ETHICALLY OPTIMISED DESIGNS

D.G. CLAYTON

University of Leicester, Department of Community Health, Leicester

Introduction

In his much quoted address to the Royal Society of Medicine in 1949, Pickering stated that 'therapeutics is the branch of medicine that, by its very nature, should be experimental'. This represents the view that has been widely accepted for nearly half a century. It represents the culmination of the movement, which took place largely in the 1930s, towards the development of a 'Clinical Science' in the application of the scientific method to clinical problems. Prior to this movement, the place of scientific method in medicine was strictly limited to the 'basic' sciences; anatomy, physiology and the rest. Clinical problems were, on the other hand, seen as best approached by use of the judgement of clinicians, based upon their long and varied experience.

The British Medical Research Council was particularly influential in promoting this view, and in establishing the place of the experimental method in therapeutics. In 1931, the M.R.C. set up the Therapeutics Trials Committee, and this Committee organised a series of trials of new treatments throughout the 1930s. These trials were not particularly well controlled by modern standards; sometimes alternate patients would be treated with 'test' and 'control' treatments, sometimes different participating centres would use different treatments. Interestingly, a search of 'Lancet' leading articles of the period reveals no serious concern for the ethical implications of such trials. Of course, the difficulties with such rule-of-thumb methods for allocating patients to treatments in trials are well known; they all too easily result in groups of patients that are not truly comparable. This mismatching of groups may occur as a result of differences in the type of patient referred to different hospitals, of secular trends in the prognosis for patients admitted to the trial, and of an unwanted intrusion of clinical judgement into the allocation/selection process. The exclusion of such influences from experiments is the aim of the principle of randomisation—the allocation of subjects to treatment at random. The randomised controlled trial was developed during the 1940s, and

its most famous and influential example was the M.R.C. streptomycin trial in 1946. However, the advent of randomisation and the accompanying principle of, where possible, 'blindness' of participating clinicians to the treatment individual patients receive, brought debate and doubts about ethics. Why it should be acceptable for a clinician to agree to treat alternate patients with different treatments, yet unacceptable to decide treatment upon the toss of a coin on arrival of each patient is not clear. In logic there seems to be no reason, save that with the former strategy, it is easier to cheat! The marriage of clinical medicine and the scientific method is indeed an uneasy one! It seems natural for clinicians, when forced to choose between the clinical judgement and the requirements of the scientific method, to opt for the former. Often, the statistician is cast in the role of policeman (Mainland, 1960); indeed many statisticians have actively sought such a role.

Since the 1940s the randomised controlled clinical trial has become the internationally accepted method of choice for evaluation of new therapeutic measures. Yet, doubts concerning their ethical implications have continued. Indeed, in recent years, the method seems to be under more attack than at any other period. A recent 'Lancet' leading article (Lancet, 1979) concluded that 'There is no doubt that the concept of controlled trials has been given undue deference and it would be unwise to wait for the results of these before some therapeutic measures are introduced.' The same article also stated that 'the conflict is between medical judgement and statistical significance' and, later, that 'There is no justification for many of these placebo trials apart from the speed with which the statisticians can be satisfied, and there is all the difference between statistical significance and biological significance'. The particular problems arising from the use of placebo treatments will be dealt with in another article in this series. The quotation is reproduced because it demonstrates the real danger that the current debate might degenerate into inter-professional bickering. This is disturbing since,

as we shall see, ethical considerations demand a closer marriage between medical and statistical judgements, not their dissociation.

The ethical problem

Perhaps the most succinct statement of the ethical responsibilities of doctors in clinical research is the 'Declaration of Helsinki', adopted by the 18th World Medical Assembly, Helsinki (1964). Several key quotations are given below:

- (a) 'It is the mission of the doctor to safeguard the health of the people'.
- (b) The declaration itself quotes the declaration of Geneva of the World Medical Association—'the health of my patient will be my first consideration'—and the International Code of Medical Ethics—'any act, or advice, which could weaken physical or mental resistance of a human being may be used only in his interest'. The declaration distinguishes between clinical research whose aim is therapeutic for the patient and purely scientific, non-therapeutic research. Clinical trials are in the former category, and the recommendations of the declaration of Helsinki are that:
- (c) 'In the treatment of a sick person, the doctor must be free to use a new therapeutic measure, if in his judgement it offers hope of saving life, re-establishing health or alleviating suffering'.
- (d) 'The doctor can combine clinical research with professional care . . . only to the extent that clinical research is justified by its therapeutic value for the patient'.

The declaration of Helsinki admirably sets out the ingredients for the ethical debate concerning clinical trials. The central problem is that the doctor is charged with two rather differing ethical responsibilities. These have been termed by Lellouch & Schwartz (1971) 'ethique collective' and 'ethique individuelle' respectively. These names are self-explanatory; the collective ethic is clearly set out under (a) in the principles listed above, while the remaining principles clearly concentrate on the responsibility of the doctor to an individual patient and, therefore, represent aspects of the 'individual' ethic.

The current assault upon randomised clinical trials is based upon the proposition that, within a clinical trial, these two principles are in direct conflict. It is argued that the protocol of such trials serves the collective ethic at the expense of the individual ethic

to an extent that is unjustifiable and possibly, in some countries, even illegal (Burkhardt & Keinle, 1978). Those putting forward this case have also drawn upon another argument in the statistical literature which questions whether the conventional controlled clinical trial is the best way in which the collective ethic may be served and urges the use of 'adaptive' designs (Weinstein, 1974).

In this paper I shall try to examine this apparent conflict and to discuss the extent to which various proposed strategies for therapeutic research resolve the difficulties. The task is a difficult one, for it involves controversial issues at every turn. The clinical trial, born out of the evolution of 'clinical science', is having to come to terms with the phenomenon which Fox (1980) has termed 'the evolution of medical uncertainty'. This uncertainty is evident in any discussion of medical ethics. Thus, Kass (1979) has written:

'How should we think about the ethical issues here, and in general? There are many possible ways and it is not altogether clear which way to test. For some people, ethical issues are immediately matters of right and wrong, of purity and sin, of good and evil. For others, the ethical terms are benefits and harms, risks and promises, gains and costs.'

This quotation, and much of the current debate, seems to imply that an absolute choice must be made between rival ethical standpoints. Thus, viewed from one position, clinical trials are immoral, and from another they are obligatory. Ethics, however, is usually defined as the reasoned analysis of human duty and Hill (1963) has pointed out that such analysis must be reapplied in a close and careful manner *in the specific circumstances of each proposed trial*. The analysis must be governed by broad principles of behaviour but, as in many other situations in medicine and elsewhere, we should not be surprised if these principles sometimes conflict with one another. The purpose of an ethical analysis is to better understand and delineate these areas of conflict. In any other context, these last few remarks might seem self evident, but they are worth stating here since, as we shall see, the areas of conflict between different ethical principles involve the knowledge and concepts of clinical medicine, on the one hand, and statistics on the other, and communication between these disciplines has always proved difficult.

For clarity in the following discussion, it is useful to re-state the 'individual' and 'collective' ethical principles:

(1) *Individual ethic*

It is the duty of the doctor to apply existing knowledge for the best possible treatment of each individual patient.

(2) *Collective ethic*

- (a) It is the duty of the doctor to acquire new knowledge so that, by such advance, future patients might benefit, and,
- (b) having acquired new knowledge, to accurately communicate it to other doctors.

These statements would seem to be non-controversial. In particular, principles such as (2a) and (2b) are necessary for medical advance. The latter is frequently overlooked, yet it provides an important part of the case for clinical trials. Altman (1980a) has rightly stressed that the accurate communication of research findings must be regarded as an ethical matter.

The controversy concerns the morality of the use of experimental method in pursuance of (2). Again, we may clarify the problem by isolating non-controversial areas. Firstly, the traditional justification of clinical trials (stated many times by Hill and others) is that, when the doctor is in a position of ignorance, there can be no conflict with (1) since there is no existing knowledge demanding to be applied. Thus, the doctor should be motivated primarily by (2), indeed ethical considerations dictate that his best efforts should be directed at the acquisition of new knowledge. The corollary (also stated repeatedly by Hill, 1963) is that, given adequate knowledge, there is little to gain by experimentation and the application of existing knowledge (1) must take first place. On most occasions the doctor will find himself in one of these non-problematic positions, and on others the nature of the disease will be such that there will be every opportunity for a patient involved in a trial to himself benefit from knowledge gained as a result of it. On these occasions ethical difficulties are few and we should not forget that perhaps the majority of trials are carried out under such conditions. On other occasions however, when there is some knowledge which falls far short of an adequate understanding, and when the consequences for the patient are serious, then there may be conflict between our guiding principles. It is the occurrence of such situations which has led to the search for 'ethically optimal' designs—strategies which minimise such conflict.

A crucial difficulty in the ethical analysis is the assessment of the quality and quantity of knowledge. Such assessment is particularly difficult in view of the variable nature of patients' responses to treatments. Theories of statistical inference are concerned with just this problem, but here too there are differences of opinion and of emphasis between different schools of thought. These differences have, in turn, indirectly contributed to the controversy surrounding clinical trials.

It is important, at this stage, to draw a clear distinction between the concepts of 'knowledge' and

'belief'. For example, a doctor might freely confess to having little or no *knowledge* concerning the relative efficacy of two treatments, and yet profess a strong belief in the relative merit of one of them. Equally, one of his colleagues, sharing the same knowledge base, might hold the reverse belief. Knowledge is commonly defined as 'assured belief', and beliefs solidify into knowledge only after careful appraisal of the weight of evidence. In a rational (and ethical) profession in which evidence is freely exchanged, two doctors might differ in their beliefs concerning treatment, but they should not differ in their knowledge. Consensus is, therefore, an attribute which differentiates knowledge from beliefs, and the achievement of consensus as to the efficacy of treatments was a central aim of the early protagonists of the controlled clinical trial.

Here, the guiding ethical principles have explicitly been stated in terms of knowledge rather than of belief. In this respect the discussion differs from that of Hill (1963) who wrote (of the participating clinician): '... if he thinks *even in the absence of any evidence* that, for the patient's benefit he ought to give one treatment rather than another, then the patient should not be admitted to the trial' (my italics). If clinicians were really to follow this advice to the letter, then I suspect that few patients would be admitted to clinical trials. It is in the nature of clinicians to form beliefs concerning treatments, which are unsupported by any evidence, but there should not be any *ethical* requirement to always act in accordance with such beliefs. Thus, we may add an explicit denial, which might be more controversial:

- (3) the doctor has no *ethical* responsibility to treat a patient in the manner he believes to be the best, if that belief is unsupported by any evidence or consensus.

This is not to say that to exercise personal 'clinical judgement' is anything other than a thoroughly sensible way to behave in day-to-day clinical work, but denies its elevation to the status of ethical principle. This status seems to be assumed in some of the current criticism of randomised trials, yet there seems no place in a reasoned analysis of human duty for a principle which urges slavish adherence to unsupported belief.

The relevance of statistical theory to this problem is in the process of weighing the evidence for, and against, rival beliefs. Much of this paper will attempt to clarify this theory; both in relation to the problem of deciding *how much* evidence is required, and that of deciding upon the best strategy for obtaining this evidence. Some understanding of these issues is necessary (a) to the understanding of the proposed ethically optimal designs, and (b) to the use of conventional designs in an ethically optimal manner.

Before embarking upon this task, however, there is

another element in the ethical analysis of any proposed trial which should be considered; the *type* of knowledge which is sought by the trial. Schwartz *et al.* (1980) make a clear and useful distinction between an *explanatory* aim (to seek knowledge at a biological level), and a *pragmatic* aim (to seek knowledge only of the effects of the treatments in practice). Clinical trials often seek both types of knowledge, and often there is no conflict between the two aims. Sometimes, however, there is conflict and this might have ethical implications. A recent example of this is the European coronary bypass surgery trial (European Coronary Surgery Study Group, 1979). The explanatory aim would be to determine whether patients who undergo surgery fare better than if they had been treated by conventional medical methods, but unfortunately suitability for surgery can only be assessed by an invasive and potentially dangerous procedure—coronary arteriography. For the trial to serve the explanatory aim, randomisation to medical or surgical treatment must be carried out *after* arteriography. This design allows us to answer the biological question but subjects the medical group to an unnecessary invasive procedure. Alternatively, the pragmatic aim would require a design in which randomisation is carried out *before* arteriography and in which the surgical group would carry with it the patients allocated to it but subsequently found to be unsuitable for surgery. The latter design better assesses the net benefit to patients of introducing the surgical option and avoids the ethical difficulty, but the biological interpretation is partially obscured. In this case it is impossible to design a trial which serves both aims (the trial organisers chose the former design). The distinction between explanatory and pragmatic aims is also important for the debate concerning the ethics of the use of placebos.

Schwartz and his colleagues (1980) do not urge that either aim should be automatically paramount in clinical trials, only that certain aspects of the design and interpretation depend upon it, not least the analysis of the ethical cost and benefits.

Tests of statistical significance

The most familiar statistical method in the medical literature, and that upon which the design and analysis of the conventional controlled clinical trial is based is the hypothesis test. The theory of such tests stems largely from the work of Neymann and Pearson, and is responsible for the (much abused) idea of 'statistical significance'. The principles of the method are illustrated below, using a simple two-group trial as an example.

If we are ignorant as to the relative efficacy of the treatments to be received by the two groups, then we have no evidence to suggest that one group will res-

pond any better than the other. That is, *as far as we know*, both treatments are equally efficacious. The test of statistical significance uses the concept of the null hypothesis (the hypothesis that there is no systematic difference in response between the two groups) to represent the initial state of ignorance. If the evidence of the trial can dislodge this hypothesis, then it can be said to have resolved our ignorance, and so has contributed to our knowledge. The steps in the analysis of a trial are as follows:

- (i) In the absence of data from a properly controlled experiment, adopt the 'null hypothesis', i.e. assume that there is no difference between the treatments.
- (ii) Then carry out the trial, and collect the data.
- (iii) Calculate a test statistic—a numerical measure of the discrepancy between the observed results of the trial and the null hypothesis.
- (iv) Assuming the null hypothesis to be true, calculate the probability of observing a discrepancy as large (or larger) in trials similar to the one considered.
- (v) If this probability is smaller than some arbitrary threshold, then we feel entitled to reject the null hypothesis and conclude that there is a proven difference between treatments. The size of the threshold probability is the 'significance level'. Traditionally, a 'small' probability is regarded as $P < 0.05$ or less.
- (vi) If this probability is not small, then the null hypothesis is not adequately discredited. The difference between the two treatments is then said to be 'not significant'.

These logical steps are widely known (even if not fully understood). Not so widely known are the criteria which are brought to bear in choosing a statistical test. Neymann and Pearson proposed that the choice should be made upon the grounds of the 'power' of tests to detect specified deviation from the null hypothesis. The power of the test is defined formally as the probability of detecting such a deviation. It depends upon (a) the size and nature of the deviation specified, (b) the size and precision of the trial, and (c) the statistical test chosen. The Neymann-Pearson theory advocates choice of the most powerful possible test, and this theory has led to the development of the commonly known statistical tests. Power considerations are vital for the design of trials (in particular, in the choice of the number of patients required) and to their interpretation.

After several decades of experience of statistical significance tests in medical research in general, and in clinical trials in particular, some serious difficulties have emerged. The first of these is that they have proved peculiarly liable to misinterpretation. In parti-

cular, misinterpretation of non-significant results is endemic in the medical literature (Altman, 1980b; Freiman *et al.*, 1978), and this gives considerable grounds for concern. Such results are commonly presented as proof either of lack of efficacy of a new treatment or that the two treatments compared are equally effective. The latter arises from an overliteral acceptance of the null hypothesis; as we have seen, the null hypothesis is an expression of ignorance, and failure to dislodge it means only that the weight of evidence has proved insufficient to justify firm conclusions. If the aim of the trial were explanatory, then it has failed to provide convincing evidence, and we must reserve judgement. If it were pragmatic then, on its completion, the clinician must continue to be guided by his judgement. However, if the results of the trial, although falling short of significance, provide some indication that one treatment might be superior, then this fact is relevant and should, together with other evidence, be taken into account in the formation of that judgement. The perceived conflict between statistical significance and medical judgement arises out of the fallacious belief that non-significant results should be disregarded.

The second difficulty is in the choice of significance level. In pure scientific work, it might be reasonable to fix significance levels at some agreed small figure, $P < 0.05$ or $P < 0.01$ say. The choice of these levels reflects the strength of the determination of the scientific community not to be falsely led from simplifying assumptions. In clinical trials, however, uncritical application of the 5% yardstick could be undesirable, particularly when their aims are largely pragmatic. Consider, for example, a disease which requires drug therapy over many years and for which the current therapy has good efficacy and acceptable safety. In these circumstances, it would be reasonable to require strong evidence of improved efficacy for a new treatment. On the other hand, for a disease with no current effective treatment, we might properly adopt a new treatment on less stringent criteria. The need to show flexibility in the choice of significance levels is doubly important if, as seems likely, non-significant results continue to be misunderstood.

A third difficulty with the theory is that it is based upon a model in which all the evidence arrives at once. Although it provides a method of assessing the implications of the totality of the data, it provides no way of examining intermediate results. This is not a serious difficulty for trials which have a relatively short duration, but is unacceptable in protracted trials of serious diseases; we shall return to this point later.

Likelihood and Bayes methods

Increasingly, what has been termed the 'coconut shy'

approach of the significance test is being questioned as the best model for judging the transition between beliefs and knowledge. Particularly disliked is the idea that this transition should be a sudden, rather than, a continuous (and possibly gradual) process. Two alternative approaches are advocated—the 'Likelihood' and 'Bayes' approaches. There is space here for only a brief discussion of these and an interested reader is referred to the books of Edwards (1972) and Lindley (1973) for thorough expositions of these two positions.

For our sample of a simple trial involving two parallel groups, both approaches start with the adoption of a mathematical model (which is simply a rather formal description) of the problem, in which some parameter, θ say, represents the amount of difference between treatments. For example, the model might state that the odds in favour of recovery in the first group are θ times as high as in the second group. Before the trial, the value of θ is unknown (or, at least, highly uncertain), and it is the purpose of a statistical analysis of the trial data to ascertain the extent to which our uncertainty concerning its true value has been resolved as a result of that trial.

The degree to which the observed results of the trial and some particular value of θ are mutually compatible may be assessed by calculating the probability of observing these results if θ were indeed equal to the assumed value. By recalculating this probability for a wide range of different values of θ , we obtain a function, $L(\theta)$, which is called *likelihood*. This idea is originally due to Fisher, who also pointed out that this function contains all the information, relevant to θ , provided by the trial. Figure 1 illustrates a hypothetical example. In general, one value of θ will be best supported by the data and this will correspond to the point at which the likelihood attains its maximum value (this value is called the maximum likelihood estimate of θ), but a range of values to either side are supported to lesser degrees. The relative support for different values of θ is measured by the relative *likelihood* which is simply obtained by expressing the likelihood relative to its maximum value.

Some statisticians argue that there is no need, in the analysis of the trial, to go any further than this. In particular, there is no need for any further recourse to probability. The relative likelihood expresses the ability of the data to differentiate between various values of θ . The function could be reported in full, or we might agree as a scientific community to regard all values of θ for which the relative likelihood exceeds some arbitrary value (10%, let us say) as being compatible with the evidence. If this range includes values indicating treatment A is better than treatment B, but also includes values indicating that treatment B is preferred, then the trial has failed to indicate convincingly which choice should be made. This is the analo-

gous concept to significance, and its dependence upon an arbitrary threshold value for relative likelihood is no more (and no less) disturbing than the dependence of conventional hypothesis tests upon arbitrary significance levels.

The Bayesian approach to inference takes this argument further. Just as the likelihood measures the weight of evidence for each value of θ provided by the trial, so it is possible to mathematically describe our 'degree of belief'. Degrees of belief, axiomatically, obey the laws of probability and lie between zero (certainly false) and one (certainly true). Further, the theory states, the degrees of belief for each value of θ *after* obtaining the results of the trial (which is called the 'posterior probability distribution' of θ) may be calculated, using some simple algebra (Bayes theorem), from the degrees of belief *before* the trial (the 'prior probability distribution' of θ) and from the likelihood.

This is an attractive mathematical model for the acquisition of knowledge although there are practical difficulties such as, for example, the choosing of appropriate mathematical expressions of prior beliefs. It is also a model in which those advocating the abandonment of controlled trials and a return to reliance upon clinical judgement might see some support. After all, Bayes theorem gives us a mechanism whereby each piece of evidence causes a defined modification to one's personal beliefs and, if presented with the same evidence, differences in prior beliefs will eventually be resolved. This is an attrac-

tive utopian dream, but its realisation in the immediate future seems unlikely.

These modern approaches to the theory of inference have much in common. They differ from the test of significance in two important respects. Firstly, the progress from unsupported belief to firm knowledge is seen as a gradual process rather than as a sudden transition; the likelihood (and the posterior probability) may be updated with each new item of data as it becomes available, and becomes sharper as more and more data is taken into account. Secondly, there is no undue interest in the 'null hypothesis', or in any other single value of θ . The emphasis is rather to adequately describe the range of uncertainty. If these ideas continue to gain force, their influence in the design, conduct and reporting of clinical trials might be considerable.

Randomised controlled trials

This, and the remaining sections of this paper will discuss the various proposed strategies for clinical trials and, bearing in mind the ideas outlined above, will attempt to assess their strengths and shortcomings both from an ethical and from a statistical standpoint. We shall consider, first, the most familiar type of trial, the conventional parallel-group randomised controlled trial involving a fixed, predetermined number of patients.

For the present, we shall not discuss the need

for randomisation (we shall return to this later), only its ethical implications. The conventional justification has already been advanced; namely, that when the doctor is in a state of ignorance concerning the relative merits of the two treatments to be compared, then there can be no ethical difficulty in choosing at random. The problem is whether this state of ignorance can be said to continue as the trial progresses.

In the strict logic of the significance test, around which such trials have conventionally been based, the answer is that ignorance does indeed continue until being finally resolved by analysis of the total data. This seems unsatisfactory, and is certainly at odds with the process of forming clinical judgements. As a result of this conflict, these trials are usually carried out under conditions of secrecy of intermediate results from participating clinicians. A recent example of this is the clofibrate primary prevention trial (Report from the Committee of Principal Investigators, 1978) in which this secrecy was maintained for close on a decade, and whose final results gave some cause for concern. The reason for such secrecy is that, if the intermediate results were available, participating clinicians might feel ethically unable to continue randomisation and might withdraw from the trial. In such circumstances it would prove very difficult to calculate a significance level for the trial.

At first sight, the need to withhold interim results from clinicians involved in trials seems to be incompatible with their ethical acceptability. However, this is not necessarily the case. Doctors participating in such a trial agree to be kept in ignorance of the results because they recognise that their subjective assessment of the evidence may not be as cool and rational as the statistical analysis. They are, therefore, agreeing to replace their own assessments by the proper analysis, and secrecy makes this easier to accomplish. This does not seem unethical, providing the statistical analysis really is the most rational way of resolving ignorance. Assuming the size of the trial to have been properly designed, then this analysis must be done when the results for *all* patients admitted finally become available.

This has two difficulties. Firstly, although it is a forceful argument against releasing intermediate raw results, a likelihood/Bayes approach would allow the likelihood function (and the posterior probabilities) to be updated continuously, and availability of these would allow a participating clinician to make a rational assessment of the changing ethical position. Indeed, with this viewpoint, it is difficult to see how suppression of information during a trial can be justified. Secondly, the argument relies heavily upon the correct sample size having been chosen. It is at this point that the ethical optimality of fixed sample size designs must be assessed. The choice of too large a

trial, is clearly ethically undesirable in that the required information could have been obtained using, let us say, only the first half of the patients; the remaining half could then have been treated in accordance with the knowledge so gained, rather than at random. The ethical costs of choosing too small a trial are harder to assess. In the current climate in which non-significant results are misinterpreted and also seem to act as a disincentive to further work, then these costs are high. In this climate, trials should have a high probability of obtaining a significant result, when there is a difference to be found. Schwartz *et al.* (1980) have pointed out, however, that the pragmatic aims of trials should only require a high probability that the trial should find in favour of the correct treatment, regardless of whether or not that finding achieves significance; this leads to smaller required sample sizes.

Whichever of these approaches is judged appropriate, the central difficulty remains, that the calculation of the sample size depends upon the (unknown) magnitude of the difference between treatments. If there is an unexpectedly large difference between treatments, then this could have been detected using a smaller trial.

These arguments, taken together, argue strongly the case for interim analyses. However, to carry out such analyses is to depart in principle from the fixed sample size design, since one of their consequences might prove to be the stopping of the trial. Randomised trials, involving interim analyses are called 'sequential' trials.

Sequential designs

A sequential trial is a randomised controlled trial in which, prior to randomisation of each patient, the results of the trial concerning all previous patients are scrutinised and a decision made whether or not to continue the trial. As indicated above, its purpose is to avoid the ethical difficulties which arise if the difference between treatments turns out to be more pronounced than expected; in such cases the sequential trial allows early termination. In terms of the ethical problem, the sequential design is superior to the fixed size design in that it allows the results of the trial to determine the moment at which to abandon the position of ignorance (and, therefore, when to stop random allocation), rather than at a fixed point decided *a priori*.

The ethical problem is not completely avoided since, until the moment of stopping the trial, patients are assigned to treatments at random. This can be defended on the grounds that, until the weight of evidence in favour of one treatment has exceeded some threshold value, knowledge is not yet firm. Since, as is demonstrated by Figure 1, this weight of

evidence may be measured in terms of the relative likelihood of the hypothesis of equality of the treatments, the ethical argument suggests stopping the trial when this relative likelihood falls below some agreed level. This is the stopping rule proposed by Wald in his pioneering work on sequential analysis (Wald, 1974).

Sequential trials are not, however, without their difficulties. Firstly, the problem of interpretation of significance levels is far from trivial. It is well known that, if at every stage of the trial, a statistical test were to be carried out and the trial stopped when (and if) the 5% significance is reached, then the statistical theory has been violated and the true significance of such a result is no longer 5%. This is because conventional statistical theory only admits the concept of probability in terms of the relative frequency of events in the world in the long run. In this theory it is logically inaccurate to speak of a hypothesis as being 'probably true', and statistical tests merely investigate the probability of making errors when adopting various decision-making strategies. The sequential trial is a different type of strategy, so it is not surprising that the significance levels are changed. Informally, a strategy which involves making a series of decisions must be more subject to error than one in which just one decision is made at the end of the trial. Thus, each decision in the series must be taken more conservatively if the error rates for the strategy as a whole are to be maintained. Bayesian statisticians argue that this simply doesn't make sense. They point out that it allows a situation to occur in which two statisticians might look at the same data, one just once, at the end of a trial, and the other, sequentially, as the trial progresses. The former statistician might conclude that there was a significant difference, and the latter that there was not. This does not seem sensible—two rational men faced with the same data should not be led to different conclusions according to how often they look at the data. It would be inappropriate to fully argue this debate here. A discussion is given in the book by Armitage (1975). The discrepancy arises because the logic of the Bayesian argument distinguishes clearly between the problem of belief and knowledge on the one hand, and the problem of deciding upon actions on the other; this is tackled by application of decision theory, to which we shall return later. In conventional sequential analysis, these problems are mixed. My personal sympathies are with the Bayesian position as far as the explanatory aim is concerned (i.e. for assessing the state of scientific knowledge), but the pragmatic aim of trials requires one to consider the probability that a trial might lead to an erroneous conclusion. This probability is affected by repeated looks at the data. Further discussion of this is to be found in the book of Schwartz *et al.* (1980).

Statistical methods exist for planning and operating

sequential schemes for most types of data. Again, it is not possible to describe these in any detail here and the reader is referred to Armitage (1975). In some cases, the necessary calculations may be carried out by completing a path on a special chart. Each time a result becomes available the path advances a step, its direction chosen by the nature of the result. The trial is terminated when the path intersects one of three boundaries of the chart, corresponding to decisions to, respectively, accept treatment A as the better, accept treatment B as the better, or to remain unconvinced either way. In other cases no such simple chart is possible, but modern computing techniques allows us to be undeterred by the calculations.

The remaining difficulties with sequential trials are essentially practical. The first and most important of these is that the trial should have one clear overriding measurement upon which evaluation of the therapies can be based. When the trial aims to assess the difference between treatments in respect of more than one clinical indication, the problem of formulating stopping rules and investigating the resultant properties of the trial become exceedingly difficult.

In many trials, it is desirable to try to learn something of the safety of the new treatment (indeed this is often an explicit aim of 'phase 3' trials) and this might require a certain minimum number of patients to be admitted. In certain cases it might be felt that final evaluation of the relative merits of two treatments requires the consideration of several different aspects of efficacy (and possibly of safety and acceptability), and it is difficult to imagine how this can be done in anything other than a fixed size design. However, it might also be the case that one measurement is regarded as of such importance that adverse trends should cause abandonment of the trial. Such considerations have led to sequential monitoring schemes being embedded within a conventional fixed size trial. One example of this is the European coronary-artery bypass trial which included a sequential monitoring of operative mortality. Another interesting example is the clofibrate prevention trial, in which the organising committee designed a sequential monitoring of all causes of mortality to be used, not explicitly to terminate the trial, but to release the interim results to the participating clinicians (although this would almost certainly have had the same effect!)

Another practical difficulty with sequential trials occurs when there is an appreciable time delay between the admission of a patient to the trial, and the final measurement of outcome. In this case, the decision to terminate the trial has to be made using incomplete data. This, and other problems involved in keeping the analysis of results up to date can make true sequential trials logistically difficult. These difficulties can be avoided in many cases by granulation of the design; i.e. carrying out the trial in a series of stages, each of fixed size. Such multi-stage trials are

analysed at the end of each stage, and, as for true sequential trials, rather more stringent significance levels should be applied at each stage if the correct overall significance level is to be maintained. A detailed discussion of this, and other issues in the design of multi-stage trials is given by Pocock (1980).

Decision theory and adaptive trials

The development of alternative trial designs has been suggested by application of the other wing of Bayesian statistical theory, decision theory. This deals with the choice between various courses of action in the light of belief formed after rational weighing of the evidence ('posterior belief'). This choice involves an analysis of the costs and benefits which might result from each course of action.

For example, let us consider the clinical decision as to whether to treat a patient with treatment A or treatment B. The two possible courses of action have different costs or benefits for the patient, and these depend upon the true relative efficacy of the treatments, θ . This is not known exactly, so that the best course of action is also unknown. However, by taking into account the posterior distribution of θ , we may calculate the *expected* cost or benefit of each choice, and decision theory urges us to choose the course with the most expected benefit. Thus, the results of a trial give us information upon which to base our posterior beliefs, and subsequent patients should be treated in accordance with the principles of decision theory, i.e. to maximise the expected benefit. The patients in the trial are treated in a manner which maximises the evidence to be provided, so that later decisions concerning future patients can be as well-informed as possible. This trade-off seems inescapable, but it is not clear whether the trial designs discussed so far are the best solution.

The decision theory approach to the problem argues against what it considers to be an artificial distinction between patients in the trial and future patients. An important point here is that the pace of medical innovation is such that the number of patients involved in a trial of a choice between two treatments might represent a sizeable proportion of the patients for whom that choice is relevant; the conventional statistical model of a trial population as a random sample from a hypothetical infinite population could be very misleading in pragmatic terms. Instead, decision theory sees the problem as being one in which the clinician is faced with a long series of decisions. In each one of these, he must weigh up (a) the likely costs and benefits of each course of action to the patient, and (b) the gain in knowledge (and, hence, the benefit to future patients) to be accrued from this patient. At the beginning of this series of decisions, the need for information will predominate

so that randomisation of patients will be indicated, but as information accumulates, the costs and benefits to the patient become more and more important, and will weigh in favour of treatment with the treatment currently believed to be the best. It is of interest whether decision theory makes any predictions as to how we ought to behave in a game with these rules. A best strategy could have claims to be an ethically optimal design.

Of course, the existence of such a strategy depends upon what we mean by 'best'. One possible solution is to interpret the collective ethic as requiring the greatest good for the greatest number, and to choose the decision strategy in such a way as to maximise the total number of patients treated correctly. This approach leads to *adaptive* designs, which start by randomly allocating patients to the two treatments, on a 50:50 basis, and, as evidence begins to favour one treatment, the randomisation becomes loaded towards that treatment. Eventually, all patients are being assigned to the more successful treatment. Day (1969) has shown, by computer simulation, that in terms of total number of patients correctly treated, the adaptive strategy out-performs the sequential design, which in turn is better than the fixed sample randomised design. However, the gain of adaptive designs over sequential designs seems to be relatively modest (Lellouch & Schwartz, 1971; Simon, 1977).

Adaptive designs suffer from the same practical difficulties as do sequential designs, particularly in the need for a single, definitive measure of efficacy. As for sequential designs, some of these practical difficulties might be avoided by running the trial as a series of stages, each stage using a different ratio for randomising patients to treatments. A comprehensive review of the literature concerning adaptive designs up to 1974 is given by Weinstein (1974). The topic is also discussed by Armitage (1975), who draws attention to a further difficulty, namely that, viewed as a whole, such a trial is not strictly randomised and would tend to follow secular trends in patients' prognosis. If, at the onset of a trend towards the admission of patients with a more favourable prognosis, treatment A happens to be favoured, then the secular trend will reinforce this (possibly fallacious) impression. These designs would have, therefore, some disadvantages in common with the use of historical controls (to be discussed in the next section). This difficulty does not seem insuperable, since the data could be examined *post hoc* for secular trend and the results could readily be corrected if necessary. Whether the optimality of the design would survive such a manoeuvre is another matter, and it would probably not be as convincing as would a simple randomised design.

Rather ironically, adaptive designs, although designed specifically with ethical considerations in mind, in a sense heighten the ethical difficulty. Late

on in the trial, clinicians may be required to randomise patients in a 9:1 ratio, and if one treatment is so good as to receive 90% of the patients, the ethical problems of withholding it from the other 10% seem even more acute. Against this, the adaptive design can only point to the best-for-most principle, and this has never been regarded as an acceptable medical ethic.

An alternative approach to ethical optimality in the choice of decision strategy is to be guided entirely by the individual ethical principle. That is, to treat each patient so as to maximise the expected benefit to that patient alone. This leads to the design proposed by Lellouch & Schwartz (1971) in which each new patient is treated with the treatment currently favoured by the results so far. This strategy is not particularly efficient in terms of the total number of patients correctly treated (although it does not perform too badly), but has the virtue of being the one design in which the interests of the individual are in no sense traded-off. It is, however, not a randomised design and would also suffer from a tendency to follow secular trends (although again there seems to be no reason why such trends could not be explicitly allowed for in the analysis).

A similar adaptive design is Zellen's 'play the winner' strategy (Zellen, 1969). This consists of playing the treatment with the current run of success, and changing to the alternative treatment when a failure is observed. Again, this is not a randomised design, and does not seem to have any claims to ethical optimality.

Other non-randomised designs

The designs discussed up to this point, with the exception of the last two mentioned, require random allocation of patients to treatments although, as we have seen, its effect in adaptive designs is partially lost. The need for randomisation is not, it must be conceded *universally* accepted; some Bayesian statisticians regard it as irrelevant. Space does not permit discussion of this debate here. The majority view remains that randomisation is most important in achieving consensus. The role of randomisation is, however, often misunderstood. In particular, randomisation does not ensure that two groups of patients are equally matched; it merely ensures that the probability of mismatch may be objectively calculated. Indeed, this probability is taken account of in the significance test, so that the possibility that the experiment has been nullified by extraneous influences is objectively assessed. However, failures of randomisation can and do occur. Confronted with a trial in which it turns out that the randomisation has failed to match the treatment groups sufficiently closely in respect of an important prognostic factor, it

is unthinkable that the statistician would not try to adjust for this in his analysis. But, if we are prepared to accept such *post-hoc* adjustments—and modern analytical techniques backed up by computing technology make them eminently feasible—then why bother with randomisation at all? The answer is that one can only adjust, *post-hoc*, for factors which (a) can be foreseen as relevant, and (b) can be accurately measured. The purpose of randomisation is, therefore, seen as being to address the problem of unforeseen factors. Taken to its extreme, the argument for randomisation would state that no causal connection can be formally proved while there exists a possibility of some such unforeseen factor. An example of this is Fisher's famous argument against the causal link between smoking and lung cancer.

Clearly, this extreme position is a luxury which medicine cannot afford. Epidemiology, for example, must depend predominantly upon non-randomised studies. Without randomisation, the possible influence of extraneous factors must be assessed on informal grounds of plausibility. Obviously, there will sometimes be controversy as a result, but the U.G.D.P. trial has shown that randomised trials are by no means isolated from such dissent (Committee for the Assessment of the Biometric Aspects of Controlled Trials and Hypoglaemic Agents, 1975).

Random allocation, then, although probably not absolutely necessary in every case, is certainly highly desirable. If we are to depart from it, it must be with good reason. The only strong reason which I can see is the existence of sufficient prior knowledge as to render withholding of the new treatment unethical. One suspects that such situations are probably rare, but they must be expected to occur occasionally. A situation in which there is much pressure on trialists in this direction is in conditions where there is no current treatment with any demonstrable efficacy. Thus, the investigator's prior belief might be that, whatever the merits of the new treatment, it is highly unlikely to be worse than the alternative. If this is genuinely the case, then a randomised trial is unethical, the new treatment must be offered to all patients and the results compared with a group of historical controls. As a note of caution, however, studies which have shown placebo treatment to be preferable to currently accepted active therapy are by no means unknown, and one is led to question the optimism of clinicians when confronted with new active therapies.

It is well known that the use of historical controls is risky. A particular problem is to determine whether there have been any secular trends in the type of patients admitted, or in the methods of assessment in use. Again, these problems are familiar in epidemiology, and they necessitate great attention to methodology, to standard agreed methods of recording and so forth. Without great care, not only are different series of patients not directly comparable,

but the factors which explain this are not recorded so that *post-hoc* adjustments cannot be made. In epidemiology, great attention is paid to response rates and the problems of selection bias. Wherever possible, whatever information on non-responders as can be obtained is published. In trials work, this care is usually absent and very little is known of the patients who might have been admitted to the trial but were not. There have been several discussions of the use of historical controls in clinical trials and the most important are cited by Altman (1980c).

An important aspect of this debate which is seldom mentioned is that the safety of treatments are often assessed by non-randomised, retrospective studies. To the end, Skegg & Doll (1977) have urged more attention to the routine recording of adverse events in clinical trials. The point could be made more generally. If data are to be usable in future retrospective studies, more attention has to be paid to the routine recording of data. We must also bear in mind the fact that the selection of data only from *published* studies is likely to be misleading!

Discussion

This paper has been concerned only with the relationship between *statistical* problems of design and analysis of clinical trials and the *ethical* position of trial lists. The account has been far from comprehensive but has, hopefully, illustrated the complexity of the problem both philosophically and practically. The

problem is one of assessing the relative claims of clinicians' judgement, scientific proof, and patient benefit. These are, of course, highly related to one another, but do not always coincide; it is this which makes the evaluation of medical uncertainty so difficult.

In summary, the following points of relevance to the design and conduct of trials emerge.

- (a) Randomisation is a highly desirable feature, and is ethical so long as the weight of evidence in favour of one treatment is not unduly strong.
- (b) The appropriate measure of this weight of evidence is the relative likelihood.
- (c) Sequential trials with stopping rules based upon relative likelihood are, therefore, strongly indicated.
- (d) The importance of significance levels in such trials is arguable, but their power (the probability of finding in favour of the correct treatment) is most important.
- (e) Adaptive designs seem to offer little gain over sequential designs yet bring new difficulties.
- (f) In situations of extreme ethical difficulty, the strategy of allocation of each patient to the current best treatment is a useful possibility.
- (g) Failing all these, we must make do with historical controls.

As far as the analysis and presentation of results are concerned, there should be less emphasis on testing the null hypothesis; we should be more interested in what trials do show than in what they do not.

References

- ALTMAN, D.G. (1980a). Statistics and ethics in medical research; misuse of statistics is unethical. *Br. med. J.*, **2**, 1182–1184.
- ALTMAN, D.G. (1980b). Statistics and ethics in medical research VII—Interpreting results. *Br. med. J.*, **2**, 1612–1614.
- ALTMAN, D.G. (1980c). Statistics and ethics in medical research; Study design. *Br. med. J.*, **2**, 1267–1269.
- ARMITAGE, P. (1975). *Sequential Medical Trials*. Oxford: Blackwell Publications.
- BRITISH MEDICAL JOURNAL (1978). Leading article. Interpreting clinical trials. *Br. med. J.*, **2**, 1318.
- BURKHARDT, R. & KEINLE, G. (1978). Controlled clinical trials and medical ethics. *Lancet*, **ii**, 1356–1359.
- COMMITTEE FOR THE ASSESSMENT OF THE BIOMETRIC ASPECTS OF CONTROLLED TRIALS OF HYPOLYCAEMIC AGENTS (1975). *J. Am. med. Ass.*, **231**, 583–608.
- DAY, N.E. (1969). A comparison of some sequential designs. *Biometrika*, **56**, 301–311.
- EDWARDS, A.W.F. (1972). *Likelihood*. Cambridge: Cambridge University Press.
- EUROPEAN CORONARY SURGERY STUDY GROUP (1979). Coronary artery by-pass surgery in stable angina pectoris: Survival at two years. *Lancet*, **i**, 889–893.
- FOX, R.C. (1980). The Evolution of Medical Uncertainty. *Milbank Memorial Fund Quarterly/Health and Society*, **58**, 1–49.
- FREIMAN, J.A., CHALMERS, T.C., SMITH, H. & KUEBLER, R.R. (1978). The importance of beta, the type II error and sample size in the design and interpretation of the randomised, controlled trial. *New Engl. J. Med.*, **299**, 690–694.
- HILL, A.B. (1963). Medical ethics and controlled trials. *Br. med. J.*, **1**, 1043–1049.
- KASS, L.R. (1979). 'Making babies' revised. *The Public Interest* (Winter), 32.
- LANCET (1979). Leading articles on controlled trials: Planned deception? *Lancet*, **i**, 534–535.
- LELLOUCH, J. & SCHWARTZ, D. (1971). L'essai thera-

- peutique: éthique individuelle ou éthique collective? *Revue Inst. Int. Statist.*, **39**, 127–136.
- LINDLEY, D.V. (1973). *Making decisions*. London: Wiley.
- MAINLAND, D.M. (1960). The clinical trial—some difficulties and suggestions. *J. chronic Dis.*, **11**, 484–496.
- POCOCK, P. (1980). *Proceedings of European Symposium on Medical Statistics*, Rome.
- REPORT FROM THE COMMITTEE OF PRINCIPAL INVESTIGATORS (1978). A co-operative trial in the primary prevention of ischaemic heart disease using clofibrate. *Br. Heart J.*, **60**, 1069–1118.
- SCHWARTZ, D., FLAMANT, R. & LELLOUCH, J. (1980). *Clinical Trials*, translated by M.J.R. Healy. London: Academic Press.
- SKEGG, D.C.G. & DOLL, R. (1977). The case for recording events in clinical trials. *Br. med. J.*, **2**, 1523–1524.
- SIMON, R. (1977). Adaptive treatment assignment methods and clinical trials. *Biometrics*, **33**, 743–749.
- WALD, A. (1947). *Sequential Analysis*, New York: Wiley.
- WEINSTEIN, M.C. (1974). Allocation of subjects in medical experiments. *New Engl. J. Med.*, **291**, 1278–1285.
- 18TH WORLD MEDICAL ASSEMBLY (1964). The Declaration of Helsinki. *World Medical Journal*.
- ZELLEN, M. (1969). Play the winner rule and the controlled clinical trial. *J. Am. stat. Ass.*, **64**, 131–146.