered by informed consent, are likely to keep us from ever learning the answer to many questions. Unbalanced boundaries, focusing-down designs, historical controls, and not-very-sequential designs are among the possible consequences.

The pressures of ethics and equity on clinical trials have always been severe. Today they are more vigorous than ever before. Many of us are convinced, by what seems to me to be very strong evidence, that the only source of reliable evidence about the usefulness of almost any sort of therapy or surgical intervention is that obtained from well-planned and carefully conducted randomized, and, where possible, double-blind clinical trials [see the review papers of Byar *et al.* (*1*) and Peto *et al.* (*2*)]. Dare we prevent ourselves from obtaining reliable evidence?

Surgical Intervention

The simplest, though not necessarily the easiest, special case is that of most surgical intervention, where all decisions rest with the patient and the patient's doctors. Consider then the case where we have something less than reliable evidence in favor of some form of surgical intervention. Some of the questions that arise are:

1) Ought a surgeon employ this plausible form of intervention?

2) Must he, under the penalties of malpractice?

3) Can a randomized study be ethically conducted in the hope of learning whether this form of intervention is better than its most favored competitor?

4) Can we fail to conduct such a randomized study in view of our responsibilities to future patients?

Clearly these are difficult questions.

It seems to me very obvious that we are obligated to do everything we can to make as clear as possible to the patients' doctors the strengths and weaknesses of the incomplete evidence involved. This obligation falls most strongly upon statisticians, but it must be shared by those of all professions who are in any way involved.

Once the strengths and weaknesses of the evidence have been made as clear as we know how to do, we have also to be very careful about how we interpret the evidence. At this point we may even, in particular, need to take into consideration the general experience of clinical trials. What fraction of what reputable experts thought was likely to be an improvement, and thus worthy of a clinical trial, does in fact prove to be an improvement? For further discussion of this point see Gilbert *et al.* (*3*).

New Drugs

The next case in terms of difficulty is that of a new drug, which, until a favorable regulatory decision has been made, can only be used in controlled (and approved) experimental situations. If we have less than copper-riveted evidence that its use is an improvement, some of the crucial questions that arise are:

1) Should the regulatory body approve its general use?

2) If it does not, should or will use of the drug be subject to suit or prosecution for malfeasance?

3) Dare we fail to conduct a randomized double-blind trial to strengthen the evidence?

Again we have conflicting obligations, now falling mainly on the regulatory agency. The answer to the third question, given no favorable regulatory action, is relatively clear, at least as answers to such difficult questions go. For if we do believe that therapy with the new drug is an improvement, and the law forbids us to use it, except in an approved experimental situation, the best we can do for individual patients is:

1) Offer each of as many patients as we can afford a probability of taking the new drug rather than the placebo.

2) Ensure that our patients do not know whether they are on drug or placebo, so that all will receive whatever psychological lift comes from a chance of receiving the possible improvement.

For the welfare of all future patients, we can

1) Assure that the "probability of taking" comes from a well-planned and well-conducted randomization.

2) Assure that the trial is fully "double-blind" so that we can be certain that knowledge by attending physicians of who was on drug and who was on placebo does not affect either unspoken messages to patients or general medical care.

3) Make the protocol of the study as responsive to medical knowledge and insight as possible.

By taking such precautions, and by being careful to wring as much out of the data as possible, we will have done what we can for future patients by getting evidence that is as strong as possible as soon as possible. This will tend both to decrease exposure to a real nonimprovement and to advance the date of regulatory approval for a real improvement, both of which are responses to an ethical obligation.

We may well wish to choose relative frequencies of administration of drug and placebo that are other than 50–50, either for statistical reasons or for ethical ones. We may also wish to plan for these probabilities to change as the study proceeds, perhaps along the lines suggested by Robbins (*4*) [but see also Byar *et al.* (*1*, pp. 76–78) and Peto *et al.* (*2*, p. 596)].

Again, the responsibility for making the evidence as clear as possible rests upon all those concerned, especially the statisticians. Beyond this, however, I would suggest that the regulatory agency has a continuing responsibility to make as clear as is reasonable, in advance,

The author is professor of statistics and Donner professor of science at Princeton University, Princeton, New Jersey 08540, and associate executive director, research, at Bell Laboratories, communication principles division, Murray Hill, New Jersey 07974.

how it will judge evidence for efficacy, so that the planning and conduct of clinical trials can be more effective, and hence more ethical.

Old Drug, New Disease

The problem becomes still worse when we deal with a new use for an old drug. Our laws now prevent the advertising of the old drug for the new use, but they do not prevent any physician from prescribing it. The simple excuse for the ethics of the clinical trial is now lost.

It is no longer true that we can only give the drug to individual patients in an experimental situation. Any physician who believes the old drug to be an improvement for its new use can prescribe it for any patient. If he or she believes an improvement to be likely but not demonstrated, is there an ethical obligation to prescribe? Up to what level of "likeliness" can patients be asked to enter a randomized trial? Do the physicians who believe an improvement is likely have an ethical obligation to write vigorously in the medical literature to convince other physicians?

The difficulties with the "use, don't experiment" situation, which ordinarily arises only when the regulators have not approved the new use, include one interesting one, namely: *So long as drug house advertising is not permitted, use of an old drug for a new disease will tend to be confined to patients of more literature-reading and more literature-influenced physicians; thus, if the new use is an improvement, producing (or more likely, some would say, increasing) inequality of health care between two classes of patients.*

How is this to be balanced against the likely loss to individual patients randomly assigned to the placebo if a clinical trial is decided on?

Safety and Relative Risk

Thus far, this discussion has bypassed safety, proceeding as if both surgical interventions and drug therapies were without risk. While this is often quite unreal, the only effect of safety on the issue that here concerns us is to make the problem more pressing. Not less.

One Role for a Statistician

These are not easy questions. It is not my place to suggest answers. But I do have an obligation to do whatever I can to reduce the frequency with which such questions arise—and the duration over which each difficult situation extends. What I can do as a statistician, I must. The largest feasible improvement I see my way to helping with at the moment is the sharpening of the understanding of the strength of the evidence. It is with this class of questions that we will now be concerned.

Table 1. Example of probabilities of not reaching significance.

Situation considered	Probability of *not* reaching significance at 5%
First class alone	95%
Second class alone	95%
Two classes together	$(95\%)(95\%) = (95\%)^2 = 90.2\%$
Third class alone	95%
Three classes together	$(95\%)(95\%)^2 = (95\%)^3 = 85.7\%$
Nine classes together	$(95\%)^9 = 63.0\%$
Tenth class alone	95%
Ten classes together	$(95\%)(95\%)^9 = (95\%)^{10} = 59.9\%$

Clinical Inquiries versus Focused Clinical Trials

The words "clinical trial" have a wide variety of meanings. Let us look at two extremes:

The clinical inquiry. This is where some intervention or therapy is hoped to be of help to some class of patients, not specified in advance, and where, consequently, we go in for massive data collection and for analysis of results for each of many classes of patients (by age, sex, previous medical history, prognosis, and symptoms, for example). (There also may be separate analyses for different end points.) The statistician must, I believe, call attention to the multiplicity of questions which any such inquiry poses, and he must, therefore, face up to the influence of this multiplicity on the strength of the evidence resulting from the inquiry.

The focused clinical trial. This is a trial in which both the class of patients and the end point to be considered are clearly specified in the initial protocol, and the only chance of multiplicity arises from analyses of the data at various cutoff dates during an ongoing study. (This kind of multiplicity also occurs in the clinical inquiry and its multiplicity has to be multiplied together with the other kinds of multiplicity there present.)

From a data analytic viewpoint, in particular in terms of the sort of statistical formalisms that seem to help me, these two extremes are very different.

Indeed, I do not believe that a clinical inquiry, by itself, is likely to be an ethically satisfactory means of providing definitive evidence that an intervention or therapy is an improvement. (We will come to the more technical reasons for this later.) To say this is *not* to say there should be no clinical inquiries. Quite the contrary. Clinical inquiries may often play a very crucial and very useful role. However, at a time before such an inquiry has reached trustable conclusions, it will ordinarily be best to initiate, or to embody in the continuing clinical inquiry, a single focused clinical trial (or, possibly, a few such) from which one can come more rapidly to trustable conclusions.

It is right for each physician to *want* to know about the behavior to be expected from an intervention or therapy when applied to his particular individual patient (to whom the physician has the strongest ethical obligation). It is not right, however, for a physician to *expect* to know this—except, possibly, for the most dramatically effective and time-tested interventions or therapies. Most useful interventions or therapies change, for the better, the chance of a favorable outcome—change it from a smaller chance to a larger chance. Most physicians and surgeons recognize this and do not demand (though they may rightfully ask for) detailed and reliable forecasts for individual patients.

They feel that they have better reason, as indeed they do, to ask for differential forecasts of improvement by age, sex, symptoms, or the like. Again it is right for them to ask for such forecasts, but as we shall soon see, they are not likely to get them for newly tested innovations; that it would undoubtedly be good for their patients if they had them is not a reason for them to be possible.

This feeling, quite proper for all patient-treating physicians and surgeons, has undoubtedly helped in a rather widespread misinterpretation of the role of clinical inquiries, as opposed to focused clinical trials.

Multiplicity and Significance

Let us emphasize one aspect of the analysis of clinical inquiries that is ethically necessary and is commonly observed: special attention is given to the results for whichever class or classes of patients for whom the results appear most favorable for the intervention or therapy under test.

Consider the simple arithmetic of asking multiple questions and concentrating on the most favorable answers. As just noted, once multiple questions are to be asked, there will be pressures, some ethical in nature, to concentrate upon those questions for which the results appear most favorable.

If we approach our data in terms of tests of significance (or in terms of confidence intervals) and neglect problems of multiplicity, we find ourselves in trouble. As another speaker has put it to me in private: "Even normal saline comes out significant 40 percent of the time." How can such things be?

Suppose that we are conducting a clinical inquiry about a single innovation which is perfectly neutral, which has no effect on any patient. Suppose we look at only ten classes of patients, defined by age, sex, and symptoms. For simplicity let us take these classes nonoverlapping, and let us suppose that the results for different classes are statistically independent. Then the probabilities of not reaching significance (wholly be chance) at 5% are as in Table 1, and the probability of finding at least one out of ten subgroups significant at 5% purely by chance, is

$$100\% - 59.9\% = 40.1\%$$

The moral seems to me to be abundantly clear: Knowing that, for one class of patient, a clinical inquiry has reached some specific level of significance, such as 4%, is not evidence of the same strength as knowing that a focused clinical trial, involving a single prechosen question, has reached exactly that level of significance, even if both the inquiry and the trial involved the same number of patients exposed to risk, and the same total number of end points, distributed in the same way. That ethics, and other reasonable motivations, ensure that we will look first at the results for whatever class was most promising is a vital fact, and cannot be neglected.

Once we admit that the best-appearing class will be examined first, we can see how to adjust our application of significance to the clinical inquiry. If there are k classes of patients that *would have been looked at seriously if the results for them had seemed favorable*, it suffices (for one who would ask for about 5% significance in a focused clinical trial) to ask about significance at $5\%/k$ [that is $(5/k)$ percent] for each of the classes that were indeed looked at seriously. Notice that it does not suffice for k to be only as large as the number of classes actually looked at; we need to use the larger number of classes, each of which would have been looked at seriously if the results for them had happened to look favorable.

If there are several end points, according to which we might have assessed improvements, then k has to be further increased to become the product of the number of plausible end points multiplied by the number of plausible classes. It is easy for k to become very large indeed.

I will turn later to the question of whether we can bear working to the $5\%/k$ standard, or even to one somewhere near this standard of rigor.

Multiplicity and Bayes

In my judgment, Bayes's methods do not offer us any satisfactory way to deal with problems of multiplicity. I have yet to see a Bayesian account in which there is an explicit recognition that the numbers at which we are looking are the most favorable of k. Until I do, I doubt that I will accept a Bayesian approach to questions of this sort as satisfactory.

The type of solution to which some Bayesians are led, particularly those who are likely to say that they are practitioners of "empirical Bayes" seems also unsatisfactory. The reason why physicians and surgeons are willing to consider results for specified classes is simple: experience shows that some interventions and some therapies are much more favorable to some classes of patients than to others. A procedure that assumes

1) that the classes we have actually looked at are all those that we would have looked at, even if their results had appeared favorable, and

2) that it is a reasonable approximation to treat the true improvements for the classes concerned as a sample from a nicely behaved population (one that surely does not involve two more or less separate collections of true values or, in technical language, one that is much better behaved than just being unimodal), does not seem to me to be near enough the real world to be a satisfactory and trustworthy basis for the careful assessment of strength of evidence to which the ethical issues discussed above must dedicate us.

Multiplicity and Decision Theory

Much the same remarks as those in the previous section seem to me to apply to any other "decision theory" approaches that I have seen.

Multiplicity and Phased Experiment

The most extreme case of multiple questions arises in purposive breeding, in particular in those areas where it is easy to generate many strains. Trying to pick out, for high yield of kernels, a particular hybrid line of maize is much like trying to pick out, for high yield of antibiotic, a particular radiation-induced mutant strain of microorganism. In each case, it is easy to obtain many candidates to begin with. The practical constraint is on the total amount of experimentation—acres times years for maize, total volume of, or number of, cultures for microorganisms—that we can afford to devote to our search for improvement.

The statistics of this situation have been clear for a quarter of a century [see, for example, (*5*)]. The main points are:

1) The experiments should be divided into phases, with a selective reduction in the number of strains carried forward between each phase.

2) Provided equally plausible candidates can be easily obtained in unlimited numbers, the size of the trials for the individual candidates in the first phase should be so small, because there are so many candidates, that no significant differences among strains can be expected to be established.

The simplest analogy to this solution of the breeding problem is a clinical program that begins with a clinical inquiry and closes with a focused clinical trial.

In its simplest form the clinical inquiry is regarded only as the place where we spend the effort and the dollars required as a sensible entrance fee for the focused clinical trial (or perhaps the two or three focused clinical trials). It will undoubtedly be hard for anyone familiar with the effort and expense associated with large clinical inquiries to accept the idea that all of it was just to pay the entrance fee. Yet this is ordinarily the most efficient way to regard any clinical inquiry.

We have convinced ourselves that a regulator, physician, or surgeon who demands significance at 5% for a focused clinical trial should demand significance at $5\%/k$ for the best class of patients of a clinical inquiry. Suppose we are conducting a clinical inquiry and that our "best" class has at least reached significance at 5%, if analyzed as if, contrary to fact, it was the only class that might have been considered. If the innovation is indeed favorable, we have at least two strategies before us: (i) continue the clinical inquiry until what is then the best class reaches significance at $5\%/k$ and (ii) replace the clinical inquiry by a single focused clinical trial (or a few such) and

carry out the focused clinical trial until it reaches significance at 5%.

Unless k is noticeably less than 10, we can expect the second choice to take less time, to say nothing of less effort, thus meeting one of our ethical obligations to future patients, to say nothing of costing less.

The decision between these strategies will thus tend to favor the second choice, particularly as the decision-makers become better acquainted with the quantitative facts.

What It Is Unethical to Learn

It is time for us to look at some consequences of the ethical issues that are somewhat more specific than the broad ones we have been considering.

We will never know, with any high relative precision, how much better a favorable innovation is than its current competitor. Once our clinical trial has accumulated favorable evidence for an innovation up to whatever level of significance regulators, physicians, or surgeons judge appropriate for action, we cannot, ethically, continue the trial (at least as we see the world today) *just to measure the improvement* with greater precision. Thus we will ordinarily be lucky indeed if we can distinguish among even three broad levels of improvement, say: small, medium, and large improvements.

This becomes painful whenever it is very expensive to put the innovation into practice. Those who are to pay for an expensive program are right to ask for a good idea of how much it will help; we are probably ethically right often to tell them that we cannot, in good conscience to our patients, find out.

Circumstances and costs may be such, however, as to limit the rate of introduction of the innovation, as severe limitations of foreign exchange limited British imports of streptomycin just after World War II. This limitation made a double-blind study of the efficacy of streptomycin in tuberculosis feasible, leading to the first adequate measurement of its helpful effects. I believe our ethical obligation to future patients should force us to consider randomized application during the period in which not all patients can receive the innovation.

What Comes After a Focused Clinical Trial?

Suppose that we have had a focused clinical trial, and that it has established the statistical significance of an improvement when the innovation is applied to certain other classes of patients, either to all patients or to those from a large class. Either at that time, or later, certain skilled physicians or surgeons may come to doubt the efficacy of the innovation when applied to certain classes of patients. What is their obligation to future patients, and how can they meet both this obligation and that to their individual patients?

In such a situation it should be possible to identify a class of patients for whom the doubters feel uncertain about the efficacy of the innovation (neither clearly for nor clearly against). Is it not now an ethical obligation of the doubters to join together, to plan an adequate randomized double-blind study of efficacy in this selected class, and to attempt to fund and carry out this study?

It would seem that only by such actions can we ethically learn more and more about the boundaries of efficacy of even the more important surgical interventions and medical therapies.

Table 2. Half-octave number of end points, and the approximate corresponding chances of reaching significance at an individual (with no adjustment for repeated looking) two-sided 5% point (binomial comparison).

Number of end points	Chance of reaching "5%" significance	
	50% improvement	20% improvement
25	3/8	
35	1/2	
50	3/5	
70	4/5	
100	9/10	
150		1/4
200		1/3
300		5/11
400		4/7
600		4/5
800		7/8

And What of the Cost?

Any form of clinical trial is expensive. (Well-conducted ones are likely to seem more expensive before they are begun, though they may be less expensive before they are meaningfully concluded.) Who pays the cost is, in detail, a matter for legislative and public debate. In the main, however, the costs of clinical trials, like the cost of all collective undertakings, has to come out of everyone's pocket. As we think of new requirements concerning safety, as well as efficacy, we push these costs higher and higher. The result of higher costs is, inevitably, fewer clinical trials. At what point do we lose instead of gain? I lack the information to take this question much further, but feel an obligation to raise it.

Knowledge versus Opinion

When the law asks for proof of efficacy, or when the physician or surgeon asks for publications, carefully refereed and clearly written, that seem to deserve trust, the demand is for knowledge—knowledge of a restricted sort, coming with a P value to indicate the size of residual doubt—not just for skilled professional opinion. Yet medical and surgical practice has always depended more on skilled professional opinion than on knowledge. I doubt that this will change within any of our lifetimes.

It would be wrong to focus on knowledge to the exclusion of opinion. When I seek medical or surgical care for someone near and dear to me, I want more than knowledge. Experts are usually experts by their opinion rather than their knowledge. I would hate to have had a hand in the leveling of medical practice to a uniformity based only on clearly recognized knowledge, something malpractice suits threaten us with to an unbearable degree.

It would be almost as bad, I believe, to disturb too deeply the practicing physician's or surgeon's belief in his or her own skill, much of which consists of informed professional opinion rather than knowledge.

In tightening the standards of knowledge from clinical trials, we need to do this without too greatly disturbing the dependence of practitioners, in the many areas not yet subjects of adequate clinical trials, on their own informed professional opinion. Indeed, we owe an ethical obligation to their future patients not to disturb them too greatly.

It is a difficult task to drive the nearly incompatible two-horse team: on the one hand, knowledge of a most carefully evaluated kind, where, in particular, questions of multiplicity are faced up to; and, on the other, informed professional opinion, where impressions gained from statistically inadequate numbers of cases often, and so far as we see, often should, control the treatment of individual patients. The same physician or surgeon must be concerned with both what is his knowledge and what is his informed professional opinion, often as part of treating a single patient. I wish I understood better how to help in this essentially ambivalent task.

Large-Scale Decisions and Opinion

As the importance of health care becomes more engrossing for more people, we shall have to make more and more large-scale decisions. Dare we do this on the basis of informed opinion alone? Dare we face the ethical problems and costs involved in enough clinical trials to allow most such decisions to be made on the basis of knowledge?

It is time to turn to a few questions of a more specific character, questions which, like the relation of clinical inquiries to focused clinical trials, are more in line with the statistician's narrow responsibilities.

Historical Controls

One of the most debated questions in clinical trials is that of "historical controls." It seems easy for some to argue that we know enough of how patients of variously specified ages, sexes, symptoms, and histories have responded to the old intervention or the old therapy, and that there is thus no need for a randomized trial. If our experience with the behavior of disease were different, this argument would be easy to accept. But phenomena such as the decline of tuberculosis mortality before any presumably efficacious treatments were widely used are not uncommon. [For disturbing examples more closely relevant to clinical trials, see Byar *et al.* (*1*, pp. 75–76) and Peto *et al.* (*2*, p. 592).]

How then should we reply to the cry that "historical controls are good enough"? Given both high expertise and a firm belief that the innovation will produce more than a 30% to 50% improvement, it can be hard to hold the line for randomized studies. Is there a possible compromise?

Perhaps there is, though it is not one that is likely to make those crying for historical controls entirely happy. If it is feasible to say, for example, that background changes in other aspects of intervention, therapy, or cure are certainly not going to make an improvement of more than 25% (or perhaps 50%) over the time from the historical controls to the new study (I wonder how often this is so?), then I could conceive:

1) starting a study without randomization;

2) analyzing the results of the new study by comparison with a 25% (or 50%) improvement over historical controls; and

3) planning, if this analysis seemed indecisive—and carrying out the plan—either to convert the study into a randomized study or to drop the trial of the innovation.

Beyond such an alternative, I have not seen an excuse for historical controls that seemed to me to be valid.

Not-Very-Sequential Designs

On the one hand, the dangers of unfavorable response urge us to monitor quite frequently the results of a clinical trial as they accumulate. On the other, the costs of frequent analysis are still often underestimated.

The names associated with the basic papers on the impact of steady monitoring of a simple paired comparison are those of Armitage and McPherson [see Armitage *et al.* (*6*) and McPherson and Armitage (*7*)]. They showed just how far our assessment of significance can be displaced by indefinitely repeated testing.

The most qualitative result is easy to understand. If we continually test at $P\%$, stopping only when the innovation appears significantly better or significantly worse, we can think of looking and testing after 100, 100^2, 100^3, . . . patients have accumulated. At each such look, the data collected before the previous look are only 1/100th of that now at hand. The results would be much the same if we made repeated independent tests at $P\%$. Sooner or later the innovation will be significantly better or worse. So our chance of reaching significance is 100% not $P\%$.

The real question is not "do such things happen?" but rather "do they happen soon enough to matter in practice?" Armitage *et al.* (*6*) showed us that, regrettably, the answer is "yes."

What then ought our response to be? The pressure for repeated checking comes only from an appropriate desire to avoid unduly prolonging trials. This is especially important on the "innovation significantly bad" side (herein abbreviated as the "bad side"), since we are particularly conscious of the need to limit exposure to bad innovations. Since I see no need for the probability of stopping a trial of a neutral innovation, because it seems bad, to be as low as we need to have the probability of stopping because the neutral innovation seems good, I am quite willing to look at least somewhat more frequently on the bad side, thus raising the true value of P on that side higher than for the good side.

What about the good side? Continuous looking is obviously wrong. If we allow for its effects, the trial will take longer to reach a conclusion because we have looked more often. Looking quite often is also ethically wrong because when we allow for how often we have looked, it will still take a longer time, on average, to reach a chosen level of significance, allowing for multiplicity, than if we looked less often. The only reasonable conclusion I can draw is that we ought to look only relatively infrequently.

Table 2 suggests some possible chosen values where our analysis is based on reaching a fixed end point, for instance, death. I believe I could bear to plan to look at three values from this half-octave sequence. (The table suggests that if I were to look at three values, perhaps I should space adjacent looks a full factor of 2 apart. But I fear the consequences of telling a trial manager he ought to get 100% more data before looking again.)

Since data will be coming in more or less steadily, how can we avoid looking more often? Especially since we expect to look more often on the bad side? Realistically, it is hard to believe that no one will look some, perhaps many, times more than prescribed. But we can hold down the amount of additional dilution of significance this causes by fixing either the effective date or the actual number of end points as of which the "file is to be cut" and all records brought up to that date. Even if such dates are only several end points in the future at the time of decision, such precautions will greatly reduce the dilution that would arise if we stopped at exactly the point where things first looked good.

How much will three looks (at half-octave numbers of end points) cost us? About as much as $k = 2$ would! I have computed the effect for looking at 10, 20, and 30 end points, with one-sided tail areas of 1.07%, 2.07% and 2.14% (total 5.28%), and have found a combined level of 3.84%, corresponding to $k = 2.2$. Doing the same for 10, 20, and 40 end points gives 3.97% combined, which is in the same area. (Half-octave looks would correspond to 15, 20, 30 and thus, since 15 is closer to 20 and 30, to a k a little smaller than 2.2.) More extensive calculations seem to have been carried out independently by Pocock (*8*).

And Then?

Suppose that in Harold Jeffrey's words "all the allowed principles of witchcraft" have been used:

1) We have carefully randomized the patients.

2) It has been possible to do the study double-blind.

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.