Evaluation of Sample Size and Power for Analyses of Survival with Allowance for Nonuniform Patient Entry, Losses to Follow-Up, Noncompliance, and Stratification

John M. Lachin; Mary A. Foulkes

Biometrics, Vol. 42, No. 3 (Sep., 1986), 507-519.

Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28198609%2942%3A3%3C507%3AEOSSAP%3E2.0.CO%3B2-D

Biometrics is currently published by International Biometric Society.

BIOMETRICS 42, 507–519
September 1986

Evaluation of Sample Size and Power for Analyses of Survival with Allowance for Nonuniform Patient Entry, Losses to Follow-up, Noncompliance, and Stratification

John M. Lachin

The Biostatistics Center, Department of Statistics/Computer & Information Systems, George Washington University, 7979 Old Georgetown Road, Bethesda, Maryland 20814, U.S.A.

and

Mary A. Foulkes

Biometry and Field Studies Branch, National Institute of Neurological and Communicative Disorders and Strokes, Bethesda, Maryland 20892, U.S.A.

SUMMARY

When designing a clinical trial to test the equality of survival distributions for two treatment groups, the usual assumptions are exponential survival, uniform patient entry, full compliance, and censoring only administratively at the end of the trial. Various authors have presented methods for estimation of sample size or power under these assumptions, some of which allow for an R-year accrual period with T total years of study, $T > R$. The method of Lachin (1981, *Controlled Clinical Trials* **2**, 93–113) is extended to allow for cases where patients enter the trial in a nonuniform manner over time, patients may exit from the trial due to loss to follow-up (other than administrative), other patients may continue follow-up although failing to comply with the treatment regimen, and a stratified analysis may be planned according to one or more prognostic covariates.

1. Introduction

The life-table method for the analysis of survival or time-to-event data is one of the principal statistical procedures employed for the assessment of clinical trial results. One of the more widely used tests for the equality of survivor functions is the log-rank or Mantel–Cox test which is fully efficient under the general proportional hazards model (cf. Kalbfleisch and Prentice, 1980). This method is easily applied and is often more powerful than an analysis of simple proportions (cf. Cuzick, 1982; Willan, 1983; Gail, 1985).

Various methods for the evaluation of sample size and power for a life-table analysis have been presented under various assumptions regarding the survival, entry, and censoring distributions (cf. Lachin, 1981; Donner, 1984). Pasternack (1972) considered the problem in the context of an actuarial life table with comparison of the L-year survival rates, not the survivor functions. This approach was extended by Palta and McHugh (1979, 1980) to allow for losses to follow-up using the competing-risk model of Chiang (1968), and for noncompliance using the model of Schork and Remington (1967).

George and Desu (1974) approached the evaluation of sample size and power using an

Key words: Log-rank test; Losses to follow-up; Noncompliance; Power; Sample size; Stratification; Survival analysis.

approximation to the local power function of the Cox (1964) F-test for the equality of exponential survivor functions which is based on exponential ordered scores. George and Desu (1974), Rubinstein, Gail, and Santner (1981), and Lustbader and Litwin (1981) also considered the related problem of determining the required duration of follow-up when patients enter the trial at a specified rate of n per year. The method of Rubinstein et al. (1981) also allows for patient accrual at a uniform rate (i.e., Poisson process) over R years with follow-up over a total period of $T > R$ years, and allows for losses to follow-up that are exponentially distributed. The method of George and Desu was also extended to allow for a stratified analysis by Bernstein and Lagakos (1978) and Palta and Amini (1985), and to K groups by Makuch and Simon (1982). Schoenfeld (1981) showed that the method of George and Desu provides an asymptotic approximation to the power function of the log-rank test under the assumption of proportional constant hazards, and Schoenfeld and Richter (1982) presented extensions of George and Desu's method to the case of patient accrual over R years with follow-up over T years.

Lachin (1981) approached the evaluation of sample size and power using the local power function of the asymptotic (Wald) test for equality of exponentially distributed survivor functions. He extended the methods presented in Gross and Clark (1975) to the case of uniform patient entry over R years with follow-up over T years. Freedman (1982) provided a derivation of Lachin's (1981) method based on the asymptotic conditional expectation and variance of the log-rank statistic under the assumption of constant proportional hazards. Lachin (1981) and Freedman (1982) showed empirically that this approach yields slightly larger sample size estimates than those of George and Desu (1974) and Schoenfeld (1981), respectively.

It must be noted, however, that these methods are not exact. What is actually desired is to estimate the power function of the distribution-free log-rank test. However, it is feasible to describe the power function of the log-rank test only under the assumptions of a population model. Thus, these methods have been proposed as approximations under the restrictive assumptions of proportional constant hazards, i.e., under the exponential model. In reality, however, the hazards likely will not be constant, nor exactly proportional, over time, in which case the log-rank test will still be applicable but may not be maximally efficient.

Thus, it is our view that these methods should be cautiously applied using "worst-case" assumptions, such as using the lowest plausible hazard rate for the control group, the smallest clinically relevant reduction in mortality, and appropriate adjustments for departures from the usual assumptions of uniform patient recruitment, complete follow-up, full compliance, and homogeneity over prognostic strata. In this paper, therefore, various extensions of these methods are presented which allow for nonuniform entry distributions, random censoring (other than administrative), noncompliance, and a stratified life-table analysis.

2. The General Model

Consider that the trial consists of independent treatment and control groups of sizes n_e and n_c ($n_e + n_c = N$). Patients are to be recruited over an accrual period of R years and are to be followed until T years of study have elapsed, $T > R$. At the conclusion of the study, a test of the equality of survival distributions is to be conducted at level α with power $1 - \beta$. Under the exponential (proportional constant hazards) model, this test of equality of hazards, H_0: $(\lambda_e - \lambda_c) = 0$, can be conducted using the asymptotic (Wald) statistic based on the maximum likelihood estimates (MLEs) $\hat{\lambda}_e$ and $\hat{\lambda}_c$ and their asymptotic variances $\sigma^2(\hat{\lambda}_e)$ and $\sigma^2(\hat{\lambda}_c)$ (Gross and Clark, 1975). Lachin (1981, eq. 24) has shown that the basic

equation relating *total* sample size N and power for this test is

$$\sqrt{N}\,|\,\lambda_e - \lambda_c\,| = Z_\alpha\sqrt{\phi(\bar{\lambda})(Q_e^{-1} + Q_c^{-1})} + Z_\beta\sqrt{\phi(\lambda_e)Q_e^{-1} + \phi(\lambda_c)Q_c^{-1}}, \tag{2.1}$$

where $Q_e = n_e/N$, $Q_c = n_c/N$, $\bar{\lambda} = Q_e\lambda_e + Q_c\lambda_c$, and Z_α and Z_β are the standard normal deviates at levels α and β, respectively (see also Gross and Clark, 1975, p. 264). In (2.1) the term $\phi(\lambda)$ is that component of the variance $\sigma^2(\hat{\lambda})$ independent of sample size such that $\sigma^2(\hat{\lambda}) = \phi(\lambda)/N$.

In this and the other models described herein, the variance (and thus power) is a function of the expected number of deaths to be observed. In a cohort of size n, the expected number of deaths, $E_n(D)$, in turn, is a function of the sample size, hazard rate, and entry and censoring distributions. If δ_i indicates that the ith patient does ($\delta_i = 1$) or does not ($\delta_i = 0$) die, then $E_n(D) = nE(\delta \mid \lambda)$, $\sigma^2(\lambda) = \lambda^2/nE(\delta \mid \lambda)$, and, therefore, $\phi(\lambda) = \lambda^2/E(\delta \mid \lambda)$. Under the assumption of uniform patient entry over the interval 0 to R with none other than administrative censoring, Lachin (1981, eq. 26) showed that $\phi(\lambda) = \lambda^2E(\delta \mid \lambda)^{-1}$ is simply

$$\phi(\lambda) = \lambda^2\left[1 - \frac{e^{-\lambda(T-R)} - e^{-\lambda T}}{\lambda R}\right]^{-1}. \tag{2.2}$$

Sample size or power is then evaluated by substituting (2.2) into (2.1) and solving for N or Z_β, respectively.

Note that in this model, the expected total number of deaths is $E_N(D) = E_{n_e}(D) + E_{n_c}(D)$. Under the null hypothesis, $E_N(D \mid H_0) = NE(\delta \mid H_0) = N\bar{\lambda}^2/\phi(\bar{\lambda})$, or alternatively, $E_N(D \mid H_0) = N\lambda_c^2/\phi(\lambda_c)$, depending on whether one expects the observed hazard under H_0 to be $\bar{\lambda}$ or λ_c. Under the alternative hypothesis, $E_N(D \mid H_1) = NE(\delta \mid H_1)$, which is

$$E_N(D \mid H_1) = \frac{n_e\lambda_e^2}{\phi(\lambda_e)} + \frac{n_c\lambda_c^2}{\phi(\lambda_c)}. \tag{2.3}$$

For illustration, consider the example of Lachin (1981, p. 107). If it is desired to detect a reduction in hazards from $\lambda_c = .30$ to $\lambda_e = .20$ at $\alpha = .05$ (one-sided) and $\beta = .10$, in a study with a 3-year recruitment period, $Q_e = Q_c$, and a total duration of 5 years, then (2.1) and (2.2) yield that a total of $N = 378$ patients is required. In this case, $E_N(D \mid \bar{\lambda}) = 217$ deaths under the null hypothesis, and 215 deaths under the alternative hypothesis, 94 in the experimental group, $E(\delta \mid \lambda_e) = .4959$, and 121 in the control group, $E(\delta \mid \lambda_c) = .6381$.

In the following sections, additional definitions for $\phi(\lambda)$ are presented under generalizations of the model assumptions. In all cases, exponential survival is employed.

These expressions for $\phi(\lambda)$ also provide equivalent extensions of the George and Desu (1974) method. For this method, the basic equation relating sample size and power is

$$\sqrt{N}\ln\left(\frac{\lambda_c}{\lambda_e}\right) = Z_\alpha\sqrt{E(\delta \mid \bar{\lambda})^{-1}(Q_e^{-1} + Q_c^{-1})} + Z_\beta\sqrt{E(\delta \mid \lambda_e)^{-1}Q_e^{-1} + E(\delta \mid \lambda_c)^{-1}Q_c^{-1}}. \tag{2.4}$$

Since $E(\delta \mid \lambda) = \lambda^2/\phi(\lambda)$, each of the expressions for $\phi(\lambda)$ also includes explicitly the expression for $E(\delta \mid \lambda)$. For the model described above, sample size or power is derived upon substituting $E(\delta \mid \lambda)$ from (2.2) into (2.4).

As described in the introduction, Rubinstein et al. (1981) and Schoenfeld and Richter (1982) also extended George and Desu's method to this model. However, their basic equation relating sample size and power uses the variance under the alternative hypothesis as the multiplier of both Z_α and Z_β to yield

$$\sqrt{N}\ln\left(\frac{\lambda_c}{\lambda_e}\right) = (Z_\alpha + Z_\beta)\{[E(\delta \mid \lambda_e)Q_e]^{-1} + [E(\delta \mid \lambda_c)Q_c]^{-1}\}^{1/2}. \tag{2.5}$$

A similar simplification of (2.1) can be obtained as

$$\sqrt{N}\,|\,\lambda_e - \lambda_c\,| = (Z_\alpha + Z_\beta)[\phi(\lambda_e)Q_e^{-1} + \phi(\lambda_c)Q_c^{-1}]^{1/2}. \tag{2.6}$$

For equal sample fractions ($Q_e = Q_c$), the sample size estimates resulting from (2.5) or (2.6) will be greater (or power less) than those obtained from the more accurate expressions (2.4) and (2.1), respectively.

By Monte Carlo simulation under an exponential model, Rubinstein et al. (1981) and Freedman (1982) showed that (2.5) and (2.6), respectively, provide accurate estimates of the actual power of the log-rank test for fixed sample sizes. However, from these limited simulations it is not possible to determine which underlying statistic provides a more accurate approximation to the power of the log-rank test for fixed λ_e and λ_c. Further, we have computed sample size estimates using both (2.1) and (2.4) over a range of likely values for λ_e, λ_c, R, and T. For $\alpha = .05$ (one-sided) and $\beta = .10$, the ratio (2.1)/(2.4) of sample sizes is uniformly greater than 1 and increases as the ratio λ_e/λ_c decreases. However, for large reductions in mortality ($\lambda_e/\lambda_c = .50$), the required sample size is small, and thus the absolute difference in sample sizes is small even though the ratio is on the order of 1.03–1.10. For moderate reductions in mortality at the levels clinical trials are usually designed to detect, such as $\lambda_e/\lambda_c = .80$, the ratio of sample sizes is less than 1.01. Such computations suggest that these two approaches are basically equivalent.

3. Nonuniform Patient Entry

Suppose that patient entry times are distributed as $g(z)$, where z_i is the entry time of the ith patient, $0 \leq z_i \leq R$. Under the uniform distribution, $g(z) = 1/R$, and the distribution function $G(z)$ is linear over $(0, R)$. Clearly, for fixed N, if $G(z)$ is convex (faster than expected), power will be greater, or if $G(z)$ is concave, power will be less, than is the case with uniform entry. The general form of $\phi(\lambda)$ for any entry distribution function $G(z)$ is presented in (A.1) of the Appendix.

In the following we explore the effects of a concave entry distribution on power and sample size. As an example, we consider a truncated exponential entry distribution over the interval 0 to R, with density

$$g(z) = \frac{\gamma e^{-\gamma z}}{1 - e^{-\gamma R}}, \qquad 0 \leq z \leq R, \quad \gamma \neq 0, \tag{3.1}$$

(cf. Johnson and Kotz, 1970, p. 223). For $\gamma > 0$, the entry distribution is convex, whereas for $\gamma < 0$, the entry distribution is concave. For the unit recruitment period ($R = 1$), (3.1) yields 50% recruitment at $Z = .55$ for $\gamma = -.4$, at $Z = .6$ for $\gamma = -.82$, at $Z = .7$ for $\gamma = -1.8$, at $Z = .8$ for $\gamma = -3.3$, and at $Z = .9$ for $\gamma = -7.0$.

When employed in (A.1), this entry distribution yields

$$\phi(\lambda) = \lambda^2\left\{1 + \frac{\gamma e^{-\lambda T}[1 - e^{(\lambda-\gamma)R}]}{(1 - e^{-\gamma R})(\lambda - \gamma)}\right\}^{-1}. \tag{3.2}$$

Sample size or power is then evaluated by substituting (3.2) into (2.1).

For the example above with $R = 3$ and $T = 5$, Table 1 presents the effects of concave (i.e., lagging) patient entry on power for various γ ranging from 0 to −6.0. Over this range of γ, power is reduced, but only moderately unless there is a marked concave departure from linear (uniform) entry. Table 1 also presents the effects of nonuniform entry on the sample size required to provide power of .90 to detect a reduction in hazards from $\lambda_c = .30$ to $\lambda_e = .20$. As is generally true, a substantial increase in sample size may be required to compensate for a small reduction in power, e.g., $N = 516$ to provide power of .90 for $\gamma = -6.0$, whereas power would be .801 for $N = 378$.

Table 1
Probability of death (δ) *in the experimental* (E) *and control* (C) *groups, power and sample size as a function of concave exponential patient entry with parameter* γ, $\lambda_e = .20$, $\lambda_c = .30$, $R = 3$, $T = 5$, $\alpha = .05$ (*one-sided*). *Power is that for* $N = 378$. *Sample size is that required to maintain power* .90.

γ	Probability of death, E $E(\delta \mid \lambda_e)$	Probability of death, C $E(\delta \mid \lambda_c)$	Power	Sample size
0.0	.496	.638	.900	378
−0.5	.457	.598	.879	404
−1.0	.426	.563	.859	430
−1.5	.402	.539	.844	452
−2.0	.387	.520	.832	468
−2.5	.379	.508	.824	480
−3.0	.371	.500	.820	490
−3.5	.363	.492	.816	496
−4.0	.359	.488	.813	502
−4.5	.355	.484	.809	506
−5.0	.352	.480	.805	510
−5.5	.352	.477	.805	512
−6.0	.348	.477	.801	516

4. Losses to Follow-up

In the above models, staggered patient entry over the accrual period 0 to R years results in administrative censoring times ranging from $T - R$ to T years of follow-up. However, some patients may also be lost to follow-up for various causes. Under the commonly invoked random censorship model, in each group it is assumed that these losses to follow-up are random and that the censoring distribution is independent of the survival distribution. In this case total censoring is a combination of both administrative censoring over the interval $T - R$ to T and losses to follow-up over the complete follow-up interval 0 to T. If we let δ_i again denote death during the study, and let ξ_i denote loss to follow-up, then in the Appendix are presented the general form for $\phi(\lambda)$, and thus $E(\delta)$, and also for $E(\xi)$, the latter being the probability of loss to follow-up.

We now consider the special case where losses are exponentially distributed with loss hazard rates η_e and η_c independently of mortality, which is exponentially distributed with hazard rates λ_e and λ_c in the two groups. In this case the probability of death is a function of both the death and loss hazard rates λ and η. If we also assume uniform patient entry with $g(z) = 1/R$, and exponential losses to follow-up with distribution function $H(u) = 1 - e^{-\eta u}$ in (A.2) or (A.3), we then obtain $\phi(\lambda, \eta) = \lambda^2 E(\delta \mid \lambda, \eta)^{-1}$, which becomes

$$\phi(\lambda, \eta) = \lambda^2 \left\{ \frac{\lambda}{(\eta + \lambda)} \left[1 - \frac{e^{-(T-R)(\eta+\lambda)} - e^{-T(\eta+\lambda)}}{R(\eta + \lambda)} \right] \right\}^{-1} \tag{4.1}$$

The expression for $E(\delta \mid \lambda, \eta)$ above was also obtained by Rubinstein et al. (1981). Note that (4.1) reduces to (2.2) for $\eta = 0$. The basic equation (2.1) relating sample size and power then becomes

$$\sqrt{N} \mid \lambda_e - \lambda_c \mid = Z_\alpha \sqrt{\phi(\bar{\lambda}, \eta_e) Q_e^{-1} + \phi(\bar{\lambda}, \eta_c) Q_c^{-1}} + Z_\beta \sqrt{\phi(\lambda_e, \eta_e) Q_e^{-1} + \phi(\lambda_c, \eta_c) Q_c^{-1}}. \tag{4.2}$$

Clearly power and sample size depend on the probability of death $E(\delta \mid \lambda, \eta)$ as shown above. However, it is also relevant to consider the probability of loss $E(\xi \mid \lambda, \eta)$, which is obtained simply as $(\eta/\lambda)E(\delta \mid \lambda, \eta)$. For example, Table 2 presents the probabilities of death and loss for various η in a study with $N = 378$ and equal-sized groups, where $\lambda_e = .20$ and

Table 2
Probability of death (δ) *and loss* (ξ) *in groups* E *and* C *with death hazard rates* $\lambda_e = .20$, $\lambda_c = .30$, *and various loss hazard rates* $\eta_e = \eta_c = \eta$; $R = 3$, $T = 5$

η	$E(\delta \mid \lambda_e)$	$E(\delta \mid \lambda_c)$	$E(\xi \mid \lambda_e, \eta)$	$E(\xi \mid \lambda_c, \eta)$
0	.496	.638	0	0
.025	.477	.615	.060	.051
.05	.459	.594	.115	.099
.075	.442	.573	.166	.143
.10	.425	.554	.213	.185
.125	.410	.535	.256	.223
.15	.396	.518	.297	.259
.175	.382	.501	.334	.292
.20	.369	.486	.369	.324

Table 3
Effects on power and sample size of exponential losses with hazard rates η_e *and* η_c *in groups* E *and* C;
$R = 3$, $T = 5$, $\lambda_e = .20$, $\lambda_c = .30$, $\alpha = .05$ (*one-sided*)

a. Power for $N = 378$

	η_c				
η_e	0	.05	.10	.15	.20
.00	.901	.890	.879	.867	.855
.05	.892	.881	.870	.858	.846
.10	.883	.872	.860	.849	.837
.15	.873	.862	.850	.839	.827
.20	.863	.852	.840	.829	.817

b. Sample size for power = .90

	η_c				
η_e	0	.05	.10	.15	.20
.00	378	394	410	428	444
.05	390	406	424	440	458
.10	404	420	436	454	472
.15	418	434	450	468	486
.20	432	448	466	482	500

$\lambda_c = .30$, and where $R = 3$ and $T = 5$. Clearly as η increases, $E(\xi \mid \lambda, \eta)$ increases and $E(\delta \mid \lambda, \eta)$ decreases.

The effects on power for this example are shown in Table 3a for various loss hazards η_e and η_c in each group. For the extreme case $\eta_e = \eta_c = .20$, with probabilities of loss of .369 and .324 in each group, respectively, power is reduced to .817. Table 3b shows the sample size required to still achieve power of .90 in the presence of losses. For the extreme case $\eta_e = \eta_c = .20$, sample size must be increased to $N = 500$ to allow for the reduced probability of death in the presence of losses.

Note that the effects of losses with hazards η_e and η_c on power or sample size are roughly proportional to $\eta_e + \eta_c$. For example, for $\eta_e = \eta_c = .10$, power = .860, and for $\eta_e = .20$ and $\eta_c = 0$, power = .863. Also, note that if $L = E(\xi \mid \lambda, \eta)$ is the expected proportion of losses to follow-up in both groups combined, a conservative adjustment is to employ sample size $N_L = N(1 + L)$ where N is obtained assuming no losses. For example, for $\eta = .10$

from Table 2, $E(\xi) \approx .20$ in each group. A 20% increase in sample size from 378 to 454 is slightly greater than the actual sample size required from Table 3b ($N = 436$ for $\eta_e = \eta_c = .10$).

Thus far, we have considered only the impact of losses to follow-up in the case of uniform patient entry. However, if we employ the exponential entry distribution (3.1) in conjunction with exponentially distributed losses to follow-up, then from (A.2) we obtain

$$\phi(\lambda, \eta, \gamma) = \lambda^2\left\{\frac{\lambda}{\lambda+\eta} + \frac{\lambda\gamma e^{-(\lambda+\eta)T}[1 - e^{(\lambda+\eta-\gamma)R}}{(1 - e^{-\gamma R})(\lambda+\eta)(\lambda+\eta-\gamma)}\right\}^{-1}. \tag{4.3}$$

This reduces to (3.2) for $\eta = 0$.

5. Noncompliance

Various authors have used the terms "losses to follow-up," "dropout," and "noncompliance" interchangeably. In the present context, we use noncompliance to refer to less than maximally effective treatment in a patient who *continues follow-up*; i.e., in a patient who is not lost to follow-up. Schork and Remington (1967), Halperin et al. (1968), and Wu, Fisher, and DeMets (1980) have presented adjustments for noncompliance for sample size estimates based on the comparison of simple proportions. These approaches assume that treatment effect is manifest as a specific function of time in a compliant patient and/or that it dissipates as a function of time in a noncompliant patient.

In the following, we use the simpler and slightly conservative expression (2.6). Assuming equal sample fractions ($Q_e = Q_c = .5$), the expression for sample size can be written as $N^2 = (Z_\alpha + Z_\beta)\Sigma/(\lambda_e - \lambda_c)$, where Σ is the multiplier of $(Z_\alpha + Z_\beta)$ in (2.6).

We now assume the following: (i) $\omega_e N_e$ patients in the experimental group will be noncompliant and will have the same exponential hazard as the control group; and (ii) $\omega_c N_c$ patients in the control group will be noncompliant, and will actually cross over to the experimental therapy, thus having the same exponential hazard as the experimental group. If we then stratify by compliant vs noncompliant patients, we obtain modified hazards in the two groups with expectations and variances

$$E(\hat{\lambda}_e) = \lambda_e^* = (1 - \omega_e)\lambda_e + \omega_e\lambda_c, \tag{5.1}$$

$$\phi(\hat{\lambda}_e^*) = (1 - \omega_e)\phi(\lambda_e) + \omega_e\phi(\lambda_c), \tag{5.2}$$

with similar expressions for λ_c^* and $\phi(\lambda_c^*)$. Substituting expressions (5.1) and (5.2) into (2.6), we then obtain

$$\Sigma_\omega = 2\phi(\lambda_e) + 2\phi(\lambda_c) + 2(\omega_e - \omega_c)[\phi(\lambda_c) - \phi(\lambda_e)]. \tag{5.3}$$

However, for local alternatives the last term is near zero, and for $\omega_e = \omega_c$ it is exactly zero, and thus $\Sigma_\omega \approx \Sigma$.

The adjustment for noncompliance, therefore, is based principally on the dilution of the expected value of the treatment effect

$$\lambda_e^* - \lambda_c^* = (1 - \omega_e - \omega_c)(\lambda_e - \lambda_c). \tag{5.4}$$

Substituting (5.4) into the above expression based on (2.6), we then obtain

$$N_\omega = \frac{N}{(1 - \omega_e - \omega_c)^2}, \tag{5.5}$$

where N is the sample size required with full compliance in both groups and N_ω is the adjusted sample size required to allow for noncompliance.

Note that (5.5) allows for $\omega_c N_c$ *positive* noncompliers in the control group. In some cases,

however, noncompliance in a control patient would not consist of a "therapeutic crossover," as when the controls are treated with a placebo (double-masked). In such cases, the noncompliance in the control group would not affect the expected hazard and the resulting adjustment would be $N_\omega = N(1 - \omega_e)^{-2}$, as was originally suggested by Lachin (1981) and Lachin et al. (1981) to apply to such calculations.

This method will provide a conservative adjustment since a noncompliant patient in one group is assumed to be subject to the hazard rate for the other group over the entire study. However, it would also be possible to expand (5.1) and (5.2) to the case where the expected hazard rate for a noncompliant patient is a function of that fraction of a patient's study time over which the patient was noncompliant, and where the hazard rate changed as a smooth function of study time. Schork and Remington (1967), Halperin et al. (1968), and Wu et al. (1980) describe such models for an analysis based on simple proportions. For proportions, say p_e and p_c, the adjustment (5.5) is more conservative than these methods because the expected proportion p_e obtained analogously to (5.1) is closer to p_c than is the expected proportion obtained by these methods.

6. Stratified Analysis

In some cases, the trial may be designed with the intent of conducting a pooled (e.g., Mantel–Haenszel) analysis over strata. For simplicity we consider the case of two strata, say A and B, with total sample sizes N_a and N_b ($N = N_a + N_b$), where each stratum may have an associated recruitment and follow-up period (R_a, T_a, R_b, T_b) and associated hazards under the alternative hypothesis (λ_{ea}, λ_{ca}, λ_{eb}, λ_{cb}). For example, strata A and B may correspond to a prognostic factor (e.g., severity of disease), or stratum A may refer to a feasibility or pilot phase and stratum B to the main trial.

For such cases, the test statistic would employ pooled estimators of the within-stratum differences in hazard rates:

$$\hat{\Delta} = (\hat{\bar{\lambda}}_e - \hat{\bar{\lambda}}_c) = w_a(\hat{\lambda}_{ea} - \hat{\lambda}_{ca}) + w_b(\hat{\lambda}_{eb} - \hat{\lambda}_{cb}), \tag{6.1}$$

where $\hat{\lambda}_{ea}$, $\hat{\lambda}_{ca}$ are the estimated hazards from stratum A of N_a patients, and likewise for $\hat{\lambda}_{eb}$, $\hat{\lambda}_{cb}$ from stratum B of N_b patients. Further, let Q_{ea}, Q_{ca} be the sample fractions of the two treated groups in stratum A ($Q_{ea} + Q_{ca} = 1$), and likewise, Q_{eb}, Q_{cb} for stratum B. As shown in the Appendix, the basic equation relating sample size and power is

$$\sqrt{N}\,|\,\bar{\lambda}_e - \bar{\lambda}_c\,| = Z_\alpha\sqrt{\Omega^{-1}} + Z_\beta\sqrt{\Omega^{-2}(K_a\psi_{1a}/\psi_{0a}^2 + K_b\psi_{1b}/\psi_{0b}^2)} \tag{6.2}$$

where

$$\bar{\lambda}_e = w_a\lambda_{ea} + w_b\lambda_{eb},$$

$$\bar{\lambda}_c = w_a\lambda_{ca} + w_b\lambda_{cb},$$

and the weights w are given in (A.5), the ψ in (A.6), and K_a and K_b are the strata sample fractions (e.g., $K_a = N_a/N$). The above immediately generalizes to more than two strata.

Under the assumption of uniform entry with no losses to follow-up, the method of George and Desu (2.4 above) was extended to the case of a stratified analysis by Bernstein and Lagakos (1978), and that of Schoenfeld (1981) by Palta and Amini (1985). Under these conditions, these methods and (6.2) yield similar results.

We now consider three cases in which a stratified formulation might be employed: (i) strata with same study duration and stratum fractions specified a priori; (ii) strata with different durations of study and fractions specified a priori; and (iii) strata where sample size N_a for stratum A is fixed a priori and N_b for stratum B must be determined (recursively). These are illustrated in Table 4. In the examples, uniform entry is assumed and the results are presented with no losses and with losses where $\eta = .10$ for both groups in both strata.

Table 4

Sample size required to provide power = .90 to detect a $\frac{1}{3}$ reduction in the hazard rate for two prognostic strata with stratum-specific hazards; $\alpha = .05$ (one-sided), $E(\delta)$ = probability of death, $E(\xi)$ = probability of loss

a. Case (i): $K_a = .25$; $\lambda_{ca} = .21$; $\lambda_{cb} = .33$; $R = 3$; $T = 5$

	N	Power	$E(\delta \mid \lambda_e)$	$E(\delta \mid \lambda_c)$	$E(\xi \mid \lambda_e)$	$E(\xi \mid \lambda_c)$
Pooled with no losses	430	.900	.495	.604	0	0
A: Early stage	108	.188	.442	.513	0	0
B: Late stage	322	.924	.513	.672	0	0
Pooled with losses, $\eta = .10$	496	.900	.425	.549	.225	.185
A: Early stage	124	.186	.378	.440	.220	.210
B: Late stage	372	.924	.440	.585	.210	.177

b. Case (ii): $K_a = .25$; $\lambda_{ca} = \lambda_{cb} = .30$; $R_a = 1$; $T_a = 7$; $R_b = 3$; $T_b = 5$

	N	Power	$E(\delta \mid \lambda_e)$	$E(\delta \mid \lambda_c)$	$E(\xi \mid \lambda_e)$	$E(\xi \mid \lambda_c)$
Pooled with no losses	344	.900	.554	.693	0	0
A: Pilot phase	86	.507	.727	.857	0	0
B: Main phase	258	.783	.496	.638	0	0
Pooled with losses, $\eta = .10$	408	.900	.462	.569	.231	.197
A: Pilot phase	102	.489	.571	.694	.286	.231
B: Main phase	306	.791	.425	.554	.213	.185

c. Case (iii): $N_a = 100$; $N_b = 2$; $\lambda_{ca} = \lambda_{cb} = .30$; $R_a = 1$; $T_a = 7$; $R_b = 3$; $T_b = 5$

	N	Power	$E(\delta \mid \lambda_e)$	$E(\delta \mid \lambda_c)$	$E(\xi \mid \lambda_e)$	$E(\xi \mid \lambda_c)$
Pooled with no losses	338	.900	.564	.703	0	0
A: Pilot phase	100	.558	.727	.857	0	0
B: Main phase	238	.753	.496	.638	0	0
Pooled with losses, $\eta = .10$	408	.900	.461	.588	.231	.196
A: Pilot phase	100	.482	.571	.694	.286	.231
B: Main phase	308	.793	.425	.554	.213	.185

Case (i) Consider the case of two prognostic strata, such as early- versus late-stage disease, with stratum-specific hazards and relative strata sizes $K_a = .25$ and $K_b = .75$. We assume that the control group hazard in the early-stage stratum is less than that in the late-stage stratum ($\lambda_{ca} = .21$ and $\lambda_{cb} = .33$). However, we again desire to detect a $\frac{1}{3}$ reduction in the hazard for the experimental group ($\lambda_{ea} = .14$ and $\lambda_{eb} = .21$). With no allowance for losses, the optimal weights are $w_a = .3492$ and $w_b = .6508$, which yields pooled hazards $\bar{\lambda}_e = .196$ and $\bar{\lambda}_c = .288$. With allowances for losses ($\eta_e = \eta_c = .10$ in both strata), the optimal weights are approximately the same as above. The resulting sample sizes are presented in Table 4a. Although the pooled analysis provides power .90, within the smaller stratum A, with the lower hazard rates, the power to detect a $\frac{1}{3}$ reduction in hazards is only .18, whereas within the larger stratum with higher hazard rates, power is .92.

Case (ii) Often it is planned to conduct a pilot or feasibility phase before launching the main trial. The purpose of such a pilot phase is not to address the principal study question (e.g., mortality), but rather to provide a complete assessment of trial procedures and of the treatment effects on precursors of the ultimate expected outcome. If the pilot phase is successful, the patients enrolled in this initial phase are included as a separate stratum in the total study and are added to those recruited in the main phase of the trial. Pilot phase patients, therefore, are usually followed for a longer duration. In the simplest case, the pilot phase sample size is specified as a fraction of the total sample size.

Consider the case of a pilot phase to comprise $\frac{1}{4}$ of the study patients ($K_a = .25$), where patients are to be recruited over a 1-year period ($R_a = 1$) and followed for at least 1 year before it is decided whether the main trial phase should be launched. Consider that the main phase will then include a 3-year accrual period ($R_b = 3$) and a total period of study of 5 years ($T_b = 5$). Thus, the total period of study for the pilot phase patients would be 7 years ($T_a = 7$). We further assume that the experimental and control hazard rates $\lambda_e = .20$ and $\lambda_c = .30$ apply to both strata. With no losses, the optimal weights are $w_a = .31817$ and $w_b = .68183$, and the resulting sample sizes are presented in Table 4b. Due to the longer duration of follow-up, the pilot phase stratum has power .50, and due to the larger sample size, the main trial stratum has power .79.

Case (iii) Often, however, the sample size for the pilot phase is based on independent considerations, such as the power to detect a difference in a secondary (e.g., precursor) outcome. In this case N_a is fixed and it is desired to determine the additional sample size $N_b = N - N_a$ required for the main trial phase. This requires a recursive solution which is easily obtained through the method of false position. The homogeneous equation based on (6.2) is

$$e = \frac{1}{N}\left|N_a(\lambda_{ea} - \lambda_{ca}) + (N - N_a)(\lambda_{eb} - \lambda_{cb})\right| - Z_\alpha\sqrt{N_a/\psi_{0a} + (N - N_a)/\psi_{0b}} - Z_\beta\sqrt{\left[\frac{N_a\psi_{1a}}{\psi_{0a}^2} + \frac{(N - N_a)\psi_{1b}}{\psi_{0b}^2}\right]\Big/\left[\frac{N_a}{\psi_{0a}} + \frac{(N - N_a)}{\psi_{0b}}\right]^2}. \tag{6.3}$$

Note that N_a is fixed and in essence one must solve for N. The recursive equation is

$$N_{j+1} = N_j - \frac{(N_j - N_{j-1})e_j}{(e_j - e_{j-1})}, \quad j = 1, 2, \ldots, \tag{6.4}$$

where the solution is obtained at that cycle where $e_j = 0$. As starting values, N_0 can be computed using the single-stratum equations with the study times R_a, T_a for the pilot

stratum; and N_1 can be computed likewise with the study times R_b, T_b ($T_b < T_a$) for the main stratum.

For example, Table 4c presents the total sample size required for the same example as in case (ii) but where $N_a = 100$ is specified a priori and the above recursive equations are applied to obtain N and thus N_b.

Finally, in all such cases, further adjustments for noncompliance can be applied using (5.5). For example, in case (iii), if a further adjustment for $\omega_e = .20$ noncompliance in group E alone is desired, in addition to the adjustment for losses, then $N_\omega = (408)(.80)^{-2} = 638$ is required.

Listings of computer programs that provide the above calculations are available from the authors on request.

Acknowledgements

This work was partially supported by the Diabetes Control and Complications Trial under Contract No. N01-AM-2-2206B from the National Institute of Diabetes, Digestive and Kidney Diseases (NIDDK), and by the Lupus Nephritis Collaborative Study under Grant No. R01-AM-27769-02 from the NIDDK. The authors especially thank Max Halperin for many helpful comments.

Résumé

Quand on met en place un essai clinique pour tester l'égalité des distributions de survie de deux groupes de traitement, les hypothèses habituelles sont la survie exponentielle, l'entrée uniforme des patients, la conformité, et la censure et, seulement à la fin de l'essai. De nombreux auteurs ont présenté des méthodes d'estimation de la taille de l'échantillon et de la puissance sous ces hypothèses, certains permettant une période d'accumulation de R avec un total de T années d'étude, $T > R$. La méthode de Lachin (1981, *Controlled Clinical Trials* **2,** 93–113) est étendue aux cas où les patients n'entrent pas dans l'essai de façon uniforme dans le temps, les patients peuvent sortir de l'essai par perte du suivi (autre qu'administratif); d'autres patients peuvent continuer à être suivis, bien que ne se soumettant plus au régime du traitement, et une analyse stratifiée peut être prévue selon une ou plusieurs covariables de pronostic.

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.

Chiang, C. L. (1968). *Introduction to Stochastic Processes in Biostatistics.* New York: Wiley.

Cox, D. R. (1964). Some applications of exponential ordered scores. *Journal of the Royal Statistical Society, Series B* **26,** 103–110.

Cuzick, J. (1982). The efficiency of the proportions test and the log-rank test for censored survival data. *Biometrics* **38,** 1033–1039.

Donner, A. (1984). Approaches to sample size estimation in the design of clinical trials—A review. *Statistics in Medicine* **3,** 199–214.

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. (1974). Planning the size and duration of a clinical trial studying the time to some critical event. *Journal of Chronic Diseases* **27,** 15–24.

Gross, A. J. and Clark, V. A. (1975). *Survival Distributions: Reliability Applications in the Biomedical Sciences.* New York: Wiley.

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.

Johnson, N. L. and Kotz, S. (1970). *Continuous Univariate Distributions, Vol.* 1. New York: Wiley.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* New York: Wiley.

Lachin, J. M. (1981). Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials* **2,** 93–113.

Lachin, J. M., Marks, J., Schoenfield, L. J., the Protocol Committee, and the NCGS Group (1981). Design and methodological considerations in the National Cooperative Gallstone Study: A multicenter clinical trial. *Controlled Clinical Trials* **2,** 177–230.

Lustbader, E. D. and Litwin, S. (1981). Design of censored survival studies. *Biometrics* **37,** 697–704.

Makuch, R. W. and Simon, R. M. (1982). Sample size requirements for comparing time-to-failure among k treatment groups. *Journal of Chronic Diseases* **35,** 861–867.

Palta, M. and Amini, S. B. (1985). Consideration of covariates and stratification in sample size determination for survival studies. *Journal of Chronic Diseases* **38,** 801–809.

Palta, M. and McHugh, R. (1979). Adjusting for losses to follow-up in sample size determination for cohort studies. *Journal of Chronic Diseases* **32,** 315–326.

Palta, M. and McHugh, R. (1980). Planning the size of a cohort study in the presence of both losses to follow-up and noncompliance. *Journal of Chronic Diseases* **33,** 501–512.

Pasternack, B. S. (1972). Sample sizes for clinical trials designed for patient accrual by cohorts. *Journal of Chronic Diseases* **25,** 673–681.

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.

Schoenfeld, D. A. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* **68,** 316–319.

Schoenfeld, D. A. and Richter, J. R. (1982). Nomograms for calculating the number of patients needed for a clinical trial with survival as an endpoint. *Biometrics* **38,** 163–170.

Schork, M. A. and Remington, R. D. (1967). The determination of sample size in treatment–control comparisons for chronic disease studies in which dropout or nonadherence is a problem. *Journal of Chronic Diseases* **20,** 233–238.

Willan, A. R. (1983). Relative efficiency in survival analysis. *Journal of Chronic Diseases* **36,** 183–192.

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

Received July 1985; *revised June* 1986.

Appendix

The general expressions for $\phi(\lambda)$ are presented herein. In each case $\phi(\lambda)$ is expressed in the form $\phi(\lambda) = \lambda^2 E(\delta \mid \lambda)^{-1}$.

Nonuniform entry Assume that patient entry follows an arbitrary distribution function $G(z)$, where z_i is the entry time of the ith patient, $0 \leq z_i \leq R$. This implies a potential follow-up or exposure time $f_i = T - z_i$, $T - R \leq f_i \leq T$, i.e., f_i is the administrative censoring time. It follows that

$$\phi(\lambda) = \lambda^2 \left[\int_{T-R}^{T} g(T-f)(1 - e^{-\lambda f})\, df \right]^{-1}. \tag{A.1}$$

Losses to follow-up For the ith patient, z_i is again the recruitment time, with distribution function $G(z)$, which implies an exposure period $f_i = T - z_i$. Let t_i denote the death time for the ith patient, and let the corresponding indicator variable δ_i denote whether the ith patient was observed to die during the study. In addition, let u_i denote the time of loss to follow-up, which follows a loss distribution function $H(u)$ over the complete follow-up interval 0 to T. Also, let ξ_i be the corresponding indicator variable to denote whether the ith patient was lost to follow-up during the study. As before, the variance is a function of the expected number of deaths, where in this context $E(\delta)$ is evaluated with respect to $G(z)$ and $H(u)$. Clearly

$$\begin{aligned} E(\delta) &= \Pr[t < \min(f, u)] \\ &= \Pr(t \leq f \text{ and } u > f) + \Pr(t \leq u \text{ and } u \leq f). \end{aligned}$$

In terms of exponential survival, this yields

$$\phi(\lambda) = \lambda^2\left\{\int_{T-R}^{T} g(T-t)(1-e^{-\lambda t})[1-H(t)]\,dt + \int_{T-R}^{T} g(T-t)\int_0^t (1-e^{-\lambda u})h(u)\,du\,dt\right\}^{-1}. \quad \text{(A.2)}$$

$E(\delta)$ can also be expressed as $\Pr(t \leqslant f \text{ and } u > t)$, which yields

$$\phi(\lambda) = \lambda^2\left\{\int_0^R \int_0^{T-z} (\lambda e^{-\lambda t})g(z)[1-H(t)]\,dt\,dz\right\}^{-1}. \quad \text{(A.3)}$$

Although sample size and power depend on $E(\delta)$ through either (A.2) or (A.3), it is also informative to examine the expected proportion of losses to follow-up, $E(\xi)$. Under either H_0 or H_1, $E(\xi)$ may be evaluated from

$$E(\xi) = \int_0^R \int_0^{T-z} g(z)h(u)e^{-\lambda u}\,du\,dz. \quad \text{(A.4)}$$

Stratification From (6.1) the MVUE is obtained using weights (w_a and w_b) inversely proportional to the variances of the within-stratum differences. Under H_0: ($\lambda_{ea} = \lambda_{ca}$) and ($\lambda_{eb} = \lambda_{cb}$),

$$w_a = \frac{1}{\sigma_{0a}^2}\left(\frac{1}{\sigma_{0a}^2} + \frac{1}{\sigma_{0b}^2}\right)^{-1}, \quad w_b = \frac{1}{\sigma_{0b}^2}\left(\frac{1}{\sigma_{0a}^2} + \frac{1}{\sigma_{0b}^2}\right)^{-1}, \quad \text{(A.5)}$$

where

$$\begin{aligned} \sigma_{0a}^2 &= \frac{\psi_{0a}}{N_a}, \\ \psi_{0a} &= \phi_a(\bar{\lambda}_a)(Q_{ea}^{-1} + Q_{ca}^{-1}), \\ \bar{\lambda}_a &= Q_{ea}\lambda_{ea} + Q_{ca}\lambda_{ca}; \end{aligned} \quad \text{(A.6)}$$

and likewise σ_{0b}^2 is a function of ψ_{0b}, $\bar{\lambda}_b$, and N_b. All terms $\phi_a(\cdot)$ are evaluated as described in Sections 3 and 4 with respect to the periods R_a, T_a and the entry and loss distributions within stratum A; similarly, terms $\phi_b(\cdot)$ are evaluated for stratum B.

Under H_0 it then follows that $\Sigma_0^2 = \text{var}(\hat{\Delta})$ is given as $\Sigma_0^2 = w_a^2\sigma_{0a}^2 + w_b^2\sigma_{0b}^2$. Under the alternative hypothesis, $\Sigma_1^2 = w_a^2\sigma_{1a}^2 + w_b^2\sigma_{1b}^2$, where

$$\begin{aligned} \sigma_{1a}^2 &= \frac{\psi_{1a}}{N_a}, \\ \psi_{1a} &= \frac{\phi_a(\lambda_{ea})}{Q_{ea}} + \frac{\phi_a(\lambda_{ca})}{Q_{ca}}, \end{aligned} \quad \text{(A.7)}$$

and likewise σ_{1b}^2 is a function of ψ_{1b}, λ_{eb}, λ_{cb}, and N_b.

If we then express the strata sample sizes as fractions of the total, $N_a = K_aN$, $N_b = K_bN$, it is easily shown that

$$\begin{aligned} \Sigma_0^2 &= (N\Omega)^{-1}, \\ \Sigma_1^2 &= \left(\frac{K_a\psi_{1a}}{\psi_{0a}^2} + \frac{K_b\psi_{1b}}{\psi_{0b}^2}\right)(N\Omega^2)^{-1}, \end{aligned} \quad \text{(A.8)}$$

where

$$\Omega = \left(\frac{K_a}{\psi_{0a}} + \frac{K_b}{\psi_{0b}}\right). \quad \text{(A.9)}$$

Substituting the above into the general equation relating sample size and power, $|\bar{\lambda}_e - \bar{\lambda}_c| = Z_\alpha\Sigma_0 + Z_\beta\Sigma_1$, then yields the expression (6.2).