STATISTICS IN MEDICINE, VOL. 1, 25–36 (1982)

ON CHOOSING THE NUMBER OF INTERIM ANALYSES IN CLINICAL TRIALS*

KLIM MCPHERSON

Department of Community Medicine and General Practice, Gibson Laboratories Building, Radcliffe Infirmary, Oxford, OX2 6HZ, England

SUMMARY

Small but important therapeutic effects of new treatments can be most efficiently detected through the study of large randomized prospective series of patients. Such large scale clinical trials are nowadays commonplace. The alternative is years of polemic and debate surrounding several trials each too small to detect plausible differences with any certainty. Such trials produce equivocal and contradictory results, which could be predicted from power calculations based upon sensible pre-trial estimates of treatment differences. Unfortunately such calculations often lead to sample sizes of several thousands.

It is not surprising that investigators tend to be over-optimistic in their estimation of treatment effects (which are necessarily uncertain) especially when the sample size requirements are so stark. In this paper a method is outlined for incorporating into the sample size calculations the uncertainty of the estimate made at the design stage of a clinical trial. In particular a formal scheme is described for deciding how many interim analyses should be performed to satisfy ethical and pragmatic requirements of large clinical trial design. Although the argument will be 'Bayesian', the criteria for assessment and comparison will be strictly of a Neyman–Pearson (i.e. significance testing) kind.

KEY WORDS Clinical trial Interim analysis Sequential analysis Prior knowledge

INTRODUCTION

Sequential methods for clinical experiments developed mainly by Armitage[1] provide a theoretical framework for the continuous assessment of data. Typically, observations are paired, either by matching patients on confounding variables or simply by matching consecutive entrants; a decision to stop or continue the trial is made on the accumulated data as each new pair of observations becomes available. Such a strategy, in principle, is extremely attractive because large therapeutic differences become obvious relatively early in the progress of the trial, with clear practical and ethical advantages. Sequential methods have another advantage over fixed-sample size designs. Initially there is uncertainty in the estimate of the expected treatment difference and consequently in the sample size requirement. This uncertainty is less critical because, at least for larger than expected treatment differences, the actual sample size will be appropriately smaller. However, for a variety of reasons, sequential trials have not often been used.

This has usually been attributed to complications of design and analysis.[2] The reasons may in fact be more deeply rooted in the wider practicalities of clinical experimentation. Sequential methods offer important advantages over fixed-size trials since fewer patients may receive inferior

* Invited paper

0277–6715/82/010025–12$01.20

Received 22 October 1981
Revised 22 December 1981

treatment and savings in expense may be substantial. However, large clinical trials are only planned and undertaken when substantial differences in efficacy are considered unlikely. If a large treatment effect is plausible a small trial will be planned for ethical and practical reasons: few clinicians would embark on a study lasting two years to detect an important therapeutic effect, when such an effect could be detected with almost equal certainty from a study lasting six months. Thus sequential methods offer major advantages mainly in large trials and then only when the estimated treatment effect turns out to have been pessimistic. In practice estimates of treatment effect, if anything, tend to be optimistic. Hence the circumstances in which a sequential design may be advantageous are likely to be rare. At the same time the costs and inconvenience of unfamiliar methods, paired allocation and assiduous analysis often appear substantial.

METHODS

The choice of design is not restricted solely to either sequential methods or a fixed sample size trial. More recently group sequential designs have been described. These were first suggested in 1966[3] as a reasonable compromise between the two extremes offering some of the theoretical advantages of sequential methods with fewer of their practical disadvantages. The notion is that the accumulated data can be analysed several times during the course of a clinical trial and if unexpected results emerge at any of these analyses then appropriate action can be taken. The question addressed in this paper is how to decide on the number of such analyses in advance of a clinical trial.

Adopting the Neyman–Pearson framework of significance testing for the analysis and interpretation of clinical trials leads immediately to the problems of optional stopping.[4] Several analyses using fixed sample size criteria will lead to inappropriate rejection of the null hypothesis more often than just one final analysis;[5] the frequency of testing must therefore be taken into account. In this paper the frequency properties of repeated significance tests are used.[6] Thus if ten interim analyses are planned at equally spaced intervals during a trial then to yield an overall probability of (say) 5 per cent of falsely rejecting a true null hypothesis the trial should be stopped as soon as any one of the analyses yields a nominally significant difference at the 1·05 per cent level.[7] If this criterion is always used for stopping a trial and claiming a significant effect, then over a large number of clinical trials the overall probability of a Type I error, of rejecting a true null hypothesis, will be 5 per cent. This argument is controversial[8,9] and alternatives which are more attractive to some involve Bayesian statistical inference or decision theory which may invoke imprecise prior distributions or utility functions. More importantly perhaps such arguments appear not to have captured the imagination of the great majority of investigators.

To maintain a particular probability (2α) of a Type I error the nominal significance level ($2\alpha^*$) at which a trial should be stopped becomes lower as the number of interim analyses increases. The probability of a Type II error (β) (accepting the null hypothesis of no treatment effect when there is a treatment effect of a stated amount), increases as more interim analyses at lower levels of nominal significance are performed. In other words if the maximum sample size of a clinical trial is fixed, the effect of increasing the number of interim analyses (hereafter called 'looks'), while imposing more stringent nominal significance levels, is to reduce the power of the trial ($1-\beta$) at all values of the treatment difference under test.

This is illustrated in Table I where results obtained by numerical integration are presented for a matched-pairs study. The difference in response within each matched pair is represented by a Gaussian variable with unit variance. The table shows the operating characteristics of eight possible designs for which the maximum sample size of each treatment group is 40 and for which the probability of a Type I error is 5 per cent. In the main part of the Table the probability ($1-\beta$) of achieving at least one result of an interim analysis which is significant at the stated nominal

Table I. Operating characteristics of eight plans with constant Type I error probability (2α) of 5 per cent.

Maximum number of patients per treatment (N)	40	40	40	40	40	40	40	40
Group size (g)	1	2	4	5	8	10	20	40
Number of looks (L)	40	20	10	8	5	4	2	1
Nominal significance level ($2\alpha^*$)	0·0056	0·0074	0·0105	0·0121	0·0155	0·0178	0·0300	0·0500
Operating characteristics (overall probability of achieving nominal significance levels)								
Size of treatment difference ($\lvert\delta\rvert$)	Type I error probability (2α)							
0·0	0·05	0·05	0·05	0·05	0·05	0·05	0·05	0·05
	Power ($1-\beta$)							
0·1	0·05	0·07	0·07	0·07	0·07	0·08	0·08	0·09
0·2	0·15	0·16	0·17	0·18	0·19	0·19	0·22	0·24
0·3	0·31	0·33	0·35	0·36	0·38	0·39	0·43	0·48
0·4	0·54	0·56	0·58	0·60	0·61	0·62	0·67	0·72
0·5	0·75	0·77	0·79	0·80	0·81	0·82	0·86	0·89
0·6	0·90	0·91	0·92	0·93	0·93	0·94	0·95	0·97
0·8	0·99	1·00	1·00	1·00	1·00	1·00	1·00	1·00
1·0	1·00	1·00	1·00	1·00	1·00	1·00	1·00	1·00

significance level ($2\alpha^*$) is tabulated. The mean (δ) of the Gaussian response represents the treatment effect (see below). A value of δ equal to zero corresponds to no treatment effect and increasing treatment differences correspond to increasing values of δ; $\delta = 1{\cdot}0$ represents a substantial effect relative to the intrinsic variability of the treatment response. It can be seen that holding the Type I error probability constant involves both more stringent nominal significance levels for more looks and a uniform reduction in power. The assumptions about the distribution of the response variable appear not to be important because qualitatively similar results are obtained for binomial and exponential response variables. If, however, the response variable is the time between diagnosis and some event such as death then results are less strictly applicable. This arises firstly from the censoring of some observations (patients will be known to be recurrence free some time after diagnosis) and secondly, from delay in the measurement of response (the number of patients available for meaningful analysis will not correspond to the number of patients admitted). However, Canner[10] has shown that the frequency properties of repeated significance tests on survival data are very similar to those illustrated in Table I. Therefore results reported here could be used as a guide to the choice of the number of interim analyses, making appropriate changes in terminology, in the design of trials with a prolonged period of observation on each patient.[11] In this particular application approximations are sufficient because the *exact* estimation of appropriate sample size or number of looks serves no useful purpose.

The results in Table I essentially apply to 40 pairs of consecutive observations. As is now explained, they also apply to the difference between the mean responses of 40 observations on each of two treatments, i.e. an unpaired study. In an unpaired design, sampling from two independent populations (one on the new treatment and the other on the old for instance) will conform to the situation of Table I if the response variable of each population has a variance of one half. The variance of the difference between two observations, one from each population, will then be unity (in analogy with the variance of the difference within each matched pair). The test statistic will be the difference between the mean response in each group. In addition, the variance of a response variable is almost always estimated from the data itself but extensive simulations using a t distribution[12] indicate that very similar results as will be described below hold for repeated t tests.

Clearly for frequent interim analyses some sort of paired allocation becomes a practical necessity to be certain of having patients on each treatment arm, but as the group size (denoted by g) increases then it is only necessary to ensure balance between treatments at each interim analysis. For example in the fourth column of Table I (group size equal to 5) the first analysis will involve the difference between the means of the first five observations on each treatment. These need not be paired but it is only necessary to ensure in the randomization that of the first (and each subsequent) group of ten patients, five are allocated to each arm.

Finally the results in Table I can be generalized to the case where the variance of the response variable is not restricted to unity: In a paired design σ^2 will be the variance of the differences of two paired observations; in an unpaired design the independent populations of response variables will have equal variances of $\sigma^2/2$. All measures of treatment effect (δ) may then be scaled by the standard deviation (σ).

Because most clinical trials are unpaired we will concentrate on an unpaired interpretation of the tables. To illustrate these ideas consider a clinical trial of antidepressant therapy with response measured by the Hamilton score for depression. The reduction in Hamilton score over a four week period attributable to a new treatment will be compared with that from conventional treatment. If the average reduction in Hamilton score on conventional treatment is 30 units with a standard deviation of 10 units, then the variance of the response is 100 units. Assuming that the patients on each treatment arm are not paired σ^2 equals 200 and σ equals 14·14. A value, $\delta = 1{\cdot}0$, would then imply an actual treatment effect of 14·14 units on the Hamilton rating scale. In other words the real response to treatment would average 44·14 (30 + 14·14) units over a four week period on the new treatment. When $\delta = 0{\cdot}5$ this average would be 37·07 units.

The expected sample size is considerably reduced by many interim analyses, particularly at large values of the treatment difference. The expected sample size and its variance for the plans of Table I, are shown in Table II. It can be seen from this table that, compared with a fixed sample size trial, a halving of the expected sample size when $\delta = 0{\cdot}6$ is obtained for 40 looks but (from Table I when $\delta = 0{\cdot}6$) at the cost of a reduction of power from 97 to 90 per cent. To maintain the power characteristics for many looks at the accumulating data the maximum sample size would have to exceed 40. The questions to be answered are firstly how much the maximum sample size need be increased to maintain power and secondly what are the consequences for the expected sample size. Having solved these problems the choice of number of looks when the probabilities of error of the two kinds (design characteristics) are held constant will be investigated.

Firstly, let us generalize the results in Tables I and II for any value of the maximum sample size. In fact we shall take 40 as a minimum number of observations because in general we shall be interested in clinical trials with more patients than 40 in each treatment arm. For instance if we wanted to compare the characteristics of clinical trials with a maximum sample size of say 200 observations we could retabulate Tables I and II by noting that the left hand column (headed by $|\delta|$) has to be divided by $\sqrt{k}$, where k is the new maximum sample size (200) divided by 40, i.e. $k = 5$ in this example. This is satisfactory for Table I, which will correctly denote the operating characteristics of the design. In Table II however, we have also to multiply all the expected sample size values by k. Similarly the variance of the sample size has to be multiplied by k^2. If we do this for $k = 5$ we will have the correct tabulations for varying numbers of looks in trials with a maximum sample size of 200 observations. The number of looks remains the same as presently tabulated but the group sizes are clearly five times greater in the new tables. Similarly the nominal significance levels remain the same as these only change as the number of looks changes to maintain a constant overall Type I error probability.

We want at the same time to maintain a constant Type II error probability as we change the number of looks and as suggested above this involves altering the maximum sample size. By

Table II. Expected sample size and variance of the sample size for the plans of Table I

Maximum number of patients per treatment (N)	40	40	40	40	40	40	40	40
Group size (g)	1	2	4	5	8	10	20	40
Number of looks (L)	40	20	10	8	5	4	2	1
Nominal significance level ($2\alpha^*$)	0·0056	0·0074	0·0105	0·0121	0·0155	0·0178	0·0300	0·0500
Size of treatment difference ($\lvert\delta\rvert$)	Expected sample size							
0·0	38·6	38·7	38·9	38·9	39·0	39·1	39·4	40
0·1	38·2	38·3	38·4	38·4	38·6	38·7	39·1	40
0·2	36·7	36·8	36·8	36·9	37·1	37·2	38·0	40
0·3	33·9	33·9	33·9	34·0	34·3	34·6	35·9	40
0·4	29·6	29·7	29·7	29·9	30·3	30·7	33·0	40
0·5	24·6	24·6	24·8	24·9	25·7	26·2	29·5	40
0·6	19·6	19·7	20·0	20·0	21·2	21·9	26·1	40
0·8	12·3	12·5	13·1	13·5	14·7	15·6	21·6	40
1·0	8·2	8·6	9·4	9·7	11·2	12·3	20·2	40
1·2	6·0	6·4	7·3	7·8	9·4	10·8	20·0	40
Size of treatment difference ($\lvert\delta\rvert$)	Variance of sample size							
0·0	42·5	37·1	31·7	30·1	24·3	21.4	11·6	0.0
0·1	51·9	46·6	41·3	38·7	33·3	30·0	17·9	0·0
0·2	80·2	75·4	69·8	65·5	59·9	55·3	36·7	0·0
0·3	122·2	117·4	110·9	108·6	98·1	92·0	64·9	0·0
0·4	158·7	152·9	144·6	139·2	129·8	123·1	91·2	0·0
0·5	165·4	158·5	149·0	145·3	135·5	129·7	99·7	0·0
0·6	139·0	132·2	123·8	120·0	114·0	110·2	84·8	0·0
0·8	65·0	61·6	57·9	55·3	54·6	52·8	31·9	0·0
1·0	27·9	26·6	25·5	24·7	23·6	21·3	4·2	0·0
1·2	14·0	13·2	12·8	12·2	10·2	7·5	0·3	0·0

interpolation from Table I we can estimate, in terms of the mean difference (appropriately scaled), the size of the alternative hypothesis (denoted by δ_1) for a stated power. These interpolated values are shown in Table III for five values of power. If, for example, we were interested in a clinical trial for which we wanted the overall significance level (2α) to be 0·05 and the power ($1-\beta$) of detecting a scaled difference of say 0·2 equal to 90 per cent we would proceed as follows. From Table III, for a given number of looks, say 8, we obtain power of 90 per cent if $k = 1$ and $\delta_1 = 0{\cdot}568$. The value of k for which we achieve a power of 90 per cent when $\delta_1 = 0{\cdot}2$ is given by $(0{\cdot}568)^2/(0{\cdot}2)^2$ or 8·07. Therefore the maximum sample size would be $40 \times 8{\cdot}07$ which is 323 and with eight looks the group size between looks would be 40·4 or approximately 40 observations. This argument is repeated for all the columns of Table III and tabulated in Table IV. From this we can now see that to maintain power a considerable increase in maximum sample size is required when the data are to be examined often. However all of these plans are now comparable in that each has the same probabilities of error for the same alternative hypothesis.

To examine the expected sample sizes for each of these plans at various values of the mean treatment effect under test (δ) we have to use the values of k given in each column of Table IV and adjust the values of δ and expected sample size given in Table II as described. To compare one plan with another at the same values of δ we have to estimate the expected sample size for a particular value of δ and number of looks by interpolation on a graph of expected sample size against δ for a given column. This has been done in Table V. It can be seen that the expected sample size is smallest for a fixed sample size trial when there are small treatment effects and for many looks when the

Table III. The size of the alternative hypothesis (δ_1) at which a given power is achieved with a maximum number (N) of 40 patients per treatment

Power $(1-\beta)$, %	Number of looks (L)							
	40	20	10	8	5	4	2	1
95	0·655	0·643	0·632	0·624	0·620	0·615	0·595	0·568
90	0·601	0·591	0·577	0·568	0·563	0·556	0·534	0·512
75	0·500	0·490	0·478	0·472	0·463	0·459	0·437	0·416
50	0·385	0·380	0·365	0·350	0·352	0·348	0·329	0·310
25	0·266	0·265	0·249	0·245	0·239	0·230	0·218	0·203

Table IV. Designs with 90 per cent power against the alternative hypotheses, $\delta_1 = 0{\cdot}2$

Number of looks (L)	40	20	10	8	5	4	2	1
δ_1'	0·601	0·591	0·577	0·568	0·563	0·556	0·534	0·512
Multiplication factor (k)	9·03	8·73	8·32	8·07	7·92	7·73	7·13	6·55
Maximum number of patients per treatment (N)	361	349	333	323	317	309	285	262
Group size (g)	9·0	17·5	33·3	40·4	63·4	77·3	142·6	262·1

δ_1' is the difference which can be detected with 90 per cent power and overall Type I error of 5 per cent when $N = 40$. (Interpolated from Table III)
See text (pp. 28–29) for definition of $k = (\delta_1'/0{\cdot}2)^2$.

Table V. Expected sample size for the plans of Table IV

Number of looks (L)	40	20	10	8	5	4	2	1
Maximum number of patients per treatment, (N)	361	349	333	323	317	309	285	252
Size of the treatment difference, ($\lvert\delta\rvert$)								
0·00	350	338	323	315	309	302	281	262
0·05	340	329	314	306	302	295	275	262
0·10	303	293	280	275	272	268	257	262
0·15	244	238	230	228	227	225	228	262
0·20	177	174	169	166	174	176	193	262
0·25	124	124	125	125	132	136	166	262
0·30	89	90	93	93	104	110	150	262
0·35	67	69	74	76	87	94	144	262
0·40	54	56	61	64	76	85	143	262

The minimum value in each row is underlined

treatment difference is large. At intermediate values of δ near the point at which the power was fixed at 90 per cent ($\delta_1 = 0{\cdot}2$) then eight looks yield the minimum expected sample size. It is worth bearing in mind that this value of δ would in practice be chosen as the minimum treatment effect of sufficient importance to warrant the experimental effort necessary to achieve a 90 per cent chance of detecting it.

We now have a basis for choosing the number of looks in a clinical trial. It is clear that clinical trials with different design characteristics could have been chosen. Indeed, as far as varying the value of δ_1 for a particular power is concerned, comparison of the expected sample size between

columns of Table V would be a matter largely of scale. In other words as the value of k in Table IV increases by 40 per cent from 1 look to 40, so the expected sample sizes in Table V will maintain a constant relationship as the number of looks changes at all values of δ_1.

RESULTS

The process of designing a clinical trial should then have two separate components. The first part consists of choosing a value of δ_1 for which a stated power is appropriate. This leads to a maximum sample size N which, as we have seen, is somewhat dependent on the proposed number of looks. The choice of number of looks depends on the uncertainty associated with the particular estimate of δ_1 and it is this second component which we will investigate now. It would seem appropriate to compare different frequencies of interim analysis by looking at the expected sample size and minimizing some appropriate function of it.

Before proceeding let us consider in more detail the choice of the critical value δ_1 itself. In principle, of course, an experimenter will weigh very carefully the losses associated with failure to demonstrate a treatment effect which is real. To do this many independent considerations must be taken into account. The treatment difference which represents a 'real' effect will be based both on clinical judgement and epidemiological assessment. At worst it may be the largest difference considered plausible but it is hoped that it will tend to be the smallest difference with clinical and public health consequences. This assessment should weigh known side effects and different costs of the treatments against possible therapeutic benefit. Moreover the losses associated with reporting a false negative result will involve estimates of the impact of this trial on clinical practice. How many subsequent patients will have their treatment decided by knowledge gained from this trial? Formal schemes for incorporating this kind of knowledge have been developed[13] but their effect on the sensible choice of δ_1 and power are qualitatively clear. A large expected impact for a negative result should mean the choice of a high power so that the possibility of a false negative against a reasonable treatment effect is small. Similarly if the disease treated is common and the extra expense and risk associated with the new treatment small then a modest treatment effect is more worthy of detection than otherwise.

Once such considerations have led to a choice of δ_1 and an associated power together with a maximum sample size we can then incorporate the uncertainty associated with this estimate of δ_1. Again qualitative distinctions can be made. Evidence about plausible treatment effects comes generally from several sources; animal experiments, clinical experience, basic biology and perhaps other clinical trials. Each of these is more or less reliable in extrapolating to the population of patients with the disease in question, and indeed the utility of each will vary markedly between diseases and treatments. Generally, however, previous randomized comparisons will provide a more secure basis for a sensible choice of δ_1 than will limited clinical experience of selected patients. In subsequent discussion we will consider four prior distributions describing approximately the knowledge about possible treatment effects at the design stage of a trial. Each is consistent with the choice of δ_1 used above, namely 0·2.

The first distribution is a uniform prior distribution between $-0{\cdot}4$ and $+0{\cdot}4$. It corresponds to almost no prior information about the possible treatment effect; for example, if the biological mechanisms are poorly understood. Large effects of the experimental treatment are dismissed but moderate effects, beneficial or not, are deemed to be equally likely. This is an extreme of uncertainty that might arise for instance for a drug which has a particular effect in a remote biological analogue of the disease but appears to perform as well as conventional therapy in the few patients, in which it has been tested for toxicity.

The second is a Gaussian distribution with a standard deviation of 0·2; this corresponds to more

prior knowledge than the first in the sense that large treatment effects are thought to be progressively less likely. The most plausible effect is thought on the basis of contemporary knowledge to be zero but the level of uncertainty is such that treatment effects twice the size of the critical value of δ remain plausible. This might represent a fairly typical degree of uncertainty where there has been little or no previous randomized comparison of the new treatments.

The third is a similar but narrower prior distribution which might arise for instance when there is a compelling biological argument for the new treatment but previous small trials have failed to demonstrate an effect. The standard deviation here is 0·1 implying that an actual treatment effect of 0·2 is considered to be unlikely but certainly worthy of detection.

The fourth prior distribution is the same as the third but centred around 0·1. In this case the most likely treatment effect is half as large as the critical value of δ_1 and on the basis of contemporary knowledge a positive rather than negative effect of the new treatment is expected. Such a characterization might be appropriate when a few trials had indicated a small treatment effect but had not achieved significance.

The four distributions are shown in Figure 1 and are meant only to illustrate the amounts of prior information which might arise in practice. However, it is clear that while each may be consistent with the same choice of δ_1 at which the power is fixed the requirements for interim analyses during a trial will be very different. This is precisely because of variation in the uncertainty of the estimated critical treatment difference.

As an example of these qualitative distinctions suppose that we wanted to test the treatment for depression mentioned before. We might be interested, in the first place, in detecting an effect which

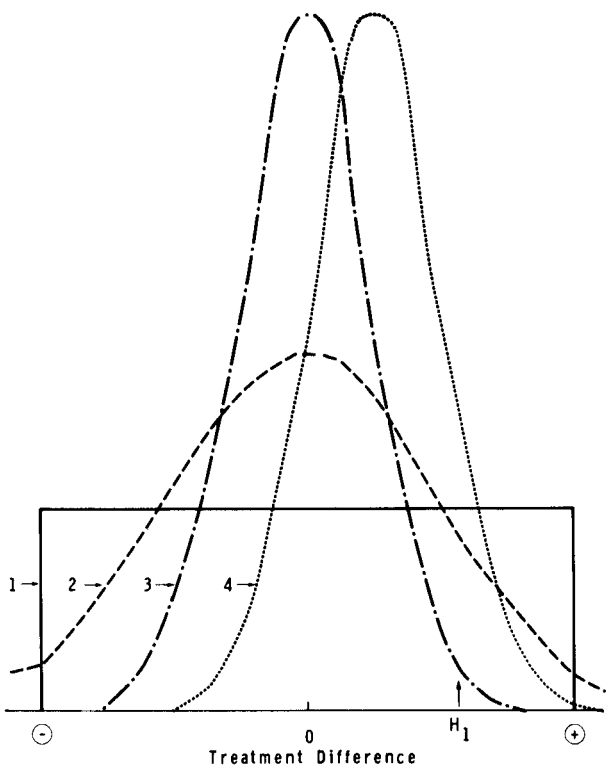

Figure 1. Four prior distributions describing prior knowledge of treatment effects

is quite small relative to the intrinsic variability of the response measurement. The mean reduction in Hamilton score on conventional therapy is 30 units and therefore a 10 per cent improvement attributable to the new treatment might be worthy of detection. That means that δ_1 would be 3 units which as a proportion of σ, the standard deviation of the therapeutic difference which equals 14·14, is roughly one fifth or 0·2. If we choose a power of 90 per cent for detecting that difference then the calculations herein would apply.

In Figure 1 therefore H_1, the magnitude of the therapeutic effect deemed to be sufficiently important to detect with a power of 90 per cent, would be set at 3 units or 0·21. Clearly we do not know the magnitude of the true therapeutic effect and the four piror distributions are intended to represent qualitatively distinct kinds of uncertainty. The first is when very little relevant information bears on the estimate. A new drug which might work but for which there is little or no clinical experience on the treatment of depression. The second prior distribution is suggestive of more concrete evidence which suggests an effect larger than 0·5 or 7 units is unlikely and that small effects are much more likely. The third is obviously more precise suggesting that an effect greater than a 10 per cent improvement is unlikely and would probably be based on previous randomized comparisons. The fourth prior distribution suggests the same precision of the estimate of size of the effect but that the balance of evidence is in favour of a positive treatment effect, although a zero effect remains plausible.

To choose the frequency of interim analysis in these four situations we will use two indices for comparative purposes. The first is simply the expected sample size in a clinical trial for a given number of looks integrated over the prior distribution; in other words the average sample size. This index expresses one attribute of group sequential testing whereby one is interested in tests with given design characteristics which involve the smallest amount of experimental effort. Another attribute concerns the desire to minimize the number of patients subjected to inferior treatment. This index is the expected sample size multiplied by the treatment difference integrated over the prior distribution. It therefore expresses a penalty at a given sample size proportional to the actual treatment difference. The first index can be regarded as an expression of a practical desire not to experiment more than is necessary and the second of an ethical desire not to treat patients with an inferior treatment for longer than is required to reach a secure conclusion.

Expressed algebraically these two indices are:

$$(1.)\quad \int_{-\infty}^{\infty} \mathrm{ASN}_x \mathrm{Pr}(x)\mathrm{d}x$$

$$(2.)\quad \int_{-\infty}^{\infty} |x| \mathrm{ASN}_x \mathrm{Pr}(x)\mathrm{d}x$$

where x is the treatment difference, ASN_x is the expected sample size for a given number of looks at this value of x and $\mathrm{Pr}(x)$ is the prior distribution of x.

The results for the first index and the clinical trial designs given in Table 5 are shown in Figure 2. Clearly the expected sample size for one look for all four prior distributions is 262 pairs of observations or 262 patients on each treatment arm. One ordinate in Figure 2 is the percentage reduction in expected sample size for each prior distribution according to the number of looks allowed in the design. From these results it can be deduced that for a disperse prior distribution (1) between 4 and 10 looks offers considerable advantage over a fixed sample size trial in terms of this index. Interestingly by extrapolation of the curve a fully sequential design, examining the accumulated data at each observation, would have a higher expected sample size than 8 looks at constant probabilities of error. The comparison with a fixed sample size trial is to be expected because large differences are deemed to be plausible and in a state of relative ignorance more looks would be expected to offer such benefits.

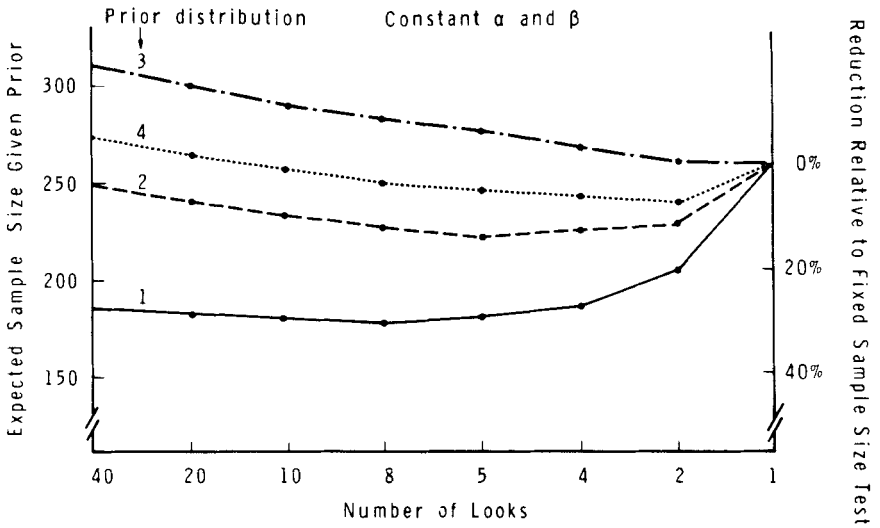

Figure 2. Index (1) for the four prior distributions (for definition of index (1) see text)

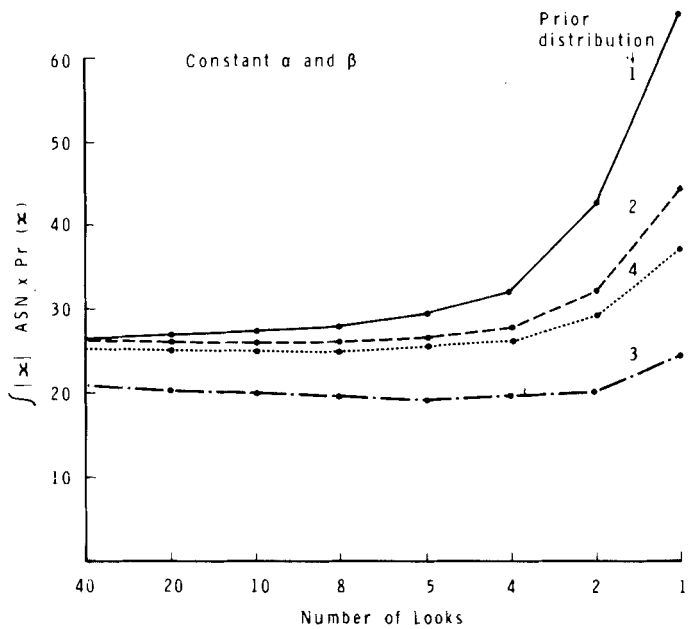

Figure 3. Index (2) for the four distributions (for definition of index (2) see text)

For the second prior distribution the same general characteristics hold but are less marked. In this case the optimum number of looks is 5 if the minimization of expected sample size is the criterion chosen. For a well defined prior distribution centred around the null point (3) a fixed sample size is optimum and interim analyses merely increase the expected sample size. For the fourth where a treatment benefit is securely expected, two or four looks during the course of the trial are optimum.

The results for the second index are plotted in Figure 3. As might be anticipated, the expected

losses attributable to a fixed sample size experiment are greatest for a dispersed prior distribution. In such a situation broadly the more looks the better because large treatment effects are weighted heavily and are plausible. As the prior distribution concentrates around zero so the benefits associated with interim analyses relative to a fixed sample size trial become less marked until for the third prior distribution 4 looks is only a little better on this index than one. For the fourth prior around 8 looks are optimum in this sense again because treatment effects larger than the critical δ are plausible.

CONCLUSIONS

The choice of sample size and number of interim analyses is in principle a complicated balance of competing objectives. The most difficult problem is one of sample size estimation and even then the decision is often a pragmatic one dictated by available patients, money or time. However, within these general constraints the choice of the number of interim analyses is often fairly open. Adopting a classic hypothesis testing framework and considering equally spaced interim analyses using repeated significance tests we have examined the effect that prior knowledge has on the optimum choice of the number of interim analyses. To make proper comparisons the probability of Type I error and the power were set constant as being, at least nominally, the most important design characteristics associated with a particular alternative hypothesis. In the example chosen here around 5–10 interim analyses were optimum for a disperse prior distribution and 2–4 for a well defined prior distribution. For the former, an average reduction in sample size of around 30 per cent could be expected compared to a fixed sample size trial. For the latter, little or no reduction in sample size can be expected but a small reduction in expected losses associated with giving patients an inferior treatment particularly when a positive treatment effect is plausibly expected.

Whether these conclusions are true for all common combinations of α, β and δ is a matter of legitimate consideration. At 90 per cent power for any alternative, as has already been noted, the above comparison is a matter largely of scale. Therefore one would expect the same conclusions in terms of optimum number of looks. For other values of power the increase in maximum sample size associated with more interim analyses is proportionately greater as power decreases and smaller as it increases from 90 per cent. Therefore for lesser power fewer looks will be generally favoured and for greater power, more looks than suggested above.

For whatever size of trial and from a pragmatic standpoint it is clear from Figure 2 that two looks do no harm but more looks are particularly indicated when one's prior knowledge is imprecise. When large effects are known to be extremely unlikely then more looks at constant design characteristics actually cause an increase in the expected sample size and therefore should be justified on ethical grounds. However, Figure 3 would indicate that, at least as measured by the second index, such justification is less forthcoming than might have been expected. On the other hand for a disperse prior distribution the ethical justification for 4 or more looks, relative to a fixed sample size trial is more persuasive. Moreover Table II tells us that the variance of the sample size increases dramatically as the number of looks increases which, in terms of the practical organisation of a multi-centre trial, represents a considerable inconvenience.

From Table II we can see that the distribution of the actual sample size for many looks in repeated trials is rather skew because the standard deviation of the sample size is of the same magnitude as the mean. This implies that in practice the average sample size is an imprecise estimate of the actual sample size and in particular that sample sizes much larger than average will not be uncommon. In large multi-centre trials it is therefore a serious organizational difficulty to obtain the active participation of many centres when the duration of the trial is so unpredictable.

Therefore the only situation where interim analyses of the data in a large clinical trial are

contraindicated (prior distribution 3) from a practical or ethical standpoint is when the critical treatment difference, (at which a high power specifies the maximum sample size) is as large as can be plausibly expected. When the balance of prior evidence suggests that either a beneficial effect is likely or that the dispersion of opinion is wide, relative to the critical treatment effect, then interim analyses offer advantages both in terms of expected sample size and expected losses associated with giving patients an inferior treatment. In practice it is suggested that once a critical treatment effect is decided and the maximum sample size calculated then these considerations can be used in deciding the number of interim analyses for a wide variety of clinical trials.

The difficulty associated with determining the nature of a prior distribution expressing uncertainty in a predicted treatment effect has not, it is hoped, been underestimated. One has to guard against the possibility of understating the dispersion of an appropriate prior distribution. One way of doing this is to ask experts to express their own betting odds on two competing hypotheses. Typically these could be no treatment effect against an effect of a particular size. Given these odds the relative plausibility, averaged over experts, could be estimated and would yield approximately the relative heights of the prior distribution at these two points. An instructive account of this process is given by Lindley.[14]

ACKNOWLEDGEMENTS

I am grateful for the benefit of many discussions with colleagues on the substance of this paper. In particular Professor Peter Armitage suggested the use of the second index with which to compare stopping rules.

REFERENCES

1. Armitage, P. *Sequential Medical Trials*, 2nd edn., Blackwell Scientific Publications, 1975.
2. Pocock, S. J. 'Group sequential methods in the design and analysis of clinical trials', *Biometrics*, **64**, 191–199 (1977).
3. Cutler, S. J. *et al.* 'The rule of hypothesis testing in clinical trials', *Biometrics Seminar, J. Chron. Dis.*, **19**, 857–882 (1966).
4. Feller, W. K. 'Statistical aspects of extra-sensory perception', *Journal Parapsychol.*, **4**, 271–298 (1940).
5. Armitage, P., McPherson, K. and Rowe, B. C. 'Repeated significance tests on accumulating data', *Journal Royal Statistics Society (A)*, **132**, 235–244 (1969).
6. McPherson, K. and Armitage, P. 'Repeated significance tests on accumulating data when the null hypothesis is not true', *Journal Royal Statistics Society, (A)*, **134**, 15–25 (1971).
7. McPherson, K. 'Statistics: the problems of examining accumulating data more than once', *New England Journal of Medicine*, **290**, 501–502 (1974).
8. Anscombe, F. J. 'Sequential medical trials', *Journal American Statistics Association*, **58**, 365–383 (1963). *Association*, **20**, 18–23 (1966).
9. Cornfield, J. 'Sequential trials–Sequential analysis and the likelihood principle', *Journal American Statistics Association*, **20**, 18–23 (1966).
10. Canner, P. L. 'Monitoring treatment differences in long-term clinical trials', *Biometrics*, **33**, 603–615 (1977).
11. Peto, R. *et al.* 'Design and analysis of randomised clinical trials requiring prolonged observation on each patient. I Introduction and design', *Br. J. Cancer*, **34**, 585–612 (1976).
12. McPherson, K. 'Some problems in sequential experimentation', *Unpublished Ph.D. Thesis*, University of London, 1971.
13. Colton, T., 'A model for selecting one of two medical treatments', *Journal American Statistics Association*, **58**, 388–400 (1963).
14. Lindley, D. V. *Making Decisions*, Wiley-Interscience, 1975.