Epidemiologic Reviews
Vol. 24, No. 1

Printed in U.S.A.

Sample Size Calculations for Randomized Controlled Trials

Janet Wittes

INTRODUCTION

Most informed consent documents for randomized controlled trials implicitly or explicitly promise the prospective participant that the trial has a reasonable chance of answering a medically important question. The medical literature, however, is replete with descriptions of trials that provided equivocal answers to the questions they addressed. Papers describing the results of such studies may clearly imply that the trial required a much larger sample size to adequately address the questions it posed. Hidden in file drawers, undoubtedly, are data from other trials whose results never saw the light of day—some, perhaps, victims of inadequate sample size. Although many inadequate-sized studies are performed in a single institution with patients who happen to be available, some are multicenter trials designed with overly optimistic assumptions about the effectiveness of therapy, too high an estimate of the event rate in the control group, or unrealistic assumptions about follow-up and compliance.

In this review, I discuss statistical considerations in the choice of sample size and statistical power for randomized controlled trials. Underlying the discussion is the view that investigators should hesitate before embarking on a trial that is unlikely to detect a biologically reasonable effect of therapy. Such studies waste both time and resources.

The number of participants in a randomized controlled trial can vary over several orders of magnitude. Rather than choose an arbitrary sample size, an investigator should allow both the variability of response to therapy and the assumed degree of effectiveness of therapy to drive the number of people to be studied in order to answer a scientific question. The more variable the response, the larger the sample size necessary to assess whether an observed effect of therapy represents a true effect of treatment or simply reflects random variation. On the other hand, the more effective or harmful the therapy, the smaller the trial required to detect that benefit or harm. As is often pointed out, only a few observations sufficed to demonstrate the dramatic benefit of penicillin; however, few therapies provide such unequivocal evidence of cure, so study of a typical medical intervention requires a large sample size. Lack of resources often constrains sample size. When they are limited by a restricted budget or a small patient pool, investigators should calculate the power of the trial to detect various outcomes of interest given the feasible sample size. A trial with very low statistical power may not be worth pursuing.

Typical first trials of a new drug include only a handful of people. Trials that study the response of a continuous variable to an effective therapy—for example, blood pressure change in response to administration of an antihypertensive agent—may include several tens of people. Controlled trials of diseases with high event rates—for example, trials of therapeutic agents for cancer—may study several hundred patients. Trials of prevention of complications of disease in slowly progressing diseases such as diabetes mellitus may enroll a few thousand people. Trials comparing agents of similar effectiveness—for instance, different thrombolytic interventions after a heart attack—may include tens of thousands of patients. The poliomyelitis vaccine trial included approximately a half-million participants (1).

This review begins with some general ideas about approaches to calculation of sample size for controlled trials. It then presents a generic formula for sample size that can be specialized to continuous, binary, and time-to-failure variables. The discussion assumes a randomized trial comparing two groups but indicates approaches to more than two groups. An example from a hypothetical controlled trial that tests the effect of a therapy on levels of high density lipoprotein (HDL) cholesterol is used to illustrate each case.

Having introduced a basic formula for sample size, the review discusses each element of the formula in relation to its applicability to controlled trials and then points to special complexities faced by many controlled trials—how the use of multiple primary endpoints, multiple treatment arms, and sequential monitoring affects the type I error rate and hence how these considerations should influence the choice of sample size; how staggered entry and lag time to effect of therapy affect statistical power in studies with binary or time-to-failure endpoints; how noncompliance with prescribed therapy attenuates the difference between treated groups and control groups; and how to adjust sample size during the course of the trial to maintain desired power. The review discusses the consequences to sample size calculation of projected rates of loss to follow-up and competing risks. It suggests strategies for determining reasonable values to assume for the different parameters in the formulas. Finally, the review addresses three special types of studies: equivalence trials, multiarm trials, and factorial designs.

Calculation of sample size is fraught with imprecision,

Received for publication November 1, 2001, and accepted for publication April 16, 2002.

Abbreviation: HDL, high density lipoprotein.

From Statistics Collaborative, Inc., 1710 Rhode Island Avenue NW, Suite 200, Washington, DC 20036 (e-mail: janet@statcollab.com). (Reprint requests to Dr. Janet Wittes at this address).

for investigators rarely have good estimates of the basic parameters necessary for the calculation. Unfortunately, the required size is often very sensitive to those unknown parameters. In planning a trial, the investigator should view the calculated sample size as an approximation to the necessary size. False precision in the choice of sample size adds no value to the design of a study.

The investigator faces the choice of sample size as one of the first practical problems in designing an actual controlled trial. Similarly, in assessing the results of a published controlled trial, the critical reader looks to the sample size to help him or her interpret the relevance of the results. Other things being equal, most people trust results from a large study more readily than those from a small one. Note that in trials with binary (yes/no) outcomes or trials that study time to some event, the word "small" refers not to the number of patients studied but rather to the number of events observed. A trial of 2,000 women on placebo and 2,000 on a new therapy who are being followed for 1 year to study the new drug's effect in preventing hospitalization for hip fracture among women aged ≥65 years is "small" in the parlance of controlled trials, because, as data from the National Center for Health Statistics suggest, only about 20 events are expected to occur in the control group. The approximately 99 percent of the sample who do not experience hip fracture provide essentially no information about the effect of the therapy.

The observation that large studies produce more widely applicable results than do small studies is neither particularly new nor startling. The participants in a small study may not be typical of the patients to whom the results are to apply. They may come from a single clinic or clinical practice, a narrow age range, or a specific socioeconomic stratum. Even if the participants represent a truly random sample from some population, the results derived from a small study are subject to the play of chance, which may have dealt a set of unusual results. Conclusions made from a large study are more likely to reflect the true effect of treatment. The operational question faced in designing controlled trials is determining whether the sample size is sufficiently large to allow an inference that is applicable in clinical practice.

The sample size in a controlled trial cannot be arbitrarily large. The total number of patients potentially available, the budget, and the amount of time available all limit the number of patients that can be included in a trial. The sample size of a trial must be large enough to allow a reasonable chance of answering the question posed but not so large that continuing randomization past the point of near-certainty will lead to ethical discomfort. A data monitoring board charged with ensuring the safety of participants might well request early stopping of a trial if a study were showing a very strong benefit of treatment. Similarly, a data monitoring board is unlikely to allow a study that is showing harm to participants to continue long enough to obtain a precise estimate of the extent of that harm. Some boards request early stopping when it is determined that the trial is unlikely to show a difference between treatments.

The literature contains some general reviews and discussions of sample size calculations, with particular reference to controlled trials (2–8).

GENERAL CONSIDERATIONS

Calculation of sample size requires precise specification of the primary hypothesis of the study and the method of analysis. In classical statistical terms, one selects a null hypothesis along with its associated type I error rate, an alternative hypothesis along with its associated statistical power, and the test statistic one intends to use to distinguish between the two hypotheses. Sample size calculation becomes an exercise in determining the number of participants required to achieve simultaneously the desired type I error rate and the desired power. For test statistics with well-known distributional properties, one may use a standard formula for sample size. Controlled trials often involve deviations from assumptions such that the test statistic has more complicated behavior than a simple formula allows. Loss to follow-up, incomplete compliance with therapy, heterogeneity of the patient population, or variability in concomitant treatment among centers of a multicenter trial may require modifications of standard formulas. Many papers in the statistical literature deal with the consequences to sample size of these common deviations. In some situations, however, the anticipated complexities of a given trial may render all available formulas inadequate. In such cases, the investigator can simulate the trial using an adequate number of randomly generated outcomes and select the sample size on the basis of those computer simulations.

Complicated studies often benefit from a three-step strategy in calculating sample size. First, one may use a simple formula to approximate the necessary size over a range of parameters of interest under a set of ideal assumptions (e.g., no loss to follow-up, full compliance, homogeneity of treatment effect). This calculation allows a rough projection of the resources necessary. Having established the feasibility of the trial and having further discussed the likely deviations from assumptions, one may then use more refined calculations. Finally, a trial that includes highly specialized features may benefit from simulation for selection of a more appropriate size.

Consider, for example, a trial comparing a new treatment with standard care in heart-failure patients. The trial uses two co-primary endpoints, total mortality and hospitalization for heart failure, with the type I error rate set at 0.04 for total mortality and 0.01 for hospitalization. In other words, the trial will declare the new treatment successful if it reduces either mortality ($p < 0.04$) or hospitalization ($p < 0.01$). This partitioning of the type I error rate preserves the overall error rate at less than 0.05. As a natural first step in calculating sample size, one would use a standard formula for time to failure and select as the candidate sample size the larger of the sizes required to achieve the desired power—for example, 80 percent—for each of the two endpoints. Suppose that sample size is 1,500 per group for hospitalization and 2,500 for mortality. Having established the feasibility of a study of this magnitude, one may then explore the effect of such complications as loss to follow-

up, intolerance to medication, or staggered entry. Suppose that these new calculations raise the sample size to 3,500. One may want to proceed further to account for the fact that the study has two primary endpoints. To achieve 80 percent power overall, one needs less than 80 percent power for each endpoint; the exact power required depends on the nature of the correlation between the two. In such a situation, one may construct a model and derive the sample size analytically, or, if the calculation is intractable, one may simulate the trial and select a sample size that yields at least 80 percent power over a range of reasonable assumptions regarding the relation between the two endpoints.

In brief, the steps for calculating sample size mirror the steps required for designing a trial.

1. Specify the null and alternative hypotheses, along with the type I error rate and the power.
2. Define the population under study.
3. Gather information relevant to the parameters of interest.
4. If the study is measuring time to failure, model the process of recruitment and choose the length of the follow-up period.
5. Consider ranges of such parameters as rates or events, loss to follow-up, competing risks, and noncompliance.
6. Calculate sample size over a range of reasonable parameters.
7. Select the sample size to use.
8. Plot power curves as the parameters range over reasonable values.

Some of these steps will be iterative. For example, one may alter the pattern of planned recruitment or extend the follow-up time to reduce the necessary sample size; one might change the entry criteria to increase event rates; or one might select clinical centers with a history of excellent retention to minimize loss to follow-up.

A BASIC FORMULA FOR SAMPLE SIZE

The statistical literature contains formulas for determining sample size in many specialized situations. In this section, I describe in detail a simple generic formula that provides a first approximation of sample size and that forms the basis of variations appropriate to specialized situations.

To understand these principles, consider a trial that aims to compare two treatments with respect to a parameter of interest. For simplicity, suppose that half of the participants will be randomized to treatment and the other half to a control group. The trial investigators may be aiming to compare mean values, proportions, odds ratios, hazard ratios, or some other statistic. Suppose that with proper mathematical transformation, the difference between the parameters in the treatment and control groups has an approximately normal distribution. These conditions allow construction of a generic formula for the required sample size. Typically, in comparing means or proportions, the difference between the sample statistics has an approximately normal distribution. In comparing odds ratios or hazard ratios, the logarithm of the differences has this property.

Consider three different trials using a new drug called "HDL-Plus" to raise HDL cholesterol levels in a study group of people without evidence of coronary heart disease whose baseline level of HDL cholesterol is below 40 mg/dl. The Veterans Affairs High-Density Lipoprotein Cholesterol Intervention Trial showed that gemfibrozil raised HDL cholesterol levels and decreased the risk of coronary events in patients with prior evidence of cardiovascular disease and low HDL cholesterol levels (9). The first hypothetical study, to be called the HDL Cholesterol Raising Trial, tests whether HDL-Plus in fact raises HDL cholesterol levels. The trial, which randomizes patients to receipt of HDL-Plus or placebo, measures HDL cholesterol levels at the end of the third month of therapy. The outcome is the continuous variable "concentration of HDL cholesterol in plasma."

The second study, to be called the Low HDL Cholesterol Prevention Trial, compares the proportions of people in the treated and control groups with HDL cholesterol levels above 45 mg/dl at the end of 1 year of treatment with HDL-Plus or placebo.

The third study, called the Myocardial Infarction Prevention Trial, follows patients for at least 5 years and compares times to fatal or nonfatal myocardial infarction in the two groups. This type of outcome is a time-to-failure variable.

The formulas for determining sample size use several statistical concepts. Throughout this paper, Greek letters denote a true or hypothesized value, while italic Roman letters denote observations.

The *null hypothesis* H_0 is the hypothesis positing the equivalence of the two interventions. The logical purpose of the trial is to disprove this null hypothesis. The HDL Cholesterol Raising Trial tests the null hypothesis that 3 months after beginning therapy with HDL-Plus, the average HDL cholesterol level in the treated group is the same as the average level in the placebo group. The Low HDL Cholesterol Prevention Trial tests the null hypothesis that the proportion of people with an HDL cholesterol level above 45 mg/dl at the end of 1 year is the same for the HDL-Plus and placebo groups. The Myocardial Infarction Prevention Trial tests the null hypothesis that the expected time to heart attack is the same in the HDL-Plus and placebo groups.

If the two treatments have identical effects (that is, if the null hypothesis is true), the group assigned to receipt of treatment is expected to respond in the same way as persons assigned to the control group. In any particular trial, however, random variation will cause the two groups to show different average responses. The *type I error rate*, α, is defined as the probability that the trial will declare two equally effective treatments "significantly" different from each other. Conventionally, controlled trials set α at 0.05, or 1 in 20. While many people express comfort with a level of $\alpha = 0.05$ as "proof" of the effectiveness of therapy, bear in mind that many common events occur with smaller probabilities. One experiences events that occur with a probability of 1 in 20 approximately twice as often as one rolls a 12 on a pair of dice (1 in 36). If you were given a pair of dice, tossed them, and rolled a pair of sixes, you would be mildly

surprised, but you would not think that the dice were loaded. A few more pairs of sixes on successive rolls of the dice would convince you that something nonrandom was happening. Similarly, a controlled trial with a p value of 0.05 should not convince you that the tested therapy truly works, but it does provide positive evidence of efficacy. Several independent replications of the results, on the other hand, should be quite convincing.

The hypothesis that the two treatment groups differ by some specified amount Δ_A is called the *alternative hypothesis*, H_A.

The *test statistic*, a number computed from the data, is the formal basis for the comparison of treatment groups. In comparing the mean values of two continuous variables when the observations are independently and identically distributed and the variance is known, the usual test statistic is the standardized difference between the means,

$$z = \frac{\bar{x} - \bar{y}}{\sigma\sqrt{2/n}}, \tag{1}$$

where $\bar{x}$ and $\bar{y}$ are the observed means of the treated group and the control group, respectively, σ is the true standard deviation of the outcome in the population, and n is the number of observations in each group. This test statistic has a standard normal distribution with mean 0 and variance 1.

In a one-tailed test, the alternative hypothesis has a direction (i.e., treatment is better than control status). The observations lead to the conclusion either that the data show no evidence of difference between the treatments or that treatment is better. In this formulation, a study that shows a higher response rate in the control group than in the treatment group provides evidence favoring the null hypothesis. Most randomized controlled trials are designed for two-tailed tests; if one-tailed testing is being used, the type I error rate is set at 0.025.

The *critical value* $\xi_{1-\alpha/2}$ is the value from a standard normal distribution that the test statistic must exceed in order to show a statistically significant result. The subscript means that the statistic must exceed the $1 - \alpha/2$'nd percentile of the distribution. In one-tailed tests, the critical value is $\xi_{1-\alpha}$.

The *difference between treatments* represents the measures of efficacy. Statistical testing refers to three types of differences. The true mean difference Δ is unknown. The mean difference under the alternative hypothesis is Δ_A. The importance of Δ_A lies in its centrality to the calculation of sample size. The observed difference at the end of the study is $\bar{d}$. Suppose that, on average, patients assigned to the control group have a true response of magnitude ω; then the hypothesized treated group has the response $\omega + \Delta_A$. For situations in which the important statistic is the ratio rather than the difference in the response, one may consider instead the logarithm of the ratio, which is the difference of the logarithms.

The *type II error rate*, or β, is the probability of failing to reject the null hypothesis when the difference between responses in the two groups is Δ_A. Typical well-designed randomized controlled trials set β at 0.10 or 0.20.

Related to β is the *statistical power* $\gamma(\Delta)$, the probability of declaring the two treatments different when the true difference is exactly Δ. A well-designed controlled trial has high power (usually at least 80 percent) to detect an important effect of treatment. At the hypothesized difference between treatments, the power $\gamma(\Delta_A)$ is $1 - \beta$. Setting power at 50 percent produces a sample size that yields a barely significant difference at the hypothesized Δ_A. One can look at the alternative that corresponds to 50 percent power as the point at which one would say, "I would kick myself if I didn't declare this difference statistically significant."

Under the above conditions, a generic formula for the total number of persons needed *in each group* to achieve the stated type I and type II error rates is

$$n = 2\sigma^2\{[\xi_{1-\alpha/2} + \xi_{1-\beta}]/\Delta_A\}^2. \tag{2}$$

The formula assumes one treatment group and one control group of equal size and two-tailed hypothesis testing. If the power is 50 percent, the formula reduces to $n = 2(\sigma\xi_{1-\alpha/2}/\Delta_A)^2$, because $\xi_{0.50} = 0$. Some people, in using sample size formulae, mistakenly interpret the "2" as meaning "two groups" and hence incorrectly use half the sample size necessary.

The derivation of formula 2, and hence the variations in it necessary when the assumptions fail, depends on two relations, one related to α and one to β.

Under the null hypothesis, the choice of type I error rate requires the probability that the absolute value of the statistic z is greater than the critical value $\xi_{1-\alpha/2}$ to be no greater than α; that is,

$$\Pr\{|z| > \xi_{1-\alpha/2}|H_0\} < \alpha. \tag{3}$$

The notation "$|H_0$" means "under the null hypothesis."

Similarly, the choice of the type II error rate restricts the distribution of z under the alternative hypothesis:

$$\Pr\{|z| > \xi_{1-\alpha/2}|H_A\} > 1 - \beta. \tag{4}$$

Under the alternative hypothesis, the expected value of $\bar{x} - \bar{y}$ is Δ_A, so formula 4 implies

$$\Pr\left\{\frac{\sqrt{n}|\bar{x} - \bar{y}|}{\sqrt{2}\sigma} > \xi_{1-\alpha/2}\middle|H_A\right\} > 1 - \beta,$$

or

$$\Pr\{|\bar{x} - \bar{y}| - \Delta_A > \sqrt{2/n}\,\sigma\,\xi_{1-\alpha/2} - \Delta_A|H_A\} > 1 - \beta.$$

Dividing both sides by $\sigma\sqrt{2/n}$,

$$\Pr\left\{\frac{\sqrt{n}(|\bar{x} - \bar{y}| - \Delta_A)}{\sqrt{2}\sigma} > \xi_{1-\alpha/2} - \frac{\sqrt{n}\Delta_A}{\sqrt{2}\sigma}\middle|H_A\right\} > 1 - \beta,$$

yields a normally distributed statistic. The definition of β and the symmetry of the normal distribution imply

$$\xi_{1-\alpha/2} - \sqrt{n}\Delta_A/(\sqrt{2}\sigma) = \xi_\beta = -\xi_{1-\beta}. \tag{5}$$

Rearranging terms and squaring both sides of the equations produces formula 2.

In some controlled trials, more participants are randomized to the treated group than to the control group. This imbalance may encourage people to participate in a trial because their chance of being randomized to the treated group is greater than one half. If the sample size n_t in the treated group is to be k times the size n_c in the control group, the sample size for the study will be

$$n_c = (1 + 1/k)\sigma^2 \frac{[\xi_{1-\alpha/2} + \xi_{1-\beta}]^2}{\Delta_A{}^2}; \; n_t = kn_c. \quad (2A)$$

Thus, the relative sample size required to maintain the power and type I error rate of a trial with two equal groups is $(2 + k + 1/k)/4$. For example, a trial that randomizes two treated participants to every control requires a sample size larger by a factor of 4.5/4 or 12.5 percent in order to maintain the same power as a trial with 1:1 randomization. A 3:1 randomization requires an increase in sample size of 33 percent. Studies investigating a new therapy in very short supply—a new device, for example—may actually randomize more participants to the control group than to the treated group. In that case, one selects n_t to be the number of devices available, sets the allocation ratio of treated to control as 1:k, and then solves for the value of k that gives adequate power. The power is limited by n_t because even arbitrarily large k's cannot make $(1 + 1/k)$ less than 1.

The derivation of the formula for sample size required a number of assumptions: the normality of the test statistic under both the null hypothesis and the alternative hypothesis, a known variance, equal variances in the two groups, equal sample sizes in the groups, and independence of the individual observations. One can modify formula 2 to produce a generic sample size formula that allows relaxation of these assumptions. Let $\eta^0_{1-\alpha/2}$ and $\eta^A_{1-\beta}$ represent the relevant percentiles of the distribution of the not-necessarily-normally-distributed test statistic, and let $\sigma_0{}^2$ and $\sigma_A{}^2$ denote the variance under the null and alternative hypotheses, respectively. Then one may generalize formula 2 to produce

$$n = \frac{[\eta^0_{1-\alpha/2}\sqrt{2}\sigma_0 + \eta^A_{1-\beta}\sigma_A]^2}{\Delta_A^2}. \quad (6)$$

Formula 6 assumes groups of equal size. To apply to the case where the allocation ratio of treated to control is k:1 rather than 1:1, the sample sizes in the control and treated groups will be $(1 + 1/k)$ and $(k + 1)$ times the sample size in formula 6, respectively.

The next three sections, which present sample sizes for normally distributed outcome variables, binomial outcomes, and time-to-failure studies, show modifications of formulas 5 and 6 needed to deal with specific situations.

CONTINUOUS VARIABLES: TESTING THE DIFFERENCE BETWEEN MEAN RESPONSES

To calculate the sample size needed to test the difference between two mean values, one makes several assumptions.

1. The responses of participants are independent of each other. The formula does not apply to studies that randomize in groups—for example, those that assign treatment by classroom, village, or clinic—or to studies that match patients or parts of the body and randomize pairwise. For randomization in groups (i.e., cluster randomization), see Donner and Klar (10). Analysis of studies with pairwise randomization focuses on the difference between the results in the two members of the pair.
2. The variance of the response is the same in both the treated group and the control group.
3. The sample size is large enough that the observed difference in means is approximately normally distributed. In practice, for reasonably symmetric distributions, a sample size of about 30 in each treatment arm is sufficient to apply normal theory. The Central Limit Theorem legitimizes the use of the standard normal distribution. For a discussion of its appropriateness in a specific application, consult any standard textbook on statistics.
4. In practice, the variance will not be known. Therefore, the test statistic under the null hypothesis replaces σ with s, the sample standard deviation. The resulting statistic has a t distribution with $2n - 2$ df. Under the alternative hypothesis, the statistic has a noncentral t distribution with noncentrality parameter $\sqrt{2n}\,\Delta_A$ and, again, $2n - 2$ df. Standard software packages for sample size calculations employ the t and noncentral t distributions (11–13). Except for small sample sizes, the difference between the normal distribution and the t distribution is quite small, so the normal approximation yields adequately close sample sizes in most situations.

BINARY VARIABLES: TESTING DIFFERENCE BETWEEN TWO PROPORTIONS

Calculation of the sample size needed to test the difference between two binary variables requires several assumptions.

1. The responses of participants are independent.
2. The probability of an event is π_c and π_t for each person in the control group and treated group, respectively. Because the sample sizes in the two groups are equal, the average event rate is $\bar{\pi} = (\pi_c + \pi_t)/2$. This assumption of constancy of proportions is unlikely to be strictly valid in practice, especially in large studies. If the proportions vary considerably in recognized ways, one may refine the sample size calculations to reflect that heterogeneity. Often, however, one hypothesizes average values for π_c and π_t and calculates sample size as if those proportions applied to each individual in the study.

Under these assumptions, the binary outcome variable has a binomial distribution, and the following simple formula provides the sample size for each of the two groups:

$$n = 2\bar{\pi}(1 - \bar{\pi})\frac{(\xi_{1-\alpha/2} + \xi_{1-\beta})^2}{(\pi_c - \pi_t)^2}. \quad (7A)$$

This simple formula, a direct application of formula 5, uses the same variance under both the null hypothesis and the

alternative hypothesis. Because the variances differ, a more accurate formula, derived from formula 6, is

$$n = \frac{[\xi_{1-\alpha/2}\sqrt{2\bar{\pi}(1-\bar{\pi})} + \xi_{1-\beta}\sqrt{\pi_c(1-\pi_c)+\pi_t(1-\pi_t)}]^2}{(\pi_c-\pi_t)^2}. \quad (7B)$$

If one will employ a correction for continuity in the final analysis, or if one will be using Fisher's exact test, one should replace n with (14)

$$n' = \frac{n}{4}\left(1+\sqrt{1+\frac{4}{n|\pi_c-\pi_t|}}\right)^2. \quad (7C)$$

All three of the above formulas use the normal distribution, which is the limiting distribution of the binomial. They become inaccurate as $n\pi_c$ and $n\pi_t$ become very small (e.g., less than 5).

My personal preference among these three formulae is formula 7C, because I believe that one should use corrected chi-squared tests or Fisher's exact test; however, not all statisticians agree with that view.

TIME-TO-FAILURE DATA WITH TREATMENTS THAT WILL BE COMPARED USING THE LOG-RANK TEST

Consider a trial that compares time to some specified event—for example, death in chronic lung disease, recurrence of tumor in a cancer study, or loss of 30 percent of baseline isometric strength in a study of degenerative nerve disease. Let π_c and π_t be the probability that a person in the control group and a person in the treated group, respectively, experiences an event during the trial. Define $\theta = \ln(1-\pi_c)/\ln(1-\pi_t)$, which is the hazard ratio, also called the relative risk.

Assume that the event rate is such that within each of the two groups every participant in a given treatment group has approximately the same probability of experiencing an event. Assume that no participant withdraws from the study.

In a study in which half of the participants will receive experimental treatment and half will be controls, Freedman (15) presents the following simple formulas.

Total number of events in both treatment groups:

$$\left(\frac{\theta+1}{\theta-1}\right)^2(\xi_{1-\alpha/2}+\xi_{1-\beta})^2. \quad (8A)$$

Sample size in each treatment group:

$$\frac{1}{(\pi_c+\pi_t)}\left(\frac{\theta+1}{\theta-1}\right)^2(\xi_{1-\alpha/2}+\xi_{1-\beta})^2. \quad (8B)$$

An even simpler formula (formula 9) is due to Bernstein and Lagakos (16), who derived it under the assumption that the time to failure has an exponential distribution, and to Schoenfeld (17), who derived it for the log-rank test without assuming an exponential model. Under their models, the total number of events required in the two treatment groups is

$$4\frac{(\xi_{1-\alpha/2}+\xi_{1-\beta})^2}{[\ln(\theta)]^2}. \quad (9A)$$

Then the total sample size required in each treatment group is

$$\frac{4}{(\pi_c+\pi_t)}\frac{(\xi_{1-\alpha/2}+\xi_{1-\beta})^2}{[\ln(\theta)]^2}. \quad (9B)$$

If the ratio of allocation to treatment and control is m:1 rather than 1:1, the "4" in formula 9A becomes $(m+1)^2/m$.

Neither formula 8 nor formula 9 explicitly incorporates time. In fact, time appears only in the calculation of the probabilities π_c and π_t of events. Below I describe how important and complicated time can be in the calculation of sample size for controlled trials that measure time to an event.

EXAMPLE

To apply the formulas given in the above sections to the three HDL cholesterol trials, one could make the following assumptions.

1. The standard deviation σ of HDL cholesterol in the population is approximately 11 mg/dl.
2. People with HDL cholesterol levels between 35 mg/dl and 40 mg/dl can expect HDL-Plus to lead to a 7-mg/dl rise in HDL cholesterol.
3. Approximately 10 percent of people with HDL cholesterol levels below 40 mg/dl have an HDL cholesterol level above 45 mg/dl 3 months later. With use of HDL-Plus, that percentage is hypothesized to increase to approximately 20 percent. Of course, these percentages will depend on the distribution of the participants' HDL cholesterol levels at entry into the study. If nearly all of the participants have an HDL cholesterol level below 35 mg/dl at baseline, the proportion of participants on placebo whose values rise to over 45 mg/dl will be very small.
4. An expected 20 percent of the people in the study will suffer a heart attack over the course of the 5 years of follow-up. Those taking HDL-Plus can expect their risk to decrease to approximately 15 percent. Averaged over the 5 years of the study, these rates translate into about 4.4 percent and 3.2 percent annually for the untreated and treated groups, respectively. (This "average" is calculated as the geometric mean—that is, under the assumption of exponential rates. For example, to calculate the annual rate for the control group, one computes $1 - \sqrt[5]{1-0.15}$.)

Before proceeding with calculation of sample size, note the vagueness of the above numbers. Words such as "approximately" or "about" modify each number. Clearly, the event rate for a specific population depends on many factors—for example, the age-sex distribution in the population recruited, other risk factors for the disease, the distri-

bution of HDL cholesterol values at baseline, and error in the measurement of HDL cholesterol. To speak of a 20 percent 5-year risk, as assumption 4 does, greatly oversimplifies reality. Calculation of an annual rate by a geometric mean makes a very strong assumption about the pattern of the event rate over time. Nonetheless, these kinds of crude data and rough approximations necessarily form the basis for many sample size calculations.

With $\alpha = 0.05$ and a power of 80 percent, the percentiles for the normal distribution are $\xi_{1-\alpha/2} = 1.96$ and $\xi_{1-\beta} = 0.84$. Plugging these numbers into the formulas yields the following sample sizes.

The HDL Cholesterol Raising Trial. Applying the formula $2\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2/\Delta^2$ yields $2 \times 11(1.96 + 0.84)^2/7^2 = 38.7$. Thus, a trial with 40 people assigned to HDL-Plus and 40 assigned to placebo will have approximately 80 percent power to show an HDL cholesterol-raising effect of 7 mg/dl. If indeed at the end of the study the observed standard deviation were 11 and the observed difference were 7 mg/dl, then the t statistic with 78 df $(80 - 2)$ would be 2.85 and the associated p value would be 0.0057. When the power was at least 80 percent, if one actually observed the hypothesized difference, the p value would be considerably less than the type I error rate. In fact, the barely significant difference in this case is 4.9.

The Low HDL Cholesterol Prevention Trial. In the Low HDL Cholesterol Prevention Trial, 10 percent of the placebo group and 20 percent of the HDL-Plus group can be expected to have HDL cholesterol levels above 45 mg/dl at the end of the 3-month study. Use of formula 5B to calculate the sample size required to observe such a difference yields a sample size of 199 in each group, for a total sample size of 398, which rounds off to 400. Use of the simpler but slightly less accurate formula 5A yields 200 people in each group, an immaterial difference. Application of formula 5C, which employs the correction for continuity, yields a sample size of 218 people per group or 436 in all. The change in endpoint from the continuous-variable level of HDL cholesterol to a dichotomous variable has led, in this case, to an approximate fivefold increase in total sample size.

The Myocardial Infarction Prevention Trial. Assume a 20 percent rate in the control group and a 15 percent rate in the treated group—that is, $\pi_c = 0.20$, $\pi_t = 0.15$, and $\theta = \ln(1 - 0.20)/\ln(1 - 0.15) = 1.3730$. Formula 8B yields a total sample size of 1,816, or 908 persons per group, to achieve the desired α level and power:

$$n = (1.96 + 0.84)^2 \frac{[(1 + 1.3730)/(1 - 1.3730)]^2}{(0.20 + 0.15)} = 908.$$

This sample size implies that 180 heart attacks would be expected in the control group and 135 in the HDL-Plus group. Formula 9B gives a sample size of 1,780, which provides nearly the same answer. Use of the binomial distribution without correction for continuity, which is a very rough approach to calculating sample size for a study that compares death rates, yields results that are nearly the same. If the proportions of people in the two groups who will experience an event are 15 percent and 20 percent, substituting the data into formula 5B yields 906 persons per group rather than 908 as calculated by formula 8B, a formula for the log-rank test. While the log-rank formula is more intellectually satisfying to use, for a wide range of scenarios it yields values very close to those of the binomial distribution.

The three above studies, all investigating the effects of HDL-Plus, ask very different questions and consequently require strikingly different resources. Under the assumptions of this section, asking whether HDL-Plus "works" in the sense of affecting levels of HDL cholesterol requires a study of approximately 80 participants followed for 3 months. Asking whether administration of HDL-Plus "works" by materially affecting the proportion of people in a high-risk stratum requires approximately 400 people followed for 1 year. However, asking the direct clinical question of whether HDL-Plus "works" in reducing the 5-year risk of heart attack by 20 percent requires 1,800 people followed for 5 years.

COMPONENTS OF SAMPLE SIZE: α AND β

Typical controlled trials set the statistical significance level at 0.05 or 0.01 and the power at 80 or 90 percent. Table 1 shows the sample sizes required for various levels of α and β relative to the sample size needed for a study with a two-sided α equal to 0.05 and 80 percent power.

Some relative sample sizes are large indeed. For example, moving from $\alpha = 0.05$ and 80 percent power to $\alpha = 0.01$ and 90 percent power almost doubles the required sample size. More modestly, raising power from 80 percent to 90 percent increases the required sample size by approximately 30 percent.

Certain features of the design of a study will affect its type I error rate. A trial that uses more than one test of significance may need to adjust the α level to preserve the true probability of observing a significant result. Multiple endpoints, multiple treatment arms, or interim analyses of the data require α-level adjustment. The basic problem that leads multiplicity to require larger sample sizes is simply stated: If the treatments under study are truly equivalent, a statistical test will reject the null hypothesis 100α percent of the time, but if a trial specifies more than a single statistical test as part of its primary outcome, the probability of rejecting at least one test is greater than α. Think of dice. The

TABLE 1. Necessary sample size as a function of power and α level, relative to the sample size required for a study with an α level of 0.05 and 80 percent power*

α	Power			
	70%	80%	90%	95%
0.05	0.8	1.0	1.3	1.7
0.01	1.2	1.5	1.9	2.3
0.001	1.8	2.2	2.7	3.1

* To read the table, choose a power and an α level. Suppose one is interested in a trial with 90 percent power and an α level of 0.01. The entry of 1.9 in the table means that such a trial would require 1.9 times the sample size required for a trial with 80 percent power and an α level of 0.05.

probability of throwing two sixes on a single throw of a pair of dice is 1/36, but the probability of *not* throwing a pair of sixes in 200 tosses of the dice is $(1 - 1/36)^{100} = 0.004$. That is, the probability of having at least one six in 200 tosses is 0.996. The more questions one asks of data, the more likely it is that the data will show statistical significance at least once, or, as some anonymous (at least to me) wag has exhorted us, "Torture the data until they confess."

If the analysis of the data is to correct for multiple testing, the sample size should account for that correction. For example, if there are r primary questions and the final analysis will use a Bonferroni correction to adjust for multiplicity, the critical value will divide the α level by r, so the factor $(\xi_{1-\alpha/2} + \xi_{1-\beta})^2$ multiplying sample size becomes $(\xi_{1-\alpha/(2r)} + \xi_{1-\beta})^2$. Table 2 shows the factor as a function of power and the number of tests performed. Methods for adjustment more sophisticated than the Bonferroni correction are available (18); the sample size calculation should account for the particular method that is planned.

A trial that includes interim monitoring of the primary endpoint with the potential for early stopping to declare efficacy should account for the final critical value when calculating sample size. This consideration usually leads to slight increases in sample size. Table 3 shows the sample size multiplier as a function of α level for the final significance test under several commonly used methods for interim analysis.

COMPONENTS OF SAMPLE SIZE: THE VARIANCE

The sample size necessary to achieve the desired α level and power is directly proportional to the variance of the outcome measure in the population under study. For normally distributed outcomes, the variance σ^2 multiplies all of the other factors and the sample variance is independent of the sample means. Therefore, calculating the required sample size requires a reasonably precise projection of the variance of the population to be studied. Several factors conspire to render the variance very difficult to project in studies of continuous outcome measures. The sample variance is a highly variable statistic, so estimating it precisely requires a large sample size. In practice, however, one often projects the variance by culling estimates of variability from small studies reported in the literature and from available case series; the entire set of published data may be too small to allow precise projection of the variance. Moreover, published studies probably underestimate variances, on average, because underestimates of variance lead to higher probabilities of finding statistically significant results and hence a higher chance of a paper's being published. Another problem stems from secular changes, some due to changes in the therapeutic milieu and some due to changes in the epidemiology and clinical course of disease. Data in the literature necessarily come from the past; estimates needed for a trial come from the as-yet-unknown future. Insofar as the past only imperfectly predicts the future, projected and actual variances may differ. Even if the milieu remains constant, the specific eligibility requirements in a study may profoundly affect variability.

For binomial outcomes and tests of time to failure, the mean and the variance are related. The usual problem in calculating sample size in those cases stems not from an imprecise prior estimate of the variance but from an inability to predict the control rates precisely. The equation for binomial variance contains the term $\pi(1 - \pi)$. Incorrectly projecting the event rate π will produce an inaccurate value for $\pi(1 - \pi)$, which leads to a sample size that is accordingly too big or too small. This part of the problem of an incorrect value of π is usually minor in practice, because $\pi(1 - \pi)$ is fairly stable over a wide range of π. The major effect of an incorrect value of π is misstating the value of $\pi_1 - \pi_2$, which, as is shown below, can lead to dramatic changes in sample size.

In planning a randomized controlled trial, an exhaustive search of the literature on the particular measure should precede the guessing of the variance that will obtain in the trial itself. One useful method is to set up a simple database that summarizes variables from published and (if available) unpublished studies. The database should record demographic characteristics of the patients, the entry and exclusion criteria used in the study, the type of institution from which the data came, and the approach to measurement. Helpful data include the number of patients excluded from the analysis and the reasons for such exclusion, because often these patients have more variable responses than those included. Comparison of this database with the composition of the projected study sample in the trial being planned allows calculation of an expected variance on the basis of the data in the studies at hand inflated by a factor that

TABLE 2. Sample size as a function of the number of significance tests, relative to the sample size required for an α level of 0.05 and a power of 90 percent (Bonferroni inequality)*

No. of significance tests	$\alpha = 0.05$			$\alpha = 0.01$		
	Power = 70%	Power = 80%	Power = 90%	Power = 70%	Power = 80%	Power = 90%
1	0.59	0.75	1.00	0.91	1.11	1.42
2	0.73	0.90	1.18	1.06	1.27	1.59
3	0.81	1.00	1.29	1.14	1.36	1.69
4	0.87	1.06	1.36	1.20	1.42	1.76
10	1.06	1.27	1.59	1.39	1.62	1.99

* To read the table, choose a power, an α level, and the number of statistical tests you intend to perform. Suppose one is interested in a trial with 90 percent power, an α level of 0.01, and four tests of significance. The entry of 1.76 in the table means that such a trial would require 1.76 times the sample size required for a trial with 90 percent power, an α level of 0.05, and one statistical test.

TABLE 3. Sample size relative to a study with no interim analysis*

Type of interim analysis (reference no.)	Critical *p* value at the final analysis	Multiplier	
		Power = 80%	Power = 90%
No interim analysis	0.05	1.00	1.00
Haybittle rule (42) or O'Brien-Fleming rule with one interim look (43)	0.049	1.01	1.01
O'Brien-Fleming rule with two interim looks	0.046	1.02	1.03
O'Brien-Fleming rule with three interim looks	0.044	1.04	1.03
Pocock rule (44)	0.030	1.16	1.13

* To read the table, choose a type of interim analysis plan and a desired power. Suppose one is interested in a trial with 90 percent power and an O'Brien-Fleming rule with two interim looks. The critical *p* value at the end of the study will be 0.046. The entry of 1.03 in the table means that such a trial would require 1.03 times the sample size needed for a trial of 90 percent power, an α level of 0.05, and no interim analysis.

represents expected extra heterogeneity. One may use the abstracted data from the available studies to develop simple mathematical models of the likely composition of the study population and the variance.

Even a careful, exhaustive search of available data may lead to incorrect projections. The section below on sample size recalculation addresses approaches one may adopt if, during the trial, one becomes aware that prior projections were seriously incorrect.

COMPONENTS OF SAMPLE SIZE: THE DIFFERENCE TO BE DETECTED

An important determinant of sample size is the difference between the parameter of interest under the null hypothesis and the parameter under the alternative hypothesis. Because sample size is inversely related to the square of that difference, even slightly misspecifying the difference can lead to a large change in the sample size. For example, a hypothesized relative risk of 80 percent and a probability in the control group of 0.2 leads to $\pi_t = 0.2(0.8) = 0.16$ and $(\pi_c - \pi_t)^2 = 0.04^2 = 0.00016$. If, however, the true rate in the placebo arm were 0.16 instead of 0.20, then $\pi_t = 0.16(0.8) = 0.128$ and $(\pi_c - \pi_t)^2 = (0.160 - 0.128)^2 = 0.0011$. The ratio of the two factors is 0.0016/0.0011 = 1.45. This slight misspecification of the placebo rate leads to a 45 percent underestimate of the necessary sample size. It is disconcerting that such large effects on sample size result from such small differences in rates, for in designing studies one rarely has available sufficient information to distinguish between rates as close to each other as 0.16 and 0.20.

Designers of controlled trials facing the problem of how to specify that difference commonly use one of two approaches. Some people select the treatment effect deemed important to detect. For example, in cancer clinical trials, a new chemotherapeutic regimen may be of interest only if it increases the probability of remission by more than 20 percent over the standard regimen. In this formulation, the investigators specify the "difference to be detected" on the basis of clinical importance without explicit regard to the likely effect of the particular intervention.

The other frequently used method is to calculate the sample size according to the best guess concerning the true effect of treatment. In a hypertension prevention trial with stroke as an endpoint, one projects the expected reduction in diastolic blood pressure, searches the epidemiologic literature to find an estimate of the number of strokes likely to be prevented if the mean diastolic blood pressure decreased by that amount, and calculates the sample size on the basis of that reduction.

WHEN THE ASSUMPTIONS FAIL

The sample size formulas presented thus far result from derivations that make simplifying assumptions about the nature of the trial and the behavior of the participants. The approaches assume that all participants in the trial are fully compliant with therapy, that all of them are followed until the end of the study, and that each participant's outcome is assessed. For trials studying time to event, follow-up times are assumed to be equal for all persons or the probability of experiencing an event after the end of follow-up is assumed to be very small. Hazard ratios are assumed to be constant in time. Some formulas assume exponentially distributed failure times. Often, these simplifying assumptions reflect reality closely enough that the methods produce reasonably accurate sample sizes. Many times, however, the complexities in the trial lead to violations of the assumptions important enough that the sample size calculations become unacceptably inaccurate. Generally, the violations occur in the direction that requires increased sample sizes to achieve the desired power.

Two general types of methods are available for calculating sample size in the presence of complicating factors. One approach posits an underlying failure-time model but allows deviations from the basic assumptions (19–21). In the following sections, in dealing with methods that specify a formal failure-time model, I use the method of Lachin and Foulkes (19), because it produces a simple yet flexible closed-form expression for sample size. The software package PASS (13) adopts this method in calculating sample size for time-to-event data.

Other approaches, developed by Halpern and Brown (22, 23), Lakatos (24, 25), and Shih (26), allow the designer considerable latitude in modeling what will happen in a trial. The methods do not require an underlying parametric survival model. Instead, they ask the designer of a trial to project in considerable detail the course of the trial under both the null hypothesis and the alternative hypothesis. All of these authors have made their programs publicly avail-

able. In the remainder of this review, I use the method described by Lakatos in 1988 (25) in dealing with this second approach. Lakatos and Lan (27) present a useful summary of approaches to sample size calculation for the log-rank test.

The sample size per group under the Lachin-Foulkes model is

$$n = \left[\xi_{1-\alpha/2} \sqrt{\phi(\bar{\lambda},\bar{\eta},\bar{\gamma})\left(\frac{1}{Q_C}+\frac{1}{Q_T}\right)} + \xi_{1-\beta} \sqrt{\frac{\phi(\lambda_C,\eta_C,\gamma_C)}{Q_C}+\frac{\phi(\lambda_T,\eta_T,\gamma_T)}{Q_T}} \right]^2 \div |\lambda_C - \lambda_T|^2,$$

where failure times are assumed to follow an exponential distribution and Q_C and Q_T are the proportions in the control group and the treated group, respectively; λ_C and λ_T are the hazard rates for the two groups; η_C and η_T are the exponential loss-to-follow-up rates in the two groups; and γ_C and γ_T are exponential parameters describing the pattern of recruitment.

This model assumes that individuals enter during an accrual period of R time periods. They are followed for an additional period of time until a total of T periods is reached. Hence, the first person entered is followed for T periods; the last person entered is followed for $T - R$ periods.

Halpern and Brown (22, 23) allow the user to specify arbitrary survival curves. Their program simulates the outcomes of a trial governed by these curves. Lakatos (24, 25) envisions a set of states and periods. In each period of a trial, a participant is in one of several states (e.g., is alive and is receiving the assigned therapy, is alive and is receiving the opposite therapy, is deceased, or has dropped out). The person then undergoes a series of transitions governed by a Markov process. Some of the states are absorbing; that is, once a person is in certain states (e.g., death), one cannot change. Others are fluid; a participant may move into and out of these states over time. The designer of the trial may choose the number of periods, the pattern of recruitment, and the transition probabilities. For example, suppose one thinks of the trial as occurring in monthly periods, and suppose the endpoint of interest is cardiovascular death or nonfatal myocardial infarction. At month i, a person in the treatment group either may be alive and still under study treatment without having experienced an event, may be alive and not under study treatment without having experienced an event, or may have already experienced a study event. Between months i and $(i + 1)$, a person who has not yet experienced an event and is still on study medication (state A) may experience an event, may experience a noncardiovascular disease event, may become lost to follow-up, or may stop taking medication. Any of the events incurred by the person in state A may also be incurred by the person who has not yet experienced a cardiovascular disease event but is not on study medication (state B), except that this person may restart active therapy. A person who was lost to follow-up (state C), who experienced a noncardiovascular death (state D), or who experienced the event of interest (state E) will remain in that state for each subsequent period. If one assigns probabilities to each of these transitions, one produces a Markov model that captures the complex of experiences that may occur. A typical matrix may look like the one below, where, for example, p_{BCi} represents the probability that a person who is not on study medication and has not experienced an outcome event will be lost to follow-up within the period $(i, i + 1)$.

$$\begin{pmatrix} p_{AAi} & p_{ABi} & p_{ACi} & p_{ADi} & p_{AEi} \\ p_{BAi} & p_{BBi} & p_{BCi} & p_{BDi} & p_{BEi} \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Shih's method (26) extends Lakatos' approach by allowing many additional states. Because of the large number of parameters at the disposal of the designer, many people become overwhelmed by this approach. If one uses the method carefully, noting which parameters have important implications for sample size in a particular setting, the method can provide valuable insights into the effect of various scenarios on sample size and power. Especially in the presence of nonproportional hazards, these methods allow considerable flexibility.

In applying either the Halpern and Brown approach or the Lakatos-type method, the user must consider carefully the meaning of the various parameters and their relation to data in the literature. The Lakatos method assumes an underlying ideal model and then perturbs that model under various failures of assumptions. Specifically, it begins with rates in the treatment and control arms that would obtain if there were perfect compliance, no drop-out or drop-in, and no competing risks. It then applies the rates of these perturbations to recalculate expected event rates. The Halpern approach, by contrast, starts with the perturbed model so that the event rates already account for the effect of drop-outs and other deviations from the ideal. Parameters derived from the literature often do not fit clearly into either approach. A large epidemiologic database has already built in some of the perturbations, because it reflects what actually happens in practice. For example, cause-specific mortality from an epidemiologic database necessarily incorporates competing risk. On the other hand, parameters from a small, tightly controlled trial may fit more appropriately into the Lakatos-type approach. The user should take the parameters from the literature and convert them as appropriate into the parameters necessary for the method to be used.

STAGGERED ENTRY

Patients enter most controlled trials not simultaneously but rather in a "staggered" fashion. Recruitment may take months or even years. In trials that study time to failure, each person may have a fixed follow-up time or the study may have a common closeout date. In the latter case, the time of follow-up varies by participant, with the first enrollee having the longest potential follow-up.

When the endpoint is binary or continuous, the time of entry into the study is immaterial to sample size. However, for studies testing time to failure, sample size is related to

the total number of person-years of follow-up and hence to the pattern of recruitment. For studies with low event rates, small shifts in patterns of recruitment can have important effects on the total amount of follow-up. One limitation of the methods of Freedman (15) and Schoenfeld (28) and related approaches is their failure to account for staggered entry. If people within the trial have very different follow-up times, these methods can produce quite inaccurate sample sizes.

By way of illustration, consider once more the Myocardial Infarction Prevention Trial. The 5-year event rates in the treated and control groups were 15 percent and 20 percent, respectively. The corresponding exponential parameters are $\lambda_C = -\ln(1 - 0.85)/5 = 0.0446$ and, analogously, $\lambda_T = 0.0325$. Different assumptions about recruitment and follow-up lead to different sample sizes. Consider, for example, three sets of assumptions all with sample size calculated by the Lachin-Foulkes method. A trial that followed all participants for exactly 5 years would require a sample size of 907 persons per group; if the recruitment period extended for 3 years and the study lasted a total of 5 years, the sample size would be 918 per group. If, however, the recruitment period lasted 3 years but the last person was followed for 6 years, only 600 people would be required in each group. Because equations for sample size are highly nonlinear, the designer of a trial should calculate, rather than guess, the consequence to sample size of a number of feasible scenarios.

NONCOMPLIANCE

For both theoretical and practical reasons, one of the most vexing problems in controlled trials is noncompliance. People who participate in trials, like patients in the ordinary practice of medicine, do not always adhere to their assigned therapeutic regimen. They may forget to take their medication; they may overdose; they may take their medication sporadically; and they may stop taking their medication, either because they suffer side effects or because they feel better. If the intervention is a nonpharmacologic treatment, such as diet or exercise, they may find adherence onerous. Rigorous statistical analysis of the data must include all people, even those who do not comply with therapy, in the group to which they were randomized.

Noncompliance in controlled trials becomes serious if its nature and extent compromise the expected difference between treated and control groups. A person assigned to placebo medication who fails to take the assigned placebo is violating the protocol; however, this type of noncompliance does not adversely affect the power of the study, for such a person, like the complier, is acting as a nontreatment control. Similarly, a person on an assigned treatment regimen who stops the assigned treatment but adopts a regimen with similar effects does not adversely affect the power of the study. The real problem comes from persons who effectively cross over to the other treatment arm. In the example of the HDL cholesterol-raising study, a person assigned to placebo who starts an HDL cholesterol-raising treatment such as niacin or gemfibrozil adopts roughly the same event rate as the rate in the treated arms, thereby reducing the overall difference between treated and control subjects. Similarly, a person assigned to HDL-Plus who stops taking the medication assumes approximately the same event rate as the rate in the placebo arm, again reducing the overall difference between treatment arms. By contrast, a person on placebo who stops taking placebo maintains roughly the same heart attack rate as the control group, while a person in the HDL-Plus group who stops taking study medication but begins to take open-label HDL-Plus or another HDL cholesterol-raising drug maintains about the same heart attack rate as the treated group.

In calculating sample size, many investigators ignore noncompliance and perform computations as if everyone will adhere to their assigned therapies. Ignoring the problem, however, invites more serious difficulties later, for if a trial has considerable noncompliance, the sample size will be insufficient to ensure the desired power.

Some researchers increase the sample size by a factor that represents the number of people who are expected not to comply. A typical approach is to inflate the sample size by the factor $1/(1 - c)$, where c is the proportion predicted not to comply. Such an approach leads to an insufficient correction. The investigator who discards the noncompliers from the analysis violates the principle that analysis should reflect randomization ("intent to treat"). Thus, if the inflation method is an admission of the intent to analyze only the compliers, it is suspect because its associated analytical method is invalid. If, however, the investigator intends to perform the as-randomized analysis, inflation by the proportion of noncompliers leads to an insufficient correction, for noncompliance reduces the effect size, which in turn increases the necessary sample size by a factor proportional to the square of the change in effect size. To see why the effect is so large, suppose that the mean response in the treated and treated control groups is μ_t and μ_c, respectively. Suppose further that a proportion ρ_t of persons in the treated group stop active therapy (the so-called "drop-outs") and a proportion ρ_c of the controls stop taking their control medication and begin using a therapy as effective as the active treatment (the "drop-ins"). Then the expected response in the treated group will be $(1 - \rho_t)\mu_t + \rho_t\mu_c$ and the expected response in the control group will be $(1 - \rho_c)\mu_c + \rho_c\mu_t$, so the expected difference between the two groups will be $(1 - \rho_c - \rho_t)(\mu_t - \mu_c)$.

Noncompliance attenuates the expected difference by the factor $(1 - \rho_c - \rho_t)$ or inflates the sample size by the square of that factor. Usually, when the drop-in rate can be considered negligible, which it will be for a trial studying a condition for which no therapy is available, the required inflation of sample size is $(1 - \rho_t)^2$. As can be seen in table 4, which shows the necessary inflation of sample size as a function of drop-in and drop-out rates, noncompliance can wreak havoc in trials, either by requiring a very large increase in sample size or by substantially reducing statistical power. The commonly used solution, simply ignoring participants who do not comply with therapy, has the potential to produce very biased results. One should instead build expected noncompliance into the overall model that describes the

TABLE 4. Sample size required relative to a trial with full compliance as a function of the proportion "dropping in" to active therapy in the control group and the proportion "dropping out" of active therapy in the treated group*

Percentage of treated participants "dropping out" of treatment	Percentage of controls "dropping in" to active therapy			
	0	5	10	15
0	1	1.11	1.23	1.38
10	1.23	1.38	1.56	1.78
20	1.56	1.78	2.04	2.37
30	2.04	2.37	2.78	3.31

* To read the table, specify the percentages of people you expect to "drop in" and "drop out" of active therapy. Suppose one expects 10 percent of the active group to drop out of active therapy and 5 percent of the control group to drop in to active therapy. Then the sample size necessary to achieve the prespecified α level and power would be 1.38 times the size needed if all participants complied with their assigned treatment.

projected experience of the cohort to be studied.

The following example shows the consequence to power of different strategies for sample size calculation in the face of noncompliance. Suppose that in our HDL cholesterol study, HDL-Plus is expected to decrease the 5-year event rate from 12 percent to 8 percent. If recruitment is uniform over the first 2 years of the study and if the study allocates half of the patients to treatment and half to the control group, a sample size of approximately 3,300 people overall is sufficient for 90 percent power. Suppose, however, that 20 percent of the people assigned to receipt of HDL-Plus are expected to stop taking their medications in the first year of the study and 5 percent of the placebo group is expected to take HDL-Plus or an equally effective HDL cholesterol-raising drug. Then, under an as-randomized analysis, the power will decrease to 70 percent. Had the sample size been calculated under these assumptions concerning crossover rates, the required sample size for 90 percent power would have been 5,400. Note that if the analysis simply ignored those who crossed over, 1,578 persons would remain in the control group and 1,320 persons would remain in the treated group. The power for this sample size would be 0.85; as stated above, however, the high power is deceptive, because it is associated with a method of analysis that does not respect the randomization.

LAG TIME

Certain interventions take time to achieve their full effect. Although cholesterol-lowering therapy, either diet or drugs, reduces the risk of heart attack, the intervention does not become fully effective for approximately 2 years after the initiation of therapy. Studies designed to investigate the effect of a preventive strategy when the outcome is a time-to-event variable should account for the lag time before the effect of therapy becomes manifest. In applying the epidemiology-to-controlled-trial paradigm to compute expected event rates, a common assumption is that the intervention will lead to an instantaneous reduction in the event rate. A more realistic approach may be to posit a certain time to full effect and then model some simple smooth function to describe the trajectory from high risk to low risk. The judgment about the length of time to effect should be based on underlying biology. In designing time-to-event studies for situations where the therapy does not achieve its effect for a while, sample size calculation should account for the lag. Follow-up time should be sufficiently long to be able to detect the treatment's beneficial effect. Trivially, a drug that does not become effective for 2 years cannot be shown to work in a 1-year study. The method of Shih (26) allows for incorporation of lag time.

LOSS TO FOLLOW-UP

Analysis of data from randomized controlled trials should strictly reflect the randomization. Ideally, the trial should have complete follow-up and complete assessment of endpoints so it will yield unbiased estimates of treatment effects. Inevitably, however, some people are lost to follow-up, such that no assessment of endpoint is possible. The protocol should include methods for handling those lost to follow-up, and the sample size calculations should be performed in accordance with those rules. Sometimes a person who has stopped returning for follow-up visits will be willing to make a final visit or provide a telephone interview. Every reasonable effort should be made to gather information germane to the final outcome. Simply excluding from the analysis persons who are lost to follow-up leads to potential bias in the inference about treatment effect. Nonetheless, both the Lachin-Foulkes and Lakatos methods permit censoring of persons who drop out of the study. The methods assume that the reason for drop-out is unrelated to treatment.

COMPETING RISKS

In time-to-event studies, some people die or are removed from the study because they experience an event that precludes measurement of the primary event. This type of event is a special form of dropping out, one that is outside of the investigators' control. Usually, analyses censor—that is, remove from further study—the person at the time of death or the competing event. This policy is reasonable as long as the competing event is independent of the outcome under investigation. For example, in a trial assessing the rate of development of cataracts, death removes the patient from further evaluation of the eye. Because death is unrelated to progression of cataracts, except insofar as loss of eyesight might be related to proneness to accidents, such censoring should not lead to bias. In the presence of competing risks, sample size calculations should adjust for the loss in person-years of follow-up attributable to the censoring.

When the censoring and the event under study are more closely linked, the censoring mechanism may differentially affect treated and control groups in such a way as to leave under study groups with unequal risks of the event being investigated. In such a situation, simply adjusting the sample size does not solve the problem.

NONPROPORTIONAL HAZARDS

Perhaps the most important practical difference between the methods spawned by Halpern and Lakatos and the other methods in the literature is that both Halpern and Lakatos allow nonproportional hazards. By permitting the user to specify the event rates for the two groups for specific periods of time, these approaches give the designer of a trial important flexibility. For diseases for which the treatment is expected to cure the patient, the assumption of proportional hazards is not reasonable: Once the patient is no longer sick, the hazard ratio should become unity. Similarly, for treatments that tend to lengthen the lives of sick patients, the risk of mortality in the treated group may become greater than the risk in the control group because the surviving patients may be sicker (29).

FACTORIAL DESIGNS

Factorial designs study more than one therapy simultaneously. The simple 2×2 factorial design has two interventions, each at two levels, which results in four treatment groups. The Post Coronary Artery Bypass Graft Trial (30), a study designed to investigate prevention of early graft closure after bypass surgery, included two lipid-lowering strategies and anticoagulant therapy (warfarin) (see table 5). The four treatment groups were: 1) moderate lipid-lowering (goal: low density lipoprotein cholesterol levels of 130–140 mg/dl) and placebo; 2) aggressive lipid-lowering (goal: low density lipoprotein cholesterol levels of 60–85 mg/dl) and placebo; 3) moderate lipid-lowering and low-dose warfarin (1–4 mg); and 4) aggressive lipid-lowering and low-dose warfarin (1–4 mg). To test the effect of warfarin, the two groups receiving warfarin—groups 3 and 4—are compared with the two groups not receiving warfarin—groups 1 and 2. If the two treatments (here, lipid-lowering strategies and warfarin) did not affect each other, the sample size could be calculated as the minimum size necessary to answer the lipid-lowering and warfarin questions.

Often factorial designs are touted as providing "two for the price of one," and, for the case of continuous variables with constant variances, factorial designs do in fact allow just that. For trials with binomial and time-to-failure endpoints, sample size calculation should account for expected interactions between the treatments and decreases in event rates (31).

EQUIVALENCE TRIALS

Equivalence and noninferiority trials deserve special mention in connection with sample size, because the considerations for inference and hence for sample size calculations differ markedly from those of conventional trials. The purpose of an equivalence trial (which I prefer to call a "not-very-different-from" trial) is to prove, or at least to indicate strongly, that two treatments have the same clinical benefit. For example, one might want to show that the effect of a new antibiotic does not differ from that of a marketed one by more than a specific amount Δ. Logically, this structure turns classical statistical analysis on its head, for one cannot "prove" the null hypothesis. The inadequacy of the classical application in this case is clear. Suppose we define two therapies as equivalent if their effects do not differ significantly. Then the smaller the sample size, the more likely we are to find the result we want. The nonsignificant result will, of course, be accompanied by a wide confidence interval.

The literature proposes two approaches, one that fixes the width of the confidence interval for the difference between the two treatments (32) and one that selects a sample size that controls not only the width but also the probability that the lower bound of that confidence interval will lie above a specified value (33, 34). The former method, which will produce smaller sample sizes than the latter, is analogous to a superiority trial with 50 percent power.

TABLE 5. The factorial design of the Post Coronary Artery Bypass Graft Trial

Warfarin (1–4 mg)	Lipid-lowering	
	Moderate*	Aggressive†
No	Group 1	Group 2
Yes	Group 3	Group 4

* Goal: low density lipoprotein cholesterol levels of 130–140 mg/dl.

† Goal: low density lipoprotein cholesterol levels of 60–85 mg/dl.

MORE THAN TWO GROUPS

If the study will examine more than two groups, the sample size calculations should reflect the primary questions under study. For example, if the study will include four groups and the null hypothesis is that the four groups are the same while the alternative hypothesis is that the four groups are not the same, the primary question is tested with a 3-df statistical test. However, usually in controlled trials the primary question or questions in multigroup studies is a set of 1-df contrasts. For example, if the study groups consist of a placebo and three different doses, the usual question is the comparison of each dose with placebo or perhaps a hypothesis concerning a dose-response relation among the groups.

OTHER SITUATIONS

This discussion has focused on basic scenarios the designers of controlled trials typically face. Other variations may occur. If the data in the final analysis will be stratified by some variables, the basic methods used will change. For continuous variables, *t* tests will become *F* tests in blocked designs; sample size for stratified designs is discussed in any standard textbook on analysis of variance. Binomial analyses will become Mantel-Haenszel tests, and time-to-event analyses will use stratified log-rank tests. For calculations of sample sizes in these cases, see Wittes and Wallenstein (35, 36).

If the data in the analysis will be adjusted for covariates, the sample size calculations can incorporate those features. For continuous variables, standard texts describe methods;

for logistic regression, the analog of binomial endpoints, the method of Hsieh (37) is applicable.

SAMPLE SIZE RECALCULATION

In the course of many trials, the investigators may become aware that the basis of the sample size calculations was incorrect. Specifically, as noted above, the variance or the event rate may have been underestimated. The consequence of this misspecification leads to a sample size that is too small. Several authors have proposed formal methods for recalculating sample size on the basis of interim data. The published approaches share some common features: They emphasize the importance of preserving the type I error rate of the trial. Some are based on recalculating variance or the event rate from the unblinded data (38) and some from the blinded data (39). Several reviews have addressed various features of these approaches (40, 41).

COMMENTS

Sample size calculation for continuous and binary variables in controlled trials does not differ from sample size calculation in other fields. Time-to-event analysis, on the other hand, poses problems peculiar to controlled trials.

For time-to-event analysis, my own approach is to calculate sample sizes from at least two very different approaches. Obtaining similar answers affords comfort that the calculations are correct. Obtaining very dissimilar answers serves as a warning that something complicated is occurring. The various methods require different parameterizations, such that the user must carefully think through how to translate values from the literature to the formulas. These translations will differ from method to method. The standard packages for sample size calculations for time-to-event analysis use different methods, and, importantly, their manuals and the actual programs may not be internally consistent. Often the programs are updated but the manual remains unchanged. I would like to have recommended one of these packages over all others, but I am reluctant to do so because the packages are changing constantly and because each one has different strengths and weaknesses.

Finally, a plea: Do not rush sample size calculation. Clinical investigators have often approached me with a request for a "quick" method of sample size calculation. The grant is due or the protocol is just about ready to be sent for internal approval; all it lacks is the sample size. When I was younger, I tried to comply with such requests, but I now refuse. Sample size is too integral a part of the design itself to be patched in at the end of the process.

REFERENCES

1. Francis TJ, Korns RF, Voight RB, et al. An evaluation of the 1954 poliomyelitis vaccine trials. Summary report. Am J Public Health 1955;45:1–51.
2. Day S, Graham D. Sample size estimation for comparing two or more treatment groups in clinical trials. Stat Med 1991;10:33–43.
3. Donner A. Approaches to sample size estimation in the design of clinical trials—a review. Stat Med 1984;3:199–214.
4. Gore SM. Statistics in question. Assessing clinical trials—trial size. Br Med J 1981;282:1687–9.
5. Johnson AF. Sample size: clues, hints or suggestions. J Chronic Dis 1985;38:721–5.
6. Lachin JM. Introduction to sample size determination and power analysis for clinical trials. Control Clin Trials 1981;2: 93–113.
7. Moussa MA. Exact, conditional, and predictive power in planning clinical trials. Control Clin Trials 1989;10:378–85.
8. Whitehead J. Sample sizes for phase II and phase III clinical trials: an integrated approach. Stat Med 1986;5:459–64.
9. Rubins H, Robins S, Collins D, et al. Gemfibrozil for the secondary prevention of coronary heart disease in men with low levels of high-density lipoprotein cholesterol. Veterans Affairs High-Density Lipoprotein Cholesterol Intervention Trial Study Group. N Engl J Med 1999;341:410–18.
10. Donner A, Klar NS. Design and analysis of cluster randomization trials in health research. London, United Kingdom: Arnold Publishers, 2000.
11. Borenstein M, Rothstein H, Cohen J, et al. Power and precision, version 2: a computer program for statistical power analysis and confidence intervals. Englewood, NJ: BioStat, Inc, 2001:287.
12. Elashoff J. nQuery Advisor version 4.0 user's guide. Los Angeles, CA: Statistical Solutions, 2000.
13. Hintze J. NCSS Trial and PASS 2000. Kaysville, UT: NCSS, 2001.
14. Fleiss J, Tytun A, Ury H. A simple approximation for calculating sample sizes for comparing independent proportions. Biometrics 1980;36:343–6.
15. Freedman LS. Tables of the number of patients required in clinical trials using the logrank test. Stat Med 1982;1:121–9.
16. Bernstein D, Lagakos SW. Sample size and power determination for stratified clinical trials. J Stat Comp Sim 1978;8: 65–73.
17. Schoenfeld D. The asymptotic properties of nonparametric tests for comparing survival distributions. Biometrika 1981; 68:316–19.
18. Hochberg Y, Tamhane A. Multiple comparison procedures. New York, NY: John Wiley and Sons, Inc, 1987.
19. Lachin J, Foulkes M. Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. Biometrics 1986;42:507–19.
20. Wu MC. Sample size for comparison of changes in the presence of right censoring caused by death, withdrawal, and staggered entry. Control Clin Trials 1988;9:32–46.
21. Wu M, Fisher M, DeMets D. Sample sizes for long-term medical trial with time-dependent dropout and event rates. Control Clin Trials 1980;1:111–23.
22. Halpern J, Brown BJ. A computer program for designing clinical trials with arbitrary survival curves and group sequential testing. Control Clin Trials 1993;14:109–22.
23. Halpern J, Brown BJ. Designing clinical trials with arbitrary specification of survival functions and for the log rank or generalized Wilcoxon test. Control Clin Trials 1987;8:177–89.
24. Lakatos E. Sample size determination in clinical trials with time-dependent rates of losses and noncompliance. Control Clin Trials 1986;7:189–99.
25. Lakatos E. Sample size based on the log-rank statistic in complex clinical trials. Biometrics 1988;44:229–41.
26. Shih J. Sample size calculation for complex clinical trials with survival endpoints. Control Clin Trials 1995;16:395–407.
27. Lakatos E, Lan K. A comparison of some methods of sample size calculation for the logrank statistic. Stat Med 1992;11: 179–91.
28. Schoenfeld D. Sample-size formula for the proportional-hazards regression model. Biometrics 1983;39:499–503.
29. Lan K, Wittes J. Data monitoring in complex clinical trials: which treatment is "better"? J Stat Planning Inference 1994; 42:241–55.
30. The Post Coronary Artery Bypass Graft Trial Investigators.

The effect of aggressive lowering of low-density lipoprotein cholesterol levels and low-dose anticoagulation on obstructive changes in saphenous-vein coronary-artery bypass grafts. N Engl J Med 1997;336:153–62.
31. Brittain E, Wittes J. Factorial designs in clinical trials: the effects of noncompliance and subadditivity. Stat Med 1989; 8:161–71.
32. Makuch R, Simon R. Sample size requirements for evaluating a conservative therapy. Cancer Treat Rep 1978;62:1037–40.
33. Blackwelder W. "Proving the null hypothesis" in clinical trials. Control Clin Trials 1982;3:345–53.
34. Blackwelder W, Chang M. Sample size graphs for "proving the null hypothesis." Control Clin Trials 1984;5:97–105.
35. Wittes J, Wallenstein S. The power of the Mantel-Haenszel test. J Am Stat Assoc 1987;82:1104–9.
36. Wallenstein S, Wittes J. The power of the Mantel-Haenszel test for grouped failure time data. Biometrics 1993;49:1077–87.
37. Hsieh F. Sample size tables for logistic regression. Stat Med 1989;8:795–802.
38. Wittes J, Brittain E. The role of internal pilot studies in increasing the efficiency of clinical trials. Stat Med 1990;9: 65–72.
39. Gould A, Shih W. Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. Commun Stat 1991;21:2833–53.
40. Betensky R, Tiernery C. An examination of methods for sample size recalculation during an experiment. Stat Med 1997;16:2587–9.
41. Zucker DM, Wittes JT, Schabenberger O, et al. Internal pilot studies II: comparison of various procedures. Stat Med 1999; 18:3493–509.
42. Haybittle JL. Repeated assessment of results in clinical trials of cancer treatment. Br J Radiol 1971;44:793–7.
43. O'Brien P, Fleming T. A multiple testing procedure for clinical trials. Biometrics 1979;35:549–56.
44. Pocock S. Group sequential methods in the design and analysis of clinical trials. Biometrika 1977;64:191–9.