STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:1969–1989 (DOI: 10.1002/sim.1193)

Designing complex group sequential survival trials

Edward Lakatos*,†

Forest Laboratories Inc, 909 Third Avenue, New York City, NY 10022-4731, U.S.A.

SUMMARY

This paper presents methodology for designing complex group sequential survival trials when the survival curves will be compared using the logrank statistic. The method can be applied to any treatment and control survival curves as long as each hazard function can be approximated by a piecewise linear function. The approach allows arbitrary accrual patterns and permits adjustment for varying rates of non-compliance, drop-in and loss to follow-up. The calendar-time–information-time transformation is derived under these complex assumptions. This permits the exploration of the operating characteristics of various interim analysis plans, including sample size and power. By using the calendar-time–information-time transformation, information fractions corresponding to desired calendar times can be determined. In this way, the interim analyses can be scheduled in information time, assuring the desired power and realization of the spending function, while the interim analyses will take place according to the desired calendar schedule. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: group sequential; survival; design; Markov; sample size; non-proportional hazards

1. INTRODUCTION

In designing a fixed-sample survival trial, there is a tension among potential endpoints, overall trial length, the relative lengths of the accrual and follow-up periods and sample size. Assessing the impact of these factors on power is all the more difficult when the hazard functions are non-linear and non-proportional, there is loss to follow-up or competing risks, non-compliance and drop-in. If a group sequential design is indicated, the statistician must additionally assess the impact of various interim monitoring plans, including potential boundaries, times of analyses, and whether the schedule for analyses will be based on the number of events or calendar time.

This paper assumes the primary analyses for the group sequential survival trial will use the logrank statistic. Over the last three decades, sample size methods for fixed sample survival trials have been developed to adapt to a variety of conditions experienced in actual trials [1–5]. The most comprehensive of these methods are flexible enough to closely model any shape

*Correspondence to: Edward Lakatos, 120 Truesdale Drive, Croton-on-Hudson, NY 10520, U.S.A.
†E-mail: elakatos@msn.com

Received October 1999
Accepted October 2001

survival curves and non-proportional hazards, to adjust for rates of non-compliance, drop-in, loss to follow-up and competing risk that may be expected to change as the trial progresses, to model any pattern of recruitment, and to model treatment lag. For group sequential survival trials, Kim and Tsiatis [6] developed an analytic approach for exponential models with uniform recruitment, but with no provision for the other factors. Some simulation approaches for sample size calculations have been developed for group sequential trials [7, 8, 10]. Halpern and Brown [7] developed a program for arbitrary survival curves including a period of follow-up after uniform accrual, but no adjustment for non-compliance, drop-in, or loss to competing risks or follow-up. Scharfstein and Tsiatis [8] address the broader question of designing group sequential trials when the interim analyses will be based on a unique parameter for which an efficient estimator will be used. They recommend a simulation-based approach, which is partially implemented in the commercial package EaST [9]. Gu and Lai [10] propose simulation for power or sample size allowing input of survival curves of varying shapes, with adjustment for non-compliance, drop-in and length of accrual. Several commercial software packages are available for group sequential design and analysis; for a recent review, see Emerson [11]. In contrast to these simulation programs which provide only sample size and power, the proposed methods provide quantification of, and thus insight into, the operating characteristics for each set of design specifications investigated. The calendar-time–information-time transformation is particularly useful in this regard.

The proposed approach is to use non-stationary Markov models to project the survival curves under the complex settings described above. These projections are then used to quantify all parameters useful in designing the group sequential trial. For example, the calendar-time–information-time transformation is generated, and this can be used to determine what the boundary will look like if, for example, interim analyses are planned every six months during the trial. Alternately, if the interim analyses are scheduled to take place after fixed numbers of events with equal increments of information, the transformation can be used to predict the calendar time when such analyses will take place. The method additionally allows exploring boundaries based on complex patterns of accruing data. The predicted values of the boundaries can be calculated once the information-time–calendar-time transformation is available. Finally, sample sizes for a variety of complex configurations are readily calculated with little investment of computer time.

In Section 2, some background for group sequential designs is presented. For analysis, both the logrank statistic and the Kaplan–Meier [22] survival curves reparameterize all times relative to time from randomization. The Markov model is constructed similarly, with the initial distribution representing time from randomization (time 0). In Section 3, the Markov model for fixed sample designs is reviewed and then the model for the group sequential setting is introduced. An example to be referenced throughout the paper is presented in Section 4. Administrative censoring, which plays a key role in group sequential survival trials, is discussed in Section 5. How the model is used for calculating sample size and for projecting the operating characteristics of the trial is found in Section 6. Section 7 provides simulation verification in a variety of situations. Section 8 treatment lag, and Section 9 the weighted logrank statistic. This is followed by a concluding discussion in Section 10. An SAS IML [12] computer program is available from the author.

2. BACKGROUND FOR GROUP SEQUENTIAL METHODS

When group sequential designs are carried out, time can be viewed in several different ways. The interrelationship of these measures of time in the survival setting is often complex, but fundamental to successful design and implementation. The non-survival setting provides a perspective useful in understanding the survival case. Assume that a trial will enrol, in sequence, up to I sets of $2n$ patients each. Of the $2n$ patients in each set, n will be assigned at random to each of groups A and B. For simplicity we assume that for each patient the endpoint is measured at baseline, the patient is randomized, treatment is administered, and the final assessment of the endpoint is made, and there is essentially no elapsed time between the baseline and follow-up assessment. It is assumed that the change from baseline is normally distributed with $X_A \sim N(\mu_A, \sigma^2)$ and $X_B \sim N(\mu_B, \sigma^2)$. After the ith set of $2n$ patients has been measured, the statistic

$$Z_i = \frac{\sum_{k=1}^{i} (\bar{x}_{A_k} - \bar{x}_{B_k})}{\sqrt{(2\hat{\sigma}^2/ni)}}$$

is calculated and compared to the ith in a prespecified sequence $b_1, b_2, \ldots, b_I$ called a boundary. If $Z_i > b_i$, then the trial is stopped and the null hypothesis is rejected. Otherwise, the process is continued with another set of $2n$ patients. If the null hypothesis is still not rejected after the testing of the Ith group, the trial is terminated and the difference declared non-significant. The boundary is chosen so that under the null hypothesis, the cumulative probability of rejection is the desired significance level.

In order to generalize this process for survival trials, the concept of 'information time' as presented by Lan and Zucker [13] is now discussed. With the Lan and Zucker approach, a single parameter is identified for monitoring efficacy, and information is defined as the reciprocal of the variance of the current estimate of this parameter. In the example just presented, the information at the ith interim analysis is $ni/2\sigma^2$. Increasing data results in increasing precision, and, in turn, increasing information. The information fraction or information time, defined as the ratio of the current information to the information were the trial to be carried out without early termination, is particularly useful in implementing group sequential analyses. In this simple example, the information fraction is the proportion $(ni/2\sigma^2)/(nI/2\sigma^2) = i/I$ of total patients currently enrolled, treated and measured, where i and I index the current and final analyses. If the rate of enrolment is reasonably predictable, then the information fraction, as a function of actual time (calendar time) is equally predictable.

The situation in most trials is more complex because the information contributed by an individual patient is not usually available at the time the patient enrols. For survival trials, information is given by

$$\left(\frac{1}{n_E} + \frac{1}{n_C}\right)^{-1} \frac{d}{(n_E + n_C)} \tag{1}$$

where d is the total deaths, and n_E and n_C are the numbers at risk in the experimental and control groups [13]. Consequently, the information is proportional to the number of deaths. As such, it is a complex function of the survival rate and pattern of accrual, as well as competing risks and other factors that complicate the trial.

Early group sequential methods required prespecification of a sequence of interim analysis times and corresponding critical values for rejecting the null hypothesis and terminating the trial. As in the above example, the times of the interim analyses were given as information fractions, not calendar times. Such trials are referred to as 'maximum information' trials, in contrast to 'maximal duration' trials in which the interim and final analyses are scheduled at calendar times. The Lan and DeMets [14] approach to group sequential monitoring provides greatly improved flexibility, allowing, among other features, DSMB meetings and interim analyses to be scheduled in calendar time rather than when prespecified numbers of events happen. With the Lan–DeMets approach, one specifies an α-spending (or use) function $\alpha^*(u)$ from which the boundary may be calculated during the course of the trial. Suppose interim analyses are performed at calendar times $t_1, t_2, \ldots, t_I$ and the corresponding information fractions at these times are $u_1, u_2, \ldots, u_I$. Then $\alpha_i = \alpha^*(u_i) - \alpha^*(u_{i-1})$ is the α to be 'used' or 'spent' at the ith interim analysis, and the boundary can be derived using numerical integration methods [16]. In designing such a trial, it is necessary to estimate the information fraction u_i corresponding to each given calendar time t_i. The Markov model presented in Section 3 provides such a link.

3. THE MARKOV MODEL

Consider a survival trial in which patients are randomized to either a control or experimental group. We begin by briefly reviewing the Markov model [3]; the extension to the group sequential setting will follow. Each treatment group is modelled separately; without loss of generality, only the model for the experimental group is presented here. Patients initially randomized to the experimental group occupy the state A_E, which designates that the patient is at risk at the experimental group rate. During the trial, patients can remain in A_E or transition to one of three other states: A_C for those who no longer comply with their treatment regimen and are at risk at the control group rate; E for patients who have the event of interest; or L for patients who are lost to follow-up, lost to competing risks, or for patients whose time of failure is censored because they are still at risk for the event of interest. Once the Markov model is derived, only the at-risk states A_E and A_C, and the event state E will be used for subsequent calculations. The loss state L is used to accumulate losses due to any possible reason: loss to follow-up, competing risks, and administrative censoring are the most common. Administrative censoring is discussed in Section 5. For this reason, except for administrative censoring, probabilites from losses due to all other causes are combined and then entered in the transition matrices. If loss rates from multiple sources are very high, additional states can be added. Assume the time points $t_0, t_1, \ldots, t_H$ equally divide the period of the trial where t_0 is the time of randomization and t_H is the end of the trial. Let $\mathscr{D}_E(t_h)$ represent the distribution, or vector (column) of occupancy probabilities for states L, E, A_E, A_C at time t_h. The model is given by

$$\mathscr{D}_E(t_h) = \left(\prod_{j=1}^{h} \mathscr{T}_j \right) * \mathscr{D}_E(t_0) \tag{2}$$

where '$*$' indicates matrix multiplication, and the 4×4 transition matrices $\mathscr{T}_j$ are given by

$$\mathscr{T}_j = \begin{bmatrix} & L & E & A_E & A_C \\ L & 1 & 0 & p_{\text{loss},j} & p_{\text{loss},j} \\ E & 0 & 1 & p_{\text{event}_E} & p_{\text{event}_C} \\ A_E & 0 & 0 & 1-\Sigma & p_{\text{dri},j} \\ A_C & 0 & 0 & p_{\text{non-comp},j} & 1-\Sigma \end{bmatrix}$$

Here, the first column and row are used to identify the states and are not part of the matrix, and Σ is the sum of the three other entries in its column. The probabilities within the transition matrix are probabilities conditional on being in the current state; the notation p_{event_E} is used to distinguish from the cumulative probabilities p_E. These transition probabilities come from assumptions about the trial, which will be discussed in the example in the next section.

4. EXAMPLE

The Randomized Aldactone Evaluation Study [16] (RALES) was a double-blind placebo-controlled trial designed to compare mortality under treatment with Aldactone to placebo in patients with congestive heart failure (CHF). In addition to this randomly assigned therapy, all patients were given ACE inhibitors and diuretics. The SOLVD [17] and CONSENSUS [18] trials which established the efficacy of ACE inhibitors in conjunction with diuretics in CHF were used to provide rates for the control group, as well as compliance rates. Table I provides conditional mortality rates calculated from published survival curves using $S(t_k | t_{k-1}) = (S(t_{k-1}) - S(t_k))/S(t_{k-1})$ where S is the cumulative survival probability, that is, $S(t) = \Pr(T \geqslant t)$, where T is a non-negative random variable representing the lifetimes of individuals in some population. Throughout this paper, all units are in months; in applications, any unit can be used. The second and third columns present cumulative mortality rates read directly from published survival curves [17, 18]. Initially, cumulative rates were obtained for each month, and for each month, the monthly conditional rates derived using the above formula. The months were combined into the periods shown in the first column because the patterns indicated similar monthly conditional rates within the periods. The monthly conditional rates shown in columns four and five were then recalculated using the above formula and an exponential assumption. For example, for CONSENSUS, for patients who have survived through the end of the sixth month, the probability of surviving to the end of the year is $S(12|6) = 1 - ((1 - 0.28) - (1 - 0.42))/(1 - 0.28) = 0.80556$. To find the hazard during the period [7, 12], solve $S(t) = e^{-\lambda t}$ for $\lambda_{12|6} = -\log(S(12|6))/6 = 0.03604$. The conditional probability of surviving for any one month period during this interval is $\exp(-1\lambda_{12|6}) = 0.96460$, so the conditional monthly failure probability is 0.03540.

In both trials, the conditional mortality rates were much higher during the first three months than during the remainder of the trial, casting doubt on the appropriateness of an exponential model. The mortality rate among CHF patients can vary substantially depending on disease severity, which was far greater in CONSENSUS than in SOLVD. The sixth column of Table I is a linear combination of the conditional rates from SOLVD and CONSENSUS (in the case, assuming 50 per cent from each trial), adjusted to take into account the disease severity of the prospective cohort in RALES. The final column converts the monthly assumed rates,

Table I. Mortality rates (per cent) from two trials of CHF.

Month	Cumulative		Conditional			
			Monthly			Annualized
	SOLVD	CONS	SOLVD	CONS	Assumed	Assumed
3	5	19	1.70	6.70	4.035	39
6	7	28	1.01	3.86	2.478	26
12	12	42	0.92	3.54	2.369	25
24	22	60	1.01	3.05	2.154	23
36	31	70	1.01	2.37	1.842	20
48	35		1.01		1.842	20

assuming an exponential model, into annualized rates, for the computer program. For example, $1-(1-0.04035)^{12}=0.39$.

To demonstrate the Markov model, first assume simultaneous entry and a five year trial; staggered entry will be discussed in the next section. In the computer program, a common time unit is used throughout for: length of trial; times of interim analyses; specifying the recruitment pattern. Consequently, the term 'month' is nominal, and can be replaced easily by any time unit. The user specifies how finely each time unit will be subdivided for the Markov process. As in Lakatos [3, 4], assumptions regarding probabilities are given in annualized rates, which are converted in the program to rates as appropriate for the specified interval length. The annualized control group event rates in this example are taken from the last column in Table I; the experimental group rates are 77.5 per cent of the control group rates. For example $0.39\times(1-0.225)=0.30222$. In this example, each month is the smallest interval, so that the program converts these two rates to 0.04035 and 0.02955 for building the transition matrix. As follow-up for mortality is essentially complete in this type of trial, it is assumed here that there is no loss to follow-up or competing risks. In other trials, where competing risks and/or losses to follow-up are possible, a fifth 'assumed' column would contain the sum of the probabilities of all such losses (see, for example, Lakatos [3]). Non-compliance is assumed to have an annual rate of 10 per cent during the first year, and 5 per cent thereafter. Drop-in is the phenomenon of patients randomized to control taking an active medication similar in effect to the experimental therapy. This may happen, for example, if such patients see their non-trial private physicians who diagnose the relevant condition and prescribe some active therapy. Only the first 24 months are shown in Table II.

The distribution for the sixth month can be derived from the distribution at the fifth month and the assumed transition probabilities using (2):

$$\begin{bmatrix} 0 \\ 0.1369 \\ 0.8190 \\ 0.0441 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0.01858 & 0.02478 \\ 0 & 0 & 0.97268 & 0.00427 \\ 0 & 0 & 0.00874 & 0.97095 \end{bmatrix} \begin{bmatrix} 0 \\ 0.1203 \\ 0.8419 \\ 0.0378 \end{bmatrix}$$

With these assumptions, the entries in column E give the expected failure curve assuming simultaneous entry.

Table II. RALES example. Markov model for *experimental* group with simultaneous entry.

Month	Assumed transition probabilities (per cent monthly)				Distribution in states			
							at risk states	
	p_{ncomp}	p_{dn}	p_{event_C}	p_{event_E}	p_{loss} L	p_{fail} E	p_{A_E} A_E	p_{A_C} A_C
rand	0.874	0.427	4.035	2.955	0	0	1	0
1	0.874	0.427	4.035	2.955	0	0.0295	0.9617	0.0087
2	0.874	0.427	4.035	2.955	0	0.0583	0.9249	0.0168
3	0.874	0.427	2.478	1.858	0	0.0863	0.8896	0.0241
4	0.874	0.427	2.478	1.858	0	0.1034	0.8654	0.0312
5	0.874	0.427	2.478	1.858	0	0.1203	0.8419	0.0378
6	0.874	0.427	2.369	1.779	0	0.1369	0.8190	0.0441
7	0.874	0.427	2.369	1.779	0	0.1525	0.7975	0.0500
8	0.874	0.427	2.369	1.779	0	0.1678	0.7766	0.0556
9	0.874	0.427	2.369	1.779	0	0.1830	0.7562	0.0608
10	0.874	0.427	2.369	1.779	0	0.1979	0.7364	0.0657
11	0.874	0.427	2.369	1.779	0	0.2125	0.7171	0.0703
12	0.874	0.427	2.369	1.779	0	0.2269	0.6984	0.0746
13	0.427	0.427	2.154	1.623	0	0.2399	0.6844	0.0757
14	0.427	0.427	2.154	1.623	0	0.2526	0.6707	0.0767
15	0.427	0.427	2.154	1.623	0	0.2652	0.6573	0.0775
16	0.427	0.427	2.154	1.623	0	0.2775	0.6442	0.0783
17	0.427	0.427	2.154	1.623	0	0.2896	0.6313	0.0791
18	0.427	0.427	2.154	1.623	0	0.3016	0.6187	0.0797
19	0.427	0.427	2.154	1.623	0	0.3133	0.6064	0.0803
20	0.427	0.427	2.154	1.623	0	0.3249	0.5943	0.0808
21	0.427	0.427	2.154	1.623	0	0.3363	0.5824	0.0813
22	0.427	0.427	2.154	1.623	0	0.3475	0.5709	0.0817
23	0.427	0.427	2.154	1.623	0	0.3585	0.5595	0.0820
24	0.427	0.427	2.154	1.623	0	0.3694	0.5484	0.0822

5. ADMINISTRATIVE CENSORING

Two types of censoring are considered. The first type is due to naturally occurring events that are mostly out of the control of the trial administrators, such as competing risks. Censoring due to competing risks occurs when, for example, we are interested in cardiac death, and the patient first dies from cancer. The other type of censoring occurs because the trial administrators may want to analyse the data before all patients die. If an analysis is performed before all patients die, then the time of death for those still alive is censored; this type of censoring is called 'administrative censoring'. In general, the later the analysis is performed, the less the censoring. This censoring occurs as part of the administration of the trial, in contrast to censoring by competing risks or loss to follow-up. Competing risks is a stochastic process similar to the failure process for the primary endpoint, and assumptions about it for the Markov model are usually based on historical data for the target population. In contrast, administrative censoring happens at a common calendar time point for all remaining patients, and, consequently, the time from randomization to administrative censoring is a function of the recruitment process.

Table III. Calculating probabilities of administrative censoring.

Recruitment							
Calendar day	May 1	+30	+60	+90	+120	+150	+180
Recruitment period j	1	2	3	4	5	6	
Randomized	50	100	75	150	125	100	0
Probability of being randomized	0.083	0.167	0.125	0.250	0.208	0.167	
Administrative censoring							
Days from randomization	30^-	60^-	90^-	120^-	150^-	180^-	
Administration censoring period m	1	2	3	4	5	6	
Still at risk at interim analysis	600	500	375	225	150	50	
Censored	100	125	150	75	100	50	
Probability of being censoring	0.167	0.250	0.400	0.333	0.667	1.00	

For group sequential trials, it is useful to think of latent survival curves, one for each treatment group. At each successive interim analysis, one is able to estimate more of the survival curves and with more data, but the underlying survival curves do not change. The approach taken here is to use one survival model per group, but different models for administrative censoring – one corresponding to each interim analysis. In the Markov model, transition probabilities are defined as functions of the time from entry rather than calendar time. The logrank statistic and Kaplan–Meier curves are calculated in the same way. To preserve the Markov assumption, the model assumes all patients are entered simultaneously, and staggered entry is accounted for by administratively censoring patients in consonance with their accrual pattern.

To understand the modelling of the recruitment/administrative censoring process, we begin with an example. If an interim analysis is performed at 180 days after study start, then a patient randomized at day 60 will have 120 days of follow-up at the time of the interim analysis and thus be administratively censored at 120 days from randomization, provided no other event occurs earlier. Similarly, a patient randomized at day 150 will be administratively censored by this interim analysis at 30 days from randomization. If a second interim analysis takes place at day 257, then the same two patients will be administratively censored at days 197 and 107, respectively, provided no other event has already occurred.

For specifying the recruitment process for this example, it is assumed that each month has 30 days with the first days of the months labelled as $0, 30, \ldots$, that all recruitment for a given month takes place on the first day of that month, and that all administrative censoring for a given month takes place just prior to the end of that month, labelled, for example, day 30^-. The recruitment process is specified in calendar time, as the trial planners expect it to take place. Administrative censoring is in the time frame of the analysis and of the Markov model, which is time from randomization.

A hypothetical recruitment process and the implied administrative censoring for a 180 day interim analysis is given in Table III. Here, study start is assumed to be 1 May. The recruitment process is specified in the upper panel, and the monthly recruitment periods in the second row of that panel. Similarly, the periods for administrative censoring appear in the second row of the lower panel. Denote the recruitment period by j, and the administrative

censoring period by m. Let t^i, $i=1,2,\ldots,I$, be the planned times of the interim analyses, where t^I is the time of the final analysis. Assume the trial is divided into k^i equal intervals or periods at the time t^i of the ith interim analysis The correspondence is given by $j \leftrightarrow k^i - j + 1 = m$, where k^i is the number of periods up to the ith interim analysis. For example, the 125 patients randomized in the fifth month from study start will be administratively censored in the second month from randomization. As with Kaplan–Meier survival curves, all 600 patients randomized during the six calendar months leading up to this interim analysis are at risk at the time of randomization. The 100 patients recruited at the beginning of the sixth calendar month will be censored at the end of the first month from randomization. Consequently, for this interim analysis, the probability of being administratively censored at the end of the first month is 100/600. The remainder of the lower panel of Table III is calculated similarly.

If n_j^i is the number expected to be randomized during the jth period, then for this interim analysis, the probability of entering during the jth period is

$$p_j^i = n_j^i \Big/ \sum_{j=1}^{k^i} n_j^i \tag{3}$$

The probability of being administratively censored during the mth period is given by

$$a_m^i = \frac{p_{k^i-m+1}^i}{\sum_{h=1}^{k^i-m+1} p_h^i} \tag{4}$$

To include administrative censoring in the Markov model, patients in the at-risk states must transition to the loss state. To accomplish this, define $\mathscr{A}_m^i$ by

$$\mathscr{A}_m^i \equiv \begin{bmatrix} & \boldsymbol{L} & \boldsymbol{E} & \boldsymbol{A}_{\mathrm{E}} & \boldsymbol{A}_{\mathrm{C}} \\ \boldsymbol{L} & 1 & 0 & a_m^i & a_m^i \\ \boldsymbol{E} & 0 & 1 & 0 & 0 \\ \boldsymbol{A}_{\mathrm{E}} & 0 & 0 & 1-a_m^i & 0 \\ \boldsymbol{A}_{\mathrm{C}} & 0 & 0 & 0 & 1-a_m^i \end{bmatrix}$$

where a_m^i is given by (4). Then the Markov model

$$\mathscr{D}_{\mathrm{E}}(t=t_h) = \left(\prod_{j=1}^{h} \mathscr{T}_j * \mathscr{A}_j^i \right) * \mathscr{D}_{\mathrm{E}}(t=0) \tag{5}$$

where $h=2,\ldots,k^i$ includes administrative censoring. The matrix $\mathscr{A}_j^i$ models transitions of patients from each of the two active states $\boldsymbol{A}_{\mathrm{E}}$ and $\boldsymbol{A}_{\mathrm{C}}$ into the loss state $\boldsymbol{L}$. Note that for the Markov model, all losses are combined into a single loss state – there is no need to differentiate between the various reasons for loss, as only the at risk and event states are used for subsequent calculations.

Although this recruitment pattern was not used in RALES, it is now applied to the Markov model with other RALES assumptions to demonstrate administrative censoring for interim analyses at months 6 and 13 of a 60 month trial. The 6 and 13 month interim analyses are in the upper and lower panels of Table IV, respectively. Here, $p_{\mathrm{event_E}}$ denotes the assumed probability of failing by the end of the designated month, conditional on being at risk at the beginning of that month.

Table IV. Modeling interim analyses at months 6 and 13, using recruitment from Table III.

Month	Assumed transition probabilities (per cent monthly)				Cens.	Distribution in states			
						Loss	Fail	At risk	
	ncomp	drop-in	p_{event_C}	p_{event_E}	a_j	L	E	A_E	A_C
rand						0	0	1	0
1	0.874	0.427	4.035	2.955	0.167	0.1634	0.0283	0.8018	0.0071
2	0.874	0.427	4.035	2.955	0.250	0.3621	0.0508	0.5768	0.0103
3	0.874	0.427	4.035	2.955	0.400	0.5917	0.0666	0.3329	0.0088
4	0.874	0.427	2.478	1.858	0.333	0.7040	0.0724	0.2159	0.0076
5	0.874	0.427	2.478	1.858	0.667	0.8517	0.0759	0.0693	0.0031
6	0.874	0.427	2.478	1.858	1	0.9230	0.0770	0.0000	0.0000
rand						0	0	1	0
1	0.874	0.427	4.035	2.955	0	0.0000	0.0296	0.9619	0.0086
2	0.874	0.427	4.035	2.955	0	0.0000	0.0584	0.9252	0.0164
3	0.874	0.427	4.035	2.955	0	0.0000	0.0864	0.8900	0.0237
4	0.874	0.427	2.478	1.858	0.125	0.1125	0.1030	0.7577	0.0268
5	0.874	0.427	2.478	1.858	0.143	0.2228	0.1172	0.6321	0.0279
6	0.874	0.427	2.478	1.858	0.167	0.3311	0.1291	0.5126	0.0272
7	0.874	0.427	2.369	1.779	0.200	0.4373	0.1384	0.3996	0.0247
8	0.874	0.427	2.369	1.779	0.167	0.5073	0.1458	0.3241	0.0229
9	0.874	0.427	2.369	1.779	0.250	0.5933	0.1517	0.2362	0.0187
10	0.874	0.427	2.369	1.779	0.400	0.6939	0.1559	0.1380	0.0122
11	0.874	0.427	2.369	1.779	0.333	0.7433	0.1584	0.0896	0.0087
12	0.874	0.427	2.369	1.779	0.667	0.8082	0.1599	0.0288	0.0030
13	0.427	0.427	2.154	1.623	1	0.8397	0.1603	0.0000	0.0000

Extending the example to show an interim analysis at month 13, the recruitment is assumed to extend to month 10 at the maximum monthly rate of 150 patients per month. This extension will effect the recruitment pattern for the interim analysis at month 13, but not at month 6.

6. USING THE MODEL TO CALCULATE SAMPLE SIZE

The Markov model will be used to calculate the expected mean and variance of the logrank statistic at each of the planned times of interim analyses. Since repeated numerical integration with these expected means and variances will be used to derive the power relative to a fixed sample size rather than applied to a sample size formula, the derivation is somewhat different from that given in Lakatos [4]. We use the logrank statistic to test $H_0: (1-F)=(1-G)$, where F and G are the failure-time distributions. The alternative under consideration is $H_1: (1-F)\neq(1-G)$. For the ith interim analysis, the weighted logrank statistic and its variance can be expressed (see Schoenfeld [19]) as

$$L_w^i = \sum_{j=1}^{d^i} w_j^i \left(X_j^i - \frac{m_j^i}{m_j^i + n_j^i} \right) \tag{6}$$

and

$$V_w^i = \sum_{j=1}^{d^i} (w_j^i)^2 \left(\frac{m_j^i n_j^i}{(m_j^i + n_j^i)^2} \right) \tag{7}$$

where X_j^i is 1 for the control group, 0 otherwise, the sum is over all deaths, w_j^i is a suitably chosen weight, and m_j^i and n_j^i are the numbers at risk just prior to the jth death in the experimental and control groups, respectively. The logrank statistic is obtained by letting $w_j^i \equiv 1$.

If $\phi_j^i = m_j^i / n_j^i$ is the ratio of patients at risk just before the jth event, and θ_j^i is the ratio of hazards just prior to the jth event, then the expectation of the logrank statistic and its variance are approximately

$$E(L_w^i) = \sum_{j=1}^{d^i} w_j^i \left(\frac{\phi_j^i \theta_j^i}{1 + \phi_j^i \theta_j^i} - \frac{\phi_j^i}{1 + \phi_j^i} \right) \tag{8}$$

$$E(V_w^i) = \sum_{j=1}^{d^i} (w_j^i)^2 \frac{\phi_j^i}{(1 + \phi_j^i)^2} \tag{9}$$

Setting up the time intervals $[t_1, t_2], \ldots, [t_{k^i - 1}, t_{k^i}]$ to coincide with those defined for the Markov process, equation (8) can be rewritten as

$$E(L_w^i) = \sum_{h=1}^{k^i} \sum_{j=1}^{d_h^i} w_j^i \left(\frac{\phi_{hj}^i \theta_{hj}^i}{1 + \phi_{hj}^i \theta_{hj}^i} - \frac{\phi_{hj}^i}{1 + \phi_{hj}^i} \right) \tag{10}$$

where the first sum is over the k^i intervals at the ith interim analysis, and the second sum is over the d_h^i events of the hth interval. By letting the length of the subintervals get small so that ϕ_{hj}^i and θ_{hj}^i can be assumed constant within each subinterval

$$E(L_w^i) = d^i \sum_{h=1}^{k^i} \rho_h^i w_h^i \gamma_h^i$$

where

$$\gamma_h^i = \frac{\phi_h^i \theta_h^i}{1 + \phi_h^i \theta_h^i} - \frac{\phi_h^i}{1 + \phi_h^i} \quad \text{and} \quad \rho_h^i = \frac{d_h^i}{\sum_j d_j^i}$$

Denote the cumulative probability of being allocated to the treatment group by Q_E, and the cumulative probability of failing in that group by the ith interim analysis by p_{Ei}^i. Then $E(d^i) = N(R(t^i)(Q_E p_{E_i}^i + Q_C p_{C_i}^i))$, where $R(t^i)$ is the proportion of patients recruited by the ith interim analysis, and

$$E(L_w^i) = N(R(t^i)(Q_E p_{Ei}^i + Q_C p_{C_i}^i)) \sum_{h=1}^{k^i} \rho_h^i w_h^i \gamma_h^i \tag{11}$$

Similarly

$$E(V_w^i) = N(R(t^i)(Q_E p_{Ei}^i + Q_C p_{Ci}^i)) \sum_{h=1}^{k^i} \rho_h^i w_h^i \xi_h^i \tag{12}$$

where

$$\zeta_h^i = \frac{\phi_h^i}{(1+\phi_h^i)^2}$$

At each interim analysis, the increment in the expected value of the logrank statistic is $E(L_w^i) - E(L_w^{i-1})$. If the weighting is independent of the recruitment process, then under the null hypothesis, Tsiatis [20] showed that the increments of the expected value of the variance of these statistics are uncorrelated and thus given by $E(V_w^i) - E(V_w^{i-1})$. Although this independence does not necessarily hold under the alternative hypothesis, for local alternatives, the independence assumption usually provides a reasonable basis for calculation (see Kim and Tsiatis [6], for example). The validity of this assumption is checked is Section 7 using simulations. Included in the simulations are a broad range of alternatives, including large departures from the local assumption as well as an extreme non-proportional hazards example. With the assumption of independent increments, the power of these hypothesis tests can be computed using numerical integration [15], where the increments for the integration are distributed $N(\mu^i, \sigma^{i2})$, $i = 1, \ldots, I$, with

$$\mu^i = E(L_w^i) - E(L_w^{i-1}) \tag{13}$$

$$\sigma^{i2} = E(V_w^i) - E(V_w^{i-1}) \tag{14}$$

and

$$E(L_w^0) = E(V_w^0) = 0$$

Substituting (11) into (13) gives (15). Note that the right hand side of (15) is completely

$$\frac{\mu^i}{N} = (R(t^i)(Q_E P_{Ei}^i + Q_C P_{Ci}^i))\sum_{l=1}^{k^i} \gamma_l^i - (R(t^{i-1})(Q_E P_{Ei-1}^{i-1} + Q_C P_{Ci-1}^{i-1}))\sum_{h=1}^{k^{i-1}} \gamma_h^{i-1} \tag{15}$$

determined by the Markov model. A similar equation for the variance can be obtained by substituting (12) into (14). Thus, while repeated numerical integration is needed to find N corresponding to a prespecified power, the Markov model need only be evaluated once.

To use the RALES example, when the recruitment pattern is specified for the Markov model, the sample size has yet to be determined. For instance, after a preliminary sample size calculation based on uniform patient entry, one may ask the medical monitor the expected rate of enrolment. The response is usually in terms of patients, for example 50 patients the first month, 80 the next etc. These numbers reflect the relative rates of enrolment expected, but generally will not add up to the sample size which has yet to be determined. The computer program converts these to relative (recruitment probabilities) rather than absolute numbers. If there are analyses at 6, 13 and 60 months, then after applying the methods just developed, the sample size providing 90 per cent power is 582 per group, and the total deaths is 744. The number of deaths expected at the 13 month interim analysis is then $(0.1603 + 0.2080) \times 582 = 214$. (The 0.1603 is from Table IV, while 0.2080 is from the control group model which is not

Table V. Expected monthly recruitment pattern.

From start of month	0	3	6	9	12	15	24
To just before month	3	6	9	12	15	24	60
Per cent expected monthly rate	10	20	40	60	80	100	0

Table VI. Projected operating characteristics of example trial.

Month	Markov model		Group sequential calculations				
	info frac	% rcrt	alpha	boundary	power	num rcrt	deaths
6	0.0087	5.9	0.00000	5.0000	0.00	73	6
12	0.0517	25.5	0.00000	5.0000	0.00	317	37
18	0.1588	60.8	0.00000	5.0000	0.03	756	114
24	0.3358	100	0.00036	3.3797	12.02	1244	240
30	0.5021	100	0.00284	2.7820	43.40	1244	359
36	0.6359	100	0.00699	2.5147	64.36	1244	455
42	0.7481	100	0.01173	2.3604	76.29	1244	535
48	0.8427	100	0.01638	2.2641	83.06	1244	603
54	0.9253	100	0.02080	2.1935	87.24	1244	662
60	1.000	100	0.02500	2.1387	90.00	1244	716

displayed.) Since only 50 per cent of the patients are expected to enrol by the sixth month, the number of deaths expected is $(0.0770+0.1029)\times 291 = 52$. The information fractions (see (1)) at months 6 and 13 are then 0.070 ($=52/744$) and 0.288 ($=214/744$).

Continuing with the RALES example, selected results of modelling interim analyses at six-month intervals of a 60 month trial are recorded in Table VI; the assumed recruitment pattern for this example is now changed to that shown in Table V, where randomization is at the start of month 0.

With recruitment now extending beyond the times of the first interim analyses, the sample size available at a given interim analysis must be based on the proportion of the total sample size expected to be recruited by the time of that interim analysis. This is calculated by forming the monthly cumulative totals to be recruited based on assumptions of the recruitment pattern, and dividing all of these by the maximum (column 3).

The information fractions in column 2 are derived directly from the Markov model as in the example just presented. The next steps in the procedure are to calculate the expected boundary using the methods of Lan and DeMets. First a spending function is selected (the one corresponding to the O'Brien–Fleming [21] boundary for this example). Evaluating this spending function at each of the information fractions gives the cumulative alpha projected for each of the interim analyses (column 4). The repeated numerical integration methods of Armitage, MacPherson and Rowe (AMR) [15] are then used to calculate the boundary (column 5). In the derivation of this boundary, the null hypothesis is assumed, and the variance is the information fraction. This gives the projected boundary z-values expected to be used at the interim analyses.

Table VII. Comparing the operating characteristics of example with three boundaries.

Month	O'Brien–Fleming use fcn			Pocock use fcn			Exact Pocock		
	Sample size = 1244			Sample size = 1401			Sample size = 1511		
	alpha	bndry	power	alpha	bndry	power	bndry	power	alpha
6	0.00000	5.000	0.00	0.00037	3.380	0.15	2.619	1.5	0.0045
12	0.00000	5.000	0.00	0.00213	2.915	2.67	2.619	6.4	0.0086
18	0.00000	5.000	0.03	0.00603	2.636	17.11	2.619	21.09	0.0124
24	0.00036	3.380	12.02	0.01139	2.479	46.92	2.619	46.83	0.0157
30	0.00284	2.782	43.40	0.01555	2.460	66.38	2.619	65.44	0.0183
36	0.00699	2.515	64.36	0.01846	2.499	76.38	2.619	75.74	0.0203
42	0.01173	2.360	76.29	0.02066	2.477	82.23	2.619	81.85	0.0218
48	0.01638	2.264	83.06	0.02238	2.483	85.84	2.619	85.66	0.0231
54	0.02080	2.194	87.24	0.02379	2.487	88.26	2.619	88.19	0.0241
60	0.02500	2.139	90.00	0.02500	2.488	90.00	2.619	90.00	0.0250

To calculate power, the numerical integration methods [15] are applied again, this time using the boundary just derived under the null hypothesis, but using (13) to specify the alternative. This projected quantification of the alternative, or Brownian drift, comes from the Markov model which includes assumptions regarding the treatment effect. The projected accumulating probability of exceeding the boundary, or power, is given in column 6. In order to evaluate (13), a sample size N must be specified, but the sample size is unknown at this point. A search for the desired sample size which gives the desired power is performed, with the minimum value specified in the computer program being the fixed design sample size for the logrank statistic. For the maximum sample size, the final boundary z-value (2.14 in the example above) replaces the nominal significance level (typically 1.96) in the fixed design sample size calculation for the logrank. A binary search between these two values is used to find the sample size corresponding to the desired power. Since the group sequential sample size may be smaller than the fixed design sample size in some non-proportional hazards situations (see Section 7), the default minimum value for the search can be changed by the user.

Table VII displays results of using different boundaries for the example in Table VI. The boundary for the Pocock [23] use function is not constant here. Pocock's original approach had a boundary of constant z-values with interim analyses occurring at equal increments of information. The Lan–DeMets use function corresponding to Pocock's approach was intended to provide an approximately constant boundary with equal increments of information. When analyses are performed at unequal increments of information, the departure from constancy can be quite large. The 'exact Pocock' boundary provides a constant boundary given the projected information fractions. To achieve this, a constant boundary value is selected, and numerical integration methods are applied, using the projected information fractions. If the overall significance level is too high, a higher constant boundary value is selected and the process repeated. Note that this approach bypasses the spending function; the increments of alpha are never specified. The spending-function approach is *not* needed in the above procedure. Any boundary that can be specified at the potentially unequal increments of information can

Table VIII. Comparing sample sizes for logrank statistic.

	Exponential	Markov
Assumptions		
Recruitment	Uniform in 2 years	As specified in Table V
Non-compliance	0.0	10% first year; 5% years 2–5
Drop-in	0.0	5% years 1–5
Loss-to-follow-up	0.0	0.0
p_C	27% years 1–5	As specified in last column Table I.
p_E	$(1 - 0.225)p_C$	$(1 - 0.225)p_C$
Calculated sample sizes		
Fixed	743	1221
O'Brien–Fleming	783	1244
Pocock	911	1401

be used. Pocock's specification is in terms of the constancy of the z-values, and does not involve the increments of alpha. In the process of applying the methods of AMR, the increments of alpha are produced (last column of Table VII), and these can be used to create a spending function with the desired constancy if interim analyses are performed at the increments of information just used. To do this, plot the cumulative alphas against the cumulative increments of information; to create a spending function, connect the nodes, either with straight lines or smoothed curves. This procedure can be used with the O'Brien–Fleming boundary as well which requires constant $z\sqrt{\text{info frac}}$, in contrast to the constant z in Pocock's procedure. The spending function produced will not be exactly as given in Lan and DeMets, which is derived based on theoretical considerations of Brownian motion.

Table VIII compares sample sizes obtained using the Markov model versus the exponential model. Both the original designers of RALES and the designers of a more recent trial of carvedilol in CHF [24] assumed exponential models. The Markov model used the control group piecewise exponential event rate as presented in the last column of Table I.

The original planners of RALES used 27 per cent, a combination of the first-year rates of SOLVD and CONCENSUS, while the trial of carvedilol used 28 per cent [24]; 27 per cent was used as the comparable yearly control-group failure rate for the exponential model. The increase in sample size due to taking the interim analyses into account is considerably larger for the Pocock boundary than the O'Brien–Fleming. In this example, the use of group-specific piecewise exponential models when modelling survival, recruitment, non-compliance and drop-in has a much larger impact on the sample sizes than does adjustment for the group sequential factor.

7. SIMULATIONS

The performance of the proposed methods is investigated through simulations. First, a rather extreme example of non-proportional hazards is examined. This is followed by a more routine range of clinical trial assumptions in an exponential setting.

Table IX. Sample sizes, simulated and calculated powers for example 2.

	Sample size		Power			
	Trial 1	Trial 2	Trial 1		Trial 2	
			Calculated	Simulated	Calculated	Simulated
Fixed	608	838				
O'Brien–Fleming	516	890	0.896	0.904	0.900	0.903
Pocock	430	1099	0.901	0.921	0.902	0.904

7.1. Example 2

Consider two hypothetical trials in which accrual takes place uniformly during the first year, and interim analyses are planned every year of these five-year trials. The control group failure rate is constant at 0.09 per year, while the treatment rate is non-constant in both trials. In trial 1, the yearly treatment rate is 0.03 for each of the first two years and 0.08 for each of the remaining three years. In trial 2, the yearly treatment rate is 0.08 for the first two years, 0.056 for year three, and 0.03 for each of the remaining two years. Table IX gives the results of using the methods presented in this paper along with simulation verification.

The overall failure rates in the control group in both trials are identical, as are the treatment group rates. Thus, the phenomenon giving rise to the different sample sizes relative to the fixed design must be the non-proportionality of the hazards. Thus, where there is expected non-proportionality, such non-proportionality is important not only in the sample size for the fixed design, but also in determining the relative effect of the group sequential design. The explanation for the sample size being smaller for the group sequential design as compared to the fixed design is as follows. Events occurring during time intervals in which the hazards are close together contribute to the variance, but little to the mean of the statistic, introducing noise, but no signal. Consequently, in trial 1, where the hazards come together after early separation, the logrank statistic is larger at the end of the period of large separation than at the end of the trial. By reallocating some of the alpha to the early part of the trial, the group sequential test takes advantage of the larger test statistic.

7.2. Additional simulations

The use of the Markov model for estimation of the mean and variance of the logrank statistic has been verified extensively by simulation (Lakatos and Lan [31]). A wide range of exponential, proportional and non-proportional hazards were evaluated. The purpose of the simulations in this section is to test the assumption of independent increments. Because the increments are independent under the null hypothesis (Tsiatis [20]), the focus of these simulations is to test the robustness as the alternatives become less local. Exponential models are used because the degree of departure from the null is easily quantified, and a non-proportional hazards model whose most extreme hazard ratio is near one of the tested exponential models can be considered at least as local.

The parameters for the simulation are defined as follows: each trial is 10 years long with the number of years of accrual given in the table; P_C is the ten-year control group failure rate; θ is the hazard ratio, and accrual is given in years. Since P_C is a ten-year rate, and

Table X. Evaluation of robustness of independent increments assumption.

Parameters			Sample size			Simulated power		
P_C	θ	Accrual	Fixed	Pocock	O'Brien–Fleming	Fixed	Pocock	O'Brien–Fleming
0.2	0.67	1	1614	1916	1664	90.2	89.9	90.3
0.2	0.67	5	2013	2390	2060	90.6	90.4	90.7
0.2	0.67	9	2715	3223	2718	90.3	90.3	90.4
0.2	0.50	1	636	755	655	90.1	90.3	90.2
0.2	0.50	5	795	943	819	90.7	90.5	90.7
0.2	0.50	9	1076	1277	1076	90.1	90.2	90.4
0.2	0.25	1	229	271	236	91.9	91.5	91.8
0.2	0.25	5	288	34	297	92.2	91.8	92.1
0.2	0.25	9	391	463	391	91.6	91.7	91.6
0.8	0.67	1	359	436	370	89.6	90.0	90.1
0.8	0.67	5	413	490	425	90.5	90.4	90.3
0.8	0.67	9	527	625	527	89.8	90.2	90.3
0.8	0.50	1	133	161	141	89.7	90.1	89.8
0.8	0.50	5	155	188	159	89.9	90.3	90.0
0.8	0.50	9	199	235	205	89.7	90.3	90.2
0.8	0.25	1	42	53	44	90.2	90.5	90.4
0.8	0.25	5	50	62	53	90.6	90.0	90.1
0.8	0.25	9	66	78	68	90.2	89.8	90.2

each trial is ten years long, the time unit is nominal. Only the unweighted logrank statistic is tested. The results presented in Table X demonstrate that the method gives good results under a wide range of alternatives. Each power is based on 5000 simulated trials.

8. TREATMENT LAG

Lag in treatment effect refers to situations in which the full effect of treatment does not occur immediately but increases gradually over a period of weeks, months or even years. The effect of treatment lag on sample size was investigated as far back as 1968 by Halperin *et al.* [1], and later in the Markov model approach of Lakatos [3, 4]. One way to adjust sample size for treatment lag is to modify the treatment group event rate over time. Specifically, if ϕ is the full treatment effect, and $l(t)$ is the proportion of the treatment effect achieved by time t, with $l(0)=0$ and $l(1)=1$, then

$$p_E=(\phi l(t)+(1-l(t)))p_C$$

provides for the gradual onset of treatment effect in any desired pattern. Using this adjusted experimental group rate in the Markov model incorporates lag into the treatment effect, while allowing further adjustment for the other factors. A limitation of this approach is that non-compliers lose all benefit at the time of non-compliance, rather than following a more gradual offset mirroring the onset of efficacy. Similarly, control group patients who drop-in are modelled as receiving the proportion of treatment effect $l(t_{\text{drop-in}})$ at the time of drop-in, rather than $l(0)$. The lag Markov model of Lakatos [3, 4] does not have this problem. While

the Markov model can be adapted to the group-sequential setting in exactly the same way as was done for the non-lag model in this paper, the computer program for lag is not currently available. However, use of the lag function $l(t)$ should provide a reasonable approximation.

9. WEIGHTED LOGRANK STATISTICS

When hazards are non-proportional, weighted versions of the logrank statistic can provide increased power. Both Gehan [25] and Prentice [26] proposed weighted logrank statistics that are generalizations of Wilcoxon's statistic for censored survival data. These place more weight on earlier events; at each event, Gehan weights by the proportion of patients at risk immediately prior to the event, while Prentice weights by the Kaplan–Meier estimate of survival immediately prior to the event. Although either weighting is easy to achieve with the Markov model, Gehan's has some undesirable properties (see, for example Prentice and Marek [27]); only Prentice's will be considered here. The probability of surviving at any given time in the experimental group is estimated by $1 - p_E$ (see equation (2) and Table II), so the probability pooled over the two treatment groups is estimated by $1 - (p_E - p_C)/2$. The Harrington and Fleming [28] class of weighted logrank statistics use $w_i = (1 - (p_E - p_C)/2)^\gamma$, $1 \geqslant \gamma \geqslant 0$, so that $\gamma = 1$ gives the Wilcoxon statistic and $\gamma = 0$ the logrank. Sample sizes for any member of this class are readily evaluated as an option in the computer program. More complex weightings such as those proposed by Zucker and Lakatos [29] require the user to provide IML code.

Referring to the RALES example, the sample sizes required for the fixed-sample design are 1221 for the unweighted logrank statistic and 1195 for the Wilcoxon. This is not surprising since the hazard ratio diminishes as this trial progresses. In contrast, the group-sequential sample size for the unweighted logrank is 1244 and 1249 for the Wilcoxon. This unexpected reversal can be explained in part by examining the information-time–calendar-time transformation. Information accrues faster with the Wilcoxon, leaving relatively less alpha to be spent in the last few interims. Since the last few analyses typically have a large impact on power, having less alpha available towards the end decreases power. If the O'Brien–Fleming spending function $\alpha^*_{OB\text{-}F}$ is replaced by $\alpha^{**} = (\alpha^*_{OB-F})^{1.5}/(0.025)^{0.5}$, then the spending under α^{**} with the Wilcoxon is similar to the spending under $\alpha^*_{OB\text{-}F}$ with the logrank (see Table XI). With α^{**}, the sample size for the Wilcoxon is 1225.

In this example, the impact of such factors as non-compliance and varying hazard rates is far more important than the choice of weightings of the logrank statistic (see Table VIII).

10. DISCUSSION

The importance of sample size and power calculations for clinical trials is well established, as is the need to consider factors such as non-compliance, loss to competing risks or follow-up, non-proportional hazards and the like. With group sequential designs, these factors are equally important for sample size and power. From a broader perspective, the same factors lead to important design considerations beyond sample size and power. In designing the group sequential plan, the accrual of information is of fundamental importance. The RALES trial, as originally designed, scheduled interim analyses to take place at 20, 40, 60, 80

Table XI. Predicted spending of alpha.

Look	Month	Spending function		
		$\alpha^{*}_{OB\text{-}F}$ Logrank	$\alpha^{*}_{OB\text{-}F}$ Wilcoxon	α^{**} Wilcoxon
1	6	0	0	0
2	12	0	0	0
3	18	0	0	0
4	24	0	0.001	0
5	30	0.003	0.005	0.003
6	36	0.007	0.010	0.007
7	42	0.012	0.015	0.012
8	48	0.016	0.019	0.016
9	54	0.021	0.022	0.021
10	60	0.025	0.025	0.025
Sample size		1244	1249	1225

and 100 per cent of planned total deaths. Using the calendar-time–information-time (CT–IT) transformation, the calendar time of these analyses would have been 19, 26, 34, 45 and 60 months. The long hiatus during the very crucial late period of the trial raises ethical concerns, and the irregular schedule is logistically difficult. However, there are advantages to planning the trial on an information basis. The CT–IT transformation derived in this paper can be used to design the trial on an information basis that satisfies ethical and logistic concerns. Further, in Section 9, when unexpected results arose relating to sample sizes for the Wilcoxon statistic, the CT–IT transformation revealed that the underlying cause was faster than anticipated accrual of information. This suggested a revised spending function more appropriate for this rate of accrual of information, solving the problem.

Simulations are useful in verifying sample sizes. Simulations often require substantial computing time, although this may become less of an issue as computing speed is rapidly increasing. The simulation programs for group sequential designs can be quite complex and should be verified independently. The simulation program by Gu and Lai [10] is much more flexible than that of Halpern and Brown [7], but the current version appears to have an error (Lakatos [30]).

The Markov model for fixed sample designs has been verified independently (see Lakatos and Lan [31]), and appears to be quite accurate across a broad range of designs. For the fixed design, the simulations verify the asymptotic formulae, as well as the discrete character of the Markov model. Additional verification of the independent increments assumption is needed in the group sequential case. A simple simulation program was written for this purpose, and the results calculated using the Markov model were in reasonably agreement for exponential models across a wide range of hazard ratios. The Markov model also agreed quite closely when applied to the decidedly non-proportional hazards situation of example 2.

Using the SAS IML programs for designing group sequential trials can take considerably more computing time (CPU) than the corresponding fixed-design sample size IML programs. The times given below reflect the CPU times for the computer runs for this paper, for which a Sun Ultra 450, which is quite fast, was used. In addition to the ten interim analysis trials,

programs for the same trials with five interim analyses were run to establish CPU times for more typical settings. The time required to run the Markov portion of the program, including evaluating the fixed logrank sample size and calendar-time–information-time transformation was always less than one second. The numerical integration portion requires substantially more time, and this depends on the user-specified precision as well as the number of interim analyses. Calculating the boundary usually took an additional 0.5 seconds (total about 1.3 seconds). Power calculations involve the most CPU time. Once the boundary is calculated, the default starting sample size used for power calculations is immediately available from the simple conservative approximation

$$N_{\text{gpseq}} = N_{\text{fixed}} \left(\frac{z_{\text{last}} + z_{1-\beta}}{z_{1-\alpha/2} + z_{1-\beta}} \right)^2$$

where z_{last} is the critical value at the last interim analysis, and α is the desired overall alpha. This approximation is usually quite good for the O'Brien–Fleming boundary, and is a reasonable way to explore sample size assumptions with slower computers. For precision of $\pm 1/2$ per cent (that is, 95 per cent confidence interval for power ≈ 89.5–90.5), CPU times ranged from 1.5 to 13 seconds. For this paper, precision was set so that all powers would be 90.00 (± 0.000049), and computing times ranged from 6.2 to 27 seconds. CPU times could increase substantially on slower computers, in which case, the strategy of using less precise estimates (or the above formula) during the exploratory stage is recommended.

REFERENCES

1. Halperin M, Rogot E, Gurian J, Ederer F. Sample sizes for medical trials with special reference to long-term therapy. *Journal of Chronic Diseases* 1968; **21**:13–24.
2. Wu M, Fisher M, DeMets D. Sample sizes for long-term medical trials with time-dependent non-compliance and event rates. *Controlled Clinical Trials* 1980; **1**:109–121.
3. Lakatos E. Sample sizes for clinical trials with time-dependent rates of losses and non-compliance. *Controlled Clinical Trials* 1986; **7**:189–199.
4. Lakatos E. Sample sizes based on the logrank statistic in complex clinical trials. *Biometrics* 1988; **44**:229–241.
5. Halpern J, Brown WB. Designing clinical trials with arbitrary specification of survival functions and for the logrank or generalized Wilcoxon test. *Controlled Clinical Trials* 1987; **8**:177–189.
6. Kim K, Tsiatis AA. Study duration for clinical trials with survival response and early stopping. *Biometrics* 1990; **46**:81–92.
7. Halpern J, Brown WB. Computer program for designing clinical trials with arbitrary survival curves and group sequential testing. *Controlled Clinical Trials* 1993; **14**:109–122.
8. Scharfstein DO, Tsiatis AA. The use of simulation and bootstrap in information-based group sequential studies. *Statistics in Medicine* 1998; **17**:75–87.
9. Mehta C. EaST 2000. Cytel Software Corporation, Cambridge, 2000.
10. Gu M, Lai TL. Determination of power and sample size in the design of clinical trials with failure-time endpoints and interim analyses. *Controlled Clinical Trials* 1999; **20**:423–438.
11. Emerson SS. Statistical packages for group sequential methods. *American Statistician* 1996; **50**:183–192.
12. SAS Institute Inc. *SAS/IML Software: Usage and Reference, Version 6, First Edition.* SAS Institute Inc., Cary, NC, 1989.
13. Lan KKG, Zucker DM. Sequential monitoring of clinical trials: the role of information and Brownian motion. *Statistics in Medicine* 1993; **12**:753–765.
14. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**:659–663.
15. Armitage P, McPherson CK, Rowe BC. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society, Series A* 1969; **132**:235–214.
16. Pitt B, Faiez Z, Remme WJ, Cody R, Castaigne A, Perez A, Palensky J, Wittes J. The effect of Spironolactone on morbidity and mortality in patients with severe heart failure. *New England Journal of Medicine* 1999; **341**(10):709–717.
17. The SOLVD investigators. Effect of enalapril on survival in patients with reduced left ventricular ejection fraction and congestive heart failure. *New England Journal of Medicine* 1991; **325**:293–302.

18. The CONCENSUS Trial Study Group. Effects of enalapril on mortality in severe congestive heart failure. *New England Journal of Medicine* 1987; **316**:1429–1435.
19. Schoenfeld D. The asymptotic properties of non-parametric test for comparing survival distributions. *Biometrika* 1981; **68**:316–318.
20. Tsiatis AA. Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal American Statistical Association* 1982; **77**:855–861.
21. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; **35**:549–556.
22. Kaplan EL, Meier P. Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958; **53**:457–481.
23. Pocock SJ. Group sequential methods in the design of clinical trials. *Biometrika* 1977; **64**:191–199.
24. Packer M, Coates AJS, Fowler MB, Katus HA, Krum H, Mohacsi P, Rouleau JL, Tendera M, Castaigne A, Roecker EB, Schultz MK, DeMets DL. Effect of carvedilol on survival in severe chronic heart failure. *New England Journal of Medicine* 2001; **344**:1651–1658.
25. Gehan EA. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* 1965; **52**:203–223.
26. Prentice RL. Linear rank test with right-censored data. *Biometrika* 1978; **65**:167–179.
27. Prentice RL, Marek PA. Qualitative discrepancy between censored data rank tests. *Biometrics* 1978; **35**: 861–886.
28. Harrington DP, Fleming TR. A class of rank test procedures for censored survival data. *Biometrika* 1982; **69**:133–143.
29. Zucker DL, Lakatos E. Weighted logrank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika* 1990; **77**:853–864.
30. Lakatos E. Letter to the editor. *Controlled Clinical Trials* (in press).
31. Lakatos E, Lan KKG. A comparison of sample size methods for the logrank statistic. *Statistics in Medicine* 1992; **11**:179–191.

