STATISTICS IN MEDICINE, VOL. 11, 179–191 (1992)

A COMPARISON OF SAMPLE SIZE METHODS FOR THE LOGRANK STATISTIC

EDWARD LAKATOS

Biostatistics Research Branch, National Heart, Lung, and Blood Institute, Bethesda, Maryland 20892, U.S.A.

AND

K. K. GORDON LAN

Department of Statistics/Computer and Information Systems, George Washington University, Washington, D.C. 20052, U.S.A.

SUMMARY

Several methods are available for sample size calculation for clinical trials when survival curves are to be compared using the logrank statistic. We discuss advantages and disadvantages of some of these methods, and present simulation results under exponential, proportional hazards and non-proportional hazard situations.

1. INTRODUCTION

Calculation of sample size is important in designing experiments. In clinical trials in which time to event is compared between two groups, the logrank test is generally used. In this paper, we compare several methods of evaluating sample size and discuss the effects of the assumptions made in these procedures. In practice, methods are sometimes applied even though assumptions are violated or theoretical justification is lacking. It is common to see binomial sample size calculation when the intended analysis will compare survival curves, or exponential sample size calculation when the proportional hazards assumption is not realistic. We will examine the results of using some sample size methods in situations for which they were not designed. Our focus will be on methods that compare survival curves.

We approach this investigation with three goals in mind. First, we wish to test the various methods when the assumptions of their derivations are satisfied. Of interest in this case is to determine the extent to which asymptotic approximations become less accurate as treatment differences become more extreme. Secondly, we wish to investigate the extent to which the use of methods in contexts different from those intended leads to inaccurate sample sizes. The methods which are easiest to use tend to be very restrictive in their assumptions. Thus there is a natural tendency to assume that some violation of the assumptions will not matter too much, with the consequence that methods may be applied without realization of the degree of inaccuracy generated. Finally, because the Markov method of Lakatos[1] is very general, we test it in a number of non-proportional hazards situations to determine if it can provide accurate sample sizes.

0277–6715/92/020179–13$06.50

Received January 1990
Revised April 1991

2. DESCRIPTION OF THE METHODS

Rubinstein, Gail and Santner[2] (RGS) discuss using exponentially distributed survival durations for trials 'which are to be analysed with the Mantel–Haenszel test'. They assume that: patients are accrued during the interval $[0, T]$ according to a Poisson process and total trial length is $T + \tau$; times from entry to event are independently and exponentially distributed within each treatment with parameters λ_E and λ_C; and loss to follow-up times are also independently and exponentially distributed, with parameters ϕ_E and ϕ_C. Formulas (A2) and (2) in their paper, based on the exponential maximum likelihood test, can be solved explicitly for sample size giving:

$$N = \left(\frac{z_\alpha + z_\beta}{\ln(\theta)}\right)^2 [1/E(P_C) + 1/E(P_E)]$$

where

$$E(P_i) = [1 - e^{-\lambda^* i\tau}(1 - e^{-\lambda^* i T})/\lambda_i^* T]\lambda_i/\lambda_i^*, \quad i = C, E$$

and where N is the sample size in each group, θ is the hazard ratio, and $\lambda_i^* = \lambda_i + \phi_i$. Rubinstein *et al.* note that their work builds on the work of Pasternak and Gilbert[3] and George and Desu,[4] and, that their approach has been incorporated in a computer program by Bernstein and Lagakos.[5] Lachin[6] and Lachin and Foulkes[7] also derive sample sizes under the exponential assumption. We have chosen Rubinstein *et al.* as representative of exponential methods.

Freedman[8] derives a sample size formula by considering the expected value and variance of the logrank statistic. Assuming that the hazard ratio θ is constant throughout the trial, and further that the ratio of patients at risk, just before a given event, in the two groups is constant, a formula is derived for the total number of events needed to be observed: $d = \{(\theta + 1)/(\theta - 1)\}^2 (z_\alpha + z_\beta)^2$. The number of patients per group can then be calculated as $n = d/(p_E + p_C)$, where p_E and p_C are the proportions of failures in the two groups. This computation is based on simultaneous entry. Freedman suggests approximating the sample size for extended recruitment by that derived from simultaneous entry by taking the values of p_E and p_C at a duration equal to the average length of follow-up. This may be estimated by adding one-half of the accrual period to the post accrual follow-up period.

The logrank statistic can still be used for analysis when, as in many trials, the proportional hazards assumption is violated. Lakatos[1] derives the sample size required for the logrank statistic in this general case by using a discrete non-stationary Markov process that allows any pattern of survival, non-compliance, loss to follow-up, drop-in and lag in the effectiveness of treatment during the course of a trial. The expected value and variance of the logrank statistic is calculated using the ratios of the hazards and proportions at risk at each discrete interval. By making the interval between transitions sufficiently short, and assuming the hazard ratio and ratio of patients at risk are constant within each interval, the sample size can be calculated.

In the Lakatos method, event rates, non-compliance, drop-in and loss to follow-up correspond to survival functions which can be specified in one of two ways. First, the user may supply the probability of failing in each period j, for those tho survived to the end of the previous period $j - 1$, and it is assumed that the survival function is exponential within each interval. Both the length and number of periods are chosen by the user. Secondly, the user may specify the survival function in closed form $S(t)$, so that the transition probabilities can be calculated directly as $1 - S(t_j)/S(t_{j-1})$. Time lag in the effectiveness of treatment, and varying rates of accrual may also be incorporated.

It is assumed throughout that the sample size is equal in the two groups, but any of the methods can be modified to handle unequal sample sizes.

Finally, while each of the methods yields sample sizes directly, all require iterative methods to find trial duration.

3. RELATIONSHIP BETWEEN PROPORTIONAL HAZARDS AND EXPONENTIALITY

If the survival distributions are exponential, then the proportional hazards assumption is satisfied. It can be shown that after an appropriate time transformation, the converse is true. However, the nature of how this time transformation interacts with accrual, lags, and other time-linked parameters, can substantially affect the sample size estimate. The proportional hazards model has been discussed in detail by Kalbfleisch and Prentice.[9] The explanation that follows should allow the reader to determine the circumstances under which use of the exponential model could produce poor estimates. Suppose survival time X has distribution F_C and Y has distribution F_E, and the corresponding hazard functions satisfy $\lambda_X(t)/\lambda_Y(t) = \theta$ for all t. Then $1 - F_E = (1 - F_C)^\theta$. Define $H_\gamma(x) = 1 - e^{-\gamma x}$, the distribution function of an exponentially distributed variable with parameter γ. Then $H_\gamma^{-1}(Y) = -(1/\gamma)\ln(1 - Y)$. Since $F_C(X)$ has the uniform distribution, $H_C^{-1}(F_C(X))$ has the unit exponential distribution. Now

$$H_C^{-1}F_C(Y) \sim -\ln(1 - F_C(Y)) = -(1/\theta)\ln(1 - F_E(Y)) = H_\theta^{-1}(F_E(Y))$$

which is exponentially distributed with parameter θ. Thus the function $g(t) = H_C^{-1}(F_C(t))$ transforms the time scale so that $g(X)$ and $g(Y)$ are distributed exponentially with parameters 1 and θ, respectively. The transformation $g(t)$ can be simplified as

$$\begin{aligned} g(t) = H_C^{-1}(F_C(t)) &= -\ln(1 - F_C(t)) \\ &= -\ln\exp\left(-\int_0^t \lambda_C(u)\mathrm{d}u\right) \\ &= \int_0^t \lambda_C(u)\mathrm{d}u = \Lambda_C(t) \end{aligned}$$

the cumulative hazard function of X. This expression shows that the transformation can be viewed as shrinking or expanding of various portions of the time scale t to arrive at a scale t', so that the cumulative hazard of $g(X)$ at t' is equal to t'. Since $\lambda_E(t)/\lambda_C(t) = \theta$, the cumulative hazard of $g(Y)$ at t' is $\theta t'$. The monotonicity of g assures that any rank statistic based on the failure times will be invariant under this transformation. Thus, if the proportional hazards assumption holds, there exists a transformation of the time scale such that the failure times in both treatment groups are exponentially distributed, and the sample sizes based on a rank statistic of these new distributions are equivalent to sample sizes based on the rank statistic of the original distribution.

The discussion above considers only the survival distributions, but not factors such as censoring, loss to follow-up, and duration of the trial. These factors can be very important because they determine the number of failures in the trial which in turn determines the variance used in the sample size calculation. For example, the time pattern of entry of patients into the trial affects the number of failures. If the original failure time distributions have proportional hazards, but are not exponential, then the transformation g may radically change the censoring distribution. In many trials, the hazard function is not constant. For example, in a trial with extended follow-up after surgery, there may be high early post-operative mortality. Wu *et al.*[10] give another example using data from the Framingham Heart Study.

4. COMPARISON OF METHODS

To evaluate the performance of the three methods we first use the family of survival curves defined by hazard functions of the form $\lambda(t) = 1/(at + b)$. This family, considered by Lan and Lachin,[11] is exponential when $a = 0$ and, when $a \neq 0$, the corresponding survival curve is

$$S(t) = \left(\frac{b}{at + b}\right)^{1/a}.$$

When $a < 0$, we consider only the region $at + b > 0$. To compare the sample size calculations for the curves under the various assumptions, we will compare trials of total length T in which the T year survival is fixed, but in which the hazard function can vary. Figures 1 and 2 present $S(t)$ when $S(T) = 0{\cdot}8$ and $0{\cdot}2$, respectively, for several values of $R = \lambda(T)/\lambda(0)$. The ratio R indicates the degree of departure from constant hazards. The control group survival curve is uniquely specified when the end of trial survival $S_C(T)$ and the ratio $R_C = \lambda_C(T)/\lambda_C(0)$ have been given. Under proportional hazards, the survival curve in the experimental group can be uniquely specified by the hazard ratio $\lambda_E(t)/\lambda_C(t) = \theta$ (equivalently, $S_E(T) = S_C^{\theta}(T)$) since $R_E = R_C$. If the hazards are non-proportional, then $\lambda_C(t)/\lambda_E(t) = \theta(t)$ is a function of time. In this case, the survival functions $S_C(t)$ and $S_E(t)$ are chosen as follows. For the control group, $S_C(t)$ is determined by the values $S_C(T) = 0{\cdot}2$ and $R_C = \lambda_C(T)/\lambda_C(0)$. The alternative is specified by θ so that $S_E(T) = S_C^{\theta}(T)$ and the value of $R_E = \lambda_E(T)/\lambda_E(0)$. While the total trial duration in our comparison is $T = 10$ years, the results presented apply to any value of T, with a specified proportion of T devoted to patient accrual.

To test the Lakatos method under more radical departures from proportional hazards, define the treatment hazard function $\lambda_E(t)$ by

$$\lambda_E(t) = \{\theta l(t) + (1 - l(t))\}\lambda_C(t),$$

where $0 \leqslant l(t) \leqslant 1$ for $0 \leqslant t \leqslant 1$. When $l(t) = 1$, the full treatment effect $\theta = \lambda_E(t)/\lambda_C(t)$ is experienced, while $l(t) = 0$ gives no treatment effect. Zucker and Lakatos[12] considered two forms of the function $l(t)$ designed to represent prototypical lags in the treatment effect. Here, we use the function

$$l(t) = c_0(t - t^2)^3 + c_1 \tanh(c_2(t - 1/2)) + c_3 \sin(c_4 t) + c_5$$

which was chosen to produce diverse treatment effects at least as complex as one is likely to encounter in practical applications. The c_0 term alone produces a treatment effect curve similar to the one shown in Figure 3, case 1, while the c_1 term permits the initial effect to differ from the final effect as shown in Figure 3 cases 3 and 4. In these simulations, the coefficients for cases 1 to 4 have been adjusted so that $l(t)$ attains a maximum of 1·0 (representing the full treatment effect θ) at some point t, $0 < t < 1$. The c_5 term permits different minimal treatment effects (see Figure 3, cases 1 and 2). The c_3 term is added (see Figure 3, cases $1N$–$4N$) to provide additional fluctuations.

In the comparisons presented below, we assume uniform recruitment rates.

4.1. Exponential survival

Table I gives sample sizes for 90 per cent power with two-sided significance level $\alpha = 0{\cdot}05$, comparing the three methods when the curves are exponential. In all tables, the simulated power given for the Lakatos method is based on 5000 simulated trials, so the standard error for the power estimate is approximately $\sqrt{(0{\cdot}9 \times 0{\cdot}1/5000)} \approx 0{\cdot}004$. For a given set of parameters, the

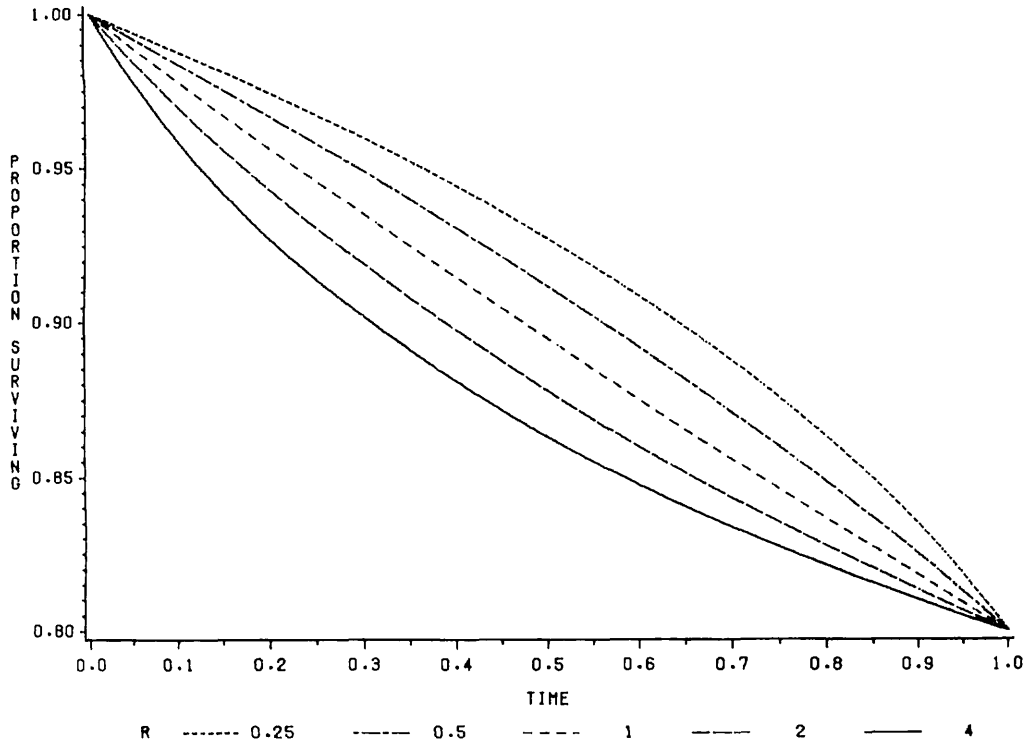

Figure 1. Lan–Lachin survival curves when 20 per cent fail by end of trial

simulation generates sets of logrank statistics which are approximately normally distributed with mean and variance μ and σ^2, say. The usual formula relating power and sample size applies, so that, for the Lakatos method, $\sqrt{N_L} = (z_\alpha + (z_\beta)_L)\sigma/\mu$, and similarly for the other methods. Thus, we can approximate the power for the other methods using the formula

$$\text{pwr}_N = \Phi\{(z_\alpha + \Phi^{-1}(\text{pwr}_L))\sqrt{(N/N_L)} - z_\alpha\}$$

where N is the sample size using the other method and Φ is the standard cumulative normal distribution function.

In Table I, the simulated powers with the exponential assumption using the Lakatos method are all very close to 90, and rise noticeably above 90 only when the hazard ratio is as extreme as 0·25. The other methods, although almost always more conservative than the Lakatos method, yield sample sizes very similar to the Lakatos method (in some cases with $S_C = 0{\cdot}8$, sample sizes using the method of Rubinstein *et al.* are considerably larger). When the sample size is less than 100, the percentage difference in the methods is sometimes large. For example, when $S_C(T) = 0{\cdot}2$, $\theta = 0{\cdot}25$ and accrual = 1, then $N_F = 53$ is 23 per cent higher than $N_L = 43$, In this case, the simulated power for N_F is 95·1 as compared to 90·2 for N_L.

4.2. Proportional hazards with non-exponential survival

Table II is similar to Table I except that Table II has an extra parameter R which determines the curve in the Lan–Lachin family. When $R = 1$, the curve is exponential.

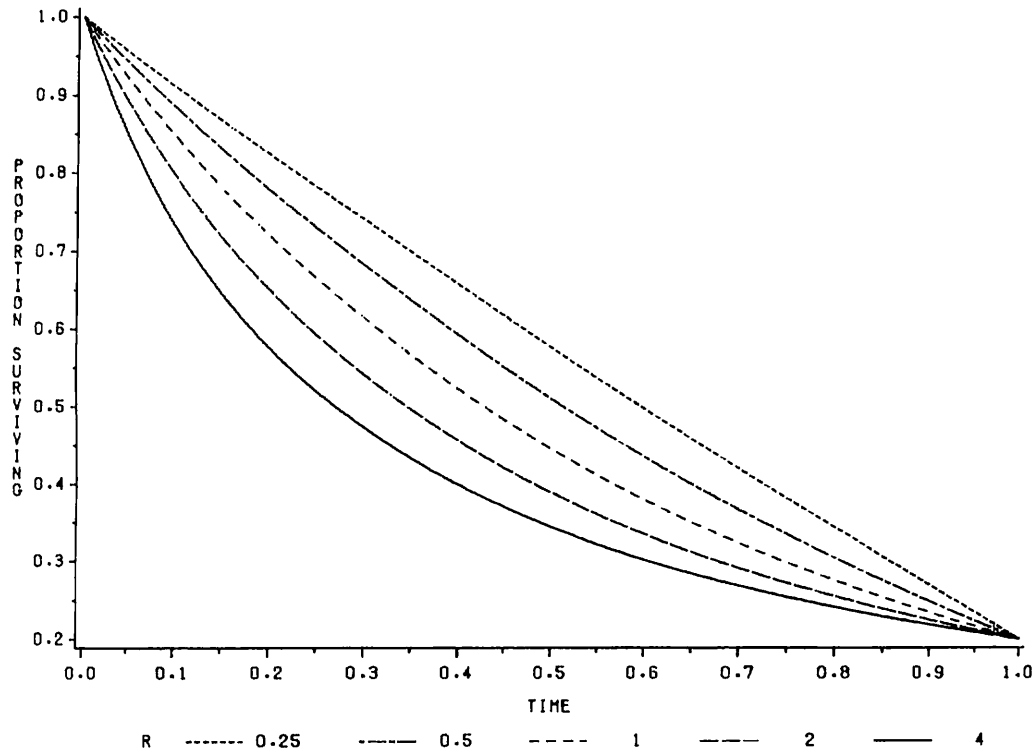

Figure 2. Lan–Lachin survival curves when 80 per cent fail by end of trial

For the methods of Lakatos and Freedman, the results are similar to those obtained using the exponential model with the Lakatos method having power close to 90, increasing modestly as the hazard ratio departs from 1. Freedman's method yields sample sizes which are almost always larger than the Lakatos method. The differential is usually inconsequential except when sample sizes are small. If one assumes an exponential model holds and uses the method of Rubinstein *et al.*, then sample sizes in Table II range from 18 per cent lower than the Lakatos method to 40 per cent higher. In Freedman's method, the approximation of average follow-up time as the sum of the post accrual follow-up time plus one-half of the accrual period may lead to inaccurate estimates of failures. Consider, for instance, row 5 of Table II. The hazard is very low during the early portion of the trial but is much higher towards the end. The use of six years' average follow-up in conjunction with very low hazard during the early portion of the trial will lead to an underestimate of the overall failure rate ($P_c \approx 0{\cdot}092$ instead of the desired 0·099) and in turn to an overestimate of the sample size.

4.3. Non-proportional hazards

If we allow different values of R for the control and experimental groups, then the hazards are no longer proportional. The proportional hazards models used for calculating the sample size table are obtained by setting $S_E(T) = S_C^{\theta}(T)$ and $R = (R_E + R_C)/2$. The results, shown in Table III, indicate that the Lakatos method is very accurate, while the other two methods can be rather inaccurate.

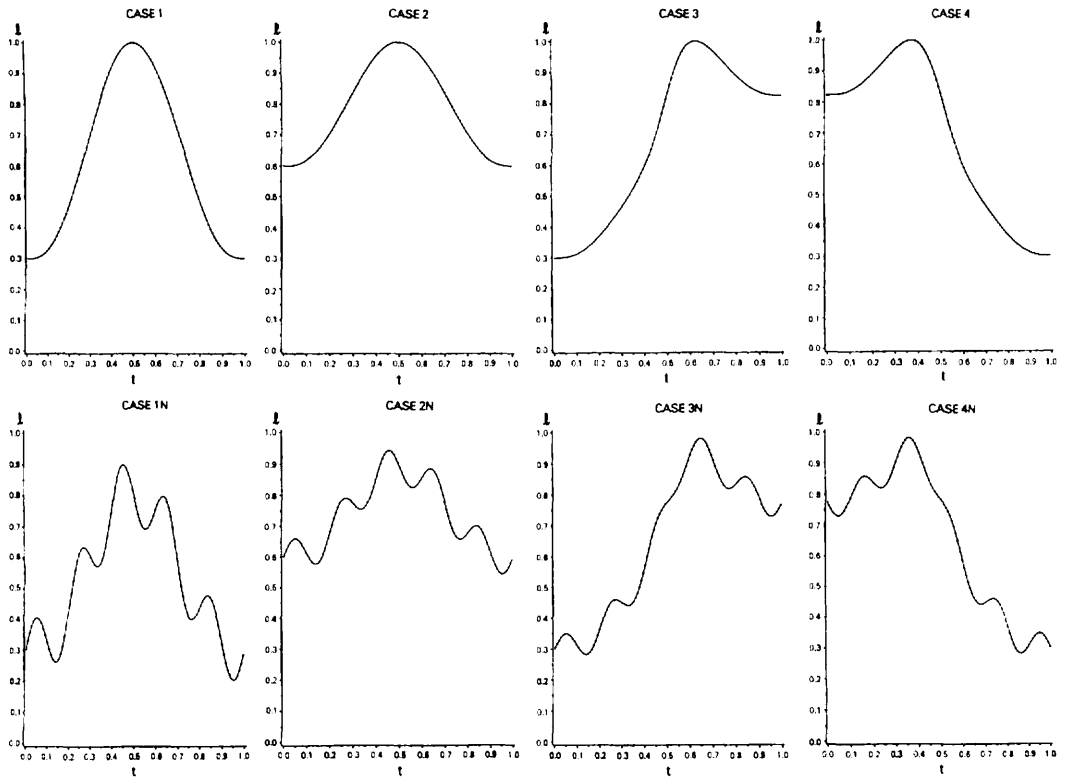

Figure 3. Some non-monotone treatment effect curves considered for evaluation of method L

4.4. Testing the Lakatos method under non-proportional hazards

We assume the effects of non-compliance, drop-in, loss to follow-up, and lag in the treatment effect are subsumed in the function $l(t)$, and $\lambda_C(t)$ is chosen to be constant. As in earlier tables, each power is based on 5000 iterations to give a standard error for the estimated power of 0·004. Sample sizes and simulated powers are given for three different weightings of the statistic: the logrank, Tarone–Ware,[13] and optimal (see Gill,[14] Section 5.2).

Table IV shows the Lakatos method to be robust to these non-proportional hazards alternatives, producing sample sizes whose simulated powers are typically within the 1 percentage point confidence interval of the desired 90 per cent power. For those situations in which the power falls outside the 95 per cent confidence interval, the estimates in these simulations are always conservative and never more than 2 percentage points above the desired 90 per cent power.

Between the results of Tables I and II, which demonstrate the accuracy of the Lakatos method over a broad range of parameter values, assuming proportional hazards, and Tables III and IV, which demonstrate its robustness to departures from proportional hazards, one can expect the Lakatos method to produce accurate results for most clinical trial situations.

Table I. Comparison of three methods of sample size calculation using an exponential model

$S_C(T)$*	θ*	Accrual*	Sample size L‡	Sample size F‡	Sample size RGS‡	Power L	Power F	Power RGS
0·8	0·667	1	1617	1628	1640	90·2	90·4	90·6
		5	2017	2024	2046	90·6	90·7	91·0
		9	2724	2709	2764	90·3	90·1	90·7
	0·50	1	638	649	664	90·1	90·6	91·2
		5	798	807	831	90·7	91·0	91·8
		9	1079	1081	1124	90·1	90·2	91·2
	0·25	1	230	241	269	91·9	93·0	95·3
		5	289	299	338	92·2	93·0	95·5
		9	392	401	459	91·6	92·2	95·1
0·2	0·667	1	360	370	363	89.6	90·4	89·8
		5	414	419	418	90·5	90·8	90·8
		9	528	509	534	89·8	88·7	90·1
	0·50	1	134	144	138	89·7	91·7	90·5
		5	156	164	161	89·9	91·3	90·8
		9	200	200	207	89·7	89·7	90·7
	0·25	1	43	53	48	90·2	95·1	93·0
		5	51	61	58	90·6	94·8	93·8
		9	66	74	76	90·2	93·1	93·7

* $S_C(T)$ is the probability of surviving to the end of the trial for patients in the control group, θ is the hazard ratio, and 'accrual' refers to the number of years of a 10 year trial during which accrual takes place
† For all simulations a two-sided alpha level of 0·05 was used
‡ The methods of Lakatos, Freedman, and Rubintein, Gail and Santner are denoted by 'L', 'F', and 'RGS', respectively

Table II. Comparison of three methods of sample size calculation using proportional hazards models

$S_C(T)$*	θ*	Accrual*	R_C†	Sample size L‡	Sample size F‡	Sample size RGS‡	Power L	Power F	Power RGS
0·8	0·667	2	4	1859	1883	1724	90·6	90·9	88·4
			2	1764	1779	1724	90·1	90·3	89·4
			0·50	1657	1667	1724	89·9	90·0	90·9
			0·25	1629	1638	1724	90·0	90·2	91·5
		8	4	3162	3416	2537	89·9	92·0	82·6
			2	2795	2885	2537	90·2	91·1	87·3
			0·50	2237	2224	2537	90·3	90·1	93·5
			0·25	2102	2043	2537	89·8	89·0	94·4
	0·50	2	4	735	751	699	89·7	90·3	88·2
			2	697	709	699	90·5	91·0	90·6
			0·50	654	664	699	91·3	91·7	92·9
			0·25	643	653	699	89·9	90·3	92·1
		8	4	1254	1364	1031	90·3	92·5	84·0
			2	1108	1152	1031	91·1	92·1	89·1
			0·50	900	887	1031	90·7	90·3	94·0
			0·25	831	815	1031	90·6	90·1	95·4

Table II. (Continued)

$S_C(T)$*	θ*	Accrual*	R_C†	Sample size L‡	Sample size F‡	Sample size RGS‡	Power L	Power F	Power RGS
	0·25	2	4	266	278	284	92·4	93·4	93·9
			2	252	263	284	91·9	92·9	94·6
			0·50	236	246	284	91·8	92·8	95·7
			0·25	232	242	284	92·4	93·4	96·4
		8	4	456	506	421	91·8	94·2	89·6
			2	403	430	421	91·8	93·3	92·9
			0·50	326	329	421	92·6	92·8	97·2
			0·25	301	302	421	92·0	92·1	97·8
0·2	0·667	2	4	391	401	374	90·4	91·1	89·1
			2	379	388	374	89·6	90·3	89·2
			0·50	365	374	374	89·9	90·6	90·6
			0·25	362	371	374	90·6	91·3	91·5
		8	4	591	604	495	89·4	90·0	83·5
			2	535	532	495	90·5	90·3	88·2
			0·50	454	445	495	90·5	89·9	92·7
			0·25	428	421	495	89·6	89·1	93·3
	0·50	2	4	147	157	142	90·5	92·2	89·5
			2	142	151	142	89·7	91·4	89·7
			0·50	136	146	142	90·1	92·0	91·3
			0·25	135	144	142	90·8	92·5	92·1
		8	4	225	238	191	90·5	92·0	85·4
			2	203	210	191	89·8	90·7	88·0
			0·50	171	174	191	90·2	90·7	93·0
			0·25	162	165	191	89·8	90·3	93·9
	0·25	2	4	47	58	50	90·4	95·2	92·0
			2	46	56	50	90·5	95·0	92·6
			0·50	44	54	50	91·0	95·5	94·0
			0·25	43	53	50	90·3	95·1	94·0
		8	4	76	88	70	90·5	94·1	88·1
			2	68	78	70	90·4	93·8	91·2
			0·50	56	65	70	90·7	94·3	95·6
			0·25	53	61	70	90·8	94·2	96·6

* See footnote * of Table I

† $R_C = \lambda_C(T)/\lambda_C(0)$ is the ratio of the hazards at the end and beginning of the trial. Under proportional hazards, $R_C = R_E$. As discussed in the text, these parameters uniquely determine the member from the family of survival curves given by $S(t) = (b/(at + b))^{1/a}$

‡ The methods of Lakatos, Freedman, and Rubintein, Gail and Santner are denoted by 'L', 'F', and 'RGS', respectively

5. DISCUSSION

A sample size which is too small can lead to an underpowered study, and is to be avoided. The situation is less straightforward when the sample size is too large. In this case, one may fail to do an important study because the cost appears too high, or, if the study is undertaken, it may cost more than it should. In addition, an appropriate balance between type I and type II errors should

Table III. Comparison of three methods of sample size calculation using non-proportional hazards models

$S_C(T)$*	θ*	Accrual*	R_C^*	R_E^*	Sample size L*	Sample size F*	Sample size RGS*	Power L	Power F	Power RGS
0·2	0·5	5	0·25	1	431	171	161	89·3	52·3	49·9
			0·50	1	240	168	161	88·6	75·4	73·7
			2	1	112	160	161	90·3	97·4	97·4
			4	1	87	155	161	90·1	99·1	99·3
			1	0·25	88	171	161	90·3	99·5	99·3
			1	0·50	115	168	161	90·5	97·7	97·2
			1	2	217	160	161	89·9	79·3	79·6
			1	4	311	155	161	89·9	62·7	64·4

* See footnotes to Table II

Table IV. Test of the Lakatos method in complex non-proportional hazards settings

Case*	C_0	C_3	C_5	r	Sample size LR†	Sample size TW†	Sample size OPT†	Power LR	Power TW	Power OPT
1‡	0	0·0	0·3	0	144	147	106	90·2	89·7	91·9
1*N*	0	0·2	0·3	0	139	141	109	90·4	90·5	91·6
2	0	0·0	0·6	0	147	149	137	89·9	90·2	90·9
2*N*	0	0·2	0·6	0	145	147	139	90·2	90·6	90·7
3	1	0·0	0·3	0	260	283	198	90·3	90·0	90·1
3*N*	1	0·2	0·3	0	250	272	195	90·7	89·8	90·0
4	1	0·0	0·3	1	185	176	156	90·2	90·1	90·0
4*N*	1	0·2	0·3	1	184	176	157	90·5	90·4	91·2

* C_0, C_3 and C_5 refer to the coefficients for the lag alternative (see Section 4). Case 4 is derived from Case 3 by reflection through the axis $t = 1/2$, as denoted by $r = 1$

† Weighting is denoted by LR for the logrank statistic, TW for Tarone and Ware's suggested weighting, and OPT for the optimal weighting

‡ Each 'case' is given by the parameters in that row and corresponds to the similarly identified case in Figure 3

be maintained. Some people feel that a conservative sample size evaluation method is desirable because it offers a degree of protection against errors in estimates of the survival and accrual patterns. We would argue, however, that the preferable procedure is to accurately determine an array of sample sizes for ranges of model assumptions (such as rates of drop-in, non-compliance, lag time, etc.) and to select a sample size which gives adequate protection against most reasonable scenarios.

Practical considerations should be taken into account when deciding which method to use. In situations such as trial planning meetings where a computer might not be available, the formula based methods of Rubinstein *et al.* and Freedman are attractive, and the choice among these could be based on the following considerations revealed in our simulations. First, when the survival distributions are exponential, and hazard ratios are close to 1, Freedman's overestimated slightly, while the method of Rubinstein *et al.* did well in most instances. For hazard ratios further from the null, Freedman's overestimated slightly, while the method of Rubinstein *et al.* was

substantially high. Secondly, if the curves are not exponential but proportional hazards is still reasonable, Freedman's was usually slightly high, and Rubinstein *et al.*'s frequently very inaccurate. Thirdly, if the hazards are non-proportional, then neither the method of Freedman nor Rubinstein *et al.* was reliable.

Where computers are available, the Lakatos method has several advantages. First, the Lakatos method did very well in the simulations, being, at worst, modestly conservative under extreme assumptions; it was usually more accurate than the formula based methods, even under the assumptions for which these formulas were derived. As deviations from the assumptions became greater, the formula based methods yielded increasingly less acceptable results. In addition, the Lakatos method goes beyond the formula based methods in that it can be adapted to a broad range of clinical trial conditions (for example, non-compliance, drop-in, staggered entry, loss to competing risks, and time lag in the effectiveness of treatment). Another attractive feature of the Lakatos method is that it can be used to derive sample sizes for all statistics in the Tarone–Ware[13] family and the G^{ρ} family of Harrington and Fleming.[15]

Since the logrank statistic is no longer optimal when the hazards are non-proportional, and the optimal weighting for a specified non-proportional hazards alternative is known, one may question the use of the unweighted logrank statistic rather than the optimally weighted version. Hazard functions are rarely known precisely. There are definite risks involved with assuming one knows a non-proportional hazards alternative and choosing the optimally weighted statistic for the *final analysis*. This is discussed in some detail in Zucker and Lakatos.[12] Briefly, weighting necessitates that some portion of the survival curve be downweighted and if unexpected adverse conditions prevail during that period, the downweighting may lead to an incorrect conclusion. Further, if the non-proportional hazards assumption is not quite right, the power may be seriously affected. However, if one is confident about the non-proportional hazards alternative, then weighting should be used. For this reason, the Lakatos method has been evaluated under optimal weighting for the non-proportional hazards alternatives in Section 4.4. We view *sample size calculation* differently. If one expects some non-proportionality to occur, either because of the nature of the treatment effect, or because of non-compliance or the like, then it is advisable to incorporate these effects in the sample size calculation to give some protection against plausible alternatives.

From Table IV, it is noteworthy that the addition of noise into the form of the hazard function does not alter the sample size much, nor disturb the calculation in the Lakatos method, provided that both the average hazard ratio and the underlying form of the hazard ratio curve remain unchanged. The sample size does vary substantially over the four different cases so that the form of the underlying hazard ratio curve is of considerable importance. In particular, if the overall form of the hazard curve is incorrectly chosen, either because of the basic shape or the level of parameters, then the sample size calculated under the Lakatos method will reflect the chosen curve rather than the unknown but desired curve. Similarly, a sample size derived through a well-designed simulation method would also reflect the chosen curve rather than the unknown desired curve. To protect against misspecifications, a range of plausible assumptions should be considered, and the sensitivity of the sample size to these assumptions explored. A sample size can then be chosen which reasonably protects against important plausible misspecifications.

It may at first seem a bit unfair to compare the methods of Rubinstein *et al.* and Freedman to that of Lakatos when the proportional hazards assumption is not satisfied, since the former were not designed for this situation. The object of the comparison is to determine whether methods such as Rubinstein *et al.*'s and Freedman's should be used when the proportional hazards assumption is not supportable. Non-proportional hazards situations occur quite frequently in clinical trials. For

example, non-compliance can continually diminish the treatment effect during the trial. In spite of this, it is common to see sample sizes based on exponential models adjusted for an overall loss due to non-compliance (see, for example, Lachin and Foulkes[7]). Another example of non-proportional hazards comes from the Physicians' Health Study,[16] in which the investigators expected pre-existing tumours to remain unaffected by treatment, but that new tumour development would be reduced. Thus, there would be a two year lag in treatment effect corresponding to the time for new tumours to become detectable. Cholesterol lowering therapies may take a year or more before the physiologic changes are sufficient to reduce the hazard.[17] Alternatively, if a therapy is thought to delay an outcome for a short time, then non-proportional hazards is likely. For example, an AIDS drug may initially lower the mortality hazard but not change the proportion of deaths within five years. In this situation, the hazard curves of the treatment and control groups must cross. Other examples were given at the end of Section 3.

Before concluding we offer some remarks regarding sample size calculation using simulation methods. The obvious advantage is that accuracy is usually limited only by the number of replications rather than asymptotic or other approximations. There are, however, a number of issues that the statistician should be aware of before embarking on this course. First, in most clinical trials, one needs some method of incorporating into the survival curves the effects of staggered entry, non-compliance, drop-in, lag time in treatment effectiveness, and loss to competing risks. Generally, this is far from straightforward, particularly if these effects are not constant over the course of the trial. One way to incorporate this information in the survival curve is to use the Markov model approach of Lakatos.[1] A second consideration is that in calculating the logrank statistic, the failure times usually need to be sorted (see, for example, Halpern and Brown[18]). Even for moderate sample sizes this can be time consuming. Lan and Lakatos[19] present a method of simulating the logrank statistic which avoids sorting and results in a substantial reduction in the use of computer time and memory. However, when non-compliance and other factors are involved, it may be difficult to apply this simulation method. A third problem with simulation methods for sample size calculation is that they do not yield sample sizes directly; rather, one must first guess the sample size and then simulate the power. An iterative process is thus required to arrive at the final estimate. Selection of a starting value can be based on one of the above procedures, and if non-compliance or the like has been incorporated in the survival curves using the Markov model of the Lakatos method, then this method would be ideal.

Considering the cost of a large clinical trial, the computer resources needed for computer simulation should not be a central consideration. By way of comparison, however, to calculate the sample size 3162 in Table II on our IBM 370 mainframe took less than 1/2 second. To verify the 3162 by simulation, a considerably simpler process than finding it by simulation, took over 24 minutes on the same machine, using the efficient method of Lan and Lakatos.[19] Using an IBM PS/2-80 386, with maths coprocessor for the same task took over 50 hours (the programs were written in SAS[20]).

Sample size calculation for use with the logrank statistic involves a certain amount of guesswork regarding the underlying survival models and a number of time dependent factors, such as non-compliance. Several formula based methods are available and can be conveniently used, particularly in meetings, when a calculator but not a computer is available. Because the current formula based methods are quite restrictive regarding the form of the survival curves and other time dependent factors, the derived sample sizes may not be accurate. When a computer is available, we believe it is neither necessary nor desirable to make such restrictive assumptions. Accurate sample sizes can be obtained under very general clinical trial conditions using the Lakatos method. Since it is unlikely that such conditions are precisely known, we recommend

exploring a variety of reasonable scenarios, in order to arrive at a sample size with adequate protection. While the prospect of specifying hazard models which need not be proportional may strike fear into the hearts of some users, the availability of similar methods for the binomial test for comparing survival curves at the NHLBI over many years (Halperin *et al.*,[21] Wu *et al.*,[10] Lakatos[22]) has led to such specification becoming routine.

REFERENCES

1. Lakatos, E. 'Sample sizes based on the log-rank statistic in complex clinical trials', *Biometrics*, **44**, 229–241 (1988).
2. Rubinstein, L. V., Gail, M. H. and Santner, T. J. 'Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation', *Journal of Chronic Diseases*, **34**, 469–479 (1981).
3. Pasternak, B. S. and Gilbert, H. S. 'Planning the duration of long-term survival time studies designed for accrual by cohorts', *Journal of Chronic Diseases*, **27**, 681–700 (1971).
4. George, S. L. and Desu, M. M. 'Planning the size and duration of a clinical trial studying the time to some critical event', *Journal of Chronic Diseases*, **27**, 15–24 (1973).
5. Bernstein, D. and Lagakos, S. W. 'Sample size and power determination for stratified clinical trials', *Journal of Statistical Computation and Simulation*, **8**, 65–73 (1978).
6. Lachin, J. M. 'Introduction to sample size determination and power analysis for clinical trials', *Controlled Clinical Trials*, **2**, 93–113 (1981).
7. Lachin, J. M. and Foulkes, M. A. 'Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification', *Biometrics*, **42**, 507–519 (1986).
8. Freedman, L. S. 'Tables of the number of patients required in clinical trials using the logrank test', *Statistics in Medicine*, **1**, 121–129 (1982).
9. Kalbfleisch, J. D. and Prentice, R. L. The Statistical Analysis of Failure Time Data Wiley, New York, 1980.
10. Wu, M., Fisher, M. and DeMets, D. 'Sample sizes for long-term medical trial with time dependent noncompliance and event rates', *Controlled Clinical Trials*, **1**, 109–121 (1980).
11. Lan, K. K. G. and Lachin, J. M. 'Group sequential logrank tests in a maximum duration trial', *Biometrics*, **46**, 759–781 (1990).
12. Zucker, D. M. and Lakatos, E. 'Weighted log rank-type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment', *Biometrika*, **77**, 853–864 (1990).
13. Tarone, R. E. and Ware, J. 'On distribution-free tests for equality of survival distributions', *Biometrika*, **64**, 156–160 (1977).
14. Gill, R. D. *Censoring and Stochastic Integrals, Mathematical Centre Tract 124*, Mathematisch Centrum, Amsterdam, 1980.
15. Harrington, D. P. and Fleming, T. R. 'A class of rank test procedures for censored survival data', *Biometrika*, **69**, 553–566 (1982).
16. Physicians' Health Study Steering Committee. *Physicians' Health Study Protocol*, Harvard Medical School, Department of Medicine, Brookline MA, 1983.
17. Lipid Research Clinics Program. 'The Coronary Primary Prevention Trial: design and implementation', *Journal of Chronic Diseases*, **32**, 609–631 (1979).
18. Halpern, J. and Brown, W. B. Jr. 'Designing clinical trials with arbitrary specification of survival functions and for the log rank or generalized Wilcoxon test', *Controlled Clinical Trials*, **8**, 177–189 (1987).
19. Lan, K. K. G. and Lakatos, E. 'On simulating linear rank statistics', Manuscript in preparation.
20. *SAS User's Guide: Statistics*, 1985 edition, SAS Institute, Cary, North Carolina, 1985.
21. Halperin, M., Rogot, E., Gurian, J. and Ederer, F. 'Sample sizes for medical trials with special reference to long-term therapy', *Journal of Chronic Diseases*, **21**, 13–24 (1968).
22. Lakatos, E. 'Sample sizes for clinical trials with time-dependent rates of losses and noncompliance', *Controlled Clinical Trials*, **7**, 189–199 (1986).