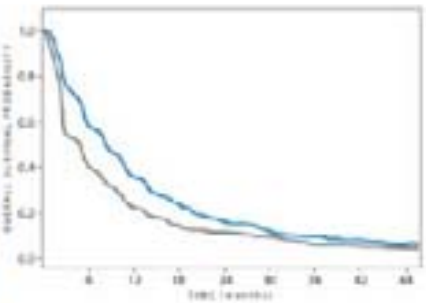
WM. KEVIN KELLY
SUSAN HALABI

ONCOLOGY CLINICAL TRIALS



Successful Design, Conduct, and Analysis

demosMEDICAL

11 Design of Phase III Clinical Trials

Stephen L. George

A phase III clinical trial is a randomized prospective controlled study designed to compare the efficacy of two or more regimens for the treatment of a specified disease or medical condition. These trials employ accepted scientific principles of good experimental design including, among other things, specification of eligibility criteria (types of patients appropriate for study), explicit statements of primary and secondary objectives, details of the treatment regimens to be compared, and statistical considerations (hypotheses tested, sample size and expected duration of the trial, statistical procedures, interim analysis plans, and related topics). A properly designed and executed phase III clinical trial provides the best available scientific evidence on the relative efficacy of the treatments being studied and the most reliable information for evidence-based medicine.

The adoption and wide-spread use of phase III clinical trials in the latter half of the twentieth century and the early twenty-first century represents one of the more important contributions to the practice of scientific medicine during the last 60 years. The statistical aspects of the design, conduct, and analysis of clinical trials have been extensively studied during this time and there are now many textbooks (including this one) covering this material at various levels of statistical sophistication. Other chapters in the current text cover important topics in the design and analysis of phase III clinical trials, including selecting endpoints, randomization and stratification, interim analysis, adaptive design, and Bayesian designs. The focus in the present chapter will be on determining the required sample size (number of patients or number of events) and duration of a phase III clinical trial in many commonly encountered practical situations (1, 2). In an attempt to provide maximum clarity for the underlying concepts, free of unnecessary complexity, the situations considered are elementary ones. References to papers covering more complex scenarios are provided where appropriate.

CANONICAL SAMPLE SIZE FORMULAE

Testing Hypotheses

The sample size considerations in this chapter are derived from a statistical hypothesis testing perspective, usually involving a test of a *null* hypothesis, H_0, against an alternative hypothesis, H_1. In the simplest case, suppose the outcome variable (endpoint) of a clinical trial comparing two treatments is some continuous random variable, X, and we wish to compare the mean value of X for the two treatments. The usual null and alternative hypotheses in this case may be expressed as

$$\begin{aligned} &H_0 : \mu_1 = \mu_2 \\ &\quad \text{vs.} \\ &H_1 : \mu_1 \neq \mu_2 \end{aligned} \qquad \text{(Eqn. 11-1)},$$

where μ_i is the mean for treatment i (= 1, 2).

The statistical inference in this case is a decision rule, based on the observed values of X in the two treatment groups, for deciding between the two competing hypotheses. In the standard statistical approach, the trial is designed to control the rates for the two possible types of error:

- Type I—rejecting H_0 (in favor of H_1) when H_0 is true
- Type II—not rejecting H_0 when H_1 is true

The error rates for these two types of errors are conventionally denoted by α and β, respectively, with the *power* of the test defined as $1-\beta$, the complement of the type II error rate. That is, the power of the test is the probability of correctly rejecting H_0 when H_1 is true. The usual approach to determining the required sample size is to set the type I error rate α at some acceptable level, often 0.05 or 0.01, and then to find the minimum required sample size to achieve a power of at least some specified value $1-\beta$, often 0.80 or 0.90, at some specified alternative value (i.e., some particular value in the alternative hypothesis space when H_1 is a composite space).

Suppose (for simplicity) that the endpoint X_i in the ith treatment group ($i = 1, 2$) has a normal statistical distribution with mean μ_i and known variance σ^2, denoted, $X_i \sim N(\mu_i, \sigma^2)$, and we plan to enter an equal number of patients, n, on each treatment. In this case, the usual test statistic, Z, used in testing H_0 versus H_1 may be written as:

$$Z = \sqrt{\frac{n}{2}}\frac{|\bar{X}_1 - \bar{X}_2|}{\sigma} \qquad \text{(Eqn. 11-2)},$$

where $\bar{X}_i$ is the sample mean of X in the ith treatment group. The hypothesis $H_0 : \mu_1 = \mu_2$ is rejected in favor of $H_1 : \mu_1 \neq \mu_2$ if $Z \geq z_{\alpha/2}$, where z_x is the upper 100 $(1 - x)$ percentile of the standard normal distribution. To determine the required common sample size, we need to solve the following equation for n:

$$P\left(Z \geq z_{\alpha/2} \middle| \delta\right) \geq 1-\beta \qquad \text{(Eqn. 11-3)},$$

where $P(X|Y)$ denotes the probability of X given Y and $\delta = \mu_1 - \mu_2 \neq 0$ is some prespecified value in the alternative hypothesis space. That is, we want the power to be at least $1-\beta$ when the true difference between the means is δ. Some straight-forward algebraic manipulation of (11-3) yields the following sample size formula for the approximate number of patients, n, required on each treatment group:

$$n = \left[\frac{2\sigma^2\left(z_{\alpha/2} + z_\beta\right)^2}{\delta^2}\right] \qquad \text{(Eqn. 11-4)},$$

where $[x]$ denotes the smallest integer $\geq x$. Equation (11-4) is the canonical sample size formula for the hypothesis-testing scenario considered here. If the variances in the two groups are not equal, the formula becomes

$$n = \left[\frac{\left(\sigma_1^2 + \sigma_2^2\right)\left(z_{\alpha/2} + z_\beta\right)^2}{\delta^2}\right] \qquad \text{(Eqn. 11-5)},$$

These formulae can also be used to determine the power for a given sample size by solving for z_β when n is given.

Although the above formulae are strictly applicable only for the assumptions underlying their derivation, they are approximately correct in many other settings. Often the test statistic or some simple transformation of the test statistic is approximately normally distributed for reasonably large sample sizes. Also, the formulae nicely illustrate several general points about the required sample size in phase III clinical trials:

- The required sample size n increases as the variance σ^2 increases. The size of σ^2 is a feature of the population under study.
- The required sample size n increases as the error rates decrease. For example, an increase in power (decrease in β) requires an increase in sample size.
- The required sample size n increases as the detectable effect size δ decreases.

The required sample size calculated from (11-4) for some common values of α and β as a function of the standardized effect size, δ/σ, is given in Table 11.1. If the variances are not equal, the standardized effect size may be defined as $\delta/\bar{\sigma}$ where $\bar{\sigma} = \sqrt{\left(\sigma_1^2 + \sigma_2^2\right)/2}$.

For example, if $\alpha = 0.05$, $1-\beta = 0.90$, and $\delta/\sigma = 0.50$, the number of patients required on each treatment is $n = 85$ and the total required sample size is $2n = 170$.

Unknown Variances

If we relax the assumption that sigma is known, the appropriate test statistic is not (11-2) but the t-statistic

$$T = \sqrt{\frac{n}{2}}\frac{|\bar{X}_1 - \bar{X}_2|}{s} \qquad \text{(Eqn. 11-6)},$$

where s is the pooled estimate of the common, but unknown, standard deviation σ. In this case, a good approximation (3) to the required sample size n^* is:

$$n^* = n + \left[\frac{z_{\alpha/2}^2}{4}\right] \qquad \text{(Eqn. 11-7)},$$

TABLE 11.1
Required Sample Size on Each Treatment Arm to Test $H_0: \mu_1 = \mu_2$ vs. $H_0: \mu_1 \neq \mu_2$.

		δ/σ				
α	1—β	**0.10**	**0.25**	**0.50**	**0.75**	**1.00**
0.01	0.80	2336	374	94	42	24
	0.90	2976	477	120	53	30
0.05	0.80	1570	252	63	28	16
	0.90	2102	337	85	38	22

where, as before, $[x]$ is the smallest integer $\geq x$ and n is defined in (11.4). Although n* is always greater than n, the difference is not large. For example, $n^* - n = 1$ when $\alpha = 0.05$ and $n^* - n = 2$ when $\alpha = 0.01$. Thus, in most practical situations, the required number of patients when the variance is unknown is only one or two more patients per treatment group than the number required when the variance is known.

Unequal Sample Sizes

To allow for different sample sizes n_i on the two arms, (11-4) may be written as:

$$\left(\frac{1}{n_1}+\frac{1}{n_2}\right)^{-1} = \left[\frac{\sigma^2\left(z_{\alpha/2}+z_\beta\right)^2}{\delta^2}\right] \quad \text{(Eqn. 11-8)},$$

Any pair of values (n_1, n_2) that satisfy (11-8) will work. However, the required total sample size, $n_1 + n_2$, is minimized when $n_1 = n_2$. If we randomize patients to the two treatments in the ratio r:1, for some $r > 1$, rather than in the usual balanced 1:1 ratio (i.e., $n_1 = rn_2$), the required sample sizes are

$$n_1 = \left[\frac{r+1}{2}n\right]$$
$$n_2 = \left[\frac{r+1}{2r}n\right] \quad \text{(Eqn. 11-9)},$$

and the required total sample size is approximately

$$n_1 + n_2 = 2n\left(r+1\right)^2 \Big/ 4r \quad \text{(Eqn. 11-10)},$$

where n is determined by (11-4). The inflation factor of $(r+1)^2/4r$ in the required total number of patients is the primary reason that balanced randomization is generally preferred to unbalanced randomization. For example, if $r = 2$ (i.e., twice as many patients are entered on treatment 1 than on treatment 2), we would require $(2n)(9/8)$ patients, a 12.5% increase over the $2n$ required in the balanced case. However, there may be other reasons for an unbalanced randomization (such as wanting more patients on one of the treatments to increase the precision of the estimated outcomes for that treatment). If so, an unbalanced allocation might be preferable to a balanced one even with the resultant sample size inflation factor.

More Than Two Treatment Groups

In many phase III clinical trials there are k > 2 treatments (4). Unfortunately, in order to control the error rates in this setting, the above sample size formulae cannot be extended simply by entering n patients on each of the k treatments (a total of kn patients). More than kn patients are required. Three important types of phase III clinical trials with more than two arms are considered below.

Testing Equality Among k Treatment Arms. In the simplest type of k arm clinical trial, there is a randomization to one of the k arms and the primary objective is to test a global null hypothesis. With an obvious extension of the notation in (11-1), the hypotheses being tested are

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$$
vs.
$$H_1: \mu_i \neq \mu_j \text{ for some } i \neq j \quad \text{(Eqn. 11-11)},$$

If σ^2 is known then the test statistic, analogous to (11-2) in the two-sample case, is

$$X^2 = \frac{n\sum_{i=1}^{k}\left(\bar{x}_i - \bar{\bar{x}}\right)^2}{\sigma^2} \quad \text{(Eqn. 11-12)},$$

where $\bar{x}_i$ is the sample mean in the ith treatment arm and $\bar{\bar{x}}$ is the overall sample mean. The hypothesis

H_0 is rejected in favor of H_1 if $X^2 > \chi^2_{\alpha,k-1}$, the upper 100(1–α) percentile of a chi-square distribution with $k - 1$ degrees of freedom. To determine the required sample size we need to solve an equation similar in form to (11-3):

$$P\left(X^2 \geq \chi^2_{\alpha,k-1} \middle| \Delta\right) \geq 1-\beta \qquad \text{(Eqn. 11-13)},$$

where $\Delta = \sum_{i=1}^{k} \left(\mu_i - \bar{\mu}\right)^2 \Big/ \sigma^2$ and $\bar{\mu} = \frac{1}{k}\sum_{i=1}^{k} \mu_i$. When $\Delta \neq 0$, X^2 has a noncentral chi-square distribution with noncentrality parameter $n\Delta$ and no closed-form solution for n exists analogous to (11-4). However, solutions are easily available either from computer programs or from tables of the noncentral chi-square distribution. As noted previously, the required sample size per treatment arm for $k > 2$ will be larger than that required when $k = 2$, increasing as a function of k. Table 11.2 gives the multiplication factor required for $k = 3$, 4, 5, and 6 when all means other than the largest and smallest are midway between the largest and smallest.

For example, the number of patients required for $k = 3$ is $1.23n$ per arm (i.e., 23% more patients on each arm) when $\alpha = 0.05$, $1—\beta = 0.80$. For $k = 4$, 5, and 6, the requirements are $1.39n$, $1.52n$, and $1.63n$, respectively. Although Table 11.2 represents the worst-case scenario, the one with the least favorable configuration of mean values between the two extreme means, there is generally a high price to be paid for increasing the number of treatment arms on a clinical trial.

Two or More Experimental Arms and a Control Arm. Another common k-arm design results when we wish to compare k—1 new or experimental arms with a standard or control with randomization of each patient to one of the k arms. In this case, there are k—1 comparisons of interest. If we let arm 1 be the control arm and define $\delta^* = \min_{i=2,\ldots,k}\{\mu_i - \mu_1\}$, then a simple, albeit conservative, approach in this setting is to apply equation (11-4), substituting α/2(k—1) for α/2 and δ* for δ. This approach ensures that a sufficient number of patients are entered to achieve the requisite power for all comparisons, allowing for the multiple comparisons.

A better approach, requiring fewer patients, is to adjust for the inherent multiplicity using a less conservative multiple comparison procedure (5–7). Jung et al. (8) use a Dunnett-type procedure for this purpose in the setting of survival distributions (see the "Comparing Survival Distributions" section later in the chapter for more details on survival endpoints).

TABLE 11.2
Multiplication Factors for the Number of Patients Required for k > 2 Treatment Arms.

		k = Number of Arms			
α	1—β	3	4	5	6
0.01	0.80	1.19	1.32	1.43	1.53
	0.90	1.17	1.29	1.39	1.48
0.05	0.80	1.23	1.39	1.52	1.63
	0.90	1.20	1.35	1.47	1.57

Factorial Designs. In a factorial design there are several factors (or treatment types in a clinical trial) to be tested in combination (9). In the simplest type of factorial design, referred to as a 2×2 design, there are two treatments, each given at one of two levels. For example, the treatments might refer to particular therapeutic agents (say, A and B) and the two levels might refer to the presence or absence of a specified regimen for the agent. The four possible combined treatment regimens are: A and B absent; A absent and B present; A present and B absent; A and B both present. In general, with two factors we could define a R × C factorial design, with R levels of one factor and C levels of the other factor, although in clinical applications it would be rare for R or C to exceed three. A 2 × 2 design is by far the most common factorial design.

In a factorial design it is possible to compare the effects of each of the treatments separately as well as to estimate the interaction effects among treatments. An interaction implies that the effect of a treatment depends on the presence or absence of another treatment. To make this point clear, consider a 2×2 design with two treatments either present or absent. The mean values in each of the four possible treatment combinations are given in Table 11.3.

The treatment effects in Table 11.3 are the differences in mean values when the treatment is given compared to when it is not given. The quantity ε measures the *interaction* between treatments. If $\varepsilon = 0$, the effect of treatment A is δ_0 regardless of whether treatment B is given or not, and the effect of treatment B is δ_1 regardless of whether treatment A is given or not. If $\varepsilon > 0$ (a *positive* interaction), there is synergy between the treatments; the effect of each treatment is increased in the presence of the other. If $\varepsilon < 0$ (a *negative* interaction), there is antagonism between the treatments; the effect of each treatment is decreased by the presence of the other.

TABLE 11.3
Treatment Effects in a 2 × 2 Factorial Clinical Trial.

		Treatment B		
		Absent	Present	Treatment B Effect
Treatment A	absent	μ	$\mu + \delta_1$	δ_1
	present	$\mu + \delta_0$	$\mu + \delta_0 + \delta_1 + \varepsilon$	$\delta_1 + \varepsilon$
	Treatment A effect	δ_0	$\delta_0 + \varepsilon$	

If one can assume that $\varepsilon \cong 0$, then a factorial design is highly efficient. One can design the trial to test the effect of treatment A or B without consideration of the other treatment and get "two trials for the price of one." If n_A and n_B are the required sample sizes for the two individual trials, then a single factorial trial of size $n = \max\{n_A, n_B\}$ will achieve at least the same power for each individual treatment comparison as two trials with total sample size of $n_A + n_B$. In fact, the power will be greater than required for the comparison with the smaller required n_i. However, if $\varepsilon < 0$, the power of the individual comparisons may be considerably less than that when there is no interaction. To allow for this possibility, one option is to assume a slight negative interaction in the design of the trial and increase the size of the trial accordingly. Unfortunately, if one wishes to test formally for interactions, the required size of the trial will be quite large, counteracting one of the primary advantages of a factorial design (9). It would also be possible to consider the trial as if it were a trial with k = RC treatments and use the approach outlined above for k treatment arms. However, this approach also can result in a very large trial, and in any case does not take advantage of the unique structure of the factorial design.

Factorial clinical trials can play an important role in evaluating therapies, especially in a setting where treatments are likely to be used in combination in practice. Indeed, such trials are essential to learn about the joint effects of treatments. However, if the treatment combinations are not likely to be used in practice, factorial designs are not appropriate because of the potential for negative interactions and the resultant loss of power.

COMPARING SUCCESS RATES

If the assumptions underlying the above derivations are approximately correct, the resultant sample size formulae can in principle be used directly. However, it is often necessary to consider modifications of the approach when designing actual clinical trials. One such situation requiring special consideration concerns trials in which the outcome measure is a binary variable. Such trials are considered in this section. Another important special case, trials in which the outcome measure is time to some event, is considered in the subsequent section.

In some phase III clinical trials the outcome on each patient is assessed as a *success* or a *failure* and the objective of the trial is to compare the success rates of the treatments under study. For example, a success might be defined as achieving a particular clinical status, perhaps achieving an objective response or remaining disease free for some specified time period. In these cases, the observed success rate on treatment i will have a binomial distribution with a mean p_i, the unknown probability of success, and variance $p_i(1 - p_i)/n$. The hypotheses equivalent to those in (11-1) are

$$H_0: p_1 = p_2 \quad \text{vs.} \quad H_1: p_1 = p_2 \qquad \text{(Eqn. 11-14)},$$

Even though the binomial distribution is not a normal distribution, for large samples a normal approximation is reasonable and one may use equation (11-5) directly with $\delta = p_1 - p_2$ and $\sigma_i^2 = p_i(1 - p_i)$. That is,

$$n = \left[\frac{\left(p_1(1-p_1) + p_2(1-p_2)\right)\left(z_{\alpha/2} + z_\beta\right)^2}{\left(p_1 - p_2\right)^2} \right]. \qquad \text{(Eqn. 11-15)},$$

A second approach is to apply the variance-stabilizing transformation $\arcsin\sqrt{x/n}$ to the observed proportion of success x/n. This approach yields

$$n = \left[\frac{\left(z_{\alpha/2} + z_\beta\right)^2}{2\left(\arcsin\sqrt{p_1} - \arcsin\sqrt{p_2}\right)^2} \right]. \qquad \text{(Eqn. 11-16)},$$

TABLE 11.4
Number of Patients on Each of Two Treatments to Compare Success Rates ($\alpha = 0.05$, $1-\beta = 0.80$).

	$\delta = p_2 - p_1$							
p_1	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**
0.10	725	219	113	72	51	38	30	25
0.20	1134	313	151	91	62	45	35	28
0.30	1416	376	176	103	68	49	37	29
0.40	1573	408	186	107	70	49	36	28
0.50	1605	408	183	103	66	45	33	25

A third approach, also based on the approximate normality of the sample proportions (10), yields

$$n = \left[\frac{\left(z_{\alpha/2}\sqrt{2\bar{p}\left(1-\bar{p}\right)} + z_{\beta}\sqrt{p_1\left(1-p_1\right)+p_2\left(1-p_2\right)} \right)^2}{\left(p_1 - p_2\right)^2} \right].$$

(Eqn. 11-17),

Haseman (11) showed that all of the above formulae result in values that are too small when the actual test being used is an exact test. Casagrande et al. (12) provided an improved formula in this setting and Fleiss et al (13) showed that a better approximation results from a simple modification to (11-17):

$$n = \left[\frac{\left(z_{\alpha/2}\sqrt{2\bar{p}(1-\bar{p})} + z_{\beta}\sqrt{p_1\left(1-p_1\right)+p_2\left(1-p_2\right)} \right)^2 + 2\left|p_1 - p_2\right|}{\left(p_1 - p_2\right)^2} \right].$$

(Eqn. 11-18),

Table 11.4 gives the required sample sizes based on (11-18) for some selected cases.

COMPARING SURVIVAL DISTRIBUTIONS

When the hypotheses being tested involve time-to-event, or survival data, several complications arise. The most important one is that the observations may be incomplete (or *censored*) at the time of the analysis, either because of dropouts or loss to follow-up or because the event in question (recurrence, progression, death, etc.) has not yet occurred for some patients. Censoring affects both the number of patients that need to be enrolled on trial as well as the required duration of trial. For reasons that will be made clearer below, the number of events, rather than the number of patients on trial, is the key quantity to be determined and the duration of the trial must be planned to achieve the desired number of events. There has been a vast literature on the design of clinical trials with survival as the major endpoint (8, 14–32), mostly at a more advanced statistical level than the level of this chapter.

Required Number of Events

For a random variable T representing the time to some event, the key probability functions are the survival distribution or probability of surviving beyond time t, $S(t) = P(T > t)$, and the hazard function $\lambda(t) = f(t)/S(t)$, where $f(t)$ is the probability density function. The hazard function may be thought of as the instantaneous failure rate at time t for a patient who has survived up to time t. Each function may be derived from the other if the other is fully specified. The simplest type of survival distribution is the exponential distribution, for which the hazard function is constant over time.

George and Desu (15) provided a framework for determining both the required number of events and the required duration of study when the survival distributions under study follow an exponential distribution. In this case, the survival function, the probability of surviving beyond time t, is $S_i(t) = \exp(-\lambda_i t)$ in the ith treatment group, where λ_i is the hazard rate in the ith treatment group. The hypotheses being tested, analogous to those in equation (11-1), are

$$H_0 : \lambda_1 = \lambda_2 \quad \text{vs.} \quad H_1 : \lambda_1 \neq \lambda_2 \qquad \text{(Eqn. 11-19)},$$

Or equivalently, in terms of the hazard ratio, $\Delta = \lambda_1/\lambda_2$,

$$H_0 : \Delta = 1 \quad \text{vs.} \quad H_1 : \Delta \neq 1 \qquad \text{(Eqn. 11-20)},$$

The ratio of the estimated hazard rates has an F distribution, so the required sample size can in principle be derived by solving an equation analogous to (11-3) for the F distribution. But these equations do not yield a closed form expression for the sample size. A much simpler and quite accurate approximation for a reasonably large number of events is based on the approximate normality of the natural logarithm of the estimated hazard rate in each treatment group:

$$\ln \hat{\lambda}_i \sim N\left(\ln \lambda_i, \frac{1}{d_i}\right),$$

where d_i is the number of observed events. Thus, the distribution of the log of the estimated hazard ratio can be approximated as:

$$\ln \hat{\Delta} = \ln\left(\frac{\hat{\lambda}_1}{\hat{\lambda}_2}\right) \sim N\left(\ln \Delta, \left(\frac{1}{d_1} + \frac{1}{d_2}\right)\right).$$

The required number of events on the ith treatment group, d_i, can be obtained from the following equation, directly analogous to equation (11-8):

$$\left(\frac{1}{d_1} + \frac{1}{d_2}\right)^{-1} = \left[\frac{\left(z_{\alpha/2} + z_\beta\right)^2}{\left(\ln \Delta_0\right)^2}\right]. \text{(Eqn. 11-21)},$$

where $\Delta_0 \neq 1$ is the specified hazard ratio for which we desire the power of the test to be $1 - \beta$. Table 11.5 gives values of $d_1 + d_2$, for some common design situations. Here we assume $d_1 \cong d_2$, yielding the minimum required total number of events. If the d_i are expected to differ substantially, an inflation factor similar to equation (11-10) should be applied.

The exponential assumption is not as restrictive as it might first appear since the calculations are approximately correct for the general proportional hazards case, in which the ratio of the hazard functions is independent of time, even though the individual hazards are not. The log-rank statistic is available as the score statistic from the maximum likelihood fitting of the proportional hazards model (33). Schoenfeld (34) showed that the method of George and Desu approximates the power of the log-rank test as long as the assumption of proportional hazards holds. Rubinstein et al. (19) show via simulations that trial lengths calculated using the statistic of George and Desu and assuming exponential failure times give valid powers for the log-rank test when the underlying survival distributions are exponential and Weibull. Under a proportional hazards model, the distribution of the log of the estimated hazard ratio, $\hat{\Delta}$, can be approximated by the same approximate normal distribution as in the exponential case. Thus, although the exponential distribution represents a simple special case of proportional hazards, the required number of events defined by (11-21) applies directly to the more general proportional hazards case. A more precise formulation is given in two papers by Lakatos (24, 26). If the proportional hazards assumption is not correct, the sample size formulae based on the assumption can produce erroneous results (27).

Required Duration of Study

The sample size approximation formula (11-21) and the tabulated values in Table 11.5 are for the required number of events at the time of the final analysis. In order to observe the requisite number of events, it is necessary to follow patients over time until the events are observed. At one extreme, we could enter exactly $2d$ patients on trial and follow all until failure; at the other extreme, we could enter patients continuously until $2d$ patients have failed. The former approach will require the maximum duration of study; the latter will yield the shortest duration of study but at the cost of entering the maximum number of patients. Either approach will yield the requisite power. However, some intermediate approach would be more appropriate in most cases.

TABLE 11.5
Total Number of Events Required to Compare Two Survival Distributions.

		Δ = Hazard Ratio								
α	$1-\beta$	0.90	0.85	0.80	0.75	0.70	0.65	0.60	0.55	0.50
0.01	0.80	4209	1769	939	565	368	252	180	131	98
	0.90	5362	2254	1196	720	468	321	229	167	124
0.05	0.80	2829	1189	631	380	247	170	121	98	66
	0.90	3787	1592	845	508	331	227	162	118	88

We assume that we will enter a sufficient number of patients (at least $2d$) on study during some accrual period, each randomized to one of the two treatment arms. After this accrual period, there will be an additional follow-up period (T, $T + \tau$) for all patients who have not failed before T in order to obtain the necessary $2d$ events. It is shown in George and Desu (15) that the expected number of events at time t, denoted $D(t)$, can be written as

$$E\left[D(t)\right] = \frac{\gamma t^*}{2}\left(p_1(t) + p_2(t)\right) \qquad \text{(Eqn. 11-22)},$$

where γ is the average accrual rate, $t^* = \min(T, t)$, and $p_i(t) = 1 - (\lambda_i t^*)^{-1} \exp(-\lambda_i t)\left[\exp(\lambda_i t^*) - 1\right]$. To find an appropriate accrual period T and follow-up time τ, we may require that the expected number of events at time $T + \tau$ be at least $2d$:

$$E\left[D(T+\tau)\right] \geq 2d \qquad \text{(Eqn. 11-23)},$$

As noted earlier, the minimum $T + \tau$ occurs when $\tau = 0$ (i.e., enter patients continuously until the end of the study with no follow-up), although this also results in the maximum number of patients entered on study. Table 11.6 gives the minimum length of study for the case $\tau = 0$ for some selected cases.

In this table, the median time in the control group is assumed to be one year. For other median times, the times in Table 11.6 must be adjusted accordingly by multiplying by the correct control median. If the control median is in fact t years, the required length of study is t times the values given in Table 11.6. In addition, it should be noted that the time at which the required number of deaths occurs is a random variable. The above formulation in terms of the expected number of deaths yields a result that can provide a rough approximation to the required length of study. Even if the assumptions are exactly correct, the *actual* time at which the required number of events will occur on any given clinical trial might be considerably different from the *expected* time.

If we enter only the minimum number of patients ($2d$), we will require the maximum length of study $T + \tau$. Indeed, the expected time until the last patient fails (this is required if we enter only the minimal number of patients) will be approximately $2d/\gamma + (1.44M_1 \ln(2d))/\Delta$, where M_1 is the median in the control group. For example, if we consider $\alpha = 0.05$, $1—\beta = 0.80$, and $\Delta = 0.70$ then from Table 11.5, $2d = 248$. If the entry rate (γ) is 200 per year, then from Table 11.6, the required minimum duration of study is 2.3 years when $M_1 = 1$. However, if we enter only 248 patients and follow all of them to failure, the expected length of study would be approximately 12.6 years ($T \cong 1.2$ years, $\tau \cong 11.4$ years).

Obviously, some kind of compromise approach is needed between the two extremes discussed above. Although we desire a reasonably short time until the study is completed, it is also desirable to keep the excess number of patients entered over the required number of events to be relatively small. One practical approach is to define a maximum proportionate increase, p, in patients entered over the required number of events, set $T = 2d(1 + p)/\gamma$, and solve (Eqn. 11-23) for τ.

The duration of study calculations in this section, in contrast to the required number of events calculations in the previous section, depend heavily on the exponential assumption. For example, if the hazard rates decrease sharply over time, additional follow-up will not yield sufficient numbers of events as quickly as in the constant hazard rate model. In designing actual clinical trials, it is important to make realistic assumptions about the anticipated hazard rates. The

TABLE 11.6

Required Minimum Duration of Study (In Years) to Compare Two Survival Distributions ($\alpha = 0.05$, $\beta = 0.80$).

		γ = Annual Accrual Rate				
Δ	$2d$	50	100	150	200	250
0.90	2830	58	30	20	16	13
0.80	632	14	7.6	5.5	4.4	3.8
0.70	248	6.2	3.7	2.8	2.3	2.0
0.60	122	3.6	2.2	1.7	1.4	1.3
0.50	68	2.3	1.5	1.2	1.0	0.9