9 The stopping rule

It would be reckless to set the ponderous machinery of a randomized clinical trial into motion without some thought about how it is to be halted. The dilemma faced by doctors has a familiar form. If a treatment trial is stopped too soon, the harmful consequences or hoped-for benefits of the innovation will be overlooked; if it is too prolonged, patients who continue to receive the old treatment are denied the new benefits, or unnecessary numbers of patients are exposed to the inferior new agent under test.

> And Elijah came unto all the people, and said,
> How long halt ye between two opinions? If the Lord *be* God follow him: but if Baal, *then* follow him. And the people answered him not a word.
>
> I *Kings* 18:21

What is needed is some sort of a stopping rule that will limit the magnitudes of these opposing risks; it is completely unrealistic to expect that we can step into the unknown with perfect safety. And the rule must be devised with what may seem like imperious vacuity reminiscent of the King of Hearts' directive to the White Rabbit in *Alice in Wonderland*: 'Begin at the beginning,' the King said, gravely, 'and go on till you come to the end, then stop.'

There are, in fact, two situations for which plans must be formulated: the stopping point of a pilot trial and the termination point in a fully mounted formal trial. The first circumstance is often neglected because the relatively formless first steps of innovation are not perceived as requiring a preplanned limit.

SIZE OF A PILOT TRIAL

The prime objective of a preliminary exercise is to rehearse a proposed trial in order to uncover unforeseen difficulties that may arise. But there is additional information that needs to be rooted out in a pilot experience: we

must have some rough estimate of the size of difference in outcomes that is to be expected under the treatments on trial. This estimate provides the reasonable basis for settling on a number-specific question which will be tested.

Setting a limit in advance

Obviously, it is impossible to make a *calculation* beforehand of the amount of experience required to hone the operations and provide a realistic estimate of the expected difference. It is necessary, nevertheless, to make a firm declaration about the dimensions of a pilot trial before it is launched. The information available from past experience and from pre-clinical tests in animals may be used to help with the judgment for setting a limit.

The exploratory phase, I said earlier, should not be open-ended because it is too easily misguided. When a run of favorable outcomes occurs under a novel treatment, there is an understandable temptation to continue and include the results of the pilot trial in the legitimate rolls of the full exercise. When the first outcomes after an innovation seem no better or are worse than the standard approach, there is strong pressure to abandon the follow-through plan for an authentic test. These attitudes act as stumbling blocks to an ordered progression of evaluation; and the obstacles are encountered particularly often when the exploration of a new treatment is conducted without concurrent controls.

> You can prove almost anything with the evidence of a small enough segment of time. How often in the search for truth, the answer of the minute is positive, the answer of the hour qualified, the answers of the year contradictory!
>
> Edwin Way Teale

A 'dramatic breakthrough' A snowball effect of early success was seen when a new treatment to support the oxygenation of newborn infants with severe respiratory distress was introduced in 1969. Slight continuous positive pressure was applied to the airway of infants with the condition called 'hyaline membrane disease' to counter the pathologic forces in this disorder that tend to collapse the lung during the exhalatory phase of each cycle of respiration. Only about one quarter of infants with severe symptoms were expected to recover. The first infant treated in this way survived; in 1970, the results of treating seven babies (six of whom recovered) were reported to the medical community; and in 1971, the results of treating 20 babies (16 of whom lived) were published in full.

This new approach was adopted by physicians quickly and widely; testimonial reports of favorable experiences appeared in medical journals throughout the world. Five years later a reviewer concluded that the weight

of confirmatory evidence (based on historical comparisons) seemed overwhelming, and this form of treatment was declared a major breakthrough in the care of distressed infants. The claims, however, were not supported by the results found in the few controlled clinical trials that were later conducted to evaluate the limits of applicability of the positive-pressure technique. The advantage over conventional methods of ventilatory support was quite modest.

The problem in conditions with high mortality

When new treatment for a highly fatal condition shows promising results in a pilot trial, an argument is often made against stopping to carry out the formalities of a randomized trial. (Such an 'unjustified-risk' argument was made after the encouraging early results of positive-pressure treatment were made known in 1970 and 1971.) There is nothing particularly illogical about this reasoning if we can accept the premises on which the argument is based. However, as we have seen, wholly regular events in medicine are quite rare. Moreover, the classification of clinical states are subject to considerable leeway in interpretation. Unlike the situation in meningeal tuberculosis before streptomycin (p 44), the definition of an 'invariably fatal condition' is often imprecise and the label is applied loosely. Uncertainty in diagnosis, particularly early in the course of illness, is common.

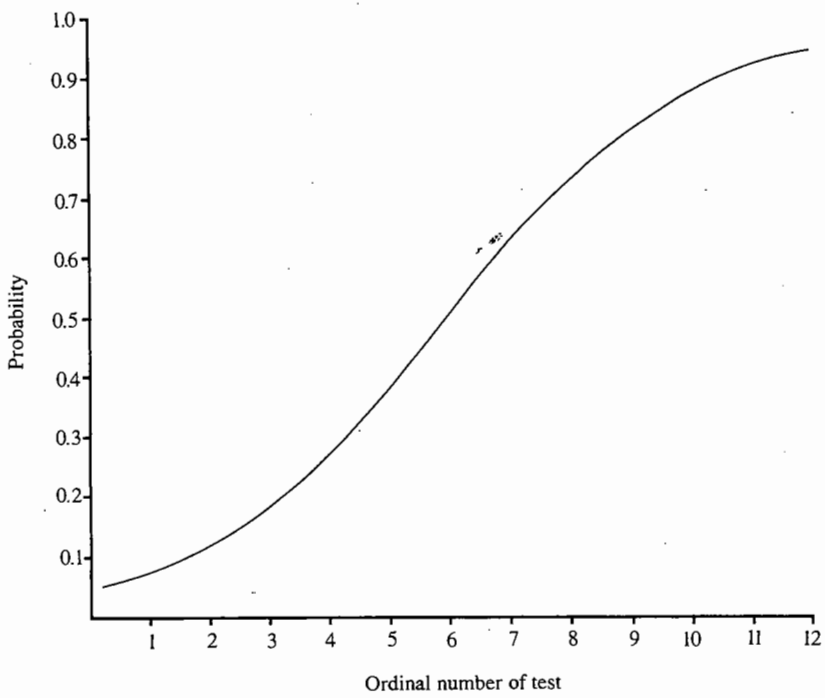
Recognition of mild cases A Boston group observed that patients with severe and fatal disease merely represent the 'tip of an iceberg'. There are almost always many more with milder forms of the same disease, and increased interest in diagnosis takes place when a new treatment is introduced. This invariably leads to the recognition of less severely involved patients who were previously overlooked (most of the unheralded patients recovered spontaneously). Consequently, outcome in currently treated persons appears improved as compared with previous experience, even when treatment is without effect or is actually deleterious.

With hindsight, a strong case could have been made for the use of a randomized control design from the very outset of the pilot trial of positive-pressure treatment. The conservative approach would have provided a hedge against unknown risks, and the concurrent control design would have guarded against the possibility that a progressive shift in severity of disease among the twenty babies enrolled over the sixteen-month period of the initial observations might distort the estimate of expected difference in outcome.

Preparation for a strong challenge

On the working assumption that there are no universally applicable, perfect treatments, the goal of critical experimentation is to flush out weakness and

An idealized model of medical diagnosis



Given certain evidence from a patient with a single disease, a doctor formulates questions and carries out a number of tests. Evidence in favor of the disease may thus be elicited at a constant rate: the probability of the presence of the disease then increases as a symmetrical S-shaped curve. W.I. Card and I.J. Good of the University of Glasgow point out that it is not known on which part of the curve a doctor usually works when he makes a diagnosis.

limits in new proposals. We have seen that it is difficult to mount a strong challenge when the question posed by an innovation is posed in general form. The numberless question, Does the new technique of positive-pressure improve the survival of distressed babies? is relatively safe from the risk of refutation until it is recast in quantitative terms.

Patients who first receive the untried treatment are asked to act as an advance party to scout the general question and bring back information for use in formulating a specific version that has an increased probability of exposing limitations. There may need to be several forays to generate information about the details of treatment (such as timing and dosage) before a vulnerable question is framed: Does the new technique, administered in the manner specified, improve survival from about 25 per cent to approximately 80 per cent?

This process of angling for a model based on observations of events as they occur is similar to the hypothesis-seeking 'fishing expedition' (p 20) using events that occurred some time in the past, and the same precautions about interpretation apply. Any preliminary search of observations makes it difficult to evaluate results of statistical tests on the same collection of information.

SIZE OF A RANDOMIZED CLINICAL TRIAL

The focus of interest in estimating the proper size of clinical experiments is not quite the same as that in the formal procedure used by statisticians for determining how many observations are required to decide whether or not to reject a hypothesis. Zelen has pointed out that many so-called hypothesis-testing situations in the world of medical events are in reality 'estimation of difference' problems.

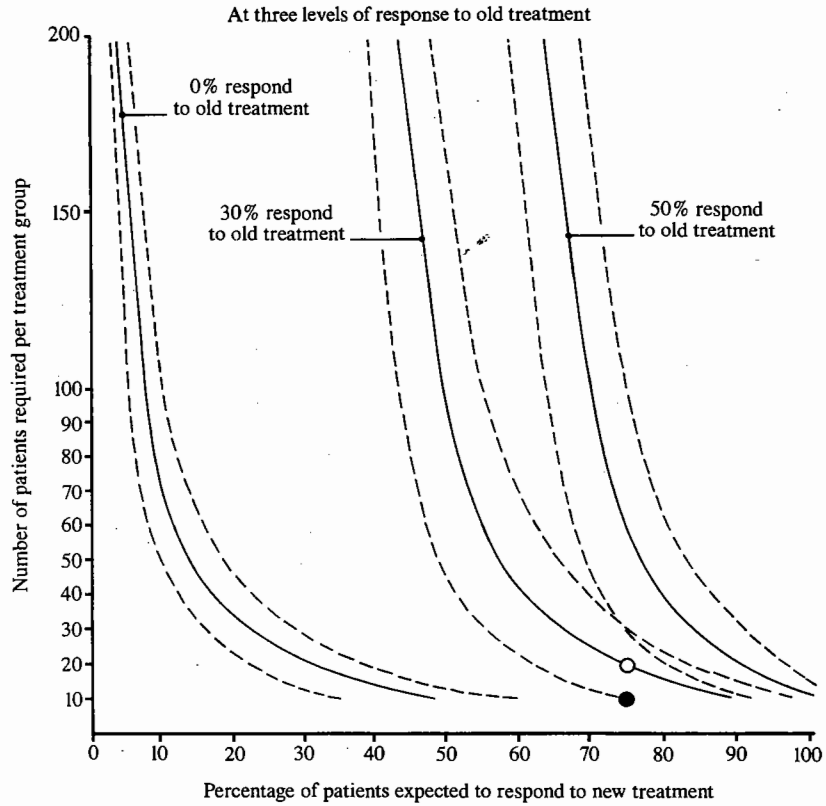
Null hypothesis versus limits of difference

In the classic hypothesis testing procedure we start with a general question structured in the form of a disclaimer, called the null hypothesis; for instance, positive-pressure treatment is neither more nor less effective than standard management. Then we specify the conditions for deciding whether or not to reject this assertion as true. In a pragmatic trial we are not primarily interested in the truth of hypotheses. When two methods of management are compared, we want to know the limits of their difference, for it is extremely unlikely that they will bring about effects that are *exactly equal*.

We are forced to concede, therefore, that in clinical matters the null hypothesis is rarely, if ever, true. In the example of the positive-pressure treatment, there was every reason to expect, from the outset, that it would not give results that were exactly the same as standard management. The goal of a formal test is to provide an estimate of some magnitude of difference that will have practical importance in general application. This target is within reach of the randomized clinical trial. The weightier questions concerning the status of hypotheses are simply left hanging—it is impractical to test them exhaustively at the bedside.

Despite this caveat, it is useful to *begin* the design of an experiment from the hypothesis-testing point of view. The reasoning used in the theoretical model allows us to make a preliminary calculation of the size of a clinical trial. The first approximation of dimensions serves as a solid point of departure for some thought about the medical and community implications of results of the proposed project. The final size, which emerges after completion of fairly wide-ranging consultations, is likely to give rise to a stopping rule that is unique for each clinical trial.

Rough estimate of the number of patients to enroll in a controlled trial



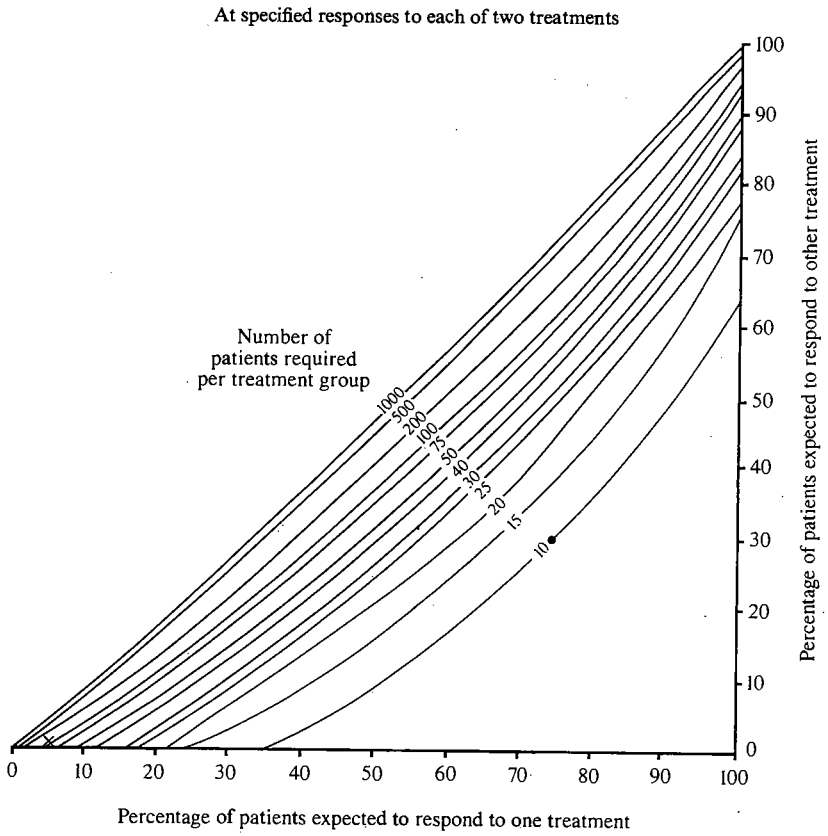
The vertical scale indicates the number of patients required in *each* treatment group, and the horizontal scale notes the proportion of patients expected to respond to the new treatment. The three sets of curves depict situations in which the proportions of responses to old treatment are 0, 30, and 50 per cent.

In each set, the solid line curve indicates the relationships when the chances* for detecting the specified differences are 4 out of 5, the dashed line to the left when the chances for successful detection are only 1 out of 2, and the dashed line to the right in each set when the chances for a decisive result are 95 out of 100.

For example (○), when 30 per cent of babies respond to old management and 75 per cent are expected to recover under the new treatment, if 20 patients are enrolled in each group (that is, a total of 40 patients in the trial), the chances are reasonably good (4 out of 5) that the difference will be detected in a trial of this size. If only 10 patients are enrolled in *each* group (●), the chances are merely 1 out of 2 that a 'statistically significant' result will be obtained. Compiled by C.J. Clark and Colin C. Downie of Imperial Chemical Industries (Britain).

*In these calculations the risk level for Type I error is set at 5 per cent. Levels for Type II error are: solid line 20 per cent (power 0.8), dashed line to the left 50 per cent (power 0.5), and dashed line to the right 5 per cent (power 0.95). There is agreement that a two-sided test of 'significance' will be performed on the results (that is, we wish to detect differences in both directions, improvement or worsening with the new treatment).

Smallest number of patients to enroll in a controlled clinical trial



The vertical and horizontal scales indicate the contrasting proportions of patients expected to respond to two compared treatments. The number of patients required in *each* treatment group for a 1 out of 2 chance* of detecting the specified difference is indicated by the set of curves. The greater of two percentage responses is always plotted against the horizontal scale when using this graph.

For example (●), when a 30 per cent versus 75 per cent response contrast is expected under two treatments, 10 patients in *each* group would be needed for 1 out of 2 chances of obtaining a 'statistically significant'* result. On the other hand (X), when a 1 per cent versus 5 per cent treatment response contrast is expected, more than 100 patients in *each* group are needed at the same risk levels.

Compiled by Clark and Downie.

*In these calculations the risk level for Type I error is set at 5 per cent. The level for Type II error is 50 per cent (relatively low power 0.5); and a two-sided test of significance will be carried out on results.

Idealized calculation of trial size

In the idealized approach we are obliged, before the trial begins, to provide arbitrary answers to three questions that will come up at the end of the exercise. First, if we say at the end of a trial, 'There is an important difference in outcomes,' how large a risk of being wrong are we willing to take? In the event the declaration is mistaken, we have committed a Type I error; the probability of this risk is reported as α, the 'significance level' adopted in the trial.

Second, how large a risk are we willing to take in missing the actual existence of an important difference by declaring 'No statistically significant difference'? A mistake of this kind is called a Type II error; the probability of the second risk is termed β.

Finally, before the trial commences, we must answer the question, What is the smallest difference we regard as important enough to find? The difference between outcome under standard treatment and the new treatment is designated as Δ.

The answers to the three questions provide the information required for a laborious calculation of trial size. Statistical texts summarize the results of the arithmetic in the form of sample-size-needed tables; these indicate the *minimal* number of patients to enroll in a clinical trial at various risk levels adopted for the two types of error and the specified difference in treatment outcomes.

Chances of detecting a specified difference The term 'power of the test' $(1-\beta)$ appears in sample size tables and graphs to indicate the probability of detecting a specified difference. Stated another way 'power' expresses the chances of avoiding a Type II error. When we adopt a relatively low risk level for β (say 0.05), the detecting 'power' of the trial is quite high ($1-0.05=0.95$). The sample-size-needed calculations make it clear that improved chances of uncovering the true difference in a clinical trial can only be achieved by increasing the number of patients enrolled.

Trial size and 'negative' results Jennie A. Freiman and collaborators at Mount Sinai School of Medicine reviewed 71 'negative' randomized clinical trials; the observed differences between the proposed and the control treatments were not large enough to satisfy a specified 'significance' level (the risk of a Type I error) and the results were declared to be 'not statistically significant'. Analysis of these clinical studies indicated that the investigators often worked with numbers of enrolled patients too small to offer a reasonable chance of avoiding the opposing mistake, a Type II error. Fifty of the trials had a greater than 10 per cent risk of missing a substantial difference (true discrepancy of 50 per cent) in treatment outcome. The reviewers

warned that many treatments labeled as 'no different from control' have not received a critical test because the trials used had insufficient 'power' to do the job intended.

PRACTICAL ASPECTS OF TRIAL SIZE ESTIMATION

When the limits imposed by the inexorable laws of probability have been determined (from the results of the statistical arithmetic displayed in sample-size-needed tables), we must turn to face the realities imposed by the everyday world. And these mundane considerations must extend to the long-term perspective of the community at large: its goals and its resources.

Magnitude of an 'important' difference

What does the statement 'clinically important difference' mean? Obviously, the notion of what we will consider 'important' is as difficult to capture as a greased pig. Unfortunately, we cannot circumvent the slippery task, for the magnitude of the 'important' difference is often the most crucial issue in planning and evaluating the clinical trial.

A Procrustean approach is often used to define the size of an 'important' difference by an inverted process. Doctors begin by estimating the limit of trial size (for example, the number of patients available in one institution over the period of time they can devote to the study) and from this number they determine the smallest difference that can be detected at various risk levels for the two types of error. This ad hoc approach to the definition of an 'important' difference should be avoided because it tends to promote trials with low 'power'. Whenever possible, multicenter trials should be carried out to overcome restrictions on numbers of patients available.

The size of difference in outcome specified in an upcoming trial should reflect a view of public gain or loss. In every case we must consider that when a medical experiment is completed and the report is published, it ceases to be an isolated decision problem. It has been said that the experimenter pays the piper and calls the tune he likes best; but the music is broadcast so that others may listen. Thus, we must ask, How much of its resources is the community willing to invest to achieve a specified improvement in outcome?

The question is not foreign to community planners who must advise about investments whenever an intervention is proposed to improve the outlook for citizens. The reasoning is the same when the issue is a new project to reduce automobile accident-related deaths and disabilities or a new treatment to reduce untoward outcomes related to premature birth.

Qualitative aspects of 'importance'

A clear distinction between pragmatic and explanatory emphasis is not always possible in a clinical trial. Under such ambiguous conditions, the definition of an 'important' difference requires some review from the point of view of the underlying theories about the pathologic process that is under study. For example, in evaluating the effect of treatment on RLF, which of two outcome indicators will provide more important practical as well as explanatory information: the frequently occurring early blood vessel changes, which often subside, or the less frequent manifestation of scarring, which is irreversible?

In order to detect a specified reduction in occurrence of RLF, we will need relatively few patients if we choose vascular changes as the indicator; many more patients must be enrolled to detect a difference in scarring, the more rarely occurring end point. But the two indicators provide different kinds of information about the disorder because the relationship between the early changes and scarring is not a simple one (for example, there is good reason to suspect that the agent which initiates the blood vessel abnormalities is unable to produce the scarring complications in the absence of an additional operative factor).

I suggest that only such far-ranging considerations by community pragmatists and by medical theorists provide a practical solution to the difficult problem of fixing the value of the 'important' difference in a randomized clinical trial.

From these remarks it will be obvious that the arbitrary values chosen for the probabilities of committing Type I and Type II errors are subject to the same review: risk levels in the formula used to calculate trial size must take into account the implications for community well-being and for medical theory. Again, the risk levels chosen exert a sharp effect on the estimate.

Multiple end points

Before leaving the considerations involved in fashioning a stopping rule, I must call attention to the fact that I have confined the arguments to a single end point of interest. A number of difficulties are introduced by multiple end points in a clinical trial. In addition to the issue of whether or not pre-trial predictions were made, there are complications that relate to the matter of a stopping rule.

Trial size in conflicting outcomes At a time when the frequency of RLF in the United States was relatively high (before 1955), a small randomized clinical trial was conducted in a single hospital to test the effect of oxygen management policy in reducing the risk of eye damage. The results sup-

High *versus* low oxygen

RLF and Survival

	Oxygen management	
	High	Low
No. enrolled	45	40
Survived	36 (80%)	28 (70%)
RLF	8 (22%)	0 (0)

A small randomized clinical trial was conducted in a single hospital in 1953–4. Babies in 'high oxygen' received this treatment for two weeks after birth; 'low oxygen' was given only for cyanosis (blue discoloration of the skin). It was concluded that the difference in survival rates in the two groups was 'not statistically significant'.

8 of 36 surviving infants (22 per cent) developed scarring RLF after 'high oxygen' treatment; there were no instances of the disorder among 28 babies in the 'low oxygen' group.

It could be said before this trial began, the chances of detecting a reduction in occurrence of RLF from 22 per cent to 0 are about 95 out of 100 with about 40 babies in each oxygen management group. However, for merely 1 out of 2 chances of detecting a 'statistically significant' difference in survival from 80 per cent to 70 per cent, well over 100 infants in each group will be required (risk level for Type I error set at 5 per cent, for Type II error at 50% = power 0.5).

ported the prediction that this risk would be reduced by a policy of oxygen restriction. Far too few infants, however, were enrolled to evaluate an adverse effect of the new policy on survival.

We can envision the mirror image of this disturbing problem of trial size in conflicting outcomes, at a time when the frequency of RLF was quite low, as indeed it was when positive pressure treatment was introduced in 1969. If this new treatment that improved the oxygenation of premature infants had been evaluated in a controlled trial, both survival *and RLF* would have been important end points of immediate interest. Here, the smallest number required for even odds of detecting an increase in survival from, say, 30 per cent to 75 per cent would be 10 infants in each group. But the number needed to detect the possibility of a five-fold increase in RLF blindness in the same babies (say from 1 per cent to 5 per cent) would have been over 100 patients enrolled in each group.

THE MULTIPLE-LOOK DILEMMA

In describing the process of determining trial size, I have made the assumption that the dimensions are rigidly fixed before the project begins. Now we need to examine some of the difficulties that arise with such fixed sample size plans.

The occurrence of temporary trends in favor of one of two compared treatments is guaranteed by the laws of probability, even when equally effective treatments are compared. (In the behavior of perfect coins, runs of 'heads' and 'tails' *must* occur in a series of tosses.) In order to guard against expectancy effects (p 65) that are introduced when physicians and nurses become aware of such trends as a trial progresses, it is literally necessary to hide the accumulating data. This is very difficult to carry out; in many circumstances, conscientious caretakers are well aware of which treatment is 'ahead'. Moreover, only an investigator with superhuman will power can refrain from 'peeking' at the results from time to time in the course of a long clinical trial.

'Peeking'

The director or overseeing group is tempted to analyze the results repeatedly as they accumulate and to stop the trial the moment 'statistical significance' is achieved (before the pre-trial requirements for a fixed sample size have been met). The motives for these actions are irreproachable; the number of patients exposed to an inferior treatment should be kept to a minimum, and patients should be protected from undertaking risks that were unforeseen when the trial was designed. But a penalty must be paid for these well-intentioned monitoring measures, and the price of 'peeking' is often overlooked.

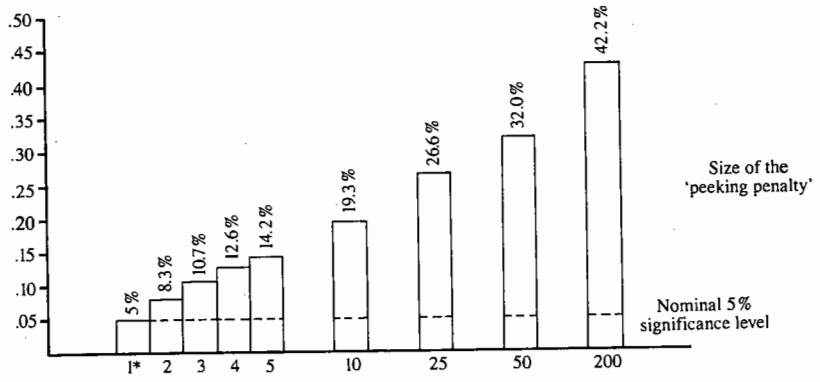
Prevention of sudden death after heart attack The problem was exemplified by the experience of a large multicenter controlled study of a new drug to prevent sudden death among patients with a recent heart attack (myocardial infarction). The researchers performed interim analyses of the results as they accumulated.

In early 1978, the outcomes appeared to favor the new agent, sulfinpyrazone, and the question of whether or not to continue had to be faced. In the absence of a predetermined stopping rule, the investigators settled their dilemma by submitting the results to a prestigious medical journal. The agony-torn researchers were quoted as saying that if the magazine accepted the findings as 'statistically valid', the experiment would be halted!

Inevitability of 'statistical significance' When accumulating data are tested repeatedly, 'statistical significance' *will be achieved* if the testing continues indefinitely. This flat statement applies no matter how improbable a 'significance' criterion we choose, and regardless of whether or not there is any difference in the effects of two treatments under test.

In essence, as an analyst provides the goddess of fortune with repeated opportunities for action during the collection of results, the likelihood of

The 'peeking' problem



* Data analyzed once at pre-determined fixed sample-size point

Change in probability (expressed as percentage) of achieving a positive result, wholly by chance, after repeated (serial) testing for 'statistical significance'. The conventional pre-trial risk level of 5 per cent for Type I error is premised on a once-and-for-all significance test. After the tenth 'peek' at accumulating results the nominal 5 per cent greatly underestimates the probability of a chance effect: it is now over 19 per cent. The chances climb to over 42 per cent by the 200th attempt to analyze accumulating data.
Calculated by McPherson.

achieving a positive result wholly by chance increases progressively. The situation is not unlike that in horse racing; all the rules of the race must be agreed to in advance, and all bets made before the starting bell. Bettors who declare that the race is over when their horse is ahead have trouble collecting their money.

Adjustment for specified 'peeks' A way of overcoming the penalty of multiple analyses while retaining the option of stopping a trial early was suggested by Klim McPherson while at the Medical Research Council (Britain): shift the nominal level of 'significance' at which the trial will be stopped. For example, it is decided *in advance of a trial* that a 5 per cent risk level of committing a Type I error will be maintained, and the accumulating data will be examined for 'statistical significance' on five occasions. McPherson calculates that under these conditions, a risk level of a bit more than $1\frac{1}{2}$ per cent will be required in any one of the five tests of 'significance', if the pre-trial Type I error risk level of 5 per cent is to be safeguarded faithfully.

Adjusted 'significance' levels after repeated tests on accumulating results

	Number of repeated 'significance' tests												
	1	2	3	4	5	6	7	8	9	10	15	20	100
Pre-trial level of 'significance' (%) 1	1	0.56	0.41	0.33	0.28	0.25	0.23	0.21	0.20	0.19	0.15	0.13	0.06
5	5	2.96	2.21	1.83	1.59	1.42	1.30	1.20	1.13	1.07	0.86	0.75	0.32
10	10	6.01	4.62	3.85	3.37	3.04	2.80	2.60	2.45	2.32	1.88	1.66	0.72

Given the pre-trial 'significance' levels of 1, 5, and 10 per cent (risk levels for Type I error), the adjusted value needed at stated numbers of repeated tests is provided in the body of the table.

By the tenth test on accumulating results, a level of 1.07 per cent must be achieved to maintain a pre-trial risk level of 5 per cent for Type I error.
Calculated by McPherson.

Sequential analysis

Another approach to the multiple-look dilemma in medical experiments is to adopt a stopping rule designed to permit 'continuous peeking' as the trial proceeds. The scheme is based on a section of statistical theory known as sequential analysis, developed largely by Abraham Wald and the Statistical Research Group at Columbia University in the 1940s. The practical problem addressed by the theorists was the sampling procedure used in industry for quality control; sequential procedures of inspection were developed that used the results of drawing articles from an assembly line to decide whether or not sampling should continue.

In sequential plans adapted for clinical trials, sample size is not fixed in advance. The results are inspected continuously and the trial halts according to predetermined rules decided by the choice of risk levels for Type I and Type II errors and specification of a clinically important difference. The method usually leads to economy in testing. On the average, fewer observations are required to reach a decision; but the 'savings', it should be noted, are not certain. The likelihood of reducing the number of patients who must be exposed to poor treatments is an attractive argument in favor of sequential medical trials.

Limitations of monitoring designs

Sequential plans (and the adaptive designs for allotment which I mentioned earlier, p 55) are limited to the kinds of clinical situations in which the outcome of treatment can be assessed fairly soon. If the diagnosis of the result must be delayed, the entire reason for the continuous monitoring strategy is lost (unnecessary enrollments continue while waiting for undecided outcomes).

Critics of sequential medical trials have pointed out that the approach allows only a single target of response in devising a stopping rule. But this drawback applies to all focused experiments. Phased approaches need to be

devised to test multiple questions, particularly when conflicting outcomes must be taken into account.

Much of the desperation felt by bedside researchers who are struggling to devise a stopping rule is summed up by the comments of one group involved in a large scale trial; 'Once you terminate this, you're up against it: You can never do it this way again!' The long-term costs of this now-or-never restriction deserve careful analysis.