Drug Information Journal, Vol. 35, 1387–1400, 2001
Printed in the USA. All rights reserved.

0092-8615/2001

PREDICTING THE DURATION OF SEQUENTIAL SURVIVAL STUDIES

JOHN WHITEHEAD, PHD
Medical and Pharmaceutical Statistics Research Unit, The University of Reading, Reading, United Kingdom

Interim analyses are a common feature of clinical trial design, especially for large trials in high mortality conditions such as cancer or cardiovascular disease in which the primary endpoint is often the survival time from randomization to death. A plan for a series of interim analyses in which the criteria for stopping are specified in advance is known as a sequential design, and can be constructed to prevent patients from being randomized to an evidently inferior treatment and avoid continuation of a trial that is obviously futile.

In this paper, methods for predicting the final sample size and total duration of a sequential survival study are described, and the play-off between speed of recruitment and length of follow-up is examined. The use of interim analyses to review the event rate, recruitment period, and model assumptions is discussed and software for the implementation of the methods is described. The approach is illustrated in the context of a trial seeking to establish noninferiority.

Key Words: Interim analysis; Noninferiority; Sequential analysis; Survival analysis

INTRODUCTION

SEQUENTIAL METHODS involve the conduct of a series of interim analyses of the accumulating data of a clinical trial. Survival studies involve the long-term observation of patients until death or some other primary event. In combination, this means that each interim analysis may contain both data on new patients, and new follow-up data on old patients.

Jones and Whitehead (1), Sellke and Siegmund (2), and Tsiatis, Rosner and Trichtler (3) describe how sequential tests of survival data can be conducted comparing an experimental treatment with a control in a series of interim analyses based on either the logrank test or Cox's regression method. These methods rely for their validity on the assumption of proportional hazards, the logrank test being a special case of Cox's regression in the absence of covariates. Descriptions of trials conducted in this way include studies in cardiovascular disease (4,5), cancer (6,7), and epilepsy (8).

In this paper, the prediction of the final sample size and the consequent duration of a sequential survival study are described. These quantities, together with the number of events at the termination of the study, are random variables, and they will be characterized through their means, medians, and ninetieth percentiles. The first two quantities give an idea of the scale of trial that is involved while the last indicates the upper range of possibilities, and all of them may vary substantially according to the true treatment difference. While computation of the properties

Presented at the DIA Workshop "Statistical Methodology in Clinical R&D," April 2–4, 2001, Vienna, Austria.

Reprint address: John Whitehead, PhD, Medical and Pharmaceutical Statistics Research Unit, The University of Reading, P.O. Box 240, Earley Gate, Reading, RG6 6FN, UK. E-mail: J.R.Whitehead@reading.ac.uk.

of the final number of events is relatively accurate, sample size and trial duration depend on the patterns of survival and recruitment, which can often be foreseen only rather vaguely. The setting and reviewing of the target maximum sample size and recruitment period are discussed.

The methods of this paper are presented in the context of a single fictitious example concerning the establishment of noninferiority of an experimental treatment relative to an effective toxic standard treatment. The choice of a noninferiority example is made to encourage the use of this methodology in a wider range of applications than has been done so far, and further to show how a sequential noninferiority study can be analyzed. All of the calculations presented are available as standard features of the software package PEST 4 (9). The principles presented here are, however, relevant to all sequential survival designs, for efficacy as well as for noninferiority, and using non-PEST 4 designs as well as those included in that package. The technical results given in the appendices could be implemented with other designs.

In the next section, the principles of determining sample size and duration of recruitment and follow-up in fixed sample trials are outlined. The review of sample size is then described in the fixed sample context, as the same principles will later be suggested for reviewing sequential designs. The main part of the paper follows, presenting the illustrative trial in detail, and the paper closes with a discussion of the methodology and its limitations.

SAMPLE SIZE CALCULATION FOR A FIXED SAMPLE SURVIVAL STUDY

Computation of the sample size for a fixed sample survival study proceeds in three stages. First a power requirement is specified. Second, the number of events required to achieve the set power is calculated. Finally, the number of patients that must be recruited to yield that number of events is deduced.

The Power Requirement

The power requirement is best specified in terms of the model that will be used in the primary efficacy analysis. Here, we suppose that a single experimental treatment (E) is to be compared with a single control treatment (C), with patients being randomized between them as they are recruited. A proportional hazards model is assumed, so that the hazards of an event at time t on the experimental and control treatments, denoted by $h_E(t)$ and $h_C(t)$ respectively, satisfy the relationship

$$\theta = -\log\{h_E(t)/h_C(t)\}, \qquad \text{for all } t > 0, \quad (1)$$

where θ is the log-hazard ratio and expresses the advantage of E over C. An alternative, equivalent, and often useful characterization of θ is in terms of the survival functions $S_E(t)$ and $S_C(t)$ on E and C:

$$\theta = -\log[-\log\{S_E(t)\}] + \log[-\log\{S_C(t)\}], \quad \text{for all } t > 0. \qquad (2)$$

The efficient score statistic for testing the null hypothesis H_0: $\theta = 0$ is the logrank statistic, which takes the form of the difference between the observed number of events in the control group and the expected number, under H_0, and conditional on the risk sets associated with each event time. Formulae for this statistic and its null variance are given in Section 2.5.2 of Collett (10), where they are denoted by U_L and V_L respectively, and in Section 3.4 of Whitehead (11), where the notation Z and V is used. Here the latter notation is adopted.

In order to express the power requirement, it is necessary first to think of some specific period (0, t_0) after recruitment, lasting, for example, six months or perhaps two years. The control probability of surviving this period, $S_C(t_0)$, has to be anticipated, perhaps from previous trials. A target value for $S_E(t_0)$ then has to be set, compromising between the smallest clinically worthwhile increase and the greatest credible improvement. Applying equation (2) for $t = t_0$, yields a value for θ, which we call the reference improvement, and denote by θ_R.

Now the power requirement can be expressed. If the model expressed by (1) and (2) is true, and if $\theta = \theta_R$, then we require a probability of $(1 - \beta)$ of rejecting H_0 against the two-sided alternative at significance level α. Notice that the anticipated value of $S_C(t_0)$ is used only to help set the value of θ_R. The power of $(1 - \beta)$ is required whenever $S_E(t)$ and $S_C(t)$ satisfy equation (2) for $\theta = \theta_R$, regardless of the two absolute survival probabilities. It should also be noted that equations (1) and (2) are certainly true under the null hypothesis that the treatments have identical effects, with $\theta = 0$ on the left-hand side. Thus, the p-value of the logrank test, which is a probability statement made under H_0, is generally valid. It is the efficiency of the test, and the validity of θ as a measure of treatment advantage, that depend upon the truth of (1) and (2).

If the final analysis is to take the form of Cox's proportional hazards regression, with allowance for some prognostic factors, then the power requirement given above, and the sample size derivation that follows, lose little accuracy. In fact, when there are no prognostic factors to adjust for, the Cox analysis and the logrank test become identical. If some other form of final analysis is anticipated, then a corresponding form of sample size calculation, differing from that presented here, should be used. A more robust approach to the analysis of survival data is described by Sooriyarachchi and Whitehead (12).

Calculating the Required Number of Events

Under the assumption of proportional hazards, and if θ_R is small, the required number of events is

$$e = 4(u_{\alpha/2} + u_\beta)^2/\theta_R^2, \tag{3}$$

where u_γ satisfies $1 - \Phi(u_\gamma) = \gamma$ for all $\gamma \in (0, 1)$, and Φ denotes the distribution function of a standard normal random variable. This can easily be deduced from the results that, when θ is small, the logrank statistic Z is approximately normally distributed with mean θV and variance V, and $V \approx e/4$ (see Section 9.2.1 of Collett [10]). For $\theta_R < 1$, equation (3) is very accurate, and it usually gives acceptable results for θ_R values up to 2.

Deducing the Number of Patients

Having chosen values for α, $1 - \beta$, and θ_R, the calculation of e from equation (3) is very precise. The weakest link in the chain is that of deducing the number of patients to recruit. In fact, it is not just a matter of choosing the sample size. Recruitment to the trial will open, and proceed at a rate of a per month for R months. After closure of recruitment, all patients will be followed for a further F months, before the final analysis is conducted. The trial design involves the choice of a, R, and F, the sample size then being $n = aR$.

A simple approach is to assume that each patient is followed up for exactly F months after recruitment. In some trials this is exactly what does happen. When the event is not death, but an endpoint that is detected by careful examination of the patient, such as some forms of disease progression, it is usual to treat each patient for a set length of time and to observe and count events only during this treatment period. In situations in which the event is death, or some other unequivocal form of failure, it is more common to count all events that have occurred up to the time of analysis. Thus, the first patient is followed up for R + F months (or until the event), while the last patient is followed up for only F months. In this case, the assumption of a common follow-up time of F months for all leads to a conservative sample size, as events after F months of follow-up are not accounted for.

Under the above supposition, a proportion of $1 - S^*(F)$ of patients can be anticipated to suffer the event, where $S^*(F) = AS_E(F) + (1 - A)S_C(F)$ and patients are allocated to the experimental and control treatments in the ratio A : 1. The recruitment proportions will be known (at least in terms of what is planned), and will usually be equal. Anticipated values for $S_E(F)$ and $S_C(F)$ will be needed to deduce a usable value for $S^*(F)$, and it is difficult to know whether to make these the null values or the desired values. Given such a value, the number of events yielded by n patients will be

$$e = n/\{1 - S^*(F)\} = aR/\{1 - S^*(F)\}. \quad (4)$$

The values of a, R, and F can then be juggled to make (4) give the same value of e as required by (3). This method is described by Freedman (13) and appears in Section 9.2.3 of Machin et al. (14).

Equation (4) is unsatisfactory when patients are followed up for differing lengths of time. It is especially unsuitable for predicting the number of events likely to be available at an interim analysis, as only partial follow-up will be available on many patients at that stage. To improve on equation (4) requires knowledge or a good prediction of the form of $S^*(t)$ for $0 < t < R + F$. Often, this is produced by anticipating the form of $S_C(t)$, deriving the corresponding form of $S_E(t)$ under the proportional hazards model with $\theta = \theta_R$, and then taking an average weighted by allocation ratio. The form of $S_C(t)$ itself may be expressed as a parametric survival function, or as a step function defined at a grid of time points $t_0 = 0, t_1, \ldots, t_k$. Equipped with such an expression for $S^*(t)$, the expected number of events available at any time during the study can be derived for any given recruitment pattern expressed in a form such as a_j per month for the period (c_{j-1}, c_j) of calendar time following the opening of the study, for $j = 1, 2, \ldots, h$; $c_0 = 0$. Details are given of the cases of exponential survival times and of expression as a step function in Appendices 1 and 2, respectively. The method is related to the approach described by Schoenfeld (15), but it uses a finer grid of times than he did, making it more reasonable to neglect the allowance for end effects that he incorporated.

SAMPLE SIZE REVIEWS FOR SURVIVAL STUDIES

A sample size review is taken to be an analysis of the accumulating data for the purpose of reassessing the chosen sample size, and in which no treatment comparison is made. In the case of survival data (as well as in other contexts), this can be done without separating the data into two treatment groups. A trial with a sample size review is not considered to be a sequential design.

Suppose that the review is performed at time T since recruitment opened, where $T < R$, so as to allow an extension of recruitment if necessary. The overall survivor function $S^*(t)$ can be estimated for $t \in (0, T)$, as a parametric function or as a step function. It may be of interest to construct survivor functions $S_C(t)$ and $S_E(t)$ which are consistent with this overall function and with the reference improvement θ_R, and this can be done using equation (2). Notice that these are illustrative survival functions, and not estimates, because θ itself has not been estimated from the data and the appropriateness of the proportional hazards model has not been assessed.

With these new estimates, and the actual recruitment pattern achieved so far, the number of events likely to be yielded by the trial can be reassessed using the same formulae as used at the beginning of the trial. Actual patterns of withdrawal and loss to follow-up can also be taken into account. It will usually be necessary to project the newly estimated overall survival function into the future, and to anticipate what future recruitment patterns are likely to be in order to achieve this. The results may predict a shortfall in terms of the number of events by the end of follow-up, which may have to be remedied by an increase in the recruitment rate or an extension of the recruitment period or both.

Methodology for sample size reviews in the context of survival studies is given in greater detail in Whitehead et al. (16). Similar reviews for trials yielding binary data were described by Wittes and Brittain (17), Gould (18,19), and for the normal case by Gould and Shih (20). Because the treatment groups do not have to be separated in order to conduct the review, it is hard to see how any appreciable effect on type I error can result: indeed, such effects have been demonstrated to be negligible for the binary case in simulations reported in the first of the Gould papers above. To date, no corresponding simulations have been published in the case of survival data, but it appears most likely that similar results would be found.

SEQUENTIAL SURVIVAL STUDIES: AN ILLUSTRATION

For illustration, a fictitious survival study that appears in section 3.4.2 of the PEST 4 Manual (9) will be used. This is a comparison between two regimens of chemotherapy for a certain form of cancer. Patients in the control arm receive a standard form of treatment that is effective in prolonging survival. Nevertheless, survival remains quite poor, and the toxic side effects of the drug are unpleasant. The experimental regimen is less aggressive, and is associated with considerably less toxicity. The objective of the trial is to establish whether it is associated with a survival pattern that is no worse than that on control. In order to reduce the incidence of toxicity within the trial itself, patients are to be randomized between the new treatment and the standard in a 2 : 1 ratio.

The Design of the Study

On the standard treatment, 20% of patients are likely to survive for 24 months. In this noninferiority trial a reduction to 10% on the experimental arm is considered to be acceptable in view of the better quality of life. Applying equation (2) for $t = 24$ shows that the corresponding reference difference is $\theta_R = -0.358$. If this value is true, then the power for observing a significant treatment difference at the 0.1 level (2-sided) is to be 0.975. When the trial is complete, a 95% confidence interval for θ will be constructed, in a manner that is appropriate for the sequential nature of the trial design. If this interval lies entirely above −0.358, then the conclusion of noninferiority will be drawn. The power requirement ensures that if $\theta = -0.358$, then noninferiority will be claimed with probability 0.025 and that if $\theta = 0$, it will be claimed with probability 0.95.

Sequential designs for equivalence trials and noninferiority trials are discussed in Whitehead (21), and the reverse triangular test is introduced as a suitable design for the latter case. This design, which will be discussed in detail here, is constructed to stop quickly if it soon becomes evident that the survival disadvantage of using the experimental treatment is greater than can be tolerated, or as soon as it becomes evident that this will not be the case. The latter stopping criterion means that, once an unacceptable survival disadvantage has been ruled out, no more patients need to be randomized to the toxic standard therapy. The conclusion about noninferiority will be made in terms of a 95% confidence interval for θ, calculated in a way that respects the sequential design, and the stopping rule will be constructed to achieve the power specification mentioned above.

Interim analyses will be conducted at intervals corresponding to approximately 50 new deaths. At each of these, the logrank statistic Z and its null variance V (as defined in the section on fixed sample studies) will be calculated, and plotted on Figure 1. Also shown in Figure 1 are the stopping boundaries, set at

$$Z = -14.153 - 0.0814V$$

and

$$Z = 14.153 - 0.2440V.$$

Once Z lies below the lower boundary, or above the upper one, the trial will be stopped, although in practice it will be usual for a Data and Safety Monitoring Board to review this recommendation before the decision is finally made. Although the definitive analysis will concern the 95% confidence interval for θ, an approximate guide is as follows. If the final value of Z lies below the lower boundary, then noninferiority will not be claimed. If the final value of Z lies above the upper boundary, then noninferiority will be claimed. If the sample path crosses the initial solid portion of the upper boundary

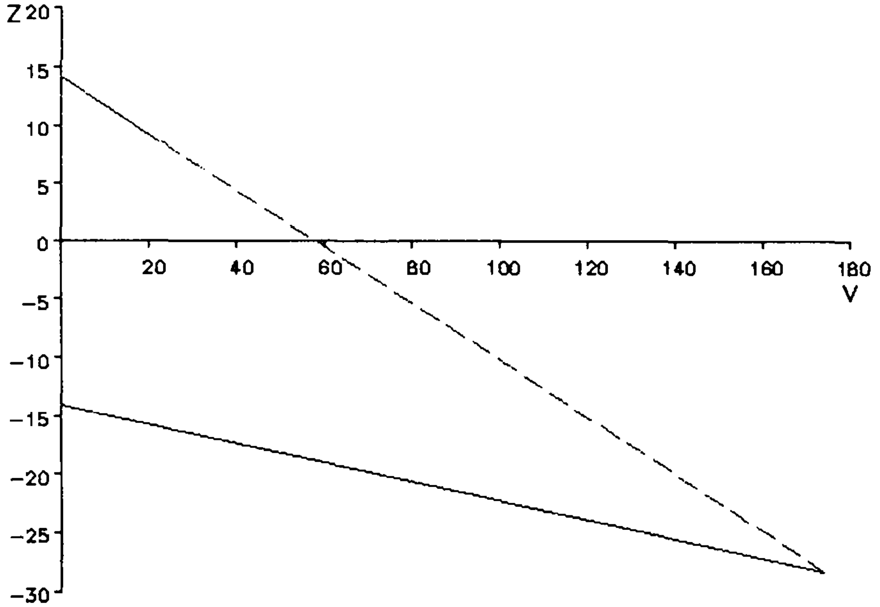


FIGURE 1. The reverse triangular test.

(for which V ≤ 22.380), then the 90% confidence interval for θ is likely to lie entirely above 0, allowing the stronger claim of superiority to be made at the 10% (2-sided) significance level. Stopping on the upper boundary even earlier could allow this claim to be made at the 5% (2-sided) level.

Table 1 shows some properties of this design for various true values of log-hazard ratio θ. Assuming that 20% of control subjects survive for 24 months, the third column shows the corresponding 24-month survival probabilities for the experimental treatment: the positive values of θ in the upper part of the table correspond to situations in which the experimental treatment is superior. Values of $c^+(\theta)$ are probabilities of finding superiority significant at the 10% (2-sided) level, while those of $c^-(\theta)$ are probabilities of finding inferiority. Taking $1 - c^-(\theta)$ gives the power for claiming noninferiority, which is almost certain for $\theta \geq 0.179$, equal to 0.95 for $\theta = 0$ and to 0.025 for $\theta = \theta_R = -0.358$ as required, and negligible when $\theta = -0.537$.

The final three columns give respectively the mean, median, and 90th percentile of e*, the number of events at termination of the study. These can be compared with 456, the number of events required for a fixed sample study with the same power requirement. The

TABLE 1
Properties of the Sequential Design

θ	$S_C(24)$	$S_E(24)$	$c^+(\theta)$	$c^-(\theta)$	E(e*)	Med(e*)	P90(e*)
0.537	0.2	0.390	0.805	0.000	94	78	114
0.358	0.2	0.325	0.509	0.000	118	100	154
0.179	0.2	0.260	0.208	0.001	163	139	232
0	0.2	0.200	0.050	0.050	258	225	407
−0.179	0.2	0.146	0.007	0.567	340	318	513
−0.358	0.2	0.100	0.000	0.975	236	203	367
−0.537	0.2	0.063	0.000	1.000	152	130	212

number required is likely to be small when the experimental treatment is associated with better or much worse survival than control, and this is a desirable feature of the design. The largest sample sizes are likely to occur when E is slightly, but not seriously, worse than C. This is the situation in which it seems ethical to randomize larger numbers of patients and it is important to estimate θ precisely. An upper bound on the number of events that will be required by the trial is approximately 783, but this value is based on continuous monitoring and the maximum under the planned schedule of interim analyses after every 50 deaths will be smaller. The values of $S_C(24)$ and $S_E(24)$ are included in Table 1 in order to interpret the value of θ, all other entries are valid under the proportional hazards model, regardless of the true survival functions, recruitment rates, or recruitment periods, provided that the trial is run until a boundary is crossed.

Table 1 gives no information about the likely duration of the study or its sample size. In order to find these, stronger assumptions have to be set. A recruitment pattern of 10 patients per month for an opening phase of 6 months, followed by 20 patients per month for a further 24 months will be assumed. These are totals for both treatments. Two survival patterns will be explored. First, an exponential model with event rates of $\lambda_C = -(1/24)\ \log(0.20) = 0.0671$ and $\lambda_E = \lambda_C \exp(-\theta)$, consistent with $S_C(24) = 0.2$ and with equation (1), will be explored. This model gives control survival times past 6, 12, 18, 24, and 36 months of 0.669, 0.447, 0.299, 0.200, and 0.089, respectively. The second model will be a step function with corresponding control survival probabilities of 0.950, 0.500, 0.210, 0.200, and 0.190. This is chosen to be a distinctly nonexponential pattern for contrast, with few early and few late deaths, but still satisfying $S_C(24) = 0.2$. The exponential and step function models are shown in Figures 2 and 3, respectively, under the assumption that proportional hazards is true with $\theta = \theta_R$. Note that these models are being used only in the prediction of study duration and sample size: they will play no role in the interim and final analyses which are to be based on the logrank statistic.

In both cases, the methods described earlier and given in detail in the appendices were used to translate the properties of e* into equivalent properties for D* and n*, denoting the duration and total sample size, respectively. This was done by ascertaining the numbers of events available at each month after the start of the study, and listing as D* the first value for which e* is exceeded; n* was then derived from this and the recruitment pattern. Tables 2 and 3 present, for the exponential and step function models, respectively, the expected value (E), median (Med), and 90th percentile (P90) of each of e*, D*, and n*.

In each case, the properties of e* are the same. Under the exponential assumption, Table 2 shows durations between 2 and 5 years, with sample sizes ranging from just short of 300 to 540. Durations and sample sizes are rather larger under the step function assumption, as shown in Table 2.

Recruitment lasts for a maximum of 30 months during which time a maximum of 540 patients enter the trial. Table 2 shows that the trial will last beyond the recruitment phase, and will require all 540 patients, with probability ≥0.1 when $\theta = 0, -0.179$, or -0.358, and with probability ≥0.5 when $\theta = -0.179$. Under the exponential model all patients will eventually die, and as all listed values of e* are smaller than 540, corresponding values for D* and n* have been found for all of them. Higher percentiles of e* will exceed 540, and so there remains a small probability that no trial boundary will be reached, even when all 540 patients have died. If this were to happen, or if no boundary was reached when some maximum trial duration had been reached, then the trial would be analyzed as a trial that has underrun (22). Thus, a confidence interval would be available on which to base the conclusion about noninferiority, but this feature means that the procedure will not quite reach the target probability of 0.95 for claiming noninferiority when $\theta = 0$. The shortfall is likely to be very small, but no work has yet been done to quantify it.

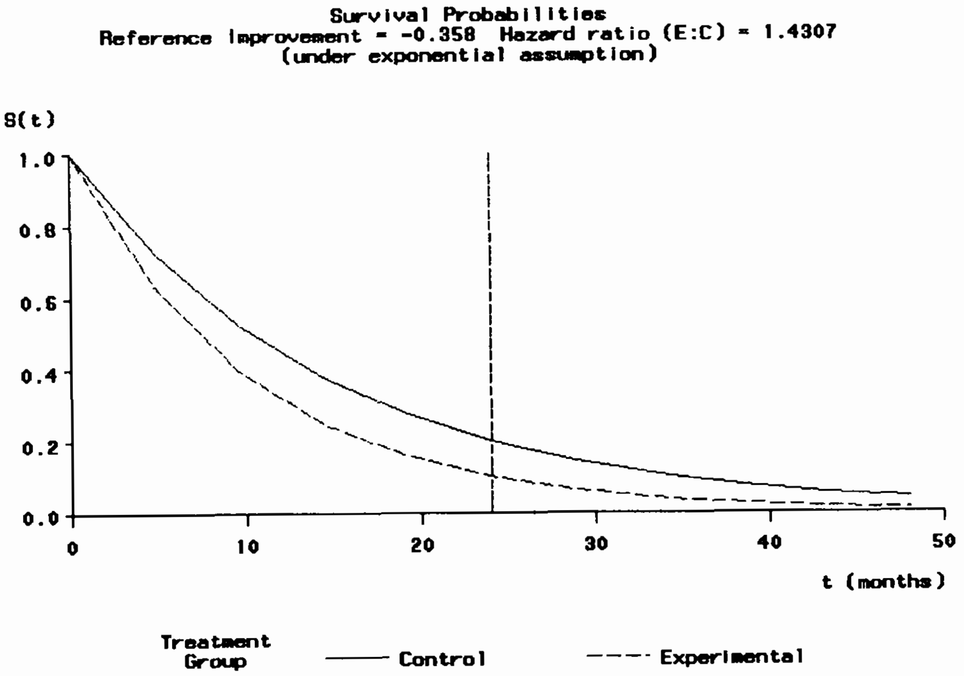


FIGURE 2. Exponential survival curves used in sample size calculation.

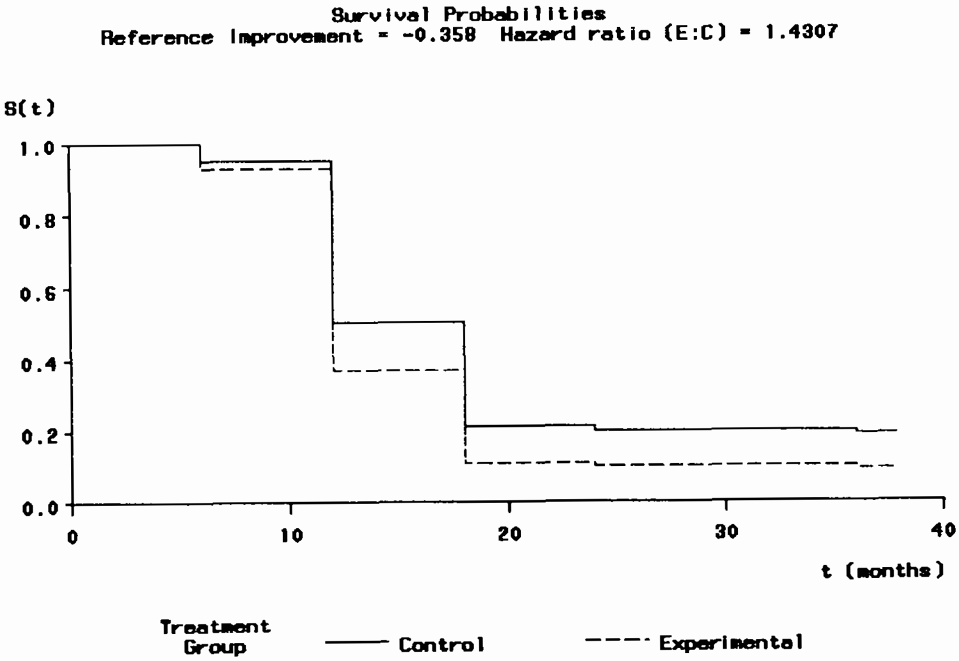


FIGURE 3. Step function survival curves used in sample size calculation.

TABLE 2
Duration and Sample Size of the Sequential Design (Exponential Calculation)

	e*			D*			n*		
θ	E	Med	P90	E	Med	P90	E	Med	P90
0.537	94	78	114	19	17.0	20.5	312	280	350
0.358	118	100	154	20	18.4	22.9	340	308	398
0.179	163	139	232	23	20.8	27.4	393	357	488
0	258	225	407	28	25.9	39.3	500	457	540
–0.179	340	318	513	32	30.4	58.5	540	540	540
–0.358	236	203	367	25	22.6	32.3	432	392	540
–0.537	152	130	212	19	17.1	22.3	311	281	387

The durations and sample sizes in Table 3 are larger than those in Table 2 partly because the calculations based on step functions are conservative. Deaths are counted at the end of each month, and the step functions decrease at the end of each time interval. Additionally, very few deaths occur after 18 months, and none at all after 36 months. This means that, according to the model, there is no point in continuing the trial once the 540th patient has been followed up for five years. With this survival pattern, the probability that no trial boundary will ever be reached actually exceeds 0.1 when θ = –0.179, and so the 90th percentile values are listed as missing. When θ = 0.179, 0, –0.179 or –0.358 the probability of requiring the maximum number of 540 patients exceeds 0.1, and when θ = 0 or –0.179 it exceeds 0.5.

As a rule of thumb, it appears reasonable to proceed with a design when (as in Table 2) the 90th percentile of duration exists for all values of θ, but to seek additional resources at the outset when (as in Table 3) this is not so. Thus, if investigators really did believe that the step function pattern of survival was likely to be realistic, then a longer recruitment period, and/or a greater entry rate per month, would be sought. Here, we shall imagine that the exponential survival pattern is more reasonable, and continue the illustration using the design explored.

Before recruitment to the trial is closed, it would be prudent to conduct a design review. A design review is essentially a sample size review for sequential studies, and methodology for this purpose has been described by Gould and Shih (23) and by Whitehead et al. (8). Existing survival and recruitment patterns are determined from blinded data, as described in the section on sample size reviews above, and then used to reevaluate the properties of duration and sample size. If the revised input now leads to missing 90th percentiles for certain values of θ, then either

TABLE 3
Duration and Sample Size of the Sequential Design (Step Function Calculation)

	e*			D*			n*		
θ	E	Med	P90	E	Med	P90	E	Med	P90
0.537	94	78	114	24	23	26	425	396	455
0.358	118	100	154	25	24	28	449	423	499
0.179	163	139	232	28	26	32	496	464	540
0	258	225	407	33	31	44	540	540	540
–0.179	340	318	513	37	36	•	540	540	•
–0.358	236	203	367	30	28	38	540	505	540
–0.537	152	130	212	25	23	28	437	410	505

recruitment should be extended or the goals of the study should be reviewed.

The Conduct and Analysis of the Study

As the data used for this illustration are simulated rather than real, little detail will be given. Table 4 below summarizes the data available at each of four interim analyses.

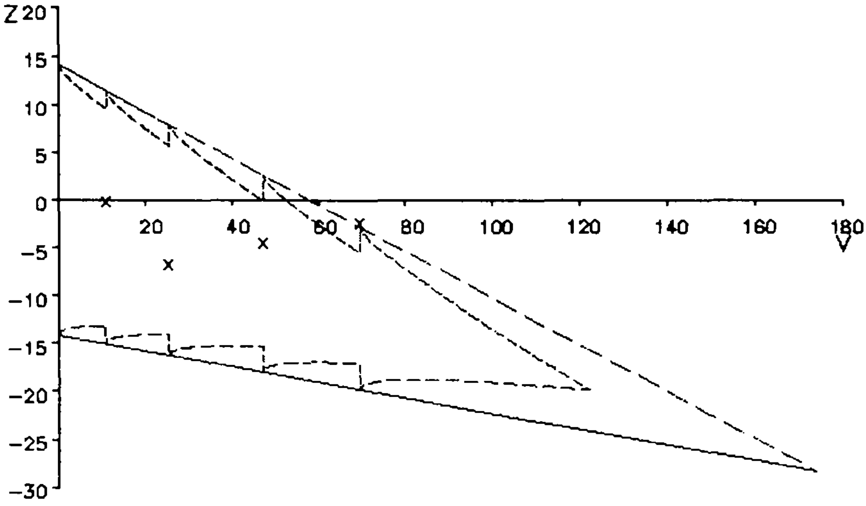
In this simulation, as is often the case in real life, recruitment has been slower than predicted. Thus, the first interim analysis, intended to take place after 12 months, was in fact delayed to 18 months. Subsequent interim analyses took place after intervals of approximately nine, nine, and six months, respectively. The sequential method adopted is robust against departures from the original schedule of looks, provided that the new timings are not influenced by the behavior of the sample path. The interim analysis on 25–05–2001 led to stopping the trial as the upper boundary had been crossed. By this time, 500 patients had been recruited, and recruitment was still open as the maximum sample size had been set to be 540. When the trial was stopped, 308 events had occurred. The fixed sample design of equal power would have required 456 events, and so the sequential design has led to a considerable reduction in data collected.

The sample path is shown in Figure 4. The inner crooked boundaries represent the Christmas tree correction, and it is sufficient for the plotted point to reach these in order for the study to be stopped. The outer boundaries are constructed for use with continuous monitoring, and the fact that looks are conducted at discrete and widely separated times means that values outside the boundaries might be missed. Looking only occasionally makes it harder to stop, and in order to compensate the boundaries must be adjusted to make it easier when looks do occur. That is the qualitative rationale for the adjustment: an indication of its quantitative aspects is given in Section 4.13 of Whitehead (11). Stallard and Facey (24) show that for the triangular test it is remarkably accurate, and by symmetry their results are equally relevant to the reverse triangular test. The critical values with which Z was actually compared are listed as ℓ (for the lower boundary) and u (for the upper) in Table 4.

Whitehead (11, Section 5.4) describes a method of analyzing data collected in a sequential clinical trial based on an ordering of all possible final data sets due to Fairbanks and Madsen (25). This method is available via the "discrete analysis" option of PEST 4, and is also the standard post-stopping analysis method in the software package EaSt 2000 (26). Applied to these data, the method gives a p-value (2-sided) of p = 0.667, a median unbiased estimate for θ of –0.054, and a corresponding 95% confidence interval of (–0.307, 0.189). As the reference value for θ of –0.358 lies below this interval, the conclusion of noninferiority can be drawn. The interval quoted has its 95% coverage probability for frequentist repetitions of this particular reverse triangular design, and so no further adjustments to it are necessary. PEST 4 also provides a more approximate "continuous analysis," for which p = 0.696 and the median unbiased estimate of θ is –0.048 with 95% confidence interval (–0.299, 0.192), which would have provided adequate interpretation in this case.

TABLE 4
Summary of Interim Analyses

Date	Month	n	e	V	Z	ℓ	u
16-01-1999	18	180	51	11.032	–0.222	–13.114	9.524
13-04-2000	27	278	116	25.637	–6.688	–14.010	5.668
05-11-2000	34	390	210	47.268	–4.540	–15.286	–0.094
25-05-2001	40	500	308	69.325	–2.405	–17.054	–5.503



STOP the study - a boundary has been crossed

FIGURE 4. The sample path at the final interim analysis.

DISCUSSION

The methodology for predicting sample size and study duration presented here is relatively simple in concept, and the replacement of unknown values by their expectations at various points in the procedure will result in loss of accuracy. On the other hand, investigation of sample size is never a particularly exact science, depending as it does on unreliable predictions of survival patterns and recruitment rates. Perhaps the accuracy of the procedure is commensurate with the task in hand.

The predictions are good enough to warrant some care in the choice of representation of survival pattern. It has been shown how false assumption of an exponential pattern could result in misleading values, and more extreme illustrations can be constructed, for example, when the true survival function falls rapidly to about 0.5 and then flattens off. The investigations of duration might even show that the sequential approach is of little value. For example, suppose that recruitment takes only six months, and that all events occur during the first half of the third year of follow-up. Interim analyses during the third year will show trial progress, but those before and after will show no change. The potential for shortening the trial, relative to a single analysis fixed for the end of the third year, will be slight.

The sequential methods considered in this paper rely on the assumption of proportional hazards for their validity. Gregory et al. (27) demonstrate how the picture emerging from a survival study can fluctuate over time, and discuss how this will be exacerbated if the assumption of proportional hazards is violated. After a fixed sample study, the assumption of proportional hazards can be checked: appropriate goodness-of-fit tests are described in Chapter 5 of Collett (10). If nonproportionality is found, then alternative presentations of the data can be given, although they are likely to be less powerful than the intended analysis. The situation is more complicated for sequential trials. Goodness-of-fit tests can accompany each interim analysis, but these reintroduce the problems of repeated significance testing eliminated from the primary comparison through the sequential design. It is also unclear how to proceed if nonproportionality of hazards is detected, and how any strategy for changing the design would effect error rates. Nevertheless, it is the approach of this author to assess proportional hazards and other assumptions at interim analyses and to react to convincing lack-of-fit as well as can be done.

Of course, sequential survival studies may stop early before data on which to assess the long-term validity of proportional hazards are available. In such a case, no extrapolation of the short-term reduction in hazard detected should be made. It may be desirable to time the first interim analysis sufficiently late to avoid this possibilty altogether. If there is reason to doubt the validity of proportional hazards in advance, and concern that interim analyses will be misleading as a consequence, then alternative sequential designs should be considered: the approach of Sooriyarachchi and Whitehead (12) is particularly robust.

REFERENCES

1. 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.
2. Sellke T, Siegmund D. Sequential analysis of the proportional hazards model. *Biometrika.* 1983;70: 315–326.
3. Tsiatis AA, Rosner GL, Tritchler DL. Group sequential tests with censored survival data adjusting for covariates. *Biometrika.* 1985;72:365–373.
4. Moss AJ, Hall WJ, Cannom DS, Daubert JP, Higgins SL, Klein H, Levine JH, Saksena S, Waldo A, Wilber D, Brown MW, Heo M. Improved survival with an implanted defibrillator in patients with coronary disease at high risk for ventricular arrythmia. *New Engl J Med.* 1996;335:1933–1940.
5. Boden WE, van Gilst WH, Scheldewaert RG, Starkey IR, Carlier MF, Julian DG, Whitehead A, Bertrand ME, Col JJ, Lederballe Pedersen O, Lie KI, Santoni J-P, Fox K. Diltiazem in acute myocardial infarction treated with thrombolytic agents: a randomised placebo-controlled trial. *Lancet.* 2000;355: 1751–1756.
6. Arriagada R, Le Chevalier T, Pignon J-P, Rivière A, Monnet I, Chomy P, Tuchais C, Tarayre M, Ruffié P. Initial chemotherapeutic doses and survival in patients with limited small-cell lung cancer. *New Engl J Med.* 1993;329:1848–1852.
7. Medical Research Council Renal Cancer Collaborators. Interferon-α and survival in metastatic renal carcinoma: early results of a randomised controlled trial. *Lancet.* 1999;353:14–17.
8. Whitehead J. Monotherapy trials: Sequential design. *Epilepsy Res.* 2001;43:81–87.
9. MPS Research Unit. *PEST 4: Operating Manual.* Reading, UK: The University of Reading; 2000.
10. Collett D. *Modelling Survival Data in Medical Research.* London: Chapman and Hall; 1994.
11. Whitehead J. *The Design and Analysis of Sequential Clinical Trials.* Revised second ed. Chichester, England: Wiley; 1997.
12. Sooriyarachchi MR, Whitehead J. The sequential analysis of survival data with non-proportional hazards. *Biometrics.* 1998;54:1072–1084.
13. Freedman LS. Tables of the numbers of patients required in clinical trials using the logrank test. *Stat Med.* 1982;1:121–129.
14. Machin D, Campbell MJ, Fayers PM, Pinol APY. *Sample Size Tables for Clinical Studies.* Second edition. Oxford: Blackwell Science Ltd.; 1997.
15. Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. *Biometrics.* 1983; 39:499–503.
16. Whitehead J, Whitehead A, Todd S, Bolland K, Sooriyarachchi MR. Mid-trial design reviews for sequential clinical trials. *Stat Med.* 2001;20:165–176.
17. Wittes J, Brittain E. The role of internal pilot studies in increasing the efficiency of clinical trials. *Stat Med.* 1990;9:65–72.
18. Gould AL. Interim analyses for monitoring clinical trials that do not materially effect the type I error rate. *Stat Med.* 1992;11:55–66.
19. Gould AL. Planning and revising the sample size for a trial. *Stat Med.* 1995;14:1039–1051.
20. Gould AL, Shih WJ. Sample size re-estimation without unblinding for normally distributed data with unknown variance. *Comm Stat—Theory Methods.* 1992;21:2833–2853.
21. Whitehead J. Sequential designs for equivalence studies. *Stat Med.* 1996;15:2703–2715.
22. Whitehead J. Overrunning and underrunning in sequential clinical trials. *Control Clin Trials.* 1992;13: 106–121.
23. Gould AL, Shih WJ. Modifying the design of ongoing trials without unblinding. *Stat Med.* 1998;17: 89–100.
24. Stallard N, Facey KM. Comparison of the spending function method and the Christmas tree correction for group sequential trials. *J Biopharmaceutical Stat.* 1996;6:361–373.
25. Fairbanks K, Madsen R. P values for tests using a repeated significance test design. *Biometrika.* 1982; 69:69–74.
26. Cytel Software Corporation. EaSt 2000: A Software Package for the Design and Interim Monitoring of Group Sequential Clinical Trials. Cambridge MA: Cytel; 2000.
27. Gregory WM, Bolland KM, Whitehead J, Souhami RL. Cautionary tales of survival analysis: conflicting

analyses from a clinical trial in breast cancer. *Br J Cancer.* 1997;76:551–558.

28. George SL, Desu MM. Planning the size and duration of a clinical trial studying the time to some critical event. *J Chronic Dis.* 1974;27:15–24.
29. Schoenfeld DA, Richter JR. Nomograms for calculating the number of patients needed for a clinical trial with survival as an endpoint. *Biometrics.* 1982; 38:163–170.

APPENDIX 1: EXPONENTIAL SURVIVAL TIMES

In the exponential model, survival times in the experimental and control arms have constant hazards λ_E and λ_C, respectively. For some value of t_0, values for $S_E(t_0)$ and $S_C(t_0)$ are anticipated. The latter will usually come from control data observed in similar trials, whereas the former might well be deduced from equation (2) for the reference improvement θ_R. As $S_E(t_0) = \exp(-\lambda_E t_0)$ and $S_C(t_0) = \exp(-\lambda_C t_0)$ in the exponential model, values for λ_E and λ_C can be deduced.

To begin with, suppose that the trial recruits at a constant rate of a patients per month, randomized A : 1 between E and C for R months, after which follow-up proceeds for a further F months. Denote the number of recruits on E and C by n_E and n_C, respectively, and the corresponding number of events by e_E and e_C. Now, conditional on n_E, the expected value of e_E is given by

$$\begin{aligned} E(e_E|n_E) &= n_E P(\text{event on E}) \\ &= n_E \int_0^R P(\text{event on E}|\text{entry at u})\ g(u)\ du \end{aligned}$$

where g(u) is the entry time distribution, taken to be uniform on (0, R). Thus,

$$E(e_E|n_E) = \frac{n_E}{R}\int_0^R \{1 - P(T_E > R + F - u)\}\ g(u)\ du$$

where T_E is the survival time of a randomly chosen patient on the experimental treatment. Using the exponential distribution,

$$\begin{aligned} E(e_E|n_E) &= \frac{n_E}{R}\int_0^R 1 - \exp\{-\lambda_E(R + F - u)\}du \\ &= n_E - (n_E/R\lambda_E)[\exp\{-\lambda_E F\} \\ &\quad - \exp\{-\lambda_E(R + F)\}], \end{aligned}$$

and a similar expression holds for $E(e_C \mid n_C)$. Now

$$E(e) = E\{E(e_E|n_E)\} + E\{E(e_C|n_C)\},$$

while $E(n_E) = RAa/(A + 1)$ and $E(n_C) = Ra/(A + 1)$. Hence

$$E(e) = a\left[R - \frac{\exp\{-\lambda_E F\} - \exp\{-\lambda_E(R + F)\}}{(A + 1)\lambda_E} - A\frac{\exp\{-\lambda_C F\} - \exp\{-\lambda_C(R + F)\}}{(A + 1)\lambda_C}\right]. \quad (A1)$$

In designing a fixed sample study, the values of a, R, and F can be juggled to make E(e) match the required value of e given by (3).

George and Desu (28) give a similar formula, simpler in that they take F to be zero, but more complicated in that they allow for the entry process to be Poisson. Machin et al. (14) use the George and Desu approach in their Section 9.2.2. Schoenfeld and Richter (29) extend the George and Desu result to the case in which F > 0, and it is their expression that is used in the software nQuery.

It is possible to generalize equation (A1) to allow for a variable entry rate. It can be supposed that, during the calendar time interval (c_{j-1}, c_j) the entry rate is a_j per month, j = 1, . . . , h, where $c_0 = 0$ and $c_k = R$. Let e(c) denote the total number of events occurring by calendar time c. Put c_j^* equal to c_j for all $c_j \le c$, and equal to c for all $c_j > c$, so that the intervals (c_{j-1}^*, c_j^*) denote intervals of constant recruitment prior to the time c, or else are empty. Now let $e_{E,j}(c)$ denote the number of events on E occurring by time c among patients entering during time interval (c_{j-1}^*, c_j^*), and define $e_{C,j}(c)$ similarly for C, j = 1, 2, . . . , h. The argument above can be generalized to show that

$$E\{e_{E,j}(c)\} = \frac{Aa_j}{A + 1}\left[(c_j^* - c_{j-1}^*) - \frac{\exp\{-\lambda_E(c - c_j^*)\} - \exp\{-\lambda_E(c - c_{j-1}^*)\}}{\lambda_E}\right], \quad (A2)$$

with a similar expression (with leading term $a_j/(A + 1)$) for $E\{e_{C,j}(c)\}$. Adding these over E and C and over j = 1, . . . , h gives E{e(c)}. This procedure allows the number of events available at any time during or after the recruitment phase to be calculated, and determination of design strategies to ensure that the required total number of events will be accumulated.

According to the exponential model, events can occur at any time after recruitment, and every patient will eventually suffer an event. Thus, in the long term, the expected number of events will approach but never quite achieve the total number of patients recruited.

APPENDIX 2: SURVIVAL FUNCTIONS EXPRESSED AS STEP FUNCTIONS

Suppose that the survivor function for the control group, $S_C(t)$ is taken to be $S_C(t) = s_{C,i}$, $t \in [t_{i-1}, t_i)$, i = 1, . . . , k + 1; $t_0 = 0$, $s_0 = 1$, $t_{k+1} = \infty$. The final value of $S_C(t)$, $s_{C,k+1}$, may be zero, indicating that all control patients are likely to experience an event before time t_k or positive, indicating the possibility of long-term survival. Literally, this means that events can take place only at the end of each time period. From the anticipated survival pattern in the control group, which might be deduced from data collected in a previous trial, the corresponding pattern in the experimental group under a proportional hazards model with log-hazard ratio equal to θ_R can be deduced from equation (2). The overall survivor function for the study will then be $S^*(t) = \{S_C(t) + AS_E(t)\}/(A + 1)$. In practice, a suitable time unit such as months can be chosen, and then the t_i can be taken to have integer values.

Now suppose that, during the mth month of recruitment, a total of a_m patients enter the trial. The expected number of these patients who experience the event during the rth month of the trial is $a_m\{S^*(r - m - 1) - S^*(r - m)\}$, for $r = m + 1, m + 2, \ldots$. Summing over m from 1 to r − 1 gives the total number of events occurring during the rth month, and then summing over r from 1 to c gives the total number of events accumulated by the cth month.

These considerations furnish an approximate but feasible method for computing the likely number of events at any stage during the trial. The step function model for survival does not admit the possibility of events occurring after time t_k. Thus, information for the trial will be complete at time $R + t_k$, which according to the model represents the maximum worthwhile duration of the study.