

RECENT METHODOLOGICAL CONTRIBUTIONS TO CLINICAL TRIALS

JEROME CORNFIELD¹

This paper is concerned with the maximum that can be achieved in clinical trials using optimal methodologies if such exist, i.e., with strategy and not tactics. It might be thought that strategy so defined is too vague and too general to be fruitfully discussed, but I hope to convince you of the contrary. Considerations of strategy will involve us in controversial questions about the foundations of statistics, but this prospect shall not inhibit us any more than it would have Reed or Frost.

A point of view about statistical methods that was once universal and is still very common is that it provides unique, unequivocal answers to problems of analysis and design—that all statisticians faced with the same problem will, or at least should, provide the same answer. R. A. Fisher's writing, in particular, reflects this view, with continual glimpses furnished by such phrases as "uniquely superior" methods of "absolute validity," "canons of valid inference," "rigorous but uncertain conclusions," etc. His *Design of Experiments* was dazzling in its self-assurance and clarity and gave statisticians in many fields the inspiration and confidence to apply these unequivocal procedures, and to develop new ones.

The extensions to clinical trials have been numerous and important, the key figure being Bradford Hill. In the opinion of the President of the Royal College of Surgeons they constitute as crucial a contribution to medicine as the discovery of penicillin (1). One would naturally have expected that continuing methodological advances would have accompanied these extensions. The statistical literature on

clinical trials has grown enormously, but a deeper, and more interesting development is that the original confidence in the existence of "uniquely superior" ways of quantifying uncertainty which sparked the clinical trial is now seen to have been overly optimistic. The empirical process of inference and decision in ongoing clinical trials is perceived to be loosely defined and structured, not because appropriate mathematical tools are lacking, but because this is the nature of the enterprise. The paradox is that a solid structure of permanent value has, nevertheless, emerged, lacking only the firm logical foundation on which it was originally thought to have been built.

In the following sections I shall give instances of important statistical issues in clinical trials in which unambiguous answers appear unattainable, shall consider why this must be so theoretically, and shall conclude with a brief consideration of what the underlying core of useful statistical methodology in such trials appears to be, despite its ambiguity.

1. DECISION MAKING IN CLINICAL TRIALS

The existence of a decision-making, or as Schwartz and Lellouch (2) put it, pragmatic, function in clinical trials was almost entirely neglected in the original formulations. Unforeseen and undesired consequences of therapy are by no means uncommon, however, and patient welfare sometimes requires fundamental changes in the design during the course of the experiment. Numerous examples could be given, but one should suffice.

Diabetes is characterized by an elevated blood sugar, and it was once widely believed that any therapy that lowered blood sugar would have a beneficial effect

¹Department of Statistics, The George Washington University, Washington, DC 20052.

Supported in part by research grant HL 15191.

upon the sequelae of diabetes, particularly its cardiovascular sequelae. The University Group Diabetes Program was designed to test this hypothesis in a randomized multiclinic comparison of insulin, oral hypoglycemic agents and diet alone in a group of mild, adult-onset diabetics. No one anticipated the possibility that an excess cardiovascular mortality might develop among those on the oral agents. When this did occur, the investigators felt obliged to discontinue their use (3). This decision did not meet with universal favor in the medical and pharmaceutical community, but the resultant controversy is beyond our scope. What is relevant is that an investigation originally designed to produce new knowledge, suddenly found itself involved in a difficult and unwanted task of decision making. From the purely formal hypothesis testing point of view that dominated the early thinking in clinical trials, what had happened was that the same body of data had been used to formulate a hypothesis and to test it. From that point of view the University Group Diabetes Program results should have been treated as suggesting a hypothesis to be tested in a new and independent trial. But to the investigators this was inappropriate. They had to decide for themselves and their patients whether the evidence available to them justified the future exposure of anyone to these agents and, as we shall see, important aspects of this kind of decision elude precise quantification.

Even without unforeseen results there are important decision-making aspects to clinical trials. In the Coronary Drug Project (4) in which three of the five drugs tested were discontinued because of unanticipated side effects, two drugs and placebo were continued to the end in accordance with the original plan. The final conclusion was no "significant" effect on the incidence of or mortality from coronary heart disease with either drug. A published criticism of this finding indicated that it confused statistical and prac-

tical significance and that the lower 95 per cent confidence limit for the treatment effect was an 18 per cent reduction in coronary mortality despite over 1000 patients on drug and over 2000 on placebo, and that the possibility of a reduction of this magnitude was of great importance (5).

The investigators replied, "we suggest that in addition to the distinction between statistical and practical significance, with which we *are* familiar, there is also a distinction between data analysis and decision... The Coronary Drug Project provides no evidence on which to recommend the use of clofibrate in the treatment of persons with coronary heart disease" (6). In this respect, clinical trials are no different from any other activity in which data are gathered to help decide what to do.

It is not universally accepted that the theoretical analysis of decision making is a useful part of statistics. The Fisherian view is that it may be fit for business and tyranny, but surely not for the high, free purposes of science—in Savage's paraphrase. Whatever truth this view may have had for Victorian or Edwardian science, it has an archaic ring today. The more usual objection is that the consequences of decisions are so complex as to defy useful quantification. But one can admit the difficulty of quantification and still argue that the development of decision, or pragmatic, points of view, in contrast to the more common inferential, or explanatory, points of view constitutes a major revision of viewpoint.

This is illustrated by the recent modification of its protocol by the National Eye Institute's Diabetic Retinopathy Study to permit each participating physician to offer photocoagulation therapy to eyes originally assigned to the control therapy (7). In one sense this is merely another instance of unforeseen results; the therapy turned out to be a lot better a lot sooner than had been expected. After two years of follow-up, the treated eyes had progressed to

blindness at less than half the rate of the control eyes and it seemed unethical to withhold the treatment from control eyes. But nothing was known about the persistence of the effect beyond two years and it could be (and was) argued that there were sound physiological reasons for expecting that the effect might shortly reverse itself and that after three or four years the untreated eyes might be better off. Nothing definite was really known and from a strict inferential point of view it was maintained that the trial should be continued unchanged until information for longer periods of follow-up became available. But a simple-minded decision point of view could admit the existence of such reversals and inquire about how large they had to be to balance the early substantial gains and lead to a net long-term increase in blindness. A simple calculation showed that even the most severe delayed harmful effects that could be envisioned were not sufficient to outweigh the observed early benefits. The decision to modify the protocol at that point became inevitable. Furthermore, subsequent follow-up of the eyes originally assigned to treatment would detect major reversals of the early results if they occurred, so that no important conflict between decision and inferential views really existed.

Decisions such as these have been largely a matter of common sense and have derived little support from formal theory, except perhaps for the philosophical comfort imparted by knowledge of its existence. The difficulty of quantifying gains and losses that are often incommensurate provides one explanation for this. But formulations now exist, which provide more natural expressions of the losses associated with incorrect decisions in a clinical trial, namely the number of present and future patients assigned to the inferior therapy. Present patients are assigned to the inferior therapy during the clinical trial stage because no one is sure which it is, while, if

the trial comes to the wrong conclusion, future patients will also be incorrectly assigned. The most general formulation of this problem, from which all the others derive, is provided by the two-armed-bandit problem which seeks a way of assigning n patients one at a time to one of two therapies in the light of past results so as to minimize the total number assigned to the inferior therapy (8). No general solution is known, but a number of approximations, all superior to present practice by that criterion, are available. These are ably reviewed by Armitage (9).

Present unfamiliarity with the procedures provides one reason why they have so far failed to influence clinical trial practice. Delayed response to therapy relative to the rate at which new patients are admitted to the study provides another. The sharp separation between patient care and research, even in countries with systems of health services delivery different from ours, provides a third. And finally, to return to incommensurability, present and future benefits, although expressible in the same units, are not as commensurate as they at first might appear. It is a common and sound practice to discount future benefits. In economics the interest rate provides an appropriate basis for this, but once one inquires about the basis for discounting future health benefits, the Pandora's box, apparently closed by the two-armed bandit, threatens to fly open again. Thus, decision making in the world of clinical trials is much the same as for any other kind of decision making, and the inability to specify utilities is an inescapable difficulty.

2. APPRAISAL OF UNCERTAINTY

A common attitude towards these problems may be paraphrased as follows: "Decisions, although important, involve non-statistical issues and should be distinguished from the purely statistical issues, which consist of asking what the data show

and how certain are the conclusions they will support. Once these are known, decisions and their costs can be considered, but preferably by someone else." But the major impact of decision theory on clinical trials, it seems, has been in the entirely new light it casts on the appraisal of certainty.

As a young man, Fisher continued in the old tradition by developing mathematically exact tests of significance, which he and everyone else regarded as providing appropriate measures of uncertainty in experimentation. But experience in clinical trials has demonstrated to all participants, biostatisticians and clinicians alike, the limited utility of the p values provided. The multiple testing problem provides one immediate indication that all is not well. Just as the Sphinx winks if you look at it too long, so, if you perform enough significance tests you are sure to find significance, even when none exists. But repeated examination of the results of a modern clinical trial is both necessary and desirable. Results must be examined periodically as they accumulate to detect early benefits or unwanted side effects. The possibility of differential effectiveness in different classes of patients requires the examination of results for subgroups; the existence of many endpoints requires the examination of many response variables; and when there is more than one therapy, each must be compared with the control or standard treatment. To do all these simultaneously, with each examination accompanied by a separate significance test, is to make it highly probable that one or more p values less than the conventional .05 or .01 will be obtained.

Much theoretical effort has been and continues to be devoted to finding multiple comparison procedures that avoid this. In one sense, Wald's original development of sequential analysis was an effort to control errors arising from multiple examination of accumulating results; the analysis of variance is a way of controlling the multiple

comparison problem for multiple treatments and multiple subgroups; the multivariate analysis of variance is a way of handling numerous response variables. If the issues were only mathematical ones, an omnibus procedure permitting any amount of multiple testing of any kind at any desired level of overall error would long since have been developed. But the issues appear deeper and not capable of mathematical resolution. How multiple should the multiple comparison be? Do we want error control over a single trial, over all the independent trials on the same agent, on the same disease, over the lifetime experience of a single investigator, etc.? Does a hypothesis rejected at the .05 level for a single comparison have the same amount of evidence against it as an omnibus hypothesis involving multiple comparisons but also rejected at the .05 level? The unanswerability of these questions suggests that they have been incorrectly posed and that the multiple comparison problem is the symptom and not the disease—the disease being the inappropriateness of the p value as a measure of uncertainty. A recent proposal that in clinical trials, all calculated p values be multiplied by 10 to take rough account of this problem (10) indicates that we may be nearing the end of the road as far as this approach is concerned.

These difficulties are all symptomatic of a deeper difficulty, which is that the pre-specification of a significance level, e.g., .05 or .01, basic to the Fisherian significance test, has no sound logical basis and remains unjustified, despite a recent attempt by Bross (11). It apparently entered statistical practice because copyrights prevented Fisher from reproducing some of the existing normal tables, but not from presenting their .05 and .01 points. It has by now achieved the status of an unstated axiom—that all hypotheses rejected at a given significance level have the same amount of evidence against them.

No one would profess belief in this axiom, but many behave as if they did. It is hard to say concisely what is wrong with it. Examples in which it leads to absurd results (12) are of some help, although they tend to be artificial. A somewhat longer but more systematic development follows, even though it is a diversion from our main theme.

3. LIKELIHOOD RATIOS AS MEASURES OF UNCERTAINTY—THE BAYESIAN VIEW

A kind of naive philosophical dualism which holds that separate languages are required for the conceptualization of scientific and of decision problems should be at least temporarily suspended. Let us start by considering two simple hypotheses about a treatment effect, such as that the odds of a favorable outcome are the same for both treatments (H_1) and as an alternative that the difference between the log odds has a particular non-zero value (H_2). We consider a fixed number of observations and three possible decisions, the acceptance of H_1 or H_2 or the suspension of judgment. The utility of each decision under each hypothesis as well as the prior probabilities of each hypothesis are specified. The data are in and a decision needs to be made. How should one proceed? The painfully simple answer is, choose the decision which maximizes the expected utility. Thus, if the observation, possibly multi-dimensional, is x , if the probability of x under the two hypotheses is $f_1(x)$ and $f_2(x)$, if the prior probabilities are g and $(1 - g)$, and if the utility of decision i when hypothesis j is true is U_{ij} , the expected utility of decision $i =$

$$k(x) [U_{i1}f_1(x)g + U_{i2}f_2(x)(1 - g)], \quad (1)$$

where $k(x)$ does not depend on i . That value of i which maximizes the expected utility is the one to choose. It is an elementary piece of algebra to show that the decisions are entirely specified by the inequality

$$A < f_2(x)/f_1(x) < B, \quad (2)$$

where A and B depend only on the utilities and prior probabilities. We accept H_2 if the likelihood ratio $f_2(x)/f_1(x)$ exceeds B , H_1 if it is less than A , and suspend judgment if it falls in between. Whether $A < B$, i.e., whether suspension of judgment is possible, depends on the utility assignment.

A hard-nosed clinician or biostatistician, confronted with this development for the first time, might very naturally point to the hypothetical and fuzzy character of the utilities and prior probabilities and insist that he was concerned only with the hard facts. But the question raised by the development is, what are the hard facts? In particular, is it the p values adjusted for multiple comparisons or is it something else? In one sense the facts are the observation x , which for anything that has been said to the contrary consist of a stack of 100,000 punchcards, every observation made on each of the patients entered in the trial. This is a matter to which we shall return. But we are concerned with two therapies, i.e., with a choice between H_1 and H_2 ; and in another sense, the facts bearing on this choice according to equation 2 are given by a single quantity, the likelihood ratio, no matter what the prior probabilities and the utilities. Thus, even if one contemplates a complete separation between the world of decision makers, willing to struggle with whatever is required to reach decisions, and a world of scientists devoted solely to the search for new knowledge, their common conceptualization of the facts bearing on the choice between H_1 and H_2 is the likelihood ratio.

This ratio provides a measure of the certainty with which the data point to H_2 rather than H_1 , values in excess of unity supporting H_2 and those between zero and unity supporting H_1 . Its numerical value has the interpretation of betting odds, or more precisely of the ratio of two sets of betting odds. A value of, say, 5 is interpreted to mean that no matter what one's odds on H_2 versus H_1 before seeing x , say $(1 - g)$ to g , they should now, after seeing x ,

be 5 times as great, i.e., $5(1 - g)$ to g . Prior opinion, which can vary from investigator to investigator, determines the odds before seeing x , but the way in which they must be modified in the light of the data, is given by the likelihood ratio and is the same for everyone.

The assumptions that lead to these conclusions, that utilities and prior probabilities exist and that one should behave in order to maximize the expected utility, goes back to at least the second Daniel Bernoulli. The modern version, which argues that the assumption is necessary and sufficient for logically consistent behavior, or as it is now called, coherence, is presented in a clear, nontechnical way by Lindley (13). Identical conclusions can be reached without the decision theoretic framework by asking for coherent measures of certainty (14).

These seemingly arcane matters are of relevance to clinical trials if only because the likelihood ratio, unlike the p value, is the same for all non-informative stopping rules. In particular, it will be noticed that the form of equation 2 is the same as that originally derived by Wald for sequential procedures (15), even though we have obtained it by assuming a fixed number of observations. Exactly the same form can be obtained by asking for the optimal sequential two-decision procedure (16). As an extreme example of the non-dependence of this measure of uncertainty on the stopping rule, consider an obdurate investigator in possession of the truth, namely that H_1 is false, who decides to convince his more obtuse colleagues by sequential observation on x . He plans to compute the likelihood ratio after each observation and to stop if and only if it exceeds B . The chance that it will do so when H_1 is true cannot exceed $1/B$. Thus, even if he suppresses information on his stopping rule, but does fairly report x , the likelihood ratio is all that is relevant. If he stops if, and only if, his fixed sample size p value falls below some value, however, the chance

that he will eventually reject H_1 when it is true is unity. It is an interesting sub-paradox that the seemingly hard-headed and objective p value approach leads to something as subjective as the conclusion that the meaning of the data depends not only on the data but also on the number of times the investigator looked at them before he stopped, while the seemingly fuzzy subjective formulation leads to the hard-headed conclusion "data are data." The latter formulation also provides theoretical support for the increasingly common and wholly necessary practice of monitoring the data as they accumulate without regard to the effect this may have on the certainty of the conclusion. The non-dependence of certainty appraisals on the stopping rule holds whether the data are accumulating because of the admission of new patients, because of longitudinal examination of the same group of patients, or because of a combination of the two. Only in the former case would Wald's and related procedures be applicable, since only then are the underlying parameters unchanging. The irrelevance of the stopping rule also holds if in the middle of the study the investigators change their decision rule, as given by equation 2, by, for example, disregarding the first violation of the inequality in the hope that it will not happen again. The most general and concise way of saying all this is that p values depend on both the x observed and on the other possible values of x that might have been observed but weren't, i.e., the sample space, while the likelihood ratio depends only on the observed x .

Conclusions derived by using p values and likelihood ratios can differ markedly, with the same data leading to strong evidence against H_1 using p values and in support of it using likelihood ratios. (For a simple example see Edwards et al. (17).) The reason is that the appraisal of uncertainty provided by likelihood ratios depends on both H_1 and H_2 , while that provided by p values depends only on H_1 .

Change H_2 and the likelihood ratio will change, but the p value will not. The likelihood ratio provides a comparison of one hypothesis against another, while the hypothesis test is an absolute appraisal. In hypothesis testing the choice of an alternative may affect the choice of test criterion but for any given test criterion, this will have no effect on p . This might suggest that no relation exists between the likelihood ratio and the probability of the two types of errors, but this suggestion would be wrong. There is a relationship which may be stated as follows: Consider any procedure for using x to reject H_1 in favor of H_2 . This procedure will entail a certain probability of rejecting H_1 when it is true, say α , and a certain probability of accepting it when it is false, say β . Then of all the possible procedures that might be used, the one which minimizes the quantity $\alpha + B\beta$ is the one which rejects when the likelihood ratio exceeds B (18).

4. PRIOR OPINION AND LIKELIHOOD RATIOS

The formulation of the previous section started by assuming the existence of two simple hypotheses, i.e., hypotheses that completely determined the distribution of the observation, $f_i(x)$. There are at least two different kinds of oversimplification involved in this assumption, each of which leads to a better conceptual definition of the problem than can be realized in practice. The first, and perhaps less important, assumes that the functional forms of the two distributions, $f_1(x)$ and $f_2(x)$ are known; otherwise the likelihood ratio could not be computed. There are probably some situations in which this is a reasonable assumption. Thus, if x is unidimensional and consists of 1's and 0's for success or failure on each patient and if independence can be assumed, $f_1(x)$ would be a hypergeometric distribution and $f_2(x)$ the non-central hypergeometric with a parameter, θ , expressing the ratio of the odds of a success with treatment 2 to that with treatment 1,

with x reducing to number of successes and failures among treated and control patients (19).² But more frequently, responses other than success and failure will be of interest; so will pre-treatment characteristics. Some of the variables will be dichotomous, some continuous, and the functional form for the complete set too complicated for anyone to be sure about. We mention this complication for completeness and shall not pursue it.

A second and major oversimplification arises from considering only simple hypotheses. Even if the observation set is so unrich as to be described by a central and non-central hypergeometric, the parameter θ will not be a single number but a range of possibilities. A cholesterol reduction of 15 per cent sustained over a five-year period may, as specified by hypothesis, lead to an increase in survival, but the hypothesis will ordinarily be silent as to whether θ is, say, 1.05 or 1.50. The likelihood ratio can be quite sensitive to this specification, however, even for large numbers of patients. If a θ of 1.50 is specified, corresponding to a decrease in mortality of about 28 per cent, and a trial with several thousand patients produces a decrease of only 2 per cent, we should regard the alternative as largely disproved. If θ is specified as 1.05, however, corresponding to a mortality decrease of about 4 per cent, an observed decrease of 2 per cent would leave the choice uncertain, no matter how large the number of patients studied.

There is no mathematical difficulty in "resolving" this issue. If the alternative admits a range of non-zero values of $\log \theta$, we can specify this range by a prior density

²The more common way of eliminating the nuisance parameter, particularly in sequential trials, is to pair each treatment with a control patient and to discard the concordant pairs. It was originally thought that this could be done with no loss in efficiency (15), but it is now clear that this holds only when the alternative approaches the null, i.e. as $\log \theta \rightarrow 0$, and that for non-zero $\log \theta$ the losses in efficiency can be substantial (20). The formulation that has been sketched out seems preferable for this reason.

on $\log \theta$, say $g(\theta)$, and take as the alternative distribution $\int f_2(x, \theta)g(\theta) d\theta$, where the dependence of $f_2(x)$ on θ has now been made explicit. As a special case we can let $g(\theta) = 1$ for one value of θ and 0 for all others, thus including the H_2 of Section 3. The likelihood ratio to which one is thus led, $\int f_2(x, \theta)g(\theta) d\theta/f_1(x)$, has been suggested by numerous authors as a way of handling uncertainty about the magnitude of the treatment effect. Good calls it the Bayes factor (21), Dickey the ratio of average likelihoods (22); I have called its reciprocal the rbo, for relative betting odds (23). Several clinical trials have tentatively adopted it as one among several possible measures of uncertainty (3, 4). For a sufficiently large number of patients it will, with high probability, be either very large if H_1 is true or very close to zero if it is false, no matter what prior probability, $g(\theta)$, is chosen. But with the number of patients often encountered in practice, the choice of $g(\theta)$ can have a marked effect on the likelihood ratio. Furthermore, if one accepts the decision theory argument, this is the measure of uncertainty one *must* adopt, since any other choice leads to incoherence.

There is nothing in the theory indicating how to select the prior density, $g(\theta)$, and only moderate help is at best obtained by talking to knowledgeable investigators about it. This ambiguity of priors is often regarded as a weakness in the Bayesian view. More cogently, however, it should be considered a strength, since it provides an appropriate explication of what in fact everyone, no matter what his behavior, seems prepared to admit theoretically, the equivocality of statistical conclusions. In any event, the conclusion to which we are led is that even when all complications are stripped away, there is a residual equivocality in appraising a null hypothesis which arises from the poorly defined nature of the alternatives to it, and that the expression of this equivocality is the prior

distribution. Whether this equivocality is a major problem in any particular instance depends on the data, but theoretically it is always present. It cannot be banished by assuming a particular prior and computing a particular likelihood ratio. Prior probabilities exist, not so much to have numerical values assigned to them, as to distinguish between coherent and incoherent ways of appraising a problem.

Another aspect of the oversimplification introduced by considering only simple hypotheses arises because of the multidimensionality of x . We cannot exclude the possibility of beneficial therapeutic effects accompanied by undesired side effects as in, say, the case of anti-hypertensive therapy. For this reason alone a more realistic formulation requires more than two hypotheses, such as, neither therapeutic benefit nor side effects, therapeutic benefit and side effects, therapeutic benefit but no side effects, or side effects but no therapeutic benefit, with each possibility indexed by one or more parameters with a range of possible values. But this is part of the multiple comparison problem, to the consideration of which from the Bayesian point of view, we now turn.

5. PATIENT SUBGROUPS

As indicated in Section 2, the multiple comparison problem in clinical trials takes many forms, one of them, multiple comparisons through time, having been touched on in Section 3. I should like to consider a very common version, the search for special treatment effects among subgroups of patients.

Dextrothyroxine (DT4), a cholesterol-lowering drug studied in the Coronary Drug Project, had been suspected to be contraindicated before the start of the study for a certain class of patients with coronary heart disease. As the study progressed, the DT4 group was observed to have a slightly elevated mortality, the elevation being most marked for those who

had been on therapy longest and with the most severe forms of coronary heart disease (23). The largest difference, a more than 90 per cent excess, was found in patients with angina pectoris, a history of one complicated myocardial infarction (MI) or multiple MI's, and a resting heart rate of 70 or more beats per minute. The group on DT4, but with none of these complications, had a 50 per cent *lower* mortality than placebo. To what extent were these subgroup effects real and how much a reflection of the technical ingenuity used to find subgroups with a large difference was the basic issue faced by those evaluating these data. It was eventually concluded that it was impossible to pinpoint precisely which subgroups were at excess risk, although such groups probably did exist. Furthermore, the loss involved in treating a patient with a harmful drug was considered greater than that entailed in not treating him with a beneficial one. The final decision reached therefore was to discontinue treatment with DT4 of all patients and not just special subgroups. Thus, a slight overall difference and a marked difference for some subgroups, with the subgroups defined not by prior hypothesis but by

computer search, led to discontinuance of the drug for all patients.

A second example comes from the Diabetic Retinopathy Study (7). A comparison of treated with untreated eyes well before the completion of planned follow-up showed that 129 out of 1375 of the latter had gone blind as compared with 56 out of 1375 of the former. The effect was most pronounced for those with the longest period of follow-up. The life table analysis showed that 28 months after treatment date, 20.0 per cent of untreated eyes and 6.4 per cent of treated eyes had gone blind, a difference equivalent to 5.9 standard errors. But the treatment did have certain side effects and not all subgroups of untreated eyes were at equal risk of blindness. It was therefore thought that a protocol change was called for, but perhaps not for all subgroups. Table 1 shows the two-year blindness rates for treated and untreated eyes for mutually exclusive subgroups defined by severity of baseline retinopathy. It was decided on the basis of these results to make treatment available to untreated eyes in subgroups *f*, *h*, *i*, and *j*, but to continue the protocol unmodified for all the other subgroups.

TABLE 1
Per cent blind at two years for treated and untreated eyes for subgroups with different baseline severity of retinopathy

Subgroup	No. of eyes at risk after 20-24 months of follow-up		Per cent blind at 2 years		(U-T)/SE
	Untreated	Treated	Untreated (U)	Treated (T)	
<i>a</i>	80	68	2.1	2.9	-0.4
<i>b</i>	7	12	5.6	0.0	0.9
<i>c</i>	51	41	5.4	2.6	0.9
<i>d</i>	6	7	0.0	0.0	
<i>e</i>	70	76	4.4	4.0	0.1
<i>f</i>	19	16	36.8	13.3	1.8
<i>g</i>	60	59	9.9	3.0	1.5
<i>h</i>	17	12	24.0	0.0	2.2
<i>i</i>	74	102	25.3	6.8	3.7
<i>j</i>	41	47	38.6	17.5	2.9
Overall*	457	477	16.3	6.4	5.5

* The information required for classification by subgroups was not available for all eyes at the time the paper was prepared.

The analysis of these data is not as simple as it might appear. A separate significance test at the .05 probability level for each subgroup did not implicate group f . But the overall error rate of these comparisons is $1-.95^{10}$ or .40, and if an overall error rate of .05 is to be preserved, the error rate for each separate comparison should be $1-.95^{1/10}$, or .0051. On this basis only the effects in subgroups i and j can be considered "established," even though inspection of the table suggests that the treatment effect, as measured by the odds ratio, is much the same for all subgroups. If the number of subgroups is increased, as it would be by using age and sex as additional variables, the numbers in each subgroup and the critical p value would decrease, and eventually there would be none for which the treatment effect could be considered established.

It is instructive to consider the Bayesian treatment of this problem. There are 10 possible odds ratios and a prior distribution, $g_i(\theta)$ ($i = 1, 2, \dots, 10$), must be selected for each. These priors can be considered as either independent and identical or independent and different. If one assigns different priors one is led immediately to an analysis in which each subgroup is analyzed separately, but without a multiple comparison adjustment. The general consequences of assigning identical priors are well known (24, 25), although the particular consequences for the non-central hypergeometric have not been studied. We may, however, consider the normal distribution, which provides an asymptotic approximation to the non-central hypergeometric (26, 27). Skipping the details, which are important but difficult to express concisely, the assignment for this case leads to 1) pooling subgroups if the differences among them appear small, 2) keeping them separate if differences appear large, and 3) providing intermediate results for intermediate situations.

The two assignments thus lead to quite

different answers, although neither leads to the multiple comparison answer. In one the possibility of a complete or partial pooling if the data suggest it is considered; in the second that possibility is rejected no matter what the data show. No objective basis exists for choosing between them. If the groups are thought possibly to have something in common, the assignment of identical priors is indicated and results for separate subgroups tend to reinforce each other, unless they are in disagreement; if not, assignment of different priors is indicated and each subgroup stands on its own; in cases of doubt, there is no statistical method that will help. These three possible attitudes were mirrored in the discussions of the decision-making bodies of the Diabetic Retinopathy Study. The fact that subgroup f , which, with only 19 and 16 eyes, could not stand on its own, was included in the same protocol change as group i , which could, indicates that some tendency equivalent to the assignment of identical priors existed. But two equally honest and statistically competent investigators could differ on the assignment and hence on the interpretation. Only more data, and not a deeper insight into the canons of inductive inference, could then produce agreement.

The first example, in which subgroups leading to maximum treatment differences were deliberately searched for, raises additional issues. These are best discussed in a closely related application, a stepwise regression analysis, which is a systematic way of searching for the independent variables with the largest linear, additive effects on a dependent variable. Caution in accepting the results of such a systematic search has been needed for some time. Thus, independent applications of the technique to two subgroups of placebo patients in the Coronary Drug Project, each of about size 800, starting with 40 variables and selecting the best 10 step-

wise, led to the remarkable finding that only two of the 10 were the same in both subgroups (28). Thus, the ST segment depression, ECG heart rate, systolic blood pressure, and New York Heart Association Class were the four factors that were most predictive of mortality in the first subgroup, but did not appear among the 10 most predictive ones in the second subgroup. Similarly, cardiomegaly on x-ray, intermittent claudication, ventricular conduction defects, and T-wave findings were the four most predictive factors in the second subgroup, but did not appear among the 10 most predictive ones in the first subgroup.

Theoretical insight into this phenomenon can be achieved by starting with a recent development in multiple regression techniques, ridge regression (29), and then considering its Bayesian version (30). In ridge regression, the diagonal elements in the variance-covariance matrix of the standardized variables are arbitrarily increased, in the hope of decreasing the expected mean square error of the regression coefficients. In the Bayesian version the standardized regression coefficients are assigned identical and independent normal priors, exactly as in the subgroup problem just discussed. The ratio of the variance of the prior to the variance of the dependent variable is a quantity formally equivalent to the arbitrary factor used to increase the diagonal elements in ridge regression. When these two quantities have the same numerical value, the Bayes and ridge regression estimates of the regression coefficients are identical. As in the subgroup problem, the variance of the prior is equivalent to the variance among the standardized regression coefficients. Both procedures pull the standardized regression coefficients towards each other, the amount of pulling back depending on the agreement among the standardized coefficients. The desirability of this pulling back is highlighted by the Coronary Drug Project experience

just cited, but it depends essentially on a willingness to consider which standardized variables are to be regarded as equally predictive a priori. An assignment of identical priors to a heterogeneous collection of variables, some of which are known to be important and some of which, like the Himalayas, are included in the analysis only because they are there, is unlikely to be productive in the absence of truly massive amounts of data.

The faith that objective, i.e., judgment-free, stepwise methodology could thread its way through the variables and emerge with the best is thus supported neither by the Coronary Drug Project (and other) experience nor this theory. Pure methodology is not enough.

6. RANDOMIZATION

One of the finest fruits of the Fisherian revolution was the idea of randomization, and statisticians who agree on few other things have at least agreed on this. But despite this agreement and despite the widespread use of randomized allocation procedures in clinical and in other forms of experimentation, its logical status, i.e., the exact function it performs, is still obscure. Does it provide the only basis by which a valid comparison can be achieved (9) or is it simply an ad hoc device to achieve comparability between treatment groups? This distinction appears important in practice. A number of questions follow, to which there can be only one answer if the first view is correct, but which must be considered on their individual merits if it is the second.

(a) Is a comparison of a new therapy with a historic control ever justified or must it always be simultaneous (31)?

(b) When randomized allocation is difficult, as in coronary bypass surgery (32) or anesthesia with halothane (33), or impossible, as in many epidemiologic investigations (34), are other methods possible or must investigation cease?

(c) When randomization is only partly successful, as with dropouts, non-adherers, or misdiagnosed patients (35), must the comparison be based solely on the originally randomized groups or can comparisons which include dropouts, non-adherers, or misdiagnosed cases also provide information?

Under the first view, the function of randomization is to generate the sample space and hence provide the basis for estimates of error and tests of significance, and the random numbers used must be generated by some physical process which is actually random. But under the second view, deterministic procedures, like the use of the decimal expansion of $\sqrt{2}$, π or e (36), or computer-generated pseudo-random numbers (37) might do equally well. Similarly, adjustments of physically-generated random numbers to yield a more nearly equal distribution of digits, as in the random numbers of Fisher and Yates (38), would appear to follow more nearly from the second than from the first view.

There are, as shown in Section 3, reasons for questioning the basic role of the sample space, i.e., of variations from sample to sample, in statistical theory. In practice, certain unusual samples would ordinarily be modified, adjusted or entirely discarded, if they in fact were obtained, even though they are part of the basic description of sampling variation. Savage reports (39) that Fisher, when asked what he would do with a randomly selected Latin Square that turned out to be a Knut Vik Square, replied that "he thought he would draw again and that, ideally, a theory explicitly excluding regular squares should be developed." But this option is not available in clinical trials and undesired baseline imbalances between treated and control groups can occur. There is often no alternative to reweighting or otherwise adjusting for these imbalances (40, 41). But both the variables showing an imbalance and the magnitude of the adjustment

will vary from sample to sample. The notion of undesired imbalance and adjustment for it, like that of regular Latin Squares, is, therefore, difficult to define in advance, making it an unattractive and perhaps unmanageable subject for theoretical investigation. It is thus an open question as to whether a more satisfactory frequency-based justification of the role of randomization can be found.

Although a subjective, decision-theoretic basis for randomization has not yet been achieved (42), it does appear consistent with the second view. A more precise expression of the idea of comparability of treated and control groups is provided by the notion that we have independent and identical priors for them, i.e., that they are exchangeable (43). If we refer to rendering priors exchangeable as "rexing," there are clearly many ways in which a sample might be rexed. Randomization is one, use of the decimal expansion of π another, use of the penultimate digit of the Social Security number still a third, etc. Your certainty that your priors are exchangeable would then depend on both the method of rexing and its outcome. It does not appear contradictory to assign a higher utility to the rexing method that gives a higher certainty of exchangeability. Although details remain to be worked out, it would appear as if such an assignment would support randomized or equivalent rexing procedures, without commitments to the acceptance of the doubtful outcomes of any of them. Such a view is thus consistent with a good deal of flexibility and with what appears to be good scientific practice. It places the emphasis on reasonable scientific judgment and the accumulation of evidence and not on dogmatic insistence on the unique validity of a particular procedure.

CONCLUSION

Despite the ambiguities involved in design, in decision making and in conclusion

reaching, it is undeniable that the clinical trial has constituted an important contribution to medicine. From the 1954 field trial on polio vaccine to this month's report on photocoagulation in diabetic retinopathy, questions have been crisply posed and definitively answered. It would be presumptuous to consider here why this is so, since it surely can be explained as simply a further triumph of experimental method as applied to clinical medicine. As such, it would not have surprised even Claude Bernard, and only the statistical participation would have puzzled him.

A clinical trial starts when interest shifts from deducing the consequences of therapy to observing them in a controlled setting. The steps from this initial conception to the completion of the first draft protocol, involving, as they do, defining the variables, the measuring instruments, the patient population, and considering many other detailed and substantively oriented matters, would astonish anyone accustomed to think of the outcome of a trial as simply the x of Section 3. It is the function of everyone engaged in the trial to assure that the x eventually observed bears a close correspondence to the x originally conceived. It is the special function of the statistician to see that this x can be and is reduced to intelligible and interpretable form. The strength of the clinical trial is a consequence of the willingness of all concerned to see that an appropriate x is conceived, observed, properly reduced and soberly interpreted. Many skills are involved in this, but for the statistician, a broad awareness of the limitations as well as the strengths of his methodology, is not the least of them.

REFERENCES

1. Atkins H: Controlled trials. *Br Med J* 1:1101, 1966
2. Schwartz D, Lellouch J: Explanatory and pragmatic attitudes in therapeutical trials. *J Chronic Dis* 20:637-648, 1967
3. The University Group Diabetes Program: A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes: II. Mortality results. *Diabetes* 19:785-830, 1970
4. The Coronary Drug Project Research Group. Clofibrate and niacin in coronary heart disease. *JAMA* 231:360-381, 1975
5. Gans DJ: Coronary drug project. (Letter) *JAMA* 234:21-22, 1975
6. The Coronary Drug Project Research Group. In reply. (Letter) *JAMA* 234:22-23, 1975
7. The Diabetic Retinopathy Study Research Group: Preliminary report on effects of photocoagulation therapy. *Am J Ophthalmol* 81:383-396, 1976
8. Robbins HE: Some aspects of the sequential design of the experiments. *Bull Am Math Soc* 55:527-535, 1952
9. Armitage P: *Sequential Medical Trials*. 2nd edition, New York, Halsted Press, 1975, Chapter 8
10. Meier P: Statistics and medical experimentation. *Biometrics* 31:511-529, 1975
11. Bross IDJ: Critical levels, statistical language, and scientific inference. *In: Foundations of Statistical Inference*. Edited by VP Godambe, DA Sprott. Toronto, Holt, Rhinehart and Winston of Canada, 1971
12. Cornfield J: Sequential trials, sequential analysis and likelihood principle. *Am Stat* 20:18-22, 1966
13. Lindley DV: *Making Decisions*. London, Wiley Interscience, 1971
14. Cornfield J: The Bayesian outlook and its application. *Biometrics* 25:617-657, 1969
15. Wald A: *Sequential Analysis*. New York, John Wiley & Sons, 1947
16. DeGroot MH: *Optimal Statistical Decisions*. New York, McGraw-Hill Book Company, 1970
17. Edwards W, Lindman H, Savage LJ: Bayesian statistical inference for psychological research. *Psychol Rev* 70:193-242, 1963
18. Cornfield J, Greenhouse SW: On certain aspects of sequential clinical trials. *In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 4:813-829. Edited by J Neyman, LM LeCam. Berkeley and Los Angeles, University of California Press, 1967
19. Fisher RA: Logic of inductive inferences. *In: Contributions to Mathematical Statistics*. Paper 26. New York, John Wiley & Sons, 1950
20. Cornfield J, Greenhouse SW: The estimation of a common odds ratio. (Manuscript)
21. Good IJ: A Bayesian significance test for multinomial distributions. *J Roy Stat Soc B* 29:399-431, 1967
22. Dickey JM: The weighted likelihood ratio, linear hypotheses on normal location parameters. *Ann Math Stat* 42:204-223, 1971
23. The Coronary Drug Project Research Group. The coronary drug project: Findings leading to further modifications of its protocol with respect to dextrothyroxine. *JAMA* 220:996-1008, 1972
24. Lindley DV: The estimation of many parameters. *In: Foundations of Statistical Inference*. Edited by VP Godambe, DA Sprott. Toronto, Holt, Rhinehart and Winston of Canada, 1971
25. Leonard T: Bayesian methods for binomial data.

- Biometrika 59:581-589, 1972
26. Cornfield J: A statistical problem arising from retrospective studies. *In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* 4:135-148. Edited by J Neyman. Berkeley and Los Angeles, University of California Press, 1956
 27. Lindley DV: The Bayesian analysis of contingency tables. *Ann Math Stat* 35:1622-1643, 1964
 28. Coronary Drug Project Research Group. Factors influencing long-term prognosis after recovery from myocardial infarction—three year findings on the Coronary Drug Project. *J Chronic Dis* 27:267-285, 1974
 29. Hoerl AE, Kennard RW: Ridge regression: Biased estimation for non-orthogonal problems, and ridge regression: Applications to non-orthogonal problems. *Technometrics* 12:55-78 and 69-82, 1970
 30. Lindley DV, Smith AFM: Bayes estimates for the linear model. *J Roy Stat Soc B* 34:1-41, 1972
 31. Gehan EA, Freireich EJ: Non-randomized controls in cancer clinical trials. *N Engl J Med* 290:190-203, 1974
 32. Cornfield J: Approaches to assessment of the efficacy of surgical revascularization. *Bull NY Acad Med* 48:1126-1134, 1972
 33. Subcommittee on the National Halothane Study, of the Committee on Anesthesia, Division of Medical Sciences, National Academy of Sciences-National Research Council: The National Halothane Study: A study of the possible association between Halothane anesthesia and postoperative hepatic necrosis. Edited by JP Bunker, WH Forrest Jr, F Mosteller, LD Vandam. National Institutes of Health, National Institute of General Medical Sciences, Washington DC, US GPO, 1969
 34. Report of the Advisory Committee to the Surgeon General of the Public Health Service: Smoking and Health. US Department of Health, Education, and Welfare, PHS Publication No 1103, Washington DC, US GPO, 1964
 35. Cooperative Study: Sodium heparin vs. sodium warfarin in acute myocardial infarction. Conclusions based on study of 798 cases at 13 hospitals. *JAMA* 189:555-562, 1964
 36. Gardner M: *Mathematical Carnival*. New York, Knopf, 1974
 37. HP-65 Stat Pac 1, Hewlett-Packard, 1974
 38. Fisher RA, Yates F: *Statistical Tables for Biological, Agricultural and Medical Research*. First Edition, London, Oliver and Boyd, 1938
 39. Savage LJ: *The Foundations of Statistical Inference*. New York, Halsted Press, 1962
 40. Cornfield J: The University Group Diabetes Program: A further statistical analysis of the mortality findings. *JAMA* 217:1676-1687, 1971
 41. Report of the Committee for the Assessment of Biometric Aspects of Controlled Trials of Hypoglycemic Agents. *JAMA* 231:583-608, 1975
 42. Savage LJ: *The Foundations of Statistics*. New York, Dover Publications, 1972
 43. DeFinetti B: *Theory of Probability: A Critical Introductory Treatment*. Vol 2, New York, John Wiley & Sons, 1975