BIOMETRICS 44, 229–241
March 1988

Sample Sizes Based on the Log-Rank Statistic in Complex Clinical Trials

Edward Lakatos

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

SUMMARY

The log-rank test is frequently used to compare survival curves. While sample size estimation for comparison of binomial proportions has been adapted to typical clinical trial conditions such as noncompliance, lag time, and staggered entry, the estimation of sample size when the log-rank statistic is to be used has not been generalized to these types of clinical trial conditions. This paper presents a method of estimating sample sizes for the comparison of survival curves by the log-rank statistic in the presence of unrestricted rates of noncompliance, lag time, and so forth. The method applies to stratified trials in which the above conditions may vary across the different strata, and does not assume proportional hazards. Power and duration, as well as sample sizes, can be estimated. The method also produces estimates for binomial proportions and the Tarone–Ware class of statistics.

1. Introduction

Sample size calculations in clinical trials are frequently complicated by the fact that the risk of event for many participants does not remain constant during the trial. Even if the effect of therapy is constant over time, noncompliance and drop-in can cause the hazard rate to vary. Often, however, the mechanisms of the treatments being compared are sufficiently different that the proportional hazards assumption is suspect. This may be exemplified by the situation in which a drug is compared to surgery, with the latter hopefully achieving a more substantial "fix" provided the patient survives the early post-operative period during which mortality is increased. Furthermore, the hazard is often time-dependent (Wu, Fisher, and DeMets, 1980; Lachin and Foulkes, 1986).

In spite of the fact that the log-rank test is usually the preferred survival test in clinical trials with discrete endpoints, the biostatistics literature on sample size calculation for failure time data is almost entirely devoted to tests based on exponential survival curves (George and Desu, 1973; Rubinstein, Gail, and Santner, 1981) or binomial populations (Halperin et al., 1968; Lakatos, 1986). A closer look at this literature reveals that while sample size for comparison of binomial populations has been derived under very general conditions, very restrictive assumptions prevail in the exponential case. This is largely due to the fact that with the more general conditions, hazard functions and ratios are no longer constant, so that the usual tests based on exponential models with constant hazard ratios no longer apply.

Schoenfeld (1981) and Freedman (1982) present methods for sample size calculation based on the asymptotic expectation and variance of the log-rank statistic. However, the conditions under which their sample size formulas are derived are also very restrictive.

Key words: Complex clinical trials; Log-rank statistic; Markov process; Noncompliance; Nonproportional hazards; Sample size; Staggered entry.

In this paper, the survival curves that could be expected under very general conditions are modelled by using a stochastic process. The asymptotic expectation and variance of the log-rank statistic applied to these curves are then used to calculate sample size.

In Section 2, a basic version of a nonstationary Markov model for clinical trials is presented, and in Section 3, the expected value of the log-rank statistic and the associated sample size formula are derived. Extensions of the basic Markov model to include lag time, accrual, and stratification appear in Section 4. Examples with assumptions typical of cardiovascular and cancer trials are considered in Section 5. Duration is also discussed.

2. The Basic Markov Model

In this nonstationary Markov process, the treatment and control groups are modelled separately. Without loss of generality we will consider only the treatment group. Assume there is no time lag in the effectiveness of treatment. Each patient randomized to the treatment group is considered to be a complier initially, with probability P_E of having an event in 1 year, say. We label this initial state A_E. As the trial progresses, a variety of circumstances can arise that would alter this probability, and thus cause a transition to a different state. If the patient no longer complies with the treatment regimen, we assume that his probability of becoming an event in 1 year is P_C, that of the placebo controls, and that he has transferred to the state A_C from his initial state A_E. The A indicates "active" trial participant, as opposed to those who can no longer be followed for the event of interest because they are lost to follow-up or competing risks (state L). Those participants who experience the primary event are transferred to the state E. Thus, at any given time, t, a person is in one of these four states with corresponding vector of occupancy probabilities $\mathbf{D}_t$. For the moment, we assume simultaneous entry at time t_0, the start of the study. If the components of the vector appear in the order L, E, A_E, and A_C, then the initial distribution of the trial population is

$$\mathbf{D}_{t_0} = \begin{bmatrix} \text{Loss} \\ \text{Event} \\ \text{Active complier} \\ \text{Active noncomplier} \end{bmatrix} = \begin{bmatrix} L \\ E \\ A_E \\ A_C \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}.$$

In general, analytic considerations determine what transitions are appropriate. For instance, when the analysis is governed by the philosophy "intention to treat," noncompliers should not be censored; rather they are still active participants but at an increased event rate. If one intends to censor these patients in the analysis, the corresponding sample size can be derived by transferring them to the censored state L. The derivation of the hazard function should also be considered, since nonadherence might already be incorporated in this function. This would happen, for example, if the source was a pilot study and the survival curve was based on both compliers and noncompliers. The sample size formula derived in the next section is based on the distributions $\mathbf{D}_{t_i}$ at intermediate times t_i. If $N(t)$ denotes the number of individuals still under treatment and subject to risk of event $r(t)$, risk of loss $l(t)$, and risk of noncompliance $\delta(t)$, then the total number of events in the treatment group at t_i is [see Halperin et al. (1968) or Lakatos (1986)]

$$N(0) \int_0^{t_i} r(u) \exp\left[-\int_0^u [\delta(x) + r(x) + l(x)]\, dx \right] d(u).$$

In the nonstationary model, the functions r, l, and δ are not constant so that, in general, complicated numerical integration programs are required. In the time-lag model given below, there is a continuum of states corresponding to event rates intermediate between the control and experimental rates. If one also allows noncompliers to return to therapy,

and incorporates staggered entry and lag times into the model, the situation is even more complicated. Finally, solutions for similar equations are needed for each state (not only the event state) at intermediate time points. While the numerical solution of this continuous-time model with a mixture of discrete and continuous states may be formidable, a discrete-time formulation leads to simple numerical computation and equivalent results (see Lakatos, 1986, Appendix 1). A computer program for the Markov model is given in Lakatos (1986). It is easily adaptable to the current setting (see the Appendix). In the discrete formulation, the transition matrices $\mathbf{T}_{i,i+1}$ are constructed so that $t_{i,i+1}(j_1, j_2)$ is the probability of transferring from state j_1 to state j_2 during the time interval $[t_i, t_{i+1}]$. For $i \leq N$,

$$\mathbf{D}_{t_i} = \mathbf{T}_{i-1,i}\mathbf{D}_{t_{i-1}}, \tag{1}$$

where t_N is the "end" of the trial. This Markov model creates for each group a sequence of distributions $\{\mathbf{D}_{t_i}, i = 1, \ldots, N\}$. To simplify notation, denote the combination of sequences from both groups by $\{\mathbf{D}_{t_i}\}$. As an example (Gail, 1985), suppose we have a 2-year trial with event rates of $1 - \exp(-1) \approx .6321$ and $1 - \exp(-\frac{1}{2}) \approx .3935$ per year in the control and treatment groups, respectively, and the yearly loss to follow-up and noncompliance rates are 3% and 4%, respectively. The rate at which patients assigned to control begin taking a medication with an efficacy similar to the experimental treatment is called the "drop-in rate" and is assumed to be 5%. In cardiovascular trials, drop-ins often occur when the private physician of a patient assigned to control detects the condition of interest, such as hypertension, and prescribes treatment. Since our analyses would include such patients, we calculate sample size assuming these drop-ins have a reduced event rate. This example assumes constant hazards, so the treatment group transition matrix for both the first and second years is

$$\mathbf{T} = \begin{array}{c} \\ L \\ E \\ A_E \\ A_C \end{array} \begin{array}{c} \begin{array}{cccc} L & E & A_E & A_C \end{array} \\ \begin{bmatrix} 1 & 0 & .03 & .03 \\ 0 & 1 & .3935 & .6321 \\ 0 & 0 & 1 - \Sigma & .05 \\ 0 & 0 & .04 & 1 - \Sigma \end{bmatrix} \end{array}.$$

Entries denoted $1 - \Sigma$ represent 1 minus the sum of the remainder of the column. The entry .05 is made with the following two assumptions: (i) the return to medication of noncompliers is the same as the drop-in rate, and (ii) those who do return to compliance are indistinguishable from those who never stopped complying. The same transition matrix $\mathbf{T}$ can be used for the control group but the initial vector of occupancy probabilities should reflect that 100% of the control group are in the state A_C at entry. With this model, patients can transfer to states at the times t_i, $i = 1, \ldots, N$. In real settings, transitions can take place at any time. If $S(t)$ is the cumulative survival distribution, then the probability of failing in the interval $[t_{k-1}, t_k]$ is $1 - S(t_k)/S(t_{k-1})$. A continuous process can be approximated by replacing each matrix $\mathbf{T}$ by $\prod_{k=1}^{K} \mathbf{T}_k$, where each year has been divided into K equal intervals, and each off-diagonal element of $\mathbf{T}_k$ is given by an appropriate term of the form $1 - S(t_k)/S(t_{k-1})$. Note that the survival curve can take any form (Weibull, Kaplan–Meier, etc.) and that each of the off-diagonal transitions can be considered as resulting from some survival distribution. It is important to recognize the distinction between two types of nonproportional hazards models: (i) lag-type models, and (ii) those in which the hazard rates depend only on the time from randomization. In the lag-type models, a control patient may "drop in" at any time, and thus a person's hazard at a given time cannot be determined solely from the time from randomization. In this case, the above model is inappropriate since the process would not be Markovian. In the lag model presented below,

there are additional states and the Markov property is satisfied. On the other hand, there are nonproportional hazards models that satisfy (ii) above, and modelling these cases with the lag model would be inappropriate. An example of this would be a drug trial in which patients are randomized immediately after surgery. Here, there is assumed to be no lag in the drug effect, but there is a high early post-operative risk that diminishes with time from surgery. This high early risk is not related to lag in the effectiveness of the drug, and is thus not experienced when a patient begins taking medication later in the trial.

When only yearly rates are given and constant hazard rate within each year is assumed, this amounts to replacing each off-diagonal entry x in $\mathbf{T}$ by $1 - (1 - x)^{1/K}$. The resulting sequence $\{\mathbf{D}_{t_i}\}$ when $K = 10$ per year for 2 years is given in columns labelled "experimental" of Table 1. The previous four columns are the corresponding control group rates. The row starting with 1.0 represents the distribution at 1 year into the trial and indicates that in the control group, 2% of the cohort has been lost, 61.9% has had events, 33.6% are still taking only placebo, and 2.4% are drop-ins. In the experimental group only 56.3% are still event-free and complying with the initial therapy. The marginal 1-year event rate .619 in the treatment group is diminished from the assumed .632 because of the losses and because drop-ins have lower event rates than those taking placebo.

Table 1
The sequence $\{\mathbf{D}_{t_i}\}$ for the example

	Control				Experimental								
t_i	L	E	A_C	A_E	L	E	A_E	A_C	γ	η	ρ	θ	ϕ
.1	.003	.095	.897	.005	.003	.049	.944	.004	.167	.222	.098	2.000	1.000
.2	.006	.181	.804	.009	.006	.095	.891	.007	.166	.226	.090	1.986	.951
.3	.008	.258	.721	.013	.009	.139	.842	.010	.166	.230	.083	1.972	.905
.4	.010	.327	.647	.016	.011	.181	.795	.013	.165	.234	.076	1.959	.862
.5	.013	.389	.580	.018	.014	.221	.750	.015	.164	.237	.070	1.945	.821
.6	.014	.445	.520	.020	.016	.259	.708	.016	.163	.240	.064	1.932	.782
.7	.016	.496	.466	.022	.018	.295	.669	.017	.162	.242	.059	1.920	.746
.8	.017	.541	.418	.023	.020	.330	.632	.018	.160	.244	.054	1.907	.711
.9	.019	.582	.375	.024	.022	.362	.596	.019	.158	.246	.050	1.894	.679
1.0	.020	.619	.336	.024	.024	.393	.563	.019	.156	.248	.046	1.882	.648
1.1	.021	.652	.302	.025	.026	.423	.532	.020	.154	.249	.043	1.870	.619
1.2	.022	.682	.271	.025	.028	.450	.502	.020	.152	.249	.039	1.857	.592
1.3	.023	.709	.243	.025	.029	.477	.474	.020	.149	.250	.036	1.845	.566
1.4	.024	.734	.218	.025	.031	.502	.448	.020	.147	.250	.034	1.833	.542
1.5	.025	.755	.195	.025	.032	.525	.423	.020	.144	.250	.031	1.820	.519
1.6	.025	.775	.175	.024	.033	.548	.399	.019	.141	.249	.029	1.808	.497
1.7	.026	.793	.157	.024	.035	.569	.377	.019	.138	.248	.027	1.796	.477
1.8	.026	.809	.141	.023	.036	.589	.356	.018	.135	.247	.025	1.783	.457
1.9	.027	.824	.127	.023	.037	.609	.336	.018	.132	.246	.023	1.771	.439
2.0	.027	.837	.114	.022	.038	.627	.318	.018	.129	.244	.021	1.758	.421

In the next section, we derive equations for sample size calculations based on the log-rank statistic using probabilities from this combined sequence of distributions.

3. Derivation of Sample Size for the Log-Rank

Since the log-rank statistic can be considered as a member of the Tarone–Ware class of statistics, we derive estimates for the latter. The Tarone–Ware statistic

can be expressed as

$$L = \frac{\sum_{k=1}^{d} w_k\left(X_k - \dfrac{m_k}{m_k + n_k}\right)}{\left[\sum_{k=1}^{d} w_k^2\left(\dfrac{m_k n_k}{(m_k + n_k)^2}\right)\right]^{1/2}}, \tag{2}$$

where the sum is over deaths, X_k is the indicator of the control group, w_k is the kth Tarone–Ware weight, and m_k and n_k are the numbers at risk, just before the kth death, in the experimental and control groups, respectively.

Consider the following notation. We first obtain a formula for d, the total number of deaths. Partition the period of the trial into N equal intervals, and let there be d_i deaths during the ith interval. Define ϕ_{i_k} to be the ratio of patients in the two treatment groups at risk just prior to the kth death in the ith interval. Define θ_{i_k} to be P_{1i_k}/P_{2i_k}, where P_{ji_k} is the hazard of dying just prior to the kth death in the ith interval in treatment group j.

Let F and G be the failure-time distributions in the treatment and control groups, respectively. We use the log-rank statistic to test H_0: $(1 - F) = (1 - G)$ versus H_a: $(1 - F) \neq (1 - G)$. Note that this makes no assumption about the form of the hazard function. Then the approximate expectation of (2) under a fixed local alternative is (see Schoenfeld, 1981; Freedman, 1982)

$$E = \frac{\sum_{i=1}^{N} \sum_{k=1}^{d_i} w_{i_k}\left[\dfrac{\phi_{i_k}\theta_{i_k}}{1 + \phi_{i_k}\theta_{i_k}} - \dfrac{\phi_{i_k}}{1 + \phi_{i_k}}\right]}{\left[\sum_{i=1}^{N} \sum_{k=1}^{d_i} \dfrac{w_{i_k}^2 \phi_{i_k}}{(1 + \phi_{i_k})^2}\right]^{1/2}}, \tag{3}$$

where the right summation of each double summation is over the d_i deaths in the ith interval, and the left summation is over the N intervals that partition the trial. When $w_{i_k} = 1$ for all i and k, the log-rank is obtained. Treating this statistic as N(E, 1), we have

$$E = z_{\alpha/2} + z_{\beta}, \tag{4}$$

where z_α is the standard normal variate. Assuming $\phi_{i_k} \equiv \phi_i$ and $w_{i_k} \equiv w_i$, constants for all k in the ith interval, and letting $\rho_i = d_i/d$, where $d = \sum d_i$, then (3) becomes

$$E = e(\mathbf{D})\sqrt{d}, \tag{5}$$

where

$$e(\mathbf{D}) = \frac{\sum_{i=1}^{N} w_i \rho_i \gamma_i}{(\sum_{i=1}^{N} w_i^2 \rho_i \eta_i)^{1/2}} \tag{6}$$

and

$$\gamma_i = \frac{\phi_i\theta_i}{1 + \phi_i\theta_i} - \frac{\phi_i}{1 + \phi_i} \quad \text{and} \quad \eta_i = \frac{\phi_i}{(1 + \phi_i)^2}.$$

Note that $e(\mathbf{D})$ is a function of parameters from the sequence $\{\mathbf{D}\}$ and is independent of d_i and d. Solving (4) and (5) for d yields

$$\sqrt{d} = (z_{\alpha/2} + z_{\beta})/e(\mathbf{D})$$

and

$$d = \frac{[z_{\alpha/2}(\sum w_i^2 \rho_i \eta_i)^{1/2} + z_\beta(\sum w_i^2 \rho_i \eta_i)^{1/2}]^2}{(\sum w_i \rho_i \gamma_i)^2}. \tag{7}$$

The quantities ρ_i, η_i, and γ_i can readily be determined using the Markov model, even under a broad range of assumptions. The last four columns of Table 1 display these parameters (with ϕ_i and θ_i) obtained by performing the appropriate arithmetic operations on the other columns of Table 1. Using (7), with power at .90 and a two-sided .05 significance level, the number of deaths is 102. Since $d = N(P_c + P_e)/2$, where P_c and P_e are the cumulative event rates, the required total sample size is obtained using the cumulative event rates from the last row of Table 1:

$$N = 2d/(P_E + P_C) = 2(102)/(.627 + .837) = 139.$$

4. Lag Times, Staggered Entry, and Stratification

Often we do not expect the reduction in hazard produced by a treatment to occur instantaneously. To model lags, consider a continuous-state Markov model in which a person assigned to treatment enters an active state A_{P_C}, which has event rate P_C of the placebo controls, and passes through a series of states $A_{P(t)}$ with successively lower event rates, eventually arriving in a state A_{P_E}, which has the risk of fully effective experiment. The risk while in $A_{P(t)}$ is $P(t)$, where $P(t)$ is intermediate between P_C and P_E and can take on any functional form. Of course, events, losses, and the like can occur at any of the intermediate states. As above, to simplify computation, we assume a discrete time and state model in which the two active states A_C and A_E have been replaced by the series of states $A_{P(t)}$. If the lag is p/q years, where p and q are integers, and there are n intervals per year, then there are $np + 1$ states of the form $A_{P(t)}$. The rates associated with each of the intermediate states are determined by the form of the lag function. The associated transition matrices are of the form

$$\begin{array}{c|cc|ccc|ccc}
 & L & E & C_1 & C_2 & \cdots\ C_{np} & D_0 & D_1 & \cdots\ D_{np-1} \\
\hline
L & 1 & 0 & l_i & l_i & \cdots\ l_i & l_i & l_i & \cdots\ l_i \\
E & 0 & 1 & e_1 & e_2 & \cdots\ e_{np} & e_0 & e_1 & \cdots\ e_{np-1} \\
\hline
C_1 & & & & & & & & \\
C_2 & & & & & & & & \\
\vdots & & \mathbf{0} & & \mathbf{A} & & & \mathbf{C} & \\
C_{np} & & & & & & & & \\
\hline
D_0 & & & & & & & & \\
D_1 & & & & & & & & \\
\vdots & & \mathbf{0} & & \mathbf{B} & & & \mathbf{D} & \\
D_{np-1} & & & & & & & &
\end{array},$$

where the states C_i denote those on active therapy with current event rate P_i and D_j denote noncompliers whose current event rate is P_j. The matrices **A**, **B**, **C**, and **D** are determined by the assumptions of the trial. The following assumptions are typical and lead to the matrices described below.

Assumption A. All actives who remain active (i.e., do not become losses, events, or noncompliers) move to the state corresponding to the next lower event rate unless the

medication has reached complete efficacy, in which case they remain at the lowest event rate P_E. In this case, the only nonzero elements of **A** are those below the diagonal and the single entry in the lower right-hand corner. Each such nonzero entry can be determined by subtracting the remainder of the column from 1.0.

Assumption B. All actives who do not comply move to the state corresponding to the next higher event rate. The probability of doing so is the current probability of noncompliance, regardless of the active state. (Note that in the T matrix, a diagonal element of **B** pairs C_1 with D_0, etc.) Thus, $\mathbf{B} = \text{diag}(d_i)$.

Assumption C. Noncompliers return to active at the drop-in rate. Thus, $\mathbf{C} = \text{diag}(c_i)$.

Assumption D. All noncompliers move to the state corresponding to the next higher event rate until the medication has worn off, in which case they remain at the highest event rate P_C. In this case, the only nonzero elements of **A** are those above the diagonal and the single entry in the upper left-hand corner. Again, each such nonzero entry can be determined by subtracting the remainder of the column from 1.0. Note that moving the above-diagonal of **D** to two above the diagonal models a decay of effectiveness after noncompliance twice as fast as assumed, without changing the rate of onset of effectiveness. Similarly, moving the above-diagonal of **D** to the first row of **D** models the medication completely losing effectiveness immediately upon withdrawal. An example of this general model is given in Lakatos (1986), and the computer programs included there account for lag time and staggered entry.

In many trials, recruitment takes place over a period of time while close-out is simultaneous. In this case not all patients are followed for the same length of time. This is generally referred to as staggered entry or extended accrual. In the Markov models described above, the transition probabilities are functions of the time from entry of the patient rather than calendar time. Thus, to preserve the Markov property, we continue to assume all patients enter simultaneously and account for staggered entry by "administratively censoring" patients in consonance with their accrual pattern. This also conforms to the calculation of the log-rank statistic under staggered entry. If the trial is divided into N equal time intervals and p_i is the probability of entering the trial during the ith interval, then conditional on being in an active state during the $(N - k + 1)$th interval, the probability of being administratively censored during this interval is $p_k / \sum_{i=1}^{k} p_i$. Thus, staggered entry can be modelled by assuming additional transitions of active patients to a censored state with these probabilities.

In the case of a stratified trial, for each stratum $j = 1, \ldots, J$, obtain a sequence $\{\mathbf{D}_j\}$. To test H_0: $(1 - F) \neq (1 - G)$, consider the statistic

$$E = \sum \tau_j E_j,$$

where τ_j is a weight to be assigned to the jth stratum, E_j is the expected value of the statistic for the jth stratum, given in (3), and p_j is the proportion of the sample in the jth stratum $(N_j = p_j N)$. The proportion of deaths in the jth stratum is d_j/d, where

$$d_j = N_j[q_1 P_E(\mathbf{D}_j) + (1 - q_1)P_C(\mathbf{D}_j)],$$

q_1 is the proportion allocated to the treatment group, and $P_E(\mathbf{D}_j)$ is the probability that an individual will die by the end of the trial in the jth stratum in the experimental group. Hence,

$$d = \sum d_j = N \sum p_j[q_1 P_E(\mathbf{D}_j) + (1 - q_1)P_C(\mathbf{D}_j)]$$

and

$$d_j = \frac{dp_j[q_1 P_E(\mathbf{D}_j) + (1 - q_1)P_C(\mathbf{D}_j)]}{\sum p_j[q_1 P_E(\mathbf{D}_j) + (1 - q_1)P_C(\mathbf{D}_j)]}.$$

Finally,

$$E = \sqrt{d} \sum \tau_j e(\mathbf{D}_j) \left\{ \frac{p_j[q_1 P_E(\mathbf{D}_j) + (1 - q_1)P_C(\mathbf{D}_j)]}{\sum p_j[q_1 P_E(\mathbf{D}_j) + (1 - q_1)P_C(\mathbf{D}_j)]} \right\}^{1/2} = C_1 \sqrt{d},$$

where C_1 is a function of $\{\mathbf{D}_j\}$, q_1, and p_j. The number of deaths d can be obtained as before, simultaneously solving $E = (z_\alpha + z_\beta)V$ and the above equation, where $V = \sum \tau_j^2$. Bernstein and Lagakos (1978) give an optimal set of weights τ_j under some proportional hazards assumptions.

5. Examples

The results of applying these methods to two trials, one with parameters typical of some cardiovascular (CVD) trials and the other typical of cancer, are presented in Tables 2 and 3. The effect of including noncompliance, drop-in, loss, and staggered entry under several alternative hypotheses is examined. An attempt is made to keep the parameters comparable: in the CVD trial, an "average" of 5 years of follow-up in a 6-year trial with 2 years of recruitment is matched with a 5-year simultaneous entry trial. In the cancer trial, a 1.5-year simultaneous entry trial is compared to a 2-year trial with 1 year of accrual.

Table 2
Sample sizes[a] for a cardiovascular trial using binomial (bin) *and log-rank* (lr) *tests*

		Proportional hazards		Lag (Halperin)		Lag Model 2	
Entry	Adjust[b]	bin	lr	bin	lr	bin	lr
Simultaneous	No	2,650	2,654	4,339	4,407	8,034	8,190
	Yes	4,914	4,880	7,898	7,965	14,573	14,953
Uniform	No	2,651	2,651	4,333	4,397	8,016	8,159
	Yes	4,941	4,903	7,949	8,009	14,700	14,953
Nonuniform	No	2,753	2,653	4,598	4,665	8,804	8,967
	Yes	5,030	4,994	8,270	8,337	15,856	16,142

[a] Two-sided test with significance level at .05 and power at .90.
[b] "Yes" indicates adjustment for noncompliance, drop-in, and competing risks, as indicated in Table 4.

Table 3
Sample[a] sizes for a cancer trial: binomial (bin) *and log-rank* (lr) *tests*

		Proportional hazards		Lag (Halperin)		Lag Model 2	
Entry	Adjust[b]	bin	lr	bin	lr	bin	lr
Simultaneous	No	149	135	408	572	1,898	5,043
	Yes	192	164	519	691	2,457	6,225
Uniform	No	156	137	437	575	2,190	5,113
	Yes	204	169	569	710	2,968	6,619
Nonuniform	No	159	141	477	629	2,946	6,862
	Yes	205	173	611	770	3,942	8,820

[a] Two-sided test with significance level at .05 and power at .90.
[b] "Yes" indicates adjustment for noncompliance, drop-in, and competing risks, as indicated in Table 4.

The rates for the CVD trial (see Table 4) are taken from the SHEP trial and are described elsewhere (see Lakatos, 1986). The yearly event rates ($P_C = .016$) are of the same order of magnitude as in the cholesterol-lowering trial of the CPPT (Lipid Research Clinics, 1984) ($P_C = .012$). The row heading "adjust" indicates adjustment for noncompliance, loss, and

Table 4
Loss, noncompliance, and drop-in rates for a clinical trial

State	Year 1	Year 2	Year 3	Year 4	Year 5
Lost	.03	.032	.034	.036	.038
Event	.0096	.0096	.0096	.0096	.0096
Noncompliance	.07	.035	.035	.035	.035
Drop-in	.09	.045	.050	.055	.060
Event	.016	.016	.016	.016	.016

drop-in. In the cancer trial we use the same rates for noncompliance and the like as the CVD trial, but event rates are taken from the example in Gail (1985). The nonuniform recruitment rate used in the CVD trial assumes that maximum recruitment is achieved during the second year but is only 30% of this rate during the first quarter-year, and 40, 60, and 80% during succeeding quarters. The corresponding nonuniform rates for the cancer trial are 40, 60, 80, and 100% for the quarters of year 1.

The column labelled "Lag (Halperin)" denotes the nonproportional hazards model hypothesized by Halperin et al. (1968): an exponential model with a hazard rate that changes linearly with time. The other lag model is motivated by the CPPT. Examination of the survival curves from that trial reveals that survival in the treatment group is no better than control in the first 2 years, after which the curves begin to diverge at an apparently constant rate. Thus, the second nonproportional hazards model assumes equal treatment and control rates (.016) for the first 2 years on therapy followed by a constant reduction (40%) in rates while therapy is maintained. A similar nonproportional hazards alternative using no reduction in rate during the first year of therapy and a 2 to 1 hazard ratio while therapy is maintained is employed in the cancer trial [here, the motivation stems from the ovarian cancer trial in Fleming et al. (1980)].

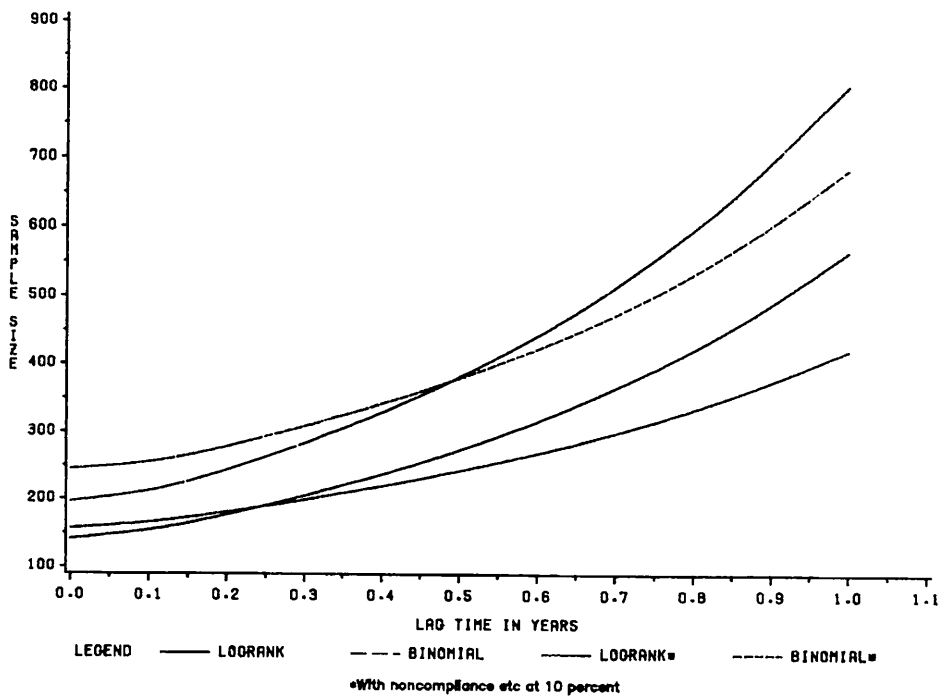

Figure 1. Sample size as a function of lag time.

While these two examples present too limited a view to draw many conclusions, it is clear that the sample size for the log-rank may be very sensitive to the specification of the nonproportional hazards alternative and that in these cases, the exponential model, as represented by the log-rank under proportional hazards, may be rather poor for estimating sample size. In these examples, the binomial fares surprisingly well.

In Figure 1, the effect of lag time (Halperin's model) on sample size in the cancer trial is plotted. As the departure from proportional hazards becomes more accentuated with increasing lag time, the advantage of the log-rank over the binomial decreases, and actually reverses. Thus, when a substantial lag time in the treatment effect is possible, as with the cholesterol-lowering trial, the binomial test may be more powerful.

With the Markov model, one can examine the effect of various clinical trial conditions on the hazard ratio. Figure 2 presents the hazard ratio as a function of time for two hypothetical clinical trials. In both cases there is a 1-year lag in the effectiveness of medication. The solid line plots the hazard ratio when there is no loss, no noncompliance, and no drop-in; the dashed line corresponds to the situation of 10% loss, noncompliance, and drop-in. Graphs such as those in Figure 2 allow investigators to determine the extent to which a proportional hazards assumption might be violated in a complex clinical trial.

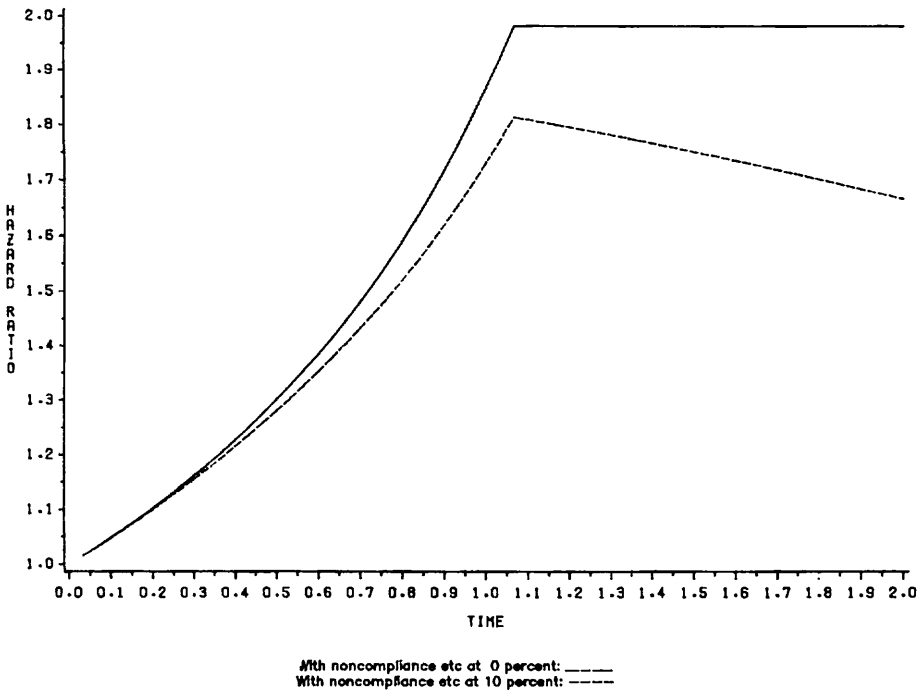

Figure 2. Hazard ratio as a function of time.

6. Duration

The necessary duration of a trial depends on the functional form of the survival and censoring distributions as well as the various parameters involved. The literature on estimation of duration generally assumes that these distributions are negative exponential, and that three parameters remain to be specified: the rate of accrual, the duration of the accrual period, and the duration of the follow-up period. Various authors fix one or more of these parameters and solve for those remaining.

In large clinical trials, not only are all three parameters variable, but the single-parameter negative exponential is often too restrictive (see Introduction). Further, the assumption of uniform accrual is violated whenever recruitment requires a phase-in period. Minimum follow-up time can depend on design considerations such as whether early or late survival is of interest. The number of patients that various clinical sites can handle and the total number of possible clinical sites introduce further constraints as well as variation into the model.

Although a simple formula or algorithm for estimating duration would be desirable, the above considerations make such a solution impossible in all but the most simple situations. Ultimately, we recommend working interactively with the trial planners, producing a variety of sample size estimates for different possible scenarios. However, the following comments may aid in the estimation of duration in situations in which restrictive assumptions are tenable.

Any method that gives sample size estimates for fixed trial lengths can be adapted in the obvious way to iteratively give numerical estimates of necessary duration. Under simultaneous entry, the Markov model approach can yield noniterative numerical duration estimates. This can be accomplished by taking advantage of the fact that once the sequence $\{\mathbf{D}\}$ has been calculated for a trial of a given length, then the sequence contains subsequences corresponding to each trial of shorter duration. Estimation of the sample sizes for the shorter-length trials can be done with little additional effort. Duration is estimated as the first time at which trial size is large enough to yield the required power, and the numerical precision of this estimate of time can be determined in advance by specifying a sufficiently short transition interval in the Markov process. While simultaneous entry is a rather severe restriction, a reasonable estimate of duration under the assumption of uniform entry can be obtained by adding one-half the accrual time to the simultaneous duration estimate. Once these rough estimates have been obtained, estimates of duration should be based on Markov models using the best available estimates of the trial parameters (i.e., event rates, accrual rates, and so forth).

7. Selection of Parameters

One of the primary advantages of the discrete Markov approach is the ease of adaptation to the many complex situations encountered in actual clinical trials. While one is not likely to encounter a trial with every feature described above, the method can be used with various combinations that do arise. Along with the freedom from specifying survival functions and the like in restricted parametric forms is the increased burden of specifying these more complex forms. One should not expect investigators to be able to supply a set of parameters as in Table 4 [see Lakatos (1986) for a description of the selection of those parameters]. Rather, in the absence of information to the contrary, one could start with exponential curves, and test the sensitivity of this and other assumptions. In one trial in which recruitment was considerably slower than anticipated, and a decision had to be made regarding extension of the recruitment period, we used actual recruitment, noncompliance, and drop-in rates for the already observed period to determine power associated with some possible recruitment extensions.

ACKNOWLEDGEMENTS

I would like to thank Drs Kent Bailey, Erica Brittain, John Lachin, and David Zucker, and the referees for valuable comments and suggestions.

Résumé

Le test du logrank est fréquemment utilisé pour comparer des courbes de survie. Alors que l'estimation des effectifs nécessaires à la comparaison de pourcentages a été adaptée aux conditions typiques des essais cliniques (abandons de traitement, inclusion échelonnée, délai de réponse), celle des effectifs nécessaires pour un test du logrank n'a pas été généralisée à ces conditions. Cet article présente une méthode d'estimation des tailles des échantillons nécessaires pour la comparaison de courbes de survie par le test du logrank, prenant en compte les conditions précitées. Cette méthode s'applique aux essais stratifiés dans lequels ces conditions peuvent varier d'une strate à l'autre et ne fait pas l'hypothèse des risques instantanés proportionnels. La puissance et la durée de l'essai peuvent être également estimées. La méthode fournit aussi des estimations dans le cas de pourcentages et la classe des statistiques de Tarone–Ware.

References

Bernstein, D. and Lagakos, S. W. (1978). Sample size and power determination for stratified clinical trials. *Journal of Statistical Computing and Simulation* **8,** 65–73.

Fleming, T. R., O'Fallon, J. R., O'Brien, P. C., and Harrington, D. P. (1980). Modified Kolmogorov–Smirnov test procedures with application to arbitrarily right-censored data. *Biometrics* **36,** 607–625.

Freedman, L. S. (1982). Tables of the number of patients required in clinical trials using the log-rank test. *Statistics in Medicine* **1,** 121–129.

Gail, M. (1985). Applicability of sample size calculations based on a comparison of proportions for use with the log-rank test. *Controlled Clinical Trials* **6,** 112–119.

George, S. L. and Desu, M. M. (1973). Planning the size and duration of a clinical trial studying the time to some critical event. *Journal of Chronic Diseases* **27,** 15–24.

Halperin, M., Rogot, E., Gurian, J., and Ederer, F. (1968). Sample sizes for medical trials with special reference to long-term therapy. *Journal of Chronic Diseases* **21,** 13–24.

Lachin, J. M. and Foulkes, M. A. (1986). 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.

Lakatos, E. (1986). Sample sizes for clinical trials with time-dependent rates of losses and noncompliance. *Controlled Clinical Trials* **7,** 189–199.

Lipid Research Clinics Program (1984). The Lipid Research Clinics Primary Prevention Trial results. *Journal of the American Medical Association* **251,** 351–364.

Rubinstein, L. V., Gail, M. H., and Santner, T. J. (1981). 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.

SAS User's Guide: Statistics, 1985 edition. Cary, North Carolina: SAS Institute.

Schoenfeld, D. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* **68,** 316–318.

Wu, M., Fisher, M., and DeMets, D. (1980). Sample sizes for long-term medical trials with time-dependent noncompliance and event rates. *Controlled Clinical Trials* **1,** 109–121.

Received March 1986; *revised March and October* 1987.

Appendix

A SAS (1985) computer program implementing the Markov models and some variations is given by Lakatos (1986). The following two lines of that program must be interchanged to obtain the intermediate distributions needed to calculate the log-rank statistic:

```
*END OF TRANSITION MATRIX LOOP; END;
DSTR_E=DSTR_E||DISTR_E; DSTR_C=DSTR_C||DISTR_C;
```

(The printing of these distributions can be suppressed by deleting the last two lines of the program which begin PRINT.) The following lines may be added to obtain the sample size for the log-rank:

```
Z_ALPHA=1.96; Z_BETA=1.28;
EVENT_C=DSTR_C(2,)-(0||DSTR_C(2,1:(NCOL(DSTR_C)-1)));
EVENT_E=DSTR_E(2,)-(0||DSTR_E(2,1:(NCOL(DSTR_E)-1)));
```

```
LOSS_C=DSTR_C(1,)-(0||DSTR_C(1,1:(NCOL(DSTR_C)-1)));
LOSS_E=DSTR_(E(1,)-(0||DSTR_E(1,1:(NCOL(DSTR_E)-1)));
ATRISK_C=DSTR_C(3 4,)(+,)+LOSS_C+EVENT_C;
ATRISK_E=DSTR_E(3 4,)(+,)+LOSS_E+EVENT_E;
PHI=ATRISK_C#/ATRISK_E;
THETA=(EVENT_C#/ATRISK_C)#/(EVENT_E#/ATRISK_E);
RHO=(EVENT_C+EVENT_E)#/((EVENT_C+EVENT_E)(,+));
GAMMA=PHI#THETA#/(1+PHI#THETA)-PHI#/(1+PHI);
ETA=PHI#/((1+PHI)##2);
SIG=SQRT((RHO#ETA)(,+));
D_LR=( (Z_ALPHA#SIG+Z_BETA#SIG)#/((RHO#GAMMA)(,+)) )##2;
N_LR=2#D_LR#/(DISTR_C+DISTR_E)(2,);
PE=DISTR_E(2,); PC=DISTR_C(2,); PBAR=(PE+PC)#/2;
SIGBINO=SQRT(2#(PBAR)#(1-PBAR)); SIGBINA=SQRT(PE#(1-PE)+PC#(1-PC));
N_BIN=2#((Z_ALPHA#SIGBINO+Z_BETA#SIGBINA)#/(PC-PE))##2;
PRINT D_LR N_LR N_BIN;
```