STATISTICS IN MEDICINE
Statist. Med. 2001; **20**:165–176

Mid-trial design reviews for sequential clinical trials

John Whitehead[1,*], Anne Whitehead[1], Susan Todd[1], Kim Bolland[1] and M. Roshini Sooriyarachchi[2]

[1] *Medical and Pharmaceutical Statistics Research Unit, The University of Reading, P.O. Box 240, Earley Gate, Reading RG6 6FN, U.K.*
[2] *Department of Statistics and Computer Science, University of Colombo, Sri Lanka*

SUMMARY

When sequential clinical trials are conducted by plotting a statistic measuring treatment difference against another measuring information, power is guaranteed regardless of nuisance parameters. However, values need to be assigned to nuisance parameters in order to gain an impression of the sample size distribution. Each interim analysis provides an opportunity to re-evaluate the relationship between sample size and information. In this paper we discuss such mid-trial design reviews. In the special cases of trials with a relatively short recruitment phase followed by a longer period of follow-up, and of normally distributed responses, mid-trial design reviews are particularly important. Examples are given of the various situations considered, and extensive simulations are reported demonstrating the validity of the review procedure in the case of normally distributed responses. Copyright © 2001 John Wiley & Sons, Ltd.

1. INTRODUCTION

Clinical trials are usually designed to detect some clinically relevant treatment difference with a specified power. Deducing the fixed sample size or the maximum and expected sample sizes of a sequential procedure from this power requirement often requires knowledge of certain nuisance parameters, such as the within-group variance of normally distributed responses or the success probabilities of binary responses. Values for these nuisance parameters are determined using expert opinion and examination of any available previous data and study reports. Inaccuracy in the selected values can lead to either loss of power or sample sizes which are in excess of those required.

Sample size reviews, also known as internal pilot studies, have been discussed in the context of fixed sample designs by Wittes and Brittain [1], Gould [2, 3] and Gould and Shih [4]. Bolland *et al.* [5] present an example of implementation in a trial in severe head injury yielding ordered categorical responses. Such reviews consist of using the data available part way through a trial to re-estimate the nuisance parameters used in the original sample size determination. This can be done from data pooled over treatments, to avoid any breaking of blindness. The sample size is then recalculated and substituted for the original value. Simulations reported in the papers cited

* Correspondence to: John Whitehead, Medical and Pharmaceutical Statistics Research Unit, The University of Reading, P.O. Box 240, Earley Gate, Reading RG6 6FN, U.K.

Received November 1998
Accepted February 2000

above show that there is negligible effect on the type I error rate, but excellent preservation of power.

The use of similar mid-trial design reviews in the context of sequential studies has been advocated by Gould and Shih [6]. In this paper we present our own approach, which has already been implemented in completed sequential trials and designs for ongoing and future studies, and discuss the similarities to and differences from Gould and Shih's method. Some complicated cases, not discussed explicitly by Gould and Shih, are considered. Here a *mid-trial design review* of a sequential clinical trial is an examination of data collected so far, principally for reassessing the distribution of sample size and study duration required to fulfil the original power specification. Such a review may lead to a 'design modification', meaning that the original power specification is altered. Often this is because it has become apparent that without modification the required expected sample size or study duration will not be feasible.

Mid-trial design reviews have a special role to play in two commonly occurring types of sequential design. In the first, patient responses occur a considerable time after randomization, and most interim analyses occur after the closure of recruitment with potential for reducing trial duration but not sample size. The review then concerns the timing of recruitment closure. The second case is trials yielding normally distributed responses in which the trial design is generated from a specification of standardized mean difference, although true interest lies in the unstandardized value. In this context, the review is conducted in order to preserve power, and our views are similar to those of Gould and Shih [6].

Section 2 concerns the relevance of mid-trial design reviews to standard information-based sequential trials: in most cases sample size predictions will be revised, but not sufficiently to warrant a design modification. In Section 3, the special case of long-term follow-up trials with rapid recruitment to a fixed number of patients is considered, and Section 4 is an account of the particular problems which occur when responses are normally distributed. The cases of two clinical trials in which these problems arose unexpectedly, and which motivated this paper, are described. A discussion of the issues raised forms Section 5.

2. SAMPLE SIZE PREDICTIONS FOR SEQUENTIAL CLINICAL TRIALS

Consider a comparative clinical trial of two treatments, an experimental (E) and a control (C). Let θ denote the true benefit of E relative to C, in terms of some primary response variable. The power to detect a significant difference at level α (two-sided) will be fixed at $(1-\beta)$ when θ takes some chosen value θ_R which represents a clinically relevant effect, known as the *reference improvement* [7]. Many sequential designs can be expressed as a plot of a statistic representing the observed advantage of E over C against a measure of information. This approach can be traced from the work of Bartlett in 1946 [8], through Cox [9] and Jones and Whitehead [10] amongst others. Whitehead [7] plots a score statistic Z, against the observed form of Fisher's information V, and uses the distributional result that Z is normally distributed with mean θV and variance V to deduce appropriate designs. Scharfstein *et al.* [11] and Jennison and Turnbull [12] follow the maximum likelihood estimate of θ relative to expected information. The spirit of these two approaches is the same, despite differences in detail. In this paper the notation of Z and V is adopted. The computer program PEST3 [13] is used in all of the examples presented.

Having chosen values for α, $(1-\beta)$ and θ_R, and fixed the shape of the design in terms of a class of boundaries (triangular, truncated sequential probability ratio test and so on, see reference [7])

or in terms of an α-spending function (perhaps recreating Pocock's or O'Brien and Fleming's procedure, see Lan and DeMets [14]), then the maximum value of V can be deduced together with the distribution of the terminal value V_T of V under any value of θ. The latter can be summarized in terms of the expected value of V_T and its percentiles. Apart from small effects due to the spacing of interim analyses, the properties of V_T are fixed by the choice of design.

The relationship between sample size n and information V can be found for each response type. For binary data, the reference improvement θ_R can be defined as a log-odds ratio. Then in large samples

$$V \approx \frac{R}{(R+1)^2}\,\bar{p}(1-\bar{p})n \tag{1}$$

where randomization to E and C is an $R:1$ ratio, n is the total sample size (both groups combined) and $\bar{p}$ is the overall proportion of successes in the trial as a whole. For ordinal data under the assumption of proportional odds, θ_R will again represent a (common) log-odds ratio, and

$$V \approx \frac{R}{3(R+1)^2}\left(1-\sum_{i=1}^{k}\bar{p}_i^3\right)n \tag{2}$$

where there are k categories with overall probabilities $\bar{p}_1,\ldots,\bar{p}_k$ of occurrence [15]. Similar relationships can be deduced for other response types. At the design stage, statements about the maximum and expected values of V_T can be translated to statements concerning the terminal sample size n_T using equations such as (1) or (2). These predictions are inaccurate to the extent $\bar{p}$ or $\bar{p}_1,\ldots,\bar{p}_k$ have been misrepresented.

The regular interim analyses which are part of a sequential design can be used to re-evaluate these nuisance parameters, allowing recalculation of the distribution of the eventual sample size. This re-evaluation of how long the design will take to complete is not a design modification. Plotting directly against information guarantees power at θ_R regardless of the values of nuisance parameters, provided that the initial design is followed to completion.

The on-line re-evaluation of likely sample sizes can lead to modification of the schedule of interim analyses as specified in terms of sample size or calendar dates. New information about nuisance parameters such as $\bar{p}$ may indicate that accrual of information is slower than anticipated, so that the schedule of interim analyses should be spaced out more widely. This is an attempt to get closer to the original interim information values $V_1, V_2,\ldots$ envisaged, and so represents only a minor change in operation of the design. More seriously, it may become apparent that expected sample sizes are far larger than previously envisaged, and the practicality of completing the trial is brought into question. In such a case, the new sample size forecasts need to be considered by the Steering Committee, who may choose to devote the extra resources likely to achieve a result. Alternatively, they may make a genuine design modification and lower the power, thereby maintaining the feasibility of the study by modifying the original sequential design. Even more radically, they may choose to abandon the study and analyse the currently available data. The Steering Committee should make their choice without knowledge of Z. Ideally, to avoid bias, the conditions under which such a design modification is to be considered should be set in advance. Simple criteria should suffice, such as informing the Steering Committee if the likely sample size increases by more than 25 per cent or the likely trial duration becomes more than one year longer than envisaged at the planning stage.

3. SEQUENTIAL TRIALS WITH LONG-TERM FOLLOW-UP

In this section trials will be considered in which the time between a patient being randomized and giving a response is long and recruitment to a fixed number of patients is relatively quick. The response may be binary, ordinal or normally distributed, but it is observed after a fixed period such as a year or 18 months after treatment. Alternatively it may be the time to an event, to be evaluated using survival methodology. The typical structure of such a trial is to recruit a fixed number of patients and then to follow them up until the last has responded or until sufficient survival data have been collected. A sequential design can be imposed on such a trial, although it will have little impact on sample size; indeed the first interim analysis may not take place until recruitment has closed. However, the stopping rule does have the potential to shorten the study duration, and in the case of long-term medication, to shorten the total exposure of patients to the inferior treatment.

The number of patients to be recruited to such a study cannot be calculated from fixed sample size formulae. Instead, a sample size large enough to ensure a high probability of reaching one of the sequential boundaries should be set. This is likely to exceed the equivalent fixed sample size, being the price paid for the early stopping option. A reasonable strategy is to require a sample size equal to the maximum value of $\mathrm{P90}(n_\mathrm{T};\theta)$, where $\mathrm{P90}(n_\mathrm{T};\theta)$ denotes the 90th percentile of sample size under the parameter value θ. The function $\mathrm{P90}(n_\mathrm{T};\theta)$ will reach its maximum for some value of θ between 0 and θ_R. This rule does admit a slight loss of power due to the remaining chance of not reaching a boundary and having to conduct an analysis with underrunning [16]. In this type of sequential study a mid-trial design review becomes relevant and important. It has to be timed to take place before the closure of recruitment, so that it has the potential to enable an extension of the recruitment period.

An example of this form of review was provided by a recent trial in head injury. Full details of this trial are as yet unpublished and confidential. The primary efficacy response was the Glasgow Outcome Score, 6 months after randomization. This response was dichotomized into binary form, with good recovery or moderate disability representing success, and severe disability, vegetative state or death being failure. The design was a truncated sequential probability ratio test [7] (SPRT) in which only a slight increase over the equivalent fixed sample size was allowed. As the follow-up between randomization and observation of the primary response was 6 months, a maximum sample size was set in advance to govern closure of recruitment. A total of three interim analyses were planned, to take place when information reached $1/4\ V_{\max}$, $1/2\ V_{\max}$ and $V_{\max}$. Patients were randomized equally between the two treatments, and a power of 0.80 was set to detect significance at level 0.05 if the experimental treatment increased the success rate from 0.45 to 0.55. This led to a reference improvement on the log-odds ratio scale of $\theta_\mathrm{R} = 0.401$. A fixed sample design for this study required information V equal to 48.8, and the SPRT was truncated slightly above this at $V_{\max} = 50.0$. The value 50.0 is also the maximum 90th percentile of V over θ in this case. Taking $\bar{p}$ to be 0.5 in equation (1) led to corresponding values of $\mathrm{P90}\ (n_\mathrm{T};\theta)$ equal to or less than 800 patients. By the second interim analysis, 700 patients had been recruited, of whom 400 had been followed up for 6 months. At this stage, $\bar{p}$ was re-estimated, and re-application of equation (1) with allowance for stratification by age and Glasgow Coma Score at entry to the trial (following Section 5 of reference [14], see also the illustration in reference [5]) now translated $V = 50.0$ to a required sample size of 920. The Steering Committee was informed and accepted this new maximum sample size.

The second example is a fictitious illustration based on a trial which was actually planned, but subsequently not carried out due to reports of adverse events in other studies. The endpoint was

Table I. Survival probabilities used in the planning of the cardiovascular trial.

	t_i (months)						
	3	6	9	12	18	24	36
$S_C(t_i)$	0.98	0.97	0.95	0.94	0.91	0.89	0.82
$S_E(t_i)$	0.985	0.978	0.963	0.956	0.933	0.918	0.865

time to event, and greater detail is presented because of the complications inherent in such a study. This sequential study concerned comparison of an experimental treatment and placebo in the prevention of cardiovascular events in diabetic patients. Patients were to be followed up for up to three years, or until the study was terminated. Assessments were to be made at baseline, months 1 and 3 then every 3 months until 36 months. The primary efficacy variable was time to cardiovascular event, as detected at one of these assessments. From various publications it was estimated that the event rate within 36 months on placebo was likely to be 18 per cent. A clinically relevant reduction was judged to be an experimental event rate of 13.5 per cent. The corresponding survival probabilities at 36 months were $S_C(36) = 0.820$ and $S_E(36) = 0.865$ for placebo and experimental, respectively, proportional hazards was assumed, and the reference improvement θ_R expressed as minus the log-hazard ratio (E:C) was 0.314. In order to reduce the duration to result of the study and to achieve a homogeneous sample of patients with regard to other treatments and care received during the trial, it was optimum to have rapid recruitment to some prespecified sample size, but to conduct interim analyses during the follow-up phase. The sequential design chosen was the triangular test. Patients were to be randomized equally between the two treatments, and a power of 0.90 was set to detect as significant at the 5 per cent level (two-sided alternative) the reference improvement of 0.314.

The 90th percentile of the duration of the trial was computed for a range of values of θ between 0 and $3\theta_R/2$. In order to achieve this, further information about the estimated survivor function for placebo patients over three years was needed. Estimates of $S_C(t_i)$, $i = 1, \ldots, 7$ where the t_i denote $3, 6, 9, 12, 18, 24$ and 36 months, were taken from various publications as shown in Table I. Also shown are corresponding experimental survivor rates $S_E(t_i)$ consistent with a minus log-hazard ratio value of $\theta = \theta_R = 0.314$. It was found that a recruitment rate of 160 patients per month for 18 months would be required to ensure a probability in excess of 0.90 of the trial reaching a boundary for all values of θ. The longest trial durations occurred for $\theta = 3\theta_R/4 = 0.236$, for which $S_E(36) = 0.865$ and for which the 90th percentile of the number of events required was 460. This recruitment pattern was adopted; it required a total sample size of 2880 patients, and led to expected durations of 28 and 37 months for $\theta = 0$ and θ_R, respectively.

In order to ensure the study would reach a boundary, a sample size review was planned, as outlined by Bolland *et al.* [5]. This was to be conducted after 15 months, whilst recruitment was still open, using a Kaplan–Meier estimate based on available data. The objective was to reassess the overall survivor rates $\bar{S}(t_i)$, $i = 1, \ldots, 7$, and use these updated estimates of the nuisance parameters to recalculate the sample size. An illustration of the planned procedure follows based on fictitious values. At 15 months estimates of $\bar{S}(18)$, $\bar{S}(24)$ and $\bar{S}(36)$ could not be made directly from the data available. Values of $\bar{S}(t_i)$ were estimated at the 15 month review for $i = 1, \ldots, 4$ (that is, for times 3, 6, 9 and 12 months); these are denoted by $\bar{S}_{new}(t_i)$. They were found by applying the Kaplan–Meier method to the pooled trial data as $\bar{S}_{new}(3) = 0.982$, $\bar{S}_{new}(6) = 0.977$,

Table II. Survival probabilities used in the mid-trial review of the cardiovascular trial.

	t_i (months)						
	3	6	9	12	18	24	36
$\bar{S}_a(t_i)$	0.983	0.974	0.957	0.948	0.922	0.904	0.843
$\bar{S}_{new}(t_i)$	0.982	0.977	0.962	0.952	0.927	0.910	0.853
$S_{C,new}(t_i)$	0.979	0.973	0.956	0.945	0.916	0.897	0.832
$S_{E,new}(t_i)$	0.985	0.981	0.968	0.959	0.938	0.923	0.874

$\bar{S}_{new}(9) = 0.962$ and $\bar{S}_{new}(12) = 0.952$. These figures were then projected to estimate $\bar{S}(t_i)$, $i = 5,\ 6$ and 7 as described below.

Let ϕ denote the average difference between the anticipated $\bar{S}_a(t_i)$ and the new $\bar{S}_{new}(t_i)$ values of $\bar{S}(t_i)$ on the complementary log–log scale for $i = 1, \ldots, 4$. That is

$$\phi = \frac{1}{4}\sum_{i=1}^{4}[-\log\{-\log \bar{S}_{new}(t_i)\} + \log\{-\log \bar{S}_a(t_i)\}]$$

The values of $\bar{S}_{new}(t_i)$ which were beyond the current duration of the trial, were found from

$$-\log\{-\log\ \bar{S}_{new}(t_i)\} = -\log\{-\log \bar{S}_a(t_i)\} + \phi, \quad i = 5,\ 6,\ 7$$

The value of ϕ was 0.068 and the values of $\bar{S}_{new}(t_i)$, $i = 5, 6, 7$ are presented in Table II. To maintain blindness the survivor rates were found for control and experimental $S_{C,new}(t_i)$ and $S_{E,new}(t_i)$, respectively, which were consistent with both the reference improvement and the observed or projected overall survivor rates $\bar{S}_{new}(t_i)$ in Table II, rather than using individual treatment survival patterns. To do this the following two equations were used:

$$S_{E,new}(t_i) = 2\bar{S}_{new}(t_i) - S_{C,new}(t_i) \tag{3}$$

$$e^{\theta_R} = \frac{\log\ S_{C,new}(t_i)}{\log\{2\bar{S}_{new}(t_i) - S_{C,new}(t_i)\}} \tag{4}$$

Equation (4) was solved iteratively to give $S_{C,new}(t_i)$ and subsequently using equation (3) values of $S_{E,new}(t_i)$ were found and these are also shown in Table II.

From the data available it was also estimated that the average monthly recruitment rate was 140 patients per month as opposed to the 160 anticipated. Recruitment for 22 months was now necessary to ensure that the chance of the trial yielding 460 events was at least 0.9. Thus, the required sample size increased by 200 patients from 2880 to 3080.

4. SEQUENTIAL TRIALS WITH NORMALLY DISTRIBUTED RESPONSES

Sequential designs based on the normal distribution present a special problem not present for binary, ordinal or survival data: the presence of the additional nuisance parameter σ^2. Suppose that responses of control patients are modelled as normally distributed with mean μ_C and variance σ^2, and those of patients on the experimental as normal with mean μ_E and the same variance σ^2. An ideal approach, from a statistical point of view, is to express the advantage of the experimental

in terms of the standardized difference in means: $\theta = (\mu_E - \mu_C)/\sigma$. This has several advantages, including being a dimensionless quantity, and completely determining $P(X_E > X_C)$ as $\Phi(\theta/\sqrt{2})$, where X_C and X_E denote typical responses in the respective treatment groups, and Φ is the standard normal distribution function. The information V about θ available from a sample of n observations, n_C on control and n_E on experimental, is given by $(n_C n_E/n) - (Z^2/2n)$. In large samples $V \approx (n_C n_E/n)$, and this has no dependence on σ^2. If power is set to be $(1 - \beta)$ when $\theta = \theta_R$, then a sequential procedure plotting against V will realize this regardless of the value of σ^2. The value of V does not depend on $(\mu_C + \mu_E)/2$ either, and so the location of the two distributions is immaterial. Mid-trial design reviews are unnecessary in this case.

However, there are often good clinical or regulatory reasons to use the absolute treatment difference $\delta = \mu_E - \mu_C$, and to fix the power at $(1 - \beta)$ when δ is equal to some clinically important value δ_R. One option is to derive the efficient score and Fisher's information for this parameterization and to use them in the sequential procedure. The large sample formula for the latter is then $\{n_C n_E/(n\sigma^2)\}$. In principle, the sequential design will guarantee power, although misspecification of σ^2 might lead to unrealistic predictions of sample size. A more serious problem is that the asymptotic results underlying the sequential theory only become accurate in very large samples; much larger than for the standardized parameter θ [17].

It is more satisfactory to proceed using the standardized parameterization, which is what Gould and Shih [6] recommend. In terms of the notation of this paper, the power is set to be $(1 - \beta)$ for $\delta = \delta_R$ and a pre-trial estimate, σ_0^2, of σ^2 is obtained. The design is formulated as one with power fixed for $\theta = \theta_R = \delta_R/\sigma_0$, and its properties are explored. In the light of the expected sample sizes for this design, the timing of the first interim analysis and the spacing of subsequent looks are chosen. As part of the first interim analysis, a mid-trial design review is conducted using the data to estimate σ^2; denote the result by $\hat{\sigma}_1^2$. If $\hat{\sigma}_1^2$ differs appreciably from σ_0^2, a design modification is made, replacing θ_R by $\theta_R^{(1)} = \delta_R/\hat{\sigma}_1$, and the study is redesigned. The first interim analysis is then performed, and the future looks are rescheduled.

Within this framework, several options are available. A conventional pooled estimate might be used for $\hat{\sigma}_1^2$, necessitating the separation of the responses into two treatment groups, but not necessarily identifying which is which. Alternatively, either a simple adjustment suggested in Section 3.3.1 of Gould [3], or a more elaborate method based on an EM algorithm [3, 4, 6] might be used to compute $\hat{\sigma}_1^2$ from the responses without any treatment labels at all. Conventions might be set prior to the start of the study limiting the extent to which the design could be shrunk or expanded, and the option of abandoning the trial might be allowed if the inflation of $\hat{\sigma}_1^2$ over σ_0^2 is so large as to render the study impractical. To be conservative, $\theta_R^{(1)}$ could be taken to be $\delta_R/\sqrt{\{\hat{\sigma}_1^2 + k\,\mathrm{SE}(\hat{\sigma}_1^2)\}}$ where k is some number typically between 0 and 2.

Tables III–V present the results of simulations conducted to verify that the mid-trial review procedure has no effect on the type I error rate, while having the desired effect on preserving power. In each case the choices $\alpha = 0.05$, $(1-\beta) = 0.90$ were made and 10 000 replicate simulations of a triangular test satisfying this specification were run. It was supposed that on control, $\mu_C = 0.0$, and that the reference improvement was $\delta_R = 1.0$. Each run was conducted twice, once under the null hypothesis with $\mu_E = 0.0$ so that δ is actually equal to zero, and once under the alternative with $\mu_E = 1.0$ so that $\delta = \delta_R = 1.0$. The simulations complement those of Gould and Shih [6] as they concern a contrasting type of design with a high probability of early stopping.

Each true standard deviation σ and its pre-trial estimate σ_0 were allowed to take the values 2, 3, 4, giving nine combinations in all. From each run, the proportion of replicates showing the experimental treatment to be significantly *better* than control (which should be 0.025 when $\delta = 0$

Table III. Properties of the sequential procedure without design review: $\alpha = 0.05$, $1 - \beta = 0.90$, $\delta_R = 1$, 10 000 replicates.

σ	σ_0	Proportion showing experimental significantly better		Average sample size		95th percentile of sample size	
		$\delta = 0$	$\delta = 1$	$\delta = 0$	$\delta = 1$	$\delta = 0$	$\delta = 1$
2.0	2.0	0.029	0.902	94	110	164	196
	3.0	0.027	0.997	211	180	374	300
	4.0	0.025	1.000	376	277	672	270
3.0	2.0	0.029	0.580	94	128	164	196
	3.0	0.026	0.900	212	247	374	448
	4.0	0.026	0.988	374	350	672	538
4.0	2.0	0.028	0.373	94	127	164	196
	3.0	0.029	0.673	213	283	374	448
	4.0	0.029	0.895	377	442	672	806

and 0.90 when $\delta = 1$), and the average and 95th percentile of sample size are presented. Following a sequential trial, the Fairbanks and Madsen ordering [18] can be used to compute a p-value. There is an approximate correspondence between the sample path crossing the upper boundary, and detecting advantage of E over C with $p < 0.05$. The approximation is due to the discrete nature of the interim analyses, and becomes less accurate as these become less frequent. The worst case occurs at the first look, especially as the misjudgement of nuisance parameters may lead to V_1 being close to or in excess of the redesigned V_{max}. Consequently, in these simulations significance was taken to be equivalent to crossing the upper boundary, unless stopping occurred at the first look, in which case the criterion $Z_1 > 1.96\sqrt{V_1}$ was adopted.

Although in practice reaction to the results of a design review can be flexible, for simulation purposes the scheme has to be specified in detail. The following procedure was adopted. For $\sigma_0 = 2, \theta_R = 0.5$ initially. The fixed sample size (n_{fix}) is 168.1. This was rounded to the nearest multiple of 5, that is 170. Interims were then planned at $2n_{fix}/5$, $3n_{fix}/5$, $4n_{fix}/5, \ldots$; that is $68, 102, 136, \ldots$ patients. The design review was performed at 68 patients, and the revised fixed sample size (n_{rev}) computed. The first comparative interim was done at 68 patients, but subsequently the schedule $2n_{rev}/5$, $3n_{rev}/5$, $4n_{rev}/5, \ldots$ was used, with the proviso that *at least* $n_{rev}/5$ *new* patients were available for the second interim. For example, if the new schedule was 100, 150, $200, \ldots$, and the first interim had already taken place at 68 patients, then the second would be at 150, skipping 100 as it is less than $68 + 50$. For $\sigma_0 = 3$, $\theta_R = 0.333$ and $n_{fix} = 378.2$, rounded to 380. A similar scheme was followed, with the review taking place at 152 patients. For $\sigma_0 = 4$, $\theta_R = 0.25$ and $n_{fix} = 672.51$, rounded to 675. The review took place at 270 patients.

Tables III, IV and V were created with, respectively: no design review; a design review using a conventional pooled estimate of σ, and a design review using an estimate of σ calculated using the EM algorithm. Table III shows adherence to the power specification when σ is anticipated correctly, but a substantial loss of power when it is underestimated. The results shown in Tables IV and V are similar to one another although sample sizes and powers are slightly smaller when the EM algorithm is used in Table V. The design review, when conducted prior to the first interim analysis, does not appear to affect type I error materially: a 95 per cent probability interval for estimating an error rate of 0.025 based on 10 000 replicates is (0.022, 0.028). Values in excess of 0.028 appear

Table IV. Properties of the sequential procedure with design review: $\alpha = 0.05$, $1 - \beta = 0.90, \delta_R = 1$, 10 000 replicates. Variance estimated using the pooled estimate (partially unblinded).

σ	σ_0	Proportion showing experimental significantly better		Average sample size		95th percentile of sample size	
		$\delta = 0$	$\delta = 1$	$\delta = 0$	$\delta = 1$	$\delta = 0$	$\delta = 1$
2.0	2.0	0.032	0.894	96	111	176	196
	3.0	0.022	0.897	153	154	152	152
	4.0	0.025	0.984	270	270	270	270
3.0	2.0	0.024	0.897	211	251	388	444
	3.0	0.027	0.891	213	249	380	440
	4.0	0.028	0.900	279	292	350	426
4.0	2.0	0.027	0.889	377	443	698	778
	3.0	0.026	0.893	376	444	674	768
	4.0	0.026	0.895	378	443	678	766

Table V. Properties of the sequential procedure with design review: $\alpha = 0.05$, $1 - \beta = 0.90, \delta_R = 1$, 10 000 replicates. Variance estimated using the EM algorithm (totally blinded).

σ	σ_0	Proportion showing experimental significantly better		Average sample size		95th percentile of sample size	
		$\delta = 0$	$\delta = 1$	$\delta = 0$	$\delta = 1$	$\delta = 0$	$\delta = 1$
2.0	2.0	0.027	0.887	92	110	164	194
	3.0	0.028	0.895	153	154	152	152
	4.0	0.024	0.984	270	270	270	270
3.0	2.0	0.032	0.878	204	247	372	428
	3.0	0.029	0.886	207	244	368	424
	4.0	0.027	0.885	278	290	344	418
4.0	2.0	0.026	0.879	361	433	662	752
	3.0	0.025	0.891	361	437	656	752
	4.0	0.028	0.890	368	433	660	750

more often in Table IV than in Table V, although they appear even without a design review in Table III, indicating that the underlying sequential test is itself not perfect in terms of error rates. In both Tables IV and V, the desired effect on power is observed to a high degree of accuracy, although this is as expected at the expense of larger sample sizes.

When σ is actually 2, but is overestimated as 3 or 4 in constructing the design, the initial interim sample size of 152 or 270 is already much larger than necessary. For $\sigma_0 = 4$, all trials stop at the first interim, so that both mean and 95th percentile of sample size are 270. For $\sigma_0 = 3$, fewer than 5 per cent continue beyond the first interim, so that the mean sample size actually exceeds the 95th percentile. This effect is not noticeable when $\sigma = 3$ or $\sigma = 4$. It can be avoided altogether by performing a sample size review early, regardless of the initial estimate σ_0. Further simulations (not shown here) of the policy of performing a design review after 50 patients, regardless of σ_0,

have been conducted. This policy produces type I error rates similar to those reported in Tables IV and V, and reduces sample sizes even when σ is equal to 2. Initial sample sizes less than 50 are likely to lead to unreliable estimation of σ, especially when the estimation is based on the EM algorithm.

Design reviews involving normally distributed data have been undertaken in two clinical trials with which we have been involved. In each case the review was undertaken during one of the interim analyses, although this had not been planned in advance. Indeed, it was these studies which prompted reconsideration of our practices and led to this paper. In the first study, a review at the time of the fourth interim analysis revealed an estimated standard deviation approximately 60 per cent larger than predicted. The estimate, $\hat{\sigma}^2$, was calculated in the usual way from pooled within-group estimates of variance by the independent statistician, that is the data were unblinded. However, the decision of the company to modify the design was taken in the absence of any information regarding the treatment difference, apart from the knowledge that a stopping boundary had not yet been crossed. The estimate then replaced σ_0^2 to give a new value for θ_R and a new design, which was used for the remainder of the analyses. A protocol amendment was prepared. Note that, this being the study that alerted us to the problem, the review in question was at the fourth interim rather than at the first as in the simulations above. The issue of deliberately performing design reviews at later interim analyses is discussed in the next section.

In the second study, an unplanned review at the first interim analysis revealed an estimated standard deviation of approximately twice its predicted value. As in the first example, the variance estimate was calculated in the usual way from pooled within-group estimates of variance. More up-to-date external information, acquired at this time, regarding the variability of patient responses in this type of trial, indicated that the predicted standard deviation was too low. The decision taken by the company was to make no change to the design, maintaining power for the original *standardized* difference.

5. DISCUSSION

The purest approach to the conduct of a clinical trial consists of setting a sample size, and not looking at the data at all until that number of responses is available. Any form of 'peeking' at the accumulating data raises suspicions of operational bias and inappropriate final analyses. The methodology of sequential analysis has been created to allow formal repeated interim treatment comparisons to be made in a way which avoids bias and allows valid inferences to be drawn when the trial stops. Sample size reviews in fixed sample studies play a quite different role, guaranteeing power by re-evaluating sample size from a blinded examination of early responses. These are being implemented in clinical trials, and if conducted properly appear to gain the approval of regulatory statisticians [19]. When sequential methods and design reviews are combined in a single trial, a wide range of options become available, and the choice of strategy must be made in the context of each individual study.

Gould and Shih [6] recommend that design reviews of sequential trials, like sample size reviews in the fixed sample case, should be conducted blind to treatment identity. However, the situation in a sequential trial is different. There will already be a statistician or statistical group that is trusted with unblinded information in order to conduct the interim treatment comparisons. They will at least be able to identify the two treatment groups, whether or not they know which is experimental and which is control. In certain cases, such as the long-term follow-up trials of

Section 3 or trials with normally distributed responses as in Section 4, it may be prudent to plan a blinded design review prior to the first interim analyses, as recommended by Gould and Shih. However, at subsequent interims, even when the blind has been broken, the statisticians preparing interim analyses will be in a position to report on progress towards achieving the desired power. Mechanisms should be available for them to report any forecasts of serious increases in likely sample size or of loss of power due to issues such as those raised in Sections 3 and 4 above. A balance has to be struck between the avoidance of bias and the maintenance of the trial's ability to achieve its set objectives.

A further potential use of interim data is the checking of modelling assumptions such as proportional hazards or proportional odds, or lack of treatment by prognostic factor interactions. Concern for patient safety and for wise use of resources does suggest that these issues also be checked at interim analyses, although such methodology is outside the scope of this paper.

The long-term follow-up design described in Section 3 can be used both with and without sequential monitoring. Fast recruitment to a number of patients will reduce the duration of a study where the follow-up is long term. In chronic conditions such as diabetes, asthma and angina, this can be a realistic option, as effectively there will be a 'queue' of potential patients. Recruitment over a short time period is also a way of achieving a homogeneous sample of patients as supportive care or concomitant treatments change over time. In any such study, review of the sample size, prior to closure of recruitment would appear to be sensible, and it would be wise to delay closure of recruitment until a sample sufficient for the purpose in both size and duration of follow-up is available.

ACKNOWLEDGEMENTS

The MPS Research Unit is grateful for financial support from Amgen Ltd., Glaxo Wellcome plc, Hoechst Marion Roussel Ltd., Pfizer Central Research, Roche Products Ltd., and AstraZeneca Pharmaceuticals.

REFERENCES

1. Wittes J, Brittain E. The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine* 1990; **9**:65–72.
2. Gould AL. Interim analyses for monitoring clinical trials that do not materially affect the type I error rate. *Statistics in Medicine* 1992; **11**:55–66.
3. Gould AL. Planning and revising the sample size for a trial. *Statistics in Medicine* 1995; **14**:1039–1051.
4. Gould AL, Shih WJ. Sample size re-estimation without unblinding for normally distributed data with unknown variance. *Communications in Statistics–Theory and Methods* 1992; **21**:2833–2853.
5. Bolland KM, Sooriyarachchi MR, Whitehead J. Sample size review in a head injury trial with ordered categorical responses. *Statistics in Medicine* 1998; **17**:2835–2847.
6. Gould AL, Shih WJ. Modifying the design of ongoing trials without unblinding. *Statistics in Medicine* 1998; **17**: 89–100.
7. Whitehead J. *The Design and Analysis of Sequential Clinical Trials*. Revised 2nd edn. Wiley: Chichester, 1997.
8. Bartlett MS. The large sample theory of sequential tests. *Proceedings of the Cambridge Philosophical Society* 1946; **42**:239–244.
9. Cox, DR. Large sample sequential tests for composite hypotheses. *Sankhyā* 1963; **25**:5–12.
10. Jones DR, Whitehead J. Sequential forms of the log rank and modified Wilcoxon tests for censored data. *Biometrika* 1979; **66**:105–113 (correction *Biometrika* 1981; **68**:576).
11. Scharfstein DO, Tsiatis AA, Robins JM. Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. *Journal of the American Statistical Association* 1997; **92**:1342–1350.
12. Jennison C, Turnbull BW. Group-sequential analysis incorporating covariate information. *Journal of the American Statistical Association* 1997; **92**:1330–1341.
13. Brunier H, Whitehead J. *PEST3.0 Operating Manual*. Reading University: Reading, 1993.
14. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**:659–663.

15. Whitehead J. Sample size calculations for ordered categorical data. *Statistics in Medicine* 1993; **12**:2257–2271.
16. Whitehead J. Overrunning and underrunning in sequential clinical trials. *Controlled Clinical Trials* 1992; **13**:106–121.
17. Facey KM. A sequential procedure for a phase II efficacy trial in hypercholesterolemia. *Controlled Clinical Trials* 1992; **13**:122–133.
18. Fairbanks K, Madsen R, P values for tests using a repeated significance test design. *Biometrika* 1982; **69**:69–74.
19. Sankoh AJ. Interim analyses: an update of an FDA reviewer's experience and perspective. *Drug Information Journal* 1999; **33**:165–176.

