Chapter 8
Sample Size

The size of the study should be considered early in the planning phase. In some instances, no formal sample size is ever calculated. Instead, the number of participants available to the investigators during some period of time determines the size of the study. Many clinical trials that do not carefully consider the sample size requirements turn out to lack the statistical power or ability to detect intervention effects of a magnitude that has clinical importance. In 1978, Freiman and colleagues [1] reviewed the power of 71 published randomized controlled clinical trials which failed to find significant differences between groups. "Sixty-seven of the trials had a greater than 10% risk of missing a true 25% therapeutic improvement, and with the same risk, 50 of the trials could have missed a 50% improvement." The situation was not much improved in 1994, when a similar survey found only 16% of negative trials had 80% power for a 25% effect, and only 36% for a 50% effect [2]. In other instances, the sample size estimation may assume an unrealistically large intervention effect. Thus, the power for more realistic intervention effects will be low or less than desired. The danger in studies with low statistical power is that interventions that could be beneficial are discarded without adequate testing and may never be considered again. Certainly, many studies do contain appropriate sample size estimates, but in spite of many years of critical review many are still too small [3, 4].

This chapter presents an overview of sample size estimation with some details. Several general discussions of sample size can be found elsewhere [5–21]. For example, Lachin [11] and Donner [9] have each written a more technical discussion of this topic. For most of the chapter, the focus is on sample size where the study is randomizing individuals. In the some sections, the concept of sample size for randomizing clusters of individuals or organs within individuals is presented.

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-3-319-18539-2_8

Fundamental Point

Clinical trials should have sufficient statistical power to detect differences between groups considered to be of clinical importance. Therefore, calculation of sample size with provision for adequate levels of significance and power is an essential part of planning.

Before a discussion of sample size and power calculations, it must be emphasized that, for several reasons, a sample size calculation provides only an estimate of the needed size of a trial [6]. First, parameters used in the calculation are estimates, and as such, have an element of uncertainty. Often these estimates are based on small studies. Second, the estimate of the relative effectiveness of the intervention over the control and other estimates may be based on a population different from that intended to be studied. Third, the effectiveness is often overestimated since published pilot studies may be highly selected and researchers are often too optimistic. Fourth, during the final planning stage of a trial, revisions of inclusion and exclusion criteria may influence the types of participants entering the trial and thus alter earlier assumptions used in the sample size calculation. Assessing the impact of such changes in criteria and the screening effect is usually quite difficult. Fifth, trial experience indicates that participants enrolled into control groups usually do better than the population from which the participants were drawn. The reasons are not entirely clear. One factor could be that participants with the highest risk of developing the outcome of interest are excluded in the screening process. In trials involving chronic diseases, because of the research protocol, participants might receive more care and attention than they would normally be given, or change their behavior because they are part of a study, thus improving their prognosis, a phenomenon sometimes called the Hawthorne or trial effect [22]. Also, secular trends toward improved care may result in risk estimates from past studies being higher than what will be found in current patient populations [23]. Participants assigned to the control group may, therefore, be better off than if they had not been in the trial at all. Finally, sample size calculations are based on mathematical models that may only approximate the true, but unknown, distribution of the response variables.

Due to the approximate nature of sample size calculations, the investigator should be as conservative as can be justified while still being realistic in estimating the parameters used in the calculation. If a sample size is drastically overestimated, the trial may be judged as unfeasible. If the sample size is underestimated, there is a good chance the trial will fall short of demonstrating any differences between study groups or be faced with the need to justify an increase in sample size or an extension of follow-up [24–26]. In general, as long as the calculated sample size is realistically obtainable, it is better to overestimate the size and possibly terminate the trial early (Chap. 16) than to modify the design of an ongoing trial, or worse, to arrive at incorrect conclusions.

Statistical Concepts

An understanding of the basic statistical concepts of hypothesis testing, significance level, and power is essential for a discussion of sample size estimation. A brief review of these concepts follows. Further discussion can be found in many basic medical statistics textbooks [27–37] as well as textbooks on sample size [17–21]. Those with no prior exposure to these basic statistical concepts might find these resources helpful.

Except where indicated, trials of one intervention group and one control group will be discussed. With some adjustments, sample size calculations can be made for studies with more than two groups [8]. For example, in the Coronary Drug Project (CDP), five active intervention arms were each compared against one control arm [38]. The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack trial (ALLHAT) compared four active intervention arms: three newer drugs to an older one as first line therapy for hypertension [39]. Both trials used the method of Dunnett [40], where the number of participants in the control group is equal to the number assigned to each of the active intervention groups times the square root of the number of active groups. The optimal size of the control arm in the CDP was determined to be 2.24 times the size of each individual active intervention arm [38]. In fact, the CDP used a factor of 2.5 in order to minimize variance. Other approaches are to use the Bonferroni adjustment to the alpha level [41]; that is, divide the overall alpha level by the number of comparisons, and use that revised alpha level in the sample size comparison.

Before computing sample size, the primary response variable used to judge the effectiveness of intervention must be identified (see Chap. 3). This chapter will consider sample size estimation for three basic kinds of outcomes: (1) dichotomous response variables, such as success and failure (2), continuous response variables, such as blood pressure level or a change in blood pressure, and (3) time to failure (or occurrence of a clinical event).

For the dichotomous response variables, the event rates in the intervention group (p_I) and the control group (p_C) are compared. For continuous response variables, the true, but unknown, mean level in the intervention group (μ_I) is compared with the mean level in the control group (μ_C). For survival data, a hazard rate, λ, is often compared for the two study groups or at least is used for sample size estimation. Sample size estimates for response variables which do not exactly fall into any of the three categories can usually be approximated by one of them.

In terms of the primary response variable, p_I will be compared with p_C or μ_I will be compared with μ_C. This discussion will use only the event rates, p_I, and p_C, although the same concepts will hold if response levels μ_I and μ_C are substituted appropriately. Of course, the investigator does not know the true values of the event rates. The clinical trial will give him only estimates of the event rates, $\widehat{p_I}$ and $\widehat{p_C}$. Typically, an investigator tests whether or not a true difference exists between the event rates of participants in the two groups. The traditional way of indicating this is in terms of a null hypothesis, denoted H_0, which states that no difference between

the true event rates exists (H_0: $p_C - p_I = 0$). The goal is to test H_0 and decide whether or not to reject it. That is, the null hypothesis is assumed to be true until proven otherwise.

Because only estimates of the true event rates are obtained, it is possible that, even if the null hypothesis is true ($p_C - p_I = 0$), the observed event rates might by chance be different. If the observed differences in event rates are large enough by chance alone, the investigator might reject the null hypothesis incorrectly. This false positive finding, or *Type I error*, should be made as few times as possible. The probability of this Type I error is called the significance level and is denoted by α. The probability of observing differences as large as, or larger than the difference actually observed given that H_0 is true is called the "p-value," denoted as p. The decision will be to reject H_0 if $p \leq \alpha$. While the chosen level of α is somewhat arbitrary, the ones used and accepted traditionally are 0.01, 0.025 or, most commonly, 0.05. As will be shown later, as α is set smaller, the required sample size estimate increases.

If the null hypothesis is not in fact true, then another hypothesis, called the alternative hypothesis, denoted by H_A, must be true. That is, the true difference between the event rates p_C and p_I is some value δ where $\delta \neq 0$. The observed difference $\widehat{p_C} - \widehat{p_I}$ can be quite small by chance alone even if the alternative hypothesis is true. Therefore, the investigator could, on the basis of small observed differences, fail to reject H_0 even when it is not true. This is called a *Type II error*, or a false negative result. The probability of a Type II error is denoted by β. The value of β depends on the specific value of δ, the true but unknown difference in event rates between the two groups, as well as on the sample size and α. The probability of correctly rejecting H_0 is denoted $1 - \beta$ and is called the power of the study. Power quantifies the potential of the study to find true differences of various values δ. Since β is a function of α, the sample size and δ, $1 - \beta$ is also a function of these parameters. The plot of $1 - \beta$ versus δ for a given sample size is called the power curve and is depicted in Fig. 8.1. On the horizontal axis, values of δ are plotted from 0 to an upper value, δ_A (0.25 in this figure). On the vertical axis, the probability or power of detecting a true difference δ is shown for a given significance level and sample size. In constructing this specific power curve, a sample size of 100 in each group, a one-sided significance level of 0.05 and a control group event rate of 0.5 (50%) were assumed. Note that as δ increases, the power to detect δ also increases. For example, if $\delta = 0.10$ the power is approximately 0.40. When $\delta = 0.20$ the power increases to about 0.90. Typically, investigators like to have a power ($1 - \beta$) of at least 0.80, but often around 0.90 or 0.95 when planning a study; that is to have an 80%, 90% or 95% chance of finding a statistically significant difference between the event rates, given that a difference, δ, actually exists.

Since the significance level α should be small, say 0.05 or 0.01, and the power ($1 - \beta$) should be large, say 0.90 or 0.95, the only quantities which are left to vary are δ, the size of the difference being tested for, and the total sample size. In planning a clinical trial, the investigator hopes to detect a difference of specified magnitude δ or larger. One factor that enters into the selection of δ is the minimum difference

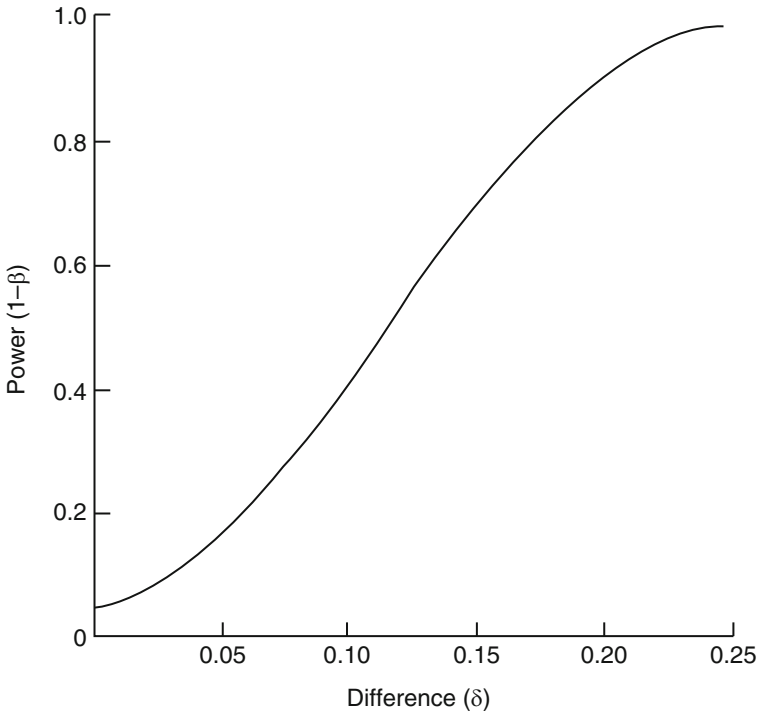

Fig. 8.1 A power curve for increasing differences (δ) between the control group rate of 0.5 and the intervention group rate with a one-sided significance level of 0.05 and a total sample size ($2N$) of 200

between groups judged to be clinically important. In addition, previous research may provide estimates of δ. This is part of the question being tested as discussed in Chap. 3. The exact nature of the calculation of the sample size, given α, $1 - \beta$ and δ is considered here. It can be assumed that the randomization strategy will allocate an equal number (N) of participants to each group, since the variability in the responses for the two groups is approximately the same; equal allocation provides a slightly more powerful design than unequal allocation. For unequal allocation to yield an appreciable increase in power, the variability needs to be substantially different in the groups [42]. Since equal allocation is usually easier to implement, it is the more frequently used strategy and will be assumed here for simplicity.

Before a sample size can be calculated, classical statistical theory says that the investigator must decide whether he is interested in differences in one direction only (one-sided test)—say improvements in intervention over control—or in differences in either direction (two-sided test). This latter case would represent testing the hypothesis that the new intervention is either better or worse than the control. In general, two-sided tests should be used unless there is a very strong justification for expecting a difference in only one direction. An investigator should always keep in mind that any new intervention could be harmful as well as helpful. However, as

discussed in Chap. 16, some investigators may not be willing to prove the intervention harmful and would terminate a study if the results are suggestive of harm. A classic example of this issue was provided by the Cardiac Arrhythmia Suppression Trial or CAST [43]. This trial was initially designed as a one-sided, 0.025 significance level hypothesis test that anti-arrhythmic drug therapy would reduce the incidence of sudden cardiac death. Since the drugs were already marketed, harmful effects were not expected. Despite the one-sided hypothesis in the design, the monitoring process used a two-sided, 0.05 significance level approach. In this respect, the level of evidence for benefit was the same for either the one-sided 0.025 or two-sided 0.05 significance level design. As it turned out, the trial was terminated early due to increased mortality in the intervention group (see Chaps. 16 and 17).

If a one-sided test of hypothesis is chosen, in most circumstances the significance level ought to be half what the investigator would use for a two-sided test. For example, if 0.05 is the two-sided significance level typically used, 0.025 would be used for the one-sided test. As done in the CAST trial, this requires the same degree of evidence or scientific documentation to declare a treatment effective, regardless of the one-sided vs. two-sided question. In this circumstance, a test for negative or harmful effects might also be done at the 0.025 level. This in effect, provides two one-sided 0.025 hypothesis tests for an overall 0.05 significance level.

As mentioned above, the total sample size $2N$ (N per arm) is a function of the significance level (α), the power ($1-\beta$) and the size of the difference in response (δ) which is to be detected. Changing either α, $1-\beta$ or δ will result in a change in $2N$. As the magnitude of the difference δ decreases, the larger the sample size must be to guarantee a high probability of finding that difference. If the calculated sample size is larger than can be realistically obtained, then one or more of the parameters in the design may need to be reconsidered. Since the significance level is usually fixed at 0.05, 0.025, or 0.01, the investigator should generally reconsider the value selected for δ and increase it, or keep δ the same and settle for a less powerful study. If neither of these alternatives is satisfactory, serious consideration should be given to abandoning the trial.

Rothman [44] argued that journals should encourage using confidence intervals to report clinical trial results instead of significance levels. Several researchers [44–46] discuss sample size formulas from this approach. Confidence intervals are constructed by computing the observed difference in event rates and then adding and subtracting a constant times the standard error of the difference. This provides an interval surrounding the observed estimated difference obtained from the trial. The constant is determined so as to give the confidence interval the correct probability of including the true, but unknown difference. This constant is related directly to the critical value used to evaluate test statistics. Trials often use a two-sided α level test (e.g., $\alpha = 0.05$) and a corresponding $(1-\alpha)$ confidence interval (e.g., 95%). If the $1-\alpha$ confidence interval excludes zero or no difference, we would conclude that the intervention has an effect. If the interval contains zero difference, no intervention effect would be claimed. However, differences of importance could exist, but might not be detected or not be statistically significant because the sample size was too small. For testing the null hypothesis of no

treatment effect, hypothesis testing and confidence intervals give the same conclusions. However, confidence intervals provide more information on the range of the likely difference that might exist. For sample size calculations, the desired confidence interval width must be specified. This may be determined, for example, by the smallest difference between two event rates that would be clinically meaningful and important. Under the null hypothesis of no treatment effect, half the desired interval width is equal to the difference specified in the alternative hypothesis. The sample size calculation methods presented here do not preclude the presentation of results as confidence intervals and, in fact, investigators ought to do so. However, unless there is an awareness of the relationship between the two approaches, as McHugh and Le [46] have pointed out, the confidence interval method might yield a power of only 50% to detect a specified difference. This can be seen later, when sample size calculations for comparing proportions are presented. Thus, some care needs to be taken in using this method.

So far, it has been assumed that the data will be analyzed only once at the end of the trial. However, as discussed in Chaps. 16 and 17, the response variable data may be reviewed periodically during the course of a study. Thus, the probability of finding significant differences by chance alone is increased [47]. This means that the significance level α may need to be adjusted to compensate for the increase in the probability of a Type I error. For purposes of this discussion, we assume that α carries the usual values of 0.05, 0.025 or 0.01. The sample size calculation should also employ the statistic which will be used in data analysis. Thus, there are many sample size formulations. Methods that have proven useful will be discussed in the rest of this chapter.

Dichotomous Response Variables

We shall consider two cases for response variables which are dichotomous, that is, yes or no, success or failure, presence or absence. The first case assumes two independent groups or samples [48–59]. The second case is for dichotomous responses within an individual, or paired responses [60–64].

Two Independent Samples

Suppose the primary response variable is the occurrence of an event over some fixed period of time. The sample size calculation should be based on the specific test statistic that will be employed to compare the outcomes. The null hypothesis H_0 ($p_C - p_I = 0$) is compared to an alternative hypothesis H_A ($p_C - p_I \neq 0$). The estimates of p_I and p_C are $\widehat{p_C} - \widehat{p}_I$ where $\widehat{p}_I = r_I/N_I$ and $\widehat{p_C} = r_C/N_C$ with r_I and r_C being the number of events in the intervention and control groups and N_I and

Table 8.1 Z_α for sample size formulas for various values of α

	Z_α	
α	One-sided test	Two-sided test
0.10	1.282	1.645
0.05	1.645	1.960
0.025	1.960	2.240
0.01	2.326	2.576

N_C being the number of participants in each group. The usual test statistic for comparing such dichotomous or binomial responses is

$$Z = (\widehat{p_C} - \widehat{p_I}) \Big/ \sqrt{\overline{p}(1-\overline{p})\,(1/N_C + 1/N_I)}$$

where $\overline{p} = (r_I + r_C)/(N_I + N_C)$. The square of the Z statistic is algebraically equivalent to the chi-square statistic, which is often employed as well. For large values of N_I and N_C, the statistic Z has approximately a normal distribution with mean 0 and variance 1. If the test statistic Z is larger in absolute value than a constant Z_α, the investigator will reject H_0 in the two-sided test.

The constant Z_α is often referred to as the critical value. The probability of a standard normal random variable being larger in absolute value than Z_α is α. For a one-sided hypothesis, the constant Z_α is chosen such that the probability that Z is greater (or less) than Z_α is α. For a given α, Z_α is larger for a two-sided test than for a one-sided test (Table 8.1). Z_α for a two-sided test with $\alpha = 0.10$ has the same value as Z_α for a one-sided test with $\alpha = 0.05$. While a smaller sample size can be achieved with a one-sided test compared to a two-sided test at the same α level, we in general do not recommend this approach as discussed earlier.

The sample size required for the design to have a significance level α and a power of $1 - \beta$ to detect true differences of at least δ between the event rates p_I and p_C can be expressed by the formula [11]:

$$2N = 2\left\{Z_\alpha\sqrt{\overline{p}(1-\overline{p})} + Z_\beta\sqrt{\overline{p_C}(1-\overline{p_C}) + \overline{p_I}(1-\overline{p_I})}\right\}^2 \Big/ (p_C - p_I)^2$$

where $2N$ = total sample size (N participants/group) with $\overline{p} = (p_C + p_I)/2$; Z_α is the critical value which corresponds to the significance level α; and Z_β is the value of the standard normal value not exceeded with probability β. Z_β corresponds to the power $1 - \beta$ (e.g., if $1 - \beta = 0.90$, $Z_\beta = 1.282$). Values of Z_α and Z_β are given in Tables 8.1 and 8.2 for several values of α and $1 - \beta$. More complete tables may be found in most introductory texts textbooks [27–29, 31, 33–37, 51], sample size texts [17–21, 65], or by using software packages and online resources [66–73]. Note that the definition of $\overline{p}$ given earlier is equivalent to the definition of $\overline{p}$ given here when $N_I = N_C$; that is, when the two study groups are of equal size. An alternative to the above formula is given by

Table 8.2 Z_β for sample size formulas for various values of power $(1-\beta)$

$1-\beta$	Z_β
0.50	0.00
0.60	0.25
0.70	0.53
0.80	0.84
0.85	1.036
0.90	1.282
0.95	1.645
0.975	1.960
0.99	2.326

$$2N = 4\left(Z_\alpha + Z_\beta\right)^2 \overline{p}(1-\overline{p}) \Big/ (p_C - p_I)^2$$

These two formulas give approximately the same answer and either may be used for the typical clinical trial.

Example: Suppose the annual event rate in the control group is anticipated to be 20%. The investigator hopes that the intervention will reduce the rate to 15%. The study is planned so that each participant will be followed for 2 years. Therefore, if the assumptions are accurate, approximately 40% of the participants in the control group and 30% of the participants in the intervention group will develop an event. Thus, the investigator sets $p_C = 0.40$, $p_I = 0.30$, and, therefore, $\overline{p} = (0.4 + 0.3)/2 = 0.35$. The study is designed as two-sided with a 5% significance level and 90% power. From Tables 8.1 and 8.2, the two-sided 0.05 critical value is 1.96 for Z_β and 1.282 for Z_β. Substituting these values into the right-hand side of the first sample size formula yields $2N$ to be

$$2\left\{1.96\sqrt{2(0.35)\,(0.65)} + 1.282\sqrt{0.4(0.6) + 0.3(0.7)}\right\}^2 \Big/ (0.4 - 0.3)^2$$

Evaluating this expression, $2N$ equals 952.3. Using the second formula, $2N$ is $4(1.96 + 1.202)2\ (0.35)(0.65)/(0.4 - 0.3)2$ or $2N = 956$. Therefore, after rounding up to the nearest ten, the calculated total sample size by either formula is 960, or 480 in each group.

Sample size estimates using the first formula are given in Table 8.3 for a variety of values of p_I and p_C, for two-sided tests, and for $\alpha = 0.01$, 0.025 and 0.05 and $1 - \beta = 0.80$ or 0.90. For the example just considered with $\alpha = 0.05$ (two-sided), $1 - \beta = 0.90$, $p_C = 0.4$ and $p_I = 0.3$, the total sample size using Table 8.3 is 960. This table shows that, as the difference in rates between groups increases, the sample size decreases.

The event rate in the intervention group can be written as $p_I = (1 - k)\, p_C$ where k represents the proportion that the control group event rate is expected to be reduced by the intervention. Figure 8.2 shows the total sample size $2N$ versus k for several values of p_C using a two-sided test with $\alpha = 0.05$ and $1 - \beta = 0.90$. In the example where $p_C = 0.4$ and $p_I = 0.3$, the intervention is expected to reduce

Table 8.3 Sample size

		2α (Two-sided)					
Alpha/power		0.01		0.025		0.05	
p_C	p_I	0.90	0.80	0.90	0.80	0.90	0.80
0.6	0.5	1470	1160	1230	940	1040	780
	0.4	370	290	310	240	260	200
	0.3	160	130	140	110	120	90
	0.20	90	70	80	60	60	50
0.5	0.40	1470	1160	1230	940	1040	780
	0.30	360	280	300	230	250	190
	0.25	220	180	190	140	160	120
	0.20	150	120	130	100	110	80
0.4	0.30	1360	1060	1130	870	960	720
	0.25	580	460	490	370	410	310
	0.20	310	250	260	200	220	170
0.3	0.20	1120	880	930	710	790	590
	0.15	460	360	390	300	330	250
	0.10	240	190	200	150	170	130
0.2	0.15	3440	2700	2870	2200	2430	1810
	0.10	760	600	630	490	540	400
	0.05	290	230	240	190	200	150
0.1	0.05	1650	1300	1380	1060	1170	870

the control rate by 25% or k = 0.25. In Fig. 8.2, locate $k = 0.25$ on the horizontal axis and move up vertically until the curve labeled $p_C = 0.4$ is located. The point on this curve corresponds to a $2N$ of approximately 960. Notice that as the control group event rate p_C decreases, the sample size required to detect the same proportional reduction increases. Trials with small event rates (e.g., $p_C = 0.1$) require large sample sizes unless the interventions have a dramatic effect.

In order to make use of the sample size formula or table, it is necessary to know something about p_C and k. The estimate for p_C is usually obtained from previous studies of similar people. In addition, the investigator must choose k based on preliminary evidence of the potential effectiveness of the intervention or be willing to specify some minimum difference or reduction that he wants to detect. Obtaining this information is difficult in many cases. Frequently, estimates may be based on a small amount of data. In such cases, several sample size calculations based on a range of estimates help to assess how sensitive the sample size is to the uncertain estimates of p_C, k, or both. The investigator may want to be conservative and take the largest, or nearly largest, estimate of sample size to be sure his study has sufficient power. The power $(1 - \beta)$ for various values of δ can be compared for a given sample size $2N$, significance level α, and control rate p_C. By examining a power curve such as in Fig. 8.1, it can be seen what power the trial has for detecting various differences in rates, δ. If the power is high, say 0.80 or larger, for the range of values δ that are of interest, the sample size is probably adequate. The power

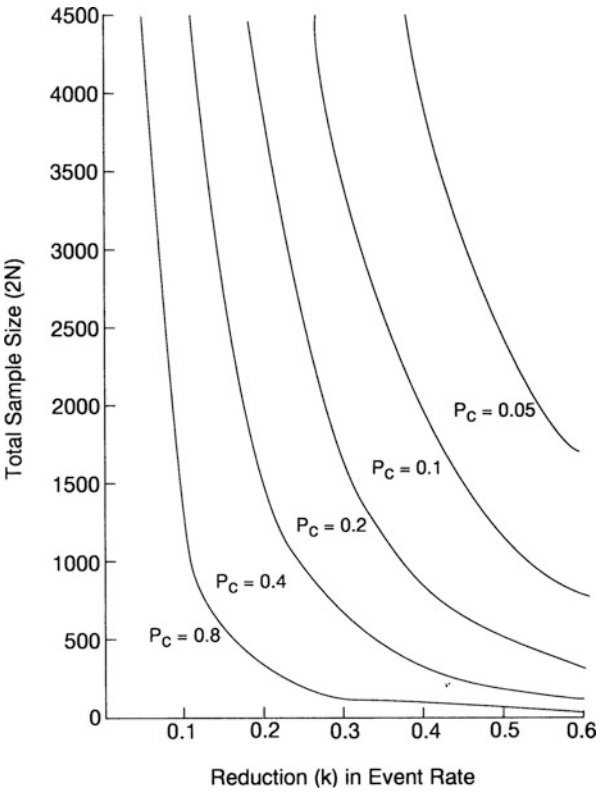

Fig. 8.2 Relationship between total sample size ($2N$) and reduction in event rate (k) for several control group event rates (p_C), with a two-sided significance level of 0.05 and power of 0.90

curve can be especially helpful if the number of available participants is relatively fixed and the investigator wants to assess the probability that the trial can detect any of a variety of reductions in event rates.

Investigators often overestimate the number of eligible participants who can be enrolled in a trial. The actual number enrolled may fall short of goal. To examine the effects of smaller sample sizes on the power of the trial, the investigator may find it useful to graph power as a function of various sample sizes. If the power falls far below 0.8 for a sample size that is very likely to be obtained, he can expand the recruitment effort, hope for a larger intervention effect than was originally assumed, accept the reduced power and its consequences or abandon the trial.

To determine the power, the second sample size equation in this section is solved for Z_β:

$$Z_\beta = \left\{-Z_\alpha\sqrt{2\overline{p}(1-\overline{p})} + \sqrt{N}(p_C - p_I)\right\} \Big/ \sqrt{p_C(1-p_C) + p_I(1-p_I)}$$

where $\overline{p}$ as before is $(p_C + p_I)/2$. The term Z_β can be translated into a power of $1-\beta$ by use of Table 8.2. For example, let $p_C = 0.4$ and $p_I = 0.3$. For a significance level of 0.05 in a two-sided test of hypothesis, Z_α is 1.96. In a previous example, it was

shown that a total sample of approximately 960 participants or 480 per group is necessary to achieve a power of 0.90. Substituting $Z_\alpha = 1.96$, $N = 480$, $p_C = 0.4$ and $p_I = 0.3$, a value for $Z_\beta = 1.295$ is obtained. The closest value of Z_β in Table 8.2 is 1.282 which corresponds to a power of 0.90. (If the exact value of $N = 476$ were used, the value of Z_β would be 1.282.) Suppose an investigator thought he could get only 350 participants per group instead of the estimated 480. Then $Z_\beta = 0.818$ which means that the power $1 - \beta$ is somewhat less than 0.80. If the value of Z_β is negative, the power is less than 0.50. For more details of power calculations, a standard text in biostatistics [27–29, 31, 33–37, 51] or sample size [17–21, 65] should be consulted.

For a given $2N$, α, $1 - \beta$, and p_C the reduction in event rate that can be detected can also be calculated. This function is nonlinear and, therefore, the details will not be presented here. Approximate results can be obtained by scanning Table 8.3, by using the calculations for several p_I until the sample size approaches the planned number, or by using a figure where sample sizes have been plotted. In Fig. 8.2, α is 0.05 and $1 - \beta$ is 0.90. If the sample size is selected as 1000, with $p_C = 0.4$, k is determined to be about 0.25. This means that the expected p_I would be 0.3. As can be seen in Table 8.3, the actual sample size for these assumptions is 960.

The above approach yields an estimate which is more accurate as the sample size increases. Modifications [49, 51–55, 58, 59, 74] have been developed which give some improvement in accuracy to the approximate formula presented for small studies. However, the availability of computer software to perform exact computations [66–73] has reduced the need for good small sample approximations. Also, given that sample size estimation is somewhat imprecise due to assumptions of intervention effects and event rates, the formulation presented is probably adequate for most clinical trials.

Designing a trial comparing proportions using the confidence interval approach, we would need to make a series of assumptions as well [6, 42, 52]. A $100(1 - \alpha)\%$ confidence interval for a treatment comparison θ would be of the general form $\hat{\theta} \pm Z_\alpha \mathrm{SE}(\hat{\theta})$, where $\hat{\theta}$ is the estimate for θ and $\mathrm{SE}(\hat{\theta})$ is the standard error of $\hat{\theta}$. In this case, the specific form would be:

$$(\widehat{p_C} - \widehat{p_I}) \pm Z_\alpha \sqrt{\overline{p}(1 - \overline{p})\,(1/N_I + 1/N_C)}$$

If we want the width of the confidence interval (CI) not to exceed W_{CI}, where W_{CI} is the difference between the upper confidence limit and the lower confidence limit, then if $N = N_I = N_C$, the width W_{CI} can be expressed simply as:

$$W_{CI} = 2\ Z_\alpha \sqrt{\overline{p}(1 - \overline{p})\,(N/2)}$$

or after solving this equation for N,

$$N = 8 Z_\alpha^2 \overline{p}(1 - \overline{p})/(W_{CI})^2$$

Thus, if α is 0.05 for a 95% confidence interval, $p_C = 0.4$ and $p_I = 0.3$ or 0.35, $N = 8(1.96)2(0.35)(0.65)/(W_{CI})^2$. If we desire the upper limit of the confidence

interval to be not more than 0.10 from the estimate or the width to be twice that, then $W_{CI} = 0.20$ and $N = 175$ or $2N = 350$. Notice that even though we are essentially looking for differences in $p_C - p_I$ to be the same as our previous calculation, the sample size is smaller. If we let $p_C - p_I = W_{CI}/2$ and substitute this into the previous sample size formula, we obtain

$$\begin{aligned} 2N &= 2\{Z_\alpha + Z_\beta\}^2 \overline{p}(1-\overline{p})/(W_{CI}/2)^2 \\ &= 8\{Z_\alpha + Z_\beta\}^2 \overline{p}(1-\overline{p})/(W_{CI})^2 \end{aligned}$$

This formula is very close to the confidence interval formula for two proportions. If we select 50% power, β is 0.50 and Z_β is 0 which would yield the confidence interval formula. Thus, a confidence interval approach gives 50% power to detect differences of $W_{CI}/2$. This may not be adequate, depending on the situation. In general, we prefer to specify greater power (e.g., 80–90%) and use the previous approach.

Analogous sample size estimation using the confidence interval approach may be used for comparing means, hazard rates, or regression slopes. We do not present details of these since we prefer to use designs which yield power greater than that obtained from a confidence interval approach.

Paired Dichotomous Response

For designing a trial where the paired outcomes are binary, the sample size estimate is based on McNemar's test [60–64]. We want to compare the frequency of success within an individual on intervention with the frequency of success on control (i.e., $p_I - p_C$). McNemar's test compares difference in discordant responses within an individual $p_I - p_C$, between intervention and control.

In this case, the number of paired observations, N_p, may be estimated by:

$$N_p = \left[Z_\alpha \sqrt{f} + Z_\beta \sqrt{f - d^2}\right]^2 \Big/ d^2$$

where $d =$ difference in the proportion of successes ($d = p_I - p_C$) and f is the proportion of participants whose response is discordant. An alternative approximate formula for N_p is

$$N_p = \left(Z_\alpha + Z_\beta\right)^2 f/d^2$$

Example: Consider an eye study where one eye is treated for loss in visual acuity by a new laser procedure and the other eye is treated by standard therapy. The failure rate on the control, p_C is estimated to be 0.40 and the new procedure is projected to reduce the failure rate to 0.20. The discordant rate f is assumed to be 0.50. Using the

latter sample size formula for a two-sided 5% significance level and 90% power, the number of pairs N_p is estimated as 132. If the discordant rate is 0.8, then 210 pairs of eyes will be needed.

Adjusting Sample Size to Compensate for Nonadherence

During the course of a clinical trial, participants will not always adhere to their prescribed intervention schedule. The reason is often that the participant cannot tolerate the dosage of the drug or the degree of intervention prescribed in the protocol. The investigator or the participant may then decide to follow the protocol with less intensity. At all times during the conduct of a trial, the participant's welfare must come first and meeting those needs may not allow some aspects of the protocol to be followed. Planners of clinical trials must recognize this possibility and attempt to account for it in their design. Examples of adjusting for nonadherence with dichotomous outcomes can be found in several clinical trials [75–82].

In the intervention group a participant who does not adhere to the intervention schedule is often referred to as a "drop-out." Participants who stop the intervention regimen lose whatever potential benefit the intervention might offer. Similarly, a participant on the control regimen may at some time begin to use the intervention that is being evaluated. This participant is referred to as a "drop-in." In the case of a drop-in a physician may decide, for example, that surgery is required for a participant assigned to medical treatment in a clinical trial of surgery versus medical care [77]. Drop-in participants from the control group who start the intervention regimen will receive whatever potential benefit or harm that the intervention might offer. Therefore, both the drop-out and drop-in participants must be acknowledged because they tend to dilute any difference between the two groups which might be produced by the intervention. This simple model does not take into account the situation in which one level of an intervention is compared to another level of the intervention. More complicated models for nonadherence adjustment can be developed. Regardless of the model, it must be emphasized that the assumed event rates in the control and intervention groups are modified by participants who do not adhere to the study protocol.

People who do not adhere should remain in the assigned study groups and be included in the analysis. The rationale for this is discussed in Chap. 18. The basic point to be made here is that eliminating participants from analysis or transferring participants to the other group could easily bias the results of the study. However, the observed δ is likely to be less than projected because of nonadherence and thus have an impact on the power of the clinical trial. A reduced δ, of course, means that either the sample size must be increased or the study will have smaller power than intended. Lachin [11] has proposed a simple formula to adjust crudely the sample size for a drop-out rate of proportion R_O. This can be generalized to adjust for drop-in rates, R_I, as well. The unadjusted sample size N should be multiplied by the factor

$\{1/(1-R_O-R_I)\}^2$ to get the adjusted sample size per arm, N^*. Thus, if $R_O = 0.20$ and $R_I = 0.05$, the originally calculated sample should be multiplied by 1/(0.75)2, or 16/9, and increased by 78%. This formula gives some quantitative idea of the effect of drop-out on the sample size:

$$N^* = N/(1-R_O-R_I)^2$$

However, more refined models to adjust sample sizes for drop-outs from the intervention to the control [83–89] and for drop-ins from the control to the intervention regimen [83] have been developed. They adjust for the resulting changes in p_I and p_C, the adjusted rates being denoted p_I^* and p_C^*. These models also allow for another important factor, which is the time required for the intervention to achieve maximum effectiveness. For example, an anti-platelet drug may have an immediate effect; conversely, even though a cholesterol-lowering drug reduces serum levels quickly, it may require years to produce a maximum effect on coronary mortality.

Example: A drug trial [76] in post myocardial infarction participants illustrates the effect of drop-outs and drop-ins on sample size. In this trial, total mortality over a 3-year follow-up period was the primary response variable. The mortality rate in the control group was estimated to be 18% ($p_C = 0.18$) and the intervention was believed to have the potential for reducing p_C by 28% ($k = 0.28$) yielding $p_I = 0.1296$. These estimates of p_C and k were derived from previous studies. Those studies also indicated that the drop-out rate might be as high as 26% over the 3 years; 12% in the first year, an additional 8% in the second year, and an additional 6% in the third year. For the control group, the drop-in rate was estimated to be 7% each year for a total drop-in rate of 21%.

Using these models for adjustment, $p_C^* = 0.1746$ and $p_I^* = 0.1375$. Therefore, instead of δ being 0.0504 (0.18 − 0.1296), the adjusted δ^* is 0.0371 (0.1746 − 0.1375). For a two-sided test with $\alpha = 0.05$ and $1-\beta = 0.90$, the adjusted sample size was 4020 participants compared to an unadjusted sample size of 2160 participants. The adjusted sample size almost doubled in this example due to the expected drop-out and drop-in experiences and the recommended policy of keeping participants in the originally assigned study groups. The remarkable increases in sample size because of drop-outs and drop-ins strongly argue for major efforts to keep nonadherence to a minimum during trials.

Sample Size Calculations for Continuous Response Variables

Similar to dichotomous outcomes, we consider two sample size cases for response variables which are continuous [9, 11, 90]. The first case is for two independent samples. The other case is for paired data.

Two Independent Samples

For a clinical trial with continuous response variables, the previous discussion is conceptually relevant, but not directly applicable to actual calculations. "Continuous" variables such as length of hospitalization, blood pressure, spirometric measures, neuropsychological scores and level of a serum component may be evaluated. Distributions of such measurements frequently can be approximated by a normal distribution. When this is not the case, a transformation of values, such as taking their logarithm, can often make the normality assumption approximately correct.

Suppose the primary response variable, denoted as x, is continuous with N_I and N_C participants randomized to the intervention and control groups respectively. Assume that the variable x has a normal distribution with mean μ and variance σ^2. The true levels of μ_I and μ_C for the intervention and control groups are not known, but it is assumed that σ^2 is known. (In practice, σ^2 is not known and must be estimated from some data. If the data set used is reasonably large, the estimate of σ^2 can be used in place of the true σ^2. If the estimate for σ^2 is based on a small set of data, it is necessary to be cautious in the interpretation of the sample size calculations.)

The null hypothesis is H_0: $\delta = \mu_C - \mu_I = 0$ and the two-sided alternative hypothesis is H_A: $\delta = \mu_C - \mu_I \neq 0$. If the variance is known, the test statistic is:

$$Z = (\overline{x}_C - \overline{x}_I)/\sigma\sqrt{(1/N_C + 1/N_I)}$$

where $\overline{x}_I$ and $\overline{x}_C$ represent mean levels observed in the intervention and control groups respectively. For adequate sample size (e.g. 50 participants per arm) this statistic has approximately a standard normal distribution. The hypothesis-testing concepts previously discussed apply to the above statistic. If $Z > Z_\alpha$, then an investigator would reject H_0 at the α level of significance. By use of the above test statistic it can be determined how large a total sample $2N$ would be needed to detect a true difference δ between μ_I and μ_C with power $(1 - \beta)$ and significance level α by the formula:

$$2N = 4\left(Z_\alpha + Z_\beta\right)^2\sigma^2/\delta^2$$

Example: Suppose an investigator wishes to estimate the sample size necessary to detect a 10 mg/dL difference in cholesterol level in a diet intervention group compared to the control group. The variance from other data is estimated to be $(50\ \text{mg/dL})^2$. For a two-sided 5% significance level, $Z_\alpha = 1.96$ and for 90% power, $Z_\beta = 1.282$. Substituting these values into the above formula, $2N = 4(1.96 + 1.282)^2(50)^2/10^2$ or approximately 1,050 participants. As δ decreases, the value of $2N$ increases, and as σ^2 increases the value of $2N$ increases. This means that the smaller the difference in intervention effect an investigator is interested in detecting and the larger the variance, the larger the study must be. As with the dichotomous case, setting a smaller α and larger $1 - \beta$ also increases the sample size. Figure 8.3 shows total sample size $2N$ as a function of δ/σ. As in the example, if $\delta = 10$ and $\sigma = 50$, then $\delta/\sigma = 0.2$ and the sample size $2N$ for $1 - \beta = 0.9$ is approximately 1,050.

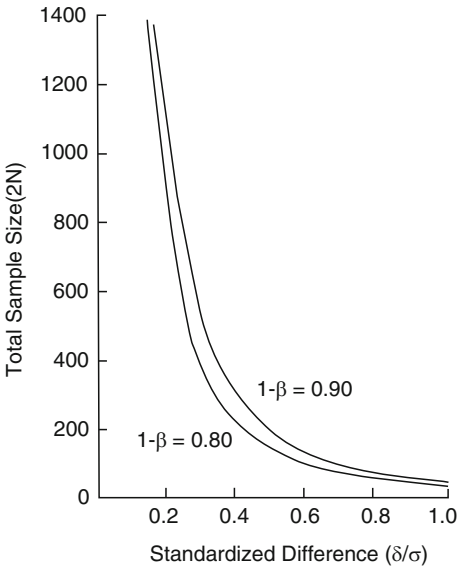

Fig. 8.3 Total sample size ($2N$) required to detect the difference (δ) between control group mean and intervention group mean as a function of the standardized difference (δ/σ) where σ is the common standard deviation, with two-sided significance level of 0.05 and power ($1-\beta$) of 0.80 and 0.90

Paired Data

In some clinical trials, paired outcome data may increase power for detecting differences because individual or within participant variation is reduced. Trial participants may be assessed at baseline and at the end of follow-up. For example, instead of looking at the difference between mean levels in the groups, an investigator interested in mean levels of change might want to test whether diet intervention lowers serum cholesterol from baseline levels when compared to a control. This is essentially the same question as asked before in the two independent sample case, but each participant's initial cholesterol level is taken into account. Because of the likelihood of reduced variability, this type of design can lead to a smaller sample size if the question is correctly posed. Assume that Δ_C and Δ_I represent the true, but unknown levels of change from baseline to some later point in the trial for the control and intervention groups, respectively. Estimates of Δ_C and Δ_I would be $\overline{d}_C = \overline{x}_{C_1} - \overline{x}_{C_2}$ and $\overline{d}_I = \overline{x}_{I_1} - \overline{x}_{I_2}$. These represent the differences in mean levels of the response variable at two points for each group. The investigator tests H_0: $\Delta_C - \Delta_I = 0$ versus H_A: $\Delta_C - \Delta_I = \delta \neq 0$. The variance σ^2 in this case reflects the variability of the change, from baseline to follow-up, and is assumed here to be the same in the control and intervention arms. This variance is likely to be smaller than the variability at a single measurement. This is the case if the correlation between the first and second measurements is greater than 0.5. Using δ and σ_Δ^2, as defined in this manner, the previous sample size formula for two independent samples and graph are applicable. That is, the total sample size $2N$ can be estimated as

$$2N = 4\left(Z_\alpha + Z_\beta\right)^2 \sigma_\Delta^2 / \delta^2$$

Another way to represent this is

$$2N = 4\left(Z_\alpha + Z_\beta\right)^2(1-\rho)\sigma^2/\delta^2$$

where $\sigma_\Delta^2 = 2\sigma^2(1-\rho)$ and σ^2 is the variance of a measurement at a single point in time, the variability is assumed to be the same at both time points (i.e. at baseline and at follow-up), and ρ is the correlation coefficient between the first and second measurement. As indicated, if the correlation coefficient is greater than 0.5, comparing the paired differences will result in a smaller sample size than just comparing the mean values at the time of follow-up.

Example: Assume that an investigator is still interested in detecting a 10 mg/dL difference in cholesterol between the two groups, but that the variance of the change is now $(20\ \text{mg/dL})^2$. The question being asked in terms of δ is approximately the same, because randomization should produce baseline mean levels in each group which are almost equal. The comparison of differences in change is essentially a comparison of the difference in mean levels of cholesterol at the second measurement. Using Fig. 8.3, where $\delta/\sigma_\Delta = 10/20 = 0.5$, the sample size is 170. This impressive reduction in sample size from 1,050 is due to a reduction in the variance from $(50\ \text{mg/dL})^2$ to $(20\ \text{mg/dL})^2$.

Another type of pairing occurs in diseases that affect paired organs such as lungs, kidneys, and eyes. In ophthalmology, for example, trials have been conducted where one eye is randomized to receive treatment and the other to receive control therapy [61–64]. Both the analysis and the sample size estimation need to take account of this special kind of stratification. For continuous outcomes, a mean difference in outcome between a treated eye and untreated eye would measure the treatment effect and could be compared using a paired t-test [9, 11], $Z = \overline{d}/S_d\sqrt{1/N}$, where $\overline{d}$ is the average difference in response and S_d is the standard deviation of the differences. The mean difference μ_d is equal to the mean response of the treated or intervention eye, for example, minus the mean response of the control eye; that is $\mu_d = \mu_I - \mu_C$. Under the null hypothesis, μ_d equals δ_d. An estimate of δ_d, $\overline{d}$, can be obtained by taking an estimate of the average differences or by calculating $\overline{x}_I - \overline{x}_C$. The variance of the paired differences σ_d^2 is estimated by S_d^2. Thus, the formula for paired continuous outcomes within an individual is a slight modification of the formula for comparison of means in two independent samples. To compute sample size, N_d, for number of pairs, we compute:

$$N_d = \left(Z_\alpha + Z_\beta\right)^2\sigma_d^2/\delta_d^2$$

As discussed previously, participants in clinical trials do not always fully adhere with the intervention being tested. Some fraction (R_O) of participants on intervention drop-out of the intervention and some other fraction (R_I) drop-in and start following the intervention. If we assume that these participants who drop-out respond as if they had been on control and those who drop-in respond as if they had been on intervention, then the sample size adjustment is the same as for the case of proportions. That is, the adjusted sample size N^* is a function of the drop-out rate, the drop-in rate, and the sample size N for a study with fully compliant participants:

$$N^* = N/(1 - R_O - R_I)^2$$

Therefore, if the drop-out rate were 0.20 and the drop-in 0.05, then the original sample size N must be increased by 16/9 or 1.78; that is, a 78% increase in sample size.

Sample Size for Repeated Measures

The previous section briefly presented the sample size calculation for trials where only two points, say a baseline and a final visit, are used to determine the effect of intervention and these two points are the same for all study participants. Often, a continuous response variable is measured at each follow-up visit. Considering only the first and last values would give one estimate of change but would not take advantage of all the available data. Many models exist for the analysis of repeated measurements and formulae [13, 91–97] as well as computer software [66, 67, 69–73] for sample size calculation are available for most. In some cases, the response variable may be categorical. We present one of the simpler models for continuous repeated measurements. While other models are beyond the scope of this book, the basic concepts presented are still useful in thinking about how many participants, how many measurements per individual, and when they should be taken, are needed. In such a case, one possible approach is to assume that the change in response variable is approximately a linear function of time, so that the rate of change can be summarized by a slope. This model is fit to each participant's data by the standard least squares method and the estimated slope is used to summarize the participant's experience. In planning such a study, the investigator must be concerned about the frequency of the measurement and the duration of the observation period. As discussed by Fitzmaurice and co-authors [98], the observed measurement x can be expressed as $x = a + bt +$ error, where $a =$ intercept, $b =$ slope, $t =$ time, and error represents the deviation of the observed measurement from a regression line. This error may be due to measurement variability, biological variability or the nonlinearity of the true underlying relationship. On the average, this error is expected to be equally distributed around 0 and have a variability denoted as $\sigma^2_{(\text{error})}$. Though it is not necessary, it simplifies the calculation to assume that $\sigma^2_{(\text{error})}$ is approximately the same for each participant.

The investigator evaluates intervention effectiveness by comparing the average slope in one group with the average slope in another group. Obviously, participants in a group will not have the same slope, but the slopes will vary around some average value which reflects the effectiveness of the intervention or control. The amount of variability of slopes over participants is denoted as σ^2_b. If D represents the total time duration for each participant and P represents the number of equally spaced measurements, σ^2_b can be expressed as:

$$\sigma_b^2 = \sigma_B^2 + \left\{12(P-1)\,\sigma_{(\text{error})}^2/\left(D^2P(P+1)\right)\right\}$$

where σ_B^2 is the component of variance attributable to differences in participants' slope as opposed to measurement error and lack of a linear fit. The sample size required to detect difference δ between the average rates of change in the two groups is given by:

$$2N = \left[4\left(Z_\alpha + Z_\beta\right)^2/\delta^2\right]\ \left[\sigma_B^2 + \left\{12(P-1)\,\sigma_{(\text{error})}^2/\left(D^2P(P+1)\right)\right\}\right]$$

As in the previous formulas, when δ decreases, $2N$ increases. The factor on the right-hand side relates D and P with the variance components σ_B^2 and $\sigma_{(\text{error})}^2$. Obviously as σ_B^2 and $\sigma_{(\text{error})}^2$ increase, the total sample size increases. By increasing P and D, however, the investigator can decrease the contribution made by $\sigma_{(\text{error})}^2$. The exact choices of P and D will depend on how long the investigator can feasibly follow participants, how many times he can afford to have participants visit a clinic and other factors. By manipulating P and D, an investigator can design a study which will be the most cost effective for his specific situation.

Example: In planning for a trial, it may be assumed that a response variable declines at the rate of 80 units/year in the control group. Suppose a 25% reduction is anticipated in the intervention group. That is, the rate of change in the intervention group would be 60 units/year. Other studies provided an estimate for $\sigma_{(\text{error})}$ of 150 units. Also, suppose data from a study of people followed every 3 months for 1 year ($D = 1$ and $P = 5$) gave a value for the standard deviation of the slopes, $\sigma_b = 200$. The calculated value of σ_B is then 63 units. Thus, for a 5% significance level and 90% power ($Z_\alpha = 1.96$ and $Z_\beta = 1.282$), the total sample size would be approximately 630 for a 3-year study with four visits per year ($D = 3$, $P = 13$). Increasing the follow-up time to 4 years, again with four measurements per year, would decrease the variability with a resulting sample size calculation of approximately 510. This reduction in sample size could be used to decide whether or not to plan a 4-year or a 3-year study.

Sample Size Calculations for "Time to Failure"

For many clinical trials, the primary response variable is the occurrence of an event and thus the proportion of events in each group may be compared. In these cases, the sample size methods described earlier will be appropriate. In other trials, the time to the event may be of special interest. For example, if the time to death or a nonfatal event can be increased, the intervention may be useful even though at some point the proportion of events in each group are similar. Methods for analysis of this type of outcome are generally referred to as life table or survival analysis methods (see Chap. 15). In this situation, other sample size approaches are more appropriate than

that described for dichotomous outcomes [99–118]. At the end of this section, we also discuss estimating the number of events required to achieve a desired power.

The basic approach is to compare the survival curves for the groups. A survival curve may be thought of as a graph of the probability of surviving, or not having an event, up to any given point in time. The methods of analysis now widely used are non-parametric; that is, no mathematical model about the shape of the survival curve is assumed. However, for the purpose of estimating sample size, some assumptions are often useful. A common model assumes that the survival curve, $S(t)$, follows an exponential distribution, $S(t) = e^{-\lambda t} = \exp(-\lambda t)$ where λ is called the hazard rate or force of mortality. Using this model, survival curves are totally characterized by λ. Thus, the survival curves from a control and an intervention group can be compared by testing H_0: $\lambda_C = \lambda_I$. An estimate of λ is obtained as the inverse of the mean survival time. If the median survival time, T_M, is known, the hazard rate λ may also be estimated by $-\ln(0.5)/T_M$. Sample size formulations have been considered by several investigators [103, 112, 119]. One simple formula is given by

$$2N = 4\left(Z_\alpha + Z_\beta\right)^2 / [\ln(\lambda_C/\lambda_I)]^2$$

where N is the size of the sample in each group and Z_α and Z_β are defined as before. As an example, suppose one assumes that the force of mortality is 0.30 in the control group and expects it to be 0.20 for the intervention being tested; that is, $\lambda_C/\lambda_I = 1.5$. If $\alpha = .05$ (two-sided) and $1 - \beta = 0.90$, then $N = 128$ or $2N = 256$. The corresponding mortality rates for 5 years of follow-up are 0.7769 and 0.6321 respectively. Using the comparison of two proportions, the total sample size would be 412. Thus, the time to failure method may give a more efficient design, requiring a smaller number of participants.

The method just described assumes that all participants will be followed to the event. With few exceptions, clinical trials with a survival outcome are terminated at time T before all participants have had an event. For those still event-free, the time to event is said to be censored at time T. For this situation, Lachin [11] gives the approximate formula:

$$2N = 2\left(Z_\alpha + Z_\beta\right)^2 [\varphi(\lambda_C) + \varphi(\lambda_I)] / (\lambda_I - \lambda_C)^2$$

where $\varphi(\lambda) = \lambda^2/(1 - e^{-\lambda T})$ and where $\varphi(\lambda_C)$ or $\varphi(\lambda_I)$ are defined by replacing λ with λ_C or λ_I, respectively. If a 5 year study were being planned ($T = 5$) with the same design specifications as above, then the sample size, $2N$ is equal to 376. Thus, the loss of information due to censoring must be compensated for by increasing the sample size. If the participants are to be recruited continually during the 5 years of the trial, the formula given by Lachin is identical but with $\varphi(\lambda) = \lambda^3 T/(\lambda T - 1 + e^{-\lambda T})$. Using the same design assumptions, we obtain $2N = 620$, showing that not having all the participants at the start requires an additional increase in sample size.

More typically participants are recruited uniformly over a period of time, T_0, with the trial continuing for a total of T years ($T > T_0$). In this situation, the sample size can be estimated as before using:

$$\phi(\lambda) = \lambda^2 / \left[1 - \left(e^{-\lambda(T-T0)} - e^{-\lambda T}\right)/(\lambda T_0)\right]$$

Here, the sample size ($2N$) of 466 is between the previous two examples suggesting that it is preferable to get participants recruited as rapidly as possible to get more follow-up or exposure time.

One of the methods used for comparing survival curves is the proportional hazards model or the Cox regression model which is discussed briefly in Chap. 15. For this method, sample size estimates have been published [101, 115]. As it turns out, the formula by Schoenfeld for the Cox model [115] is identical to that given above for the simple exponential case, although developed from a different point of view. Further models are given by Lachin [11].

All of the above methods assume that the hazard rate remains constant during the course of the trial. This may not be the case. The Beta-Blocker Heart Attack Trial [76] compared 3-year survival in two groups of participants with intervention starting one to 3 weeks after an acute myocardial infarction. The risk of death was high initially, decreased steadily, and then became relatively constant.

For cases where the event rate is relatively small and the clinical trial will have considerable censoring, most of the statistical information will be in the number of events. Thus, the sample size estimates using simple proportions will be quite adequate. In the Beta-Blocker Heart Attack Trial, the 3 year control group event rate was assumed to be 0.18. For the intervention group, the event rate was assumed to be approximately 0.13. In the situation of $\phi(\lambda) = \lambda^2(1 - e^{-\lambda T})$, a sample size $2N = 2{,}208$ is obtained, before adjustment for estimated nonadherence. In contrast, the unadjusted sample size using simple proportions is 2,160. Again, it should be emphasized that all of these methods are only approximations and the estimates should be viewed as such.

As the previous example indicates, the power of a survival analysis still is a function of the number of events. The expected number of events $E(D)$ is a function of sample size, hazard rate, recruitment rate, and censoring distribution [11, 106]. Specifically, the expected number of events in the control group can be estimated as

$$E(D) = N\lambda_C^2 / \phi(\lambda_C)$$

where $\phi(\lambda_C)$ is defined as before, depending on the recruitment and follow-up strategy. If we assume a uniform recruitment over the interval $(0,T_0)$ and follow-up over the interval $(0,T)$, then $E(D)$ can be written using the most general form for $\phi(\lambda_C)$;

$$E(D) = N\left[1 - \left(e^{-\lambda(T-T0)} - e^{-\lambda T}\right)/(\lambda T_0)\right]$$

Table 8.4 Number of expected events (in the control group) at each interim analysis given different event rates in control group

	Number of expected events			
	Calendar time into study			
Yearly event rate in control group	6 Months ($N = 138$/group)	1 Year ($N = 275$/group)	1.5 Years ($N = 412$/group)	2 Years ($N = 412$/group)
40%	16	60	124	189
35%	14	51	108	167
30%	12	44	94	146
25%	10	36	78	123

Assumptions

1. Time to event exponentially distributed
2. Uniform entry into the study over 1.5 years
3. Total duration of 2 years

This estimate of the number of events can be used to predict the number of events at various time points during the trial including the end of follow-up. This prediction can be compared to the observed number of events in the control group to determine if an adjustment needs to be made to the design. That is, if the number of events early in the trial is larger than expected, the trial may be more powerful than designed or may be stopped earlier than the planned T years of follow-up (see Chap. 16). However, more worrisome is when the observed number of events is smaller than what is expected and needed to maintain adequate power. Based on this early information, the design may be modified to attain the necessary number of events by increasing the sample size or expanding recruitment effort within the same period of time, increasing follow-up, or a combination of both.

This method can be illustrated based on a placebo-controlled trial of congestive heart failure [82]. Severe or advanced congestive heart failure has an expected 1 year event rate of 40%, where the events are all-cause mortality and nonfatal myocardial infarction. A new drug was to be tested to reduce the event rate by 25%, using a two-sided 5% significance level and 90% power. If participants are recruited over 1.5 years ($T_0 = 1.5$) during a 2 year study ($T = 2$) and a constant hazard rate is assumed, the total sample size ($2N$) is estimated to be 820 participants with congestive heart failure. The formula $E(D)$ can be used to calculate that approximately 190 events (deaths plus nonfatal myocardial infarctions) must be observed in the control group to attain 90% power. If the first year event rate turns out to be less than 40%, fewer events will be observed by 2 years than the required 190. Table 8.4 shows the expected number of control group events at 6 months and 1 year into the trial for annual event rates of 40, 35, 30, and 25%. Two years is also shown to illustrate the projected number of events at the completion of the study. These numbers are obtained by calculating the number of participants enrolled by 6 months (33% of 400) and 1 year (66% of 400) and multiplying by the bracketed term on the right hand side of the equation for $E(D)$. If the assumed annual event rate of 40% is correct, 60 control group events should be observed at 1 year.

However, if at 1 year only 44 events are observed, the annual event rate might be closer to 30% (i.e., $\lambda = 0.357$) and some design modification should be considered to assure achieving the desired 190 control group events. One year would be a sensible time to make this decision, based only on control group events, since recruitment efforts are still underway. For example, if recruitment efforts could be expanded to 1220 participants in 1.5 years, then by 2 years of follow-up the 190 events in the placebo group would be observed and the 90% power maintained. If recruitment efforts were to continue for another 6 months at a uniform rate ($T_0 = 2$ years), another 135 participants would be enrolled. In this case, $E(D)$ is $545 \times 0.285 = 155$ events which would not be sufficient without some additional follow-up. If recruitment and follow-up continued for 27 months (i.e., $T_0 = T = 2.25$), then 605 control group participants would be recruited and $E(D)$ would be 187, yielding the desired power.

Sample Size for Testing "Equivalency" or Noninferiority of Interventions

In some instances, an effective intervention has already been established and is considered the standard. New interventions under consideration may be preferred because they are less expensive, have fewer side effects, or have less adverse impact on an individual's general quality of life. This issue is common in the pharmaceutical industry where a product developed by one company may be tested against an established intervention manufactured by another company. Studies of this type are sometimes referred to as trials with positive controls or as noninferiority designs (see Chaps. 3 and 5).

Given that several trials have shown that certain beta-blockers are effective in reducing mortality in post-myocardial infarction participants [76, 120, 121], it is likely that any new beta-blockers developed will be tested against proven agents. The Nocturnal Oxygen Therapy Trial [122] tested whether the daily amount of oxygen administered to chronic obstructive pulmonary disease participants could be reduced from 24 to 12 h without impairing oxygenation. The Intermittent Positive Pressure Breathing [80] trial considered whether a simple and less expensive method for delivering a bronchodilator into the lungs would be as effective as a more expensive device. A breast cancer trial compared the tumor regression rates between subjects receiving the standard, diethylstilbestrol, or the newer agent, tamoxifen [123].

The problem in designing noninferiority trials is that there is no statistical method to demonstrate complete equivalence. That is, it is not possible to show $\delta = 0$. Failure to reject the null hypothesis is not a sufficient reason to claim two interventions to be equal but merely that the evidence is inadequate to say they are different [124]. Assuming no difference when using the previously described formulas results in an infinite sample size.

While demonstrating perfect equivalence is an impossible task, one possible approach has been discussed for noninferiority designs [125–128]. The strategy is to specify some value, δ, such that interventions with differences which are less than this might be considered "equally effective" or "noninferior" (see Chap. 5 for discussion of noninferiority designs). Specification of δ may be difficult but it is a necessary element of the design. The null hypothesis states that $p_C > p_I + \delta$ while the alternative specifies $p_C < p_I + \delta$. The methods developed require that if the two interventions really are equally effective or at least noninferior, the upper 100 $(1-\alpha)\%$ confidence interval for the intervention difference will not exceed δ with the probability of $1-\beta$. One can alternatively approach this from a hypothesis testing point of view, stating the null hypothesis that the two interventions differ by less than δ.

For studies with a dichotomous response, one might assume the event rate for the two interventions to be equal to p (i.e., $p = p_C = p_I$). This simplifies the previously shown sample size formula to

$$2N = 4p(1-p)\left(Z_\alpha + Z_\beta\right)^2/\delta^2$$

where N, Z_α and Z_β are defined as before. Makuch and Simon [127] recommend for this situation that $\alpha = 0.10$ and $\beta = 0.20$. However, for many situations, β or Type II error needs to be 0.10 or smaller in order to be sure a new therapy is correctly determined to be equivalent to an older standard. We prefer an $\alpha = 0.05$, but this is a matter of judgment and will depend on the situation. (This formula differs slightly from its analogue presented earlier due to the different way the hypothesis is stated.) The formula for continuous variables,

$$2N = 4\left(Z_\alpha + Z_\beta\right)^2/(\delta/\sigma)^2$$

is identical to the formula for determining sample size discussed earlier. Blackwelder and Chang [126] give graphical methods for computing sample size estimates for studies of equivalency.

As mentioned above and in Chap. 5, specifying δ is a key part of the design and sample size calculations of all equivalency and noninferiority trials. Trials should be sufficiently large, with enough power, to address properly the questions about equivalence or noninferiority that are posed.

Sample Size for Cluster Randomization

So far, sample size estimates have been presented for trials where individuals are randomized. For some prevention trials or health care studies, it may not be possible to randomize individuals. For example, a trial of smoking prevention strategy for teenagers may be implemented most easily by randomizing schools, some schools to be exposed to the new prevention strategy while other schools remain with a

standard approach. Individual students are grouped or clustered within each school. As Donner et al. [129] point out, "Since one cannot regard the individuals within such groups as statistically independent, standard sample size formulas underestimate the total number of subjects required for the trial." Several authors [129–133] have suggested incorporating a single inflation factor in the usual sample size calculation to account for the cluster randomization. That is, the sample size per intervention arm N computed by previous formulas will be adjusted to N^* to account for the randomization of N_m clusters, each with m individuals.

A continuous response is measured for each individual within a cluster of these components. Differences of individuals within a cluster and differences of individuals between clusters contribute to the overall variability of the response. We can separate the between-cluster variance σ_b^2 and within cluster variance σ_w^2. Estimates are denoted S_b^2 and S_w^2, respectively and can be estimated by standard analysis of variance. One measure of the relationship of these components is the intra-class correlation coefficient. The intra-class correlation coefficient ρ is $\sigma_b^2/\left(\sigma_b^2+\sigma_w^2\right)$ where $0 \le \rho \le 1$. If $\rho = 0$, all clusters respond identically so all of the variability is within a cluster. If $\rho = 1$, all individuals in a cluster respond alike so there is no variability within a cluster. Estimates of ρ are given by $r = S_b^2/\left(S_b^2+S_w^2\right)$. Intra-class correlation may range from 0.1 to 0.4 in typical clinical studies. If we computed the sample size calculations assuming no clustering, the sample size per arm would be N participants per treatment arm. Now, instead of randomizing N individuals, we want to randomize N_m clusters with m individuals each for a total of $N^* = N_m \times m$ participants per treatment arm. The inflation factor [133] is $[1+(m-1)r]$ so that

$$N^* = N_m \times m = N[1+(m-1)\rho]$$

Note that the inflation factor is a function of both cluster size m and intra-class correlation. If the intra-cluster correlation ($\rho = 0$), then each individual in one cluster responds like any individual in another cluster, and the inflation factor is unity ($N^* = N$). That is, no penalty is paid for the convenience of cluster randomization. At the other extreme, if all individuals in a cluster respond the same ($\rho = 1$), there is no added information within each cluster, so only one individual per cluster is needed, and the inflation factor is m. That is, our adjusted sample $N^* = N \times m$ and we pay a severe price for this type of cluster randomization. However, it is unlikely that ρ is either 0 or 1, but as indicated, is more likely to be in the range of 0.1–0.4 in clinical studies.

Example: Donner et al. [129] provide an example for a trial randomizing households to a sodium reducing diet in order to reduce blood pressure. Previous studies estimated the intra-class correlation coefficient to be 0.2; that is $\hat{\rho} = r = S_b^2/\left(S_b^2+S_w^2\right) = 0.2$. The average household size was estimated at 3.5 ($m = 3.5$). The sample size per arm N must be adjusted by $1+(m-1)\rho = 1+(3.5-1)(0.2) = 1.5$. Thus, the normal sample size must be inflated by 50% to account for this randomization indicating a small between cluster variability. If

$\rho = 0.1$, then the factor is $1 + (3.5 - 1)(0.1)$ or 1.25. If $\rho = 0.4$, indicating a larger between cluster component of variability, the inflation factor is 2.0 or a doubling.

For binomial responses, a similar expression for adjusting the standard sample size can be developed. In this setting, a measure of the degree of within cluster dependency or concordancy rate in participant responses is used in place of the intra-class correlation. The commonly used measure is the kappa coefficient, denoted κ, and may be thought of as an intra-class correlation coefficient for binomial responses, analogous to ρ for continuous responses. A concordant cluster with $\kappa = 1$ is one where all responses within a cluster are identical, all successes or failures, in which a cluster contributes no more than a single individual. A simple estimate for κ is provided [129]:

$$\kappa = p^* \left[p_C^m + (1 - p_C)^m\right] / \left(1 - \left[p_C^m + (1 - p_C)^m\right]\right)$$

Here p^* is the proportion of the control group with concordant clusters, and p_C is the underlying success rate in the control group. The authors then show that the inflation factor is $[1 + (m - 1)\kappa]$, or that the regular sample size per treatment arm N must be multiplied by this factor to attain the adjust sample size N^*:

$$N^* = N[1 + (m - 1)\kappa]$$

Example: Donner et al. [129] continues the sodium diet example where couples ($m = 2$) are randomized to either a low sodium or a normal diet. The outcome is the hypertension rate. Other data suggest the concordancy of hypertension status among married couples is 0.85 ($p^* = 0.85$). The control group hypertension rate is 0.15 ($p_C = 0.15$). In this case, $\kappa = 0.41$, so that the inflation factor is $1 + (2 - 1)(0.41) = 1.41$; that is, the regular sample size must be inflated by 41% to adjust for the couples being the randomization unit. If there is perfect control group concordance, $p^* = 1$ and $\kappa = 1$, in which case, $N^* = 2N$.

Cornfield proposed another adjustment procedure [130]. Consider a trial where C clusters will be randomized, each cluster of size m_i ($i = 1, 2, \ldots, C$) and each having a different success rate of p_i ($i = 1, 2, \ldots, C$). Define the average cluster size $\overline{m} = \Sigma m_i / C$ and $\overline{p} = \Sigma m_i p_i / \Sigma m_i$ as the overall success rate weighted by cluster size. The variance of the overall success rate is $\sigma_p^2 = \Sigma m_i (p_i - \overline{p})^2 / C\overline{m}^2$. In this setting, the efficiency of simple randomization to cluster randomization is $E = \overline{p}(1 - \overline{p})^2 \overline{m} \sigma_p^2$. The inflation factor (IF) for this design is $\mathrm{IF} = 1/E = \overline{m}\sigma_p^2 / (1 - \overline{p})$. Note that if the response rate varies across clusters, the normal sample size must be increased.

While cluster randomization may be logistically required, the process of making the cluster the randomization unit has serious sample size implications. It would be unwise to ignore this consequence in the design phase. As shown, the sample size adjustments can easily be factors of 1.5 or higher. For clusters which are schools or cities, the intra-class correlation is likely to be quite small. However, the cluster size

is multiplied by the intra-class correlation so the impact might still be nontrivial. Not making this adjustment would substantially reduce the study power if the analyses were done properly, taking into account the cluster effect. Ignoring the cluster effect in the analysis would be viewed critically in most cases and is not recommended.

Multiple Response Variables

We have stressed the advantages of having a single primary question and a single primary response variable, but clinical trials occasionally have more than one of each. More than one question may be asked because investigators cannot agree about which outcome is most important. As an example, one clinical trial involving two schedules of oxygen administration to participants with chronic obstructive pulmonary disease had three major questions in addition to comparing the mortality rate [122]. Measures of pulmonary function, neuro-psychological status, and quality of life were evaluated. For the participants, all three were important.

Sometimes more than one primary response variable is used to assess a single primary question. This may reflect uncertainty as to how the investigator can answer the question. A clinical trial involving participants with pulmonary embolism [134] employed three methods of determining a drug's ability to resolve emboli. They were: lung scanning, arteriography, and hemodynamic studies. Another trial involved the use of drugs to limit myocardial infarct size [135]. Precordial electrocardiogram mapping, radionuclide studies, and enzyme levels were all used to evaluate the effectiveness of the drugs. Several approaches to the design and analysis of trials with multiple endpoints have been described [136–139].

Computing a sample size for such clinical trials is not easy. One could attempt to define a single model for the multidimensional response and use one of the previously discussed formulas. Such a method would require several assumptions about the model and its parameters and might require information about correlations between different measurements. Such information is rarely available. A more reasonable procedure would be to compute sample sizes for each individual response variable. If the results give about the same sample size for all variables, then the issue is resolved. However, more commonly, a range of sample sizes will be obtained. The most conservative strategy would be to use the largest sample size computed. The other response variables would then have even greater power to detect the hoped-for reductions or differences (since they required smaller sample sizes). Unfortunately, this approach is the most expensive and difficult to undertake. Of course, one could also choose the smallest sample size of those computed. That would probably not be desirable, because the other response variables would have less power than usually required, or only larger differences than expected would be detectable. It is possible to select a middle range sample size, but there is no assurance that this will be appropriate. An alternative approach is to look at the difference between the largest and smallest sample sizes. If this difference is very

large, the assumptions that went into the calculations should be re-examined and an effort should be made to resolve the difference.

As is discussed in Chap. 18, when multiple comparisons are made, the chance of finding a significant difference in one of the comparisons (when, in fact, no real differences exist between the groups) is greater than the stated significance level. In order to maintain an appropriate significance level α for the entire study, the significance level required for each test to reject H_0 should be adjusted [41]. The significance level required for rejection (α') in a single test can be approximated by α/k where k is the number of multiple response variables. For several response variables this can make α' fairly small (e.g., $k = 5$ implies $\alpha' = 0.01$ for each of k response variables with an overall $\alpha = 0.05$). If the correlation between response variables is known, then the adjustment can be made more precisely [140, 141]. In all cases, the sample size would be much larger than if the use of multiple response variables were ignored, so that most studies have not strictly adhered to this solution of modifying the significance level. Some investigators, however, have attempted to be conservative in the analysis of results [142]. There is a reasonable limit as to how much α' can be decreased in order to give protection against false rejection of the null hypothesis. Some investigators have chosen $\alpha' = 0.01$ regardless of the number of tests. In the end, there are no easy solutions. A somewhat conservative value of α' needs to be set and the investigators need to be aware of the multiple testing problem during the analysis.

Estimating Sample Size Parameters

As shown in the methods presented, sample size estimation is quite dependent upon assumptions made about variability of the response, level of response in the control group, and the difference anticipated or judged to be clinically relevant [16, 143–148]. Obtaining reliable estimates of variability or levels of response can be challenging since the information is often based on very small studies or studies not exactly relevant to the trial being designed. Applying Bayesian methods to incorporate explicitly uncertainty in these estimated parameters has been attempted [149]. Sometimes, pilot or feasibility studies may be conducted to obtain these data. In such cases, the term external pilot has been used [148].

In some cases, the information may not exist prior to starting the trial, as was the case for early trials in AIDS; that is, no incidence rates were available in an evolving epidemic. Even in cases where data are available, other factors affect the variability or level of response observed in a trial. Typically, the variability observed in the planned trial is larger than expected or the level of response is lower than assumed. Numerous examples of this experience exist [143]. One is provided by the Physicians' Health Study [150]. In this trial, 22,000 U.S. male physicians were randomized into a 2×2 factorial design. One factor was aspirin versus placebo in reducing cardiovascular mortality. The other factor was beta-carotene versus placebo for reducing cancer incidence. The aspirin portion of the trial was

terminated early in part due to a substantially lower mortality rate than expected. In the design, the cardiovascular mortality rate was assumed to be approximately 50% of the U.S. age-adjusted rate in men. However, after 5 years of follow-up, the rate was approximately 10% of the U.S. rate in men. This substantial difference reduced the power of the trial dramatically. In order to compensate for the extremely low event rate, the trial would have had to be extended another 10 years to get the necessary number of events [150]. One can only speculate about reasons for low event rates, but screening of potential participants prior to the entry almost certainly played a part. That is, screenees had to complete a run-in period and be able to tolerate aspirin. Those at risk for other competing events were also excluded. This type of effect is referred to as a screening effect. Physicians who began to develop cardiovascular signs may have obtained care earlier than non-physicians. In general, volunteers for trials tend to be healthier than the general population, a phenomenon often referred to as the healthy volunteer effect.

Another approach to obtaining estimates for ultimate sample size determination is to design so-called internal pilot studies [148]. In this approach, a small study is initiated based the best available information. A general sample target for the full study may be proposed, but the goal of the pilot is to refine that sample size estimate based on screening and healthy volunteer effects. The pilot study uses a protocol very close if not identical to the protocol for the full study, and thus parameter estimates will reflect those effects. If the protocol for the pilot and the main study are essentially identical, then the small pilot can become an internal pilot. That is, the data from the internal pilot become part of the data for the overall study. This approach was used successfully in the Diabetes Control and Complications Trial [151]. If data from the internal pilot are used only to refine estimates of variability or control group response rates, and not changes in treatment effect, then the impact of this two-step approach on the significance level is negligible. However, the benefit is that this design will more likely have the desired power than if data from external pilots and other sources are relied on exclusively [147]. It must be emphasized that pilot studies, either external or internal, should not be viewed as providing reliable estimates of the intervention effect [152]. Because power is too small in pilot studies to be sure that no effect exists, small or no differences may erroneously be viewed as reason not to pursue the question. A positive trend may also be viewed as evidence that a large study is not necessary, or that clinical equipoise no longer exists.

Our experience indicates that both external and internal pilot studies are quite helpful. Internal pilot studies should be used if at all possible in prevention trials, when screening and healthy volunteer effects seem to cause major design problems. Design modifications based on an internal pilot are more prudent than allowing an inadequate sample size to create yield misleading results.

One approach is to specify the number of events needed for a desired power level. Obtaining the specified number of events requires a number of individuals followed for a period of time. How many participants and how long a follow-up period can be adjusted during the early part of the trial, or during an internal pilot study, but the target number of events does not change. This is also discussed in more detail in Chaps. 16 and 17.

Another approach is to use adaptive designs which modify the sample size based on an emerging trend, referred to as trend adaptive designs (see Chaps. 5 and 17). Here the sample size may be adjusted for an updated estimate of the treatment effect, δ, using the methods described in this chapter. However, an adjustment must then be made at the analysis stage which may require a substantially larger critical value than the standard one in order to maintain a prespecified α level.

References

1. Freiman JA, Chalmers TC, Smith H, Jr., Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. *N Engl J Med* 1978;299:690–694.
2. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* 1994;272:122–124.
3. Chan AW, Altman DG. Epidemiology and reporting of randomised trials published in PubMed journals. *Lancet* 2005;365:1159–1162.
4. Halpern SD, Karlawish JH, Berlin JA. The continuing unethical conduct of underpowered clinical trials. *JAMA* 2002;288:358–362.
5. Altman DG. Statistics and ethics in medical research: III How large a sample? *Br Med J* 1980;281:1336–1338.
6. Brown BW. Statistical Controversies in the Design of Clinical-Trials - Some Personal Views. *Control Clin Trials* 1980;1:13–27.
7. Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. *BMJ* 1995;311:1145–1148.
8. Day SJ, Graham DF. Sample size estimation for comparing two or more treatment groups in clinical trials. *Stat Med* 1991;10:33–43.
9. Donner A. Approaches to sample size estimation in the design of clinical trials—a review. Stat Med 1984;3:199–214.
10. Gore SM. Statistics in question. Assessing clinical trials—trial size. *Br Med J* (Clin Res Ed) 1981;282:1687–1689.
11. Lachin JM. Introduction to sample size determination and power analysis for clinical trials. *Control Clin Trials* 1981;2:93–113.
12. Phillips AN, Pocock SJ. Sample size requirements for prospective studies, with examples for coronary heart disease. *J Clin Epidemiol* 1989;42:639–648.
13. Schlesselman JJ. Planning a longitudinal study. I. Sample size determination. *J Chronic Dis* 1973;26:553–560.
14. Schouten HJ. Planning group sizes in clinical trials with a continuous outcome and repeated measures. *Stat Med* 1999;18:255–264.
15. Streiner DL. Sample size and power in psychiatric research. *Can J Psychiatry* 1990;35:616–620.
16. Whitehead J. Sample sizes for phase II and phase III clinical trials: an integrated approach. *Stat Med* 1986;5:459–464.
17. Chow SC, Shao J, Wang H. Sample size calculations in clinical research, ed 2nd ed. Boca Raton, Taylor & Francis, 2008.
18. Desu MM, Raghavarao D. Sample Size Methodology. Boston, Academic Press, 1990.
19. Julious SA. Sample Sizes for Clinical Trials. Chapman and Hall, 2009.
20. Machin D, Campbell MJ, Tan S-B, Tan S-H. Sample Size Tables for Clinical Studies, ed 3rd. Wiley-Blackwell, 2008.
21. Odeh RE, Fox M. Sample Size Choice: Charts for Experiments with Linear Models, ed 2nd. New York, Marcel Dekker, 1991.

22. Braunholtz DA, Edwards SJ, Lilford RJ. Are randomized clinical trials good for us (in the short term)? Evidence for a "trial effect". *J Clin Epidemiol* 2001;54:217–224.
23. Paynter NP, Sharrett AR, Louis TA, et al. Paired comparison of observed and expected coronary heart disease rates over 12 years from the Atherosclerosis Risk in Communities Study. *Ann Epidemiol* 2010;20:683–690.
24. Brancati FL, Evans M, Furberg CD, et al. Midcourse correction to a clinical trial when the event rate is underestimated: the Look AHEAD (Action for Health in Diabetes) Study. *Clin Trials* 2012;9:113–124.
25. McClure LA, Szychowski JM, Benavente O, Coffey CS. Sample size re-estimation in an on-going NIH-sponsored clinical trial: the secondary prevention of small subcortical strokes experience. *Contemp Clin Trials* 2012;33:1088–1093.
26. Wittes J. On changing a long-term clinical trial midstream. *Statist Med* 10-15-2002;21:2789–2795.
27. Armitage P, Berry G, Mathews J. Statistical Methods in Medical Research, ed 4th. Malden MA, Blackwell Publishing, 2002.
28. Brown BW, Hollander M. Statistics - A Biomedical Introduction. New York, John Wiley and Sons, 1977.
29. Dixon WJ, Massey FJ Jr. Introduction to Statistical Analysis, ed 3rd. New York, McGraw-Hill, 1969.
30. Fisher L, Van Belle G. Biostatistics - A Methodology for the Health Sciences. New York, John Wiley and Sons, 1993.
31. Fisher L, Van Belle G, Heagerty PL, Lumley TS. Biostatistics - A Methodology for the Health Sciences. New York, John Wiley and Sons, 2004.
32. Fleiss JL, Levin B, Paik MC. Statistical Methods for Rates and Proportions, ed 3rd. John Wiley & Sons, Inc., 2003.
33. Remington RD, Schork MA. Statistics With Applications to the Biological and Health Sciences. Englewood Cliffs, Prentice-Hall, 1970.
34. Rosner B. Fundamentals of Biostatistics, ed 3rd. Boston, PWS-Kent, 1990.
35. Schork MA, Remington RD. Statistics With Applications to the Biological and Health Sciences, ed 3rd. Englewood Cliffs, Prentice-Hall, 2000.
36. Snedecor GW, Cochran WG. Statistical Methods, ed 8th. Ames, Iowa State University Press, 1989.
37. Woolson RF, Clarke WR. Statistical Methods for the Analysis of Biomedical Data. Wiley, 2011.
38. Canner PL, Klimt CR. The Coronary Drug Project. Experimental design features. *Control Clin Trials* 1983;4:313–332.
39. Davis BR, Cutler JA, Gordon DJ, et al. Rationale and design for the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). ALLHAT Research Group. *Am J Hypertens* 1996;9:342–360.
40. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. *J Am Statist Assoc* 1955;50:1096–1121.
41. Costigan T. Bonferroni inequalities and intervals; in Armitage P, Colton T (eds): Encyclopedia of Biostatistics. John Wiley and Sons, 2007.
42. Brittain E, Schlesselman JJ. Optimal Allocation for the Comparison of Proportions. *Biometrics* 1982;38:1003–1009.
43. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *N Engl J Med* 1989;321:406–412.
44. Rothman KJ. A show of confidence. *N Engl J Med* 1978;299:1362–1363.
45. Brown BW, Hollander M. Statistics: A Biomedical Introduction. Wiley, 2009.
46. McHugh RB, Le CT. Confidence estimation and the size of a clinical trial. *Control Clin Trials* 1984;5:157–163.
47. Armitage P, McPherson CK, Rowe BC. Repeated Significance Tests on Accumulating Data. *J R Stat Soc Ser A* 1969;132:235–244.

48. Bristol DR. Sample sizes for constructing confidence intervals and testing hypotheses. *Statist Med* 1989;8:803–811.
49. Casagrande JT, Pike MC, Smith PG. An Improved Approximate Formula for Calculating Sample Sizes for Comparing Two Binomial Distributions. *Biometrics* 1978;34:483–486.
50. Day SJ. Optimal placebo response rates for comparing two binomial proportions. *Statist Med* 1988;7:1187–1194.
51. Fleiss JL, Tytun A, Ury HK. A Simple Approximation for Calculating Sample Sizes for Comparing Independent Proportions. *Biometrics* 1980;36:343–346.
52. Fu YX, Arnold J. A Table of Exact Sample Sizes for Use with Fisher's Exact Test for 2 × 2 Tables. *Biometrics* 1992;48:1103–1112.
53. Gail MH, Gart JJ. The Determination of Sample Sizes for Use with the Exact Conditional Test in 2 x 2 Comparative Trials. *Biometrics* 1973;29:441–448.
54. Gail MH. The determination of sample sizes for trials involving several independent 2 × 2 tables. *J Chronic Dis* 1973;26:669–673.
55. Haseman JK. Exact Sample Sizes for Use with the Fisher-Irwin Test for 2 × 2 Tables. *Biometrics* 1978;34:106–109.
56. Lachenbruch PA. A note on sample size computation for testing interactions. *Statist Med* 1988;7:467–469.
57. McMahon RP, Proschan M, Geller NL, et al. Sample size calculation for clinical trials in which entry criteria and outcomes are counts of events. *Statist Med* 1994;13:859–870.
58. Ury HK, Fleiss JL. On Approximate Sample Sizes for Comparing Two Independent Proportions with the Use of Yates' Correction. *Biometrics* 1980;36:347–351.
59. Wacholder S, Weinberg CR. Paired versus Two-Sample Design for a Clinical Trial of Treatments with Dichotomous Outcome: Power Considerations. *Biometrics* 1982;38:801–812.
60. Connor RJ. Sample Size for Testing Differences in Proportions for the Paired-Sample Design. *Biometrics* 1987;43:207–211.
61. Donner A. Statistical Methods in Ophthalmology: An Adjusted Chi-Square Approach. *Biometrics* 1989;45:605–611.
62. Gauderman W, Barlow WE. Sample size calculations for ophthalmologic studies. Arch Ophthalmol 1992;110:690–692.
63. Rosner B. Statistical Methods in Ophthalmology: An Adjustment for the Intraclass Correlation between Eyes. *Biometrics* 1982;38:105–114.
64. Rosner B, Milton RC. Significance Testing for Correlated Binary Outcome Data. *Biometrics* 1988;44:505–512.
65. Hayes RL, Moulton LH. Cluster Randomised Trials. Chapman and Hall, 2009.
66. Cytel: SiZ. Cambridge, MA, Cytel Software Corporation, 2011.
67. Elashoff JD. nQuery Advisor Version 7.0 User's Guide. Cork, Ireland, Statistical Solutions, 2007.
68. Pezzullo JC. Web Pages that Perform Statistical Calculations. Computer Program 2014.
69. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria, R Foundation for Statistical Computing, 2013.
70. SAS Institute: Getting Started with the SAS Power and Sample Size Application. Cary, NC, SAS Institute Inc., 2004.
71. Shiboski S. Power and Sample Size Programs. Department of Epidemiology and Biostatistics, University of California San Francisco. Computer Program 2006.
72. StataCorp: Stata: Release 13. College Station, TX, StataCorp, 2013.
73. TIBCO Software I: SPLUS. TIBCO Softward Inc., 2008.
74. Feigl P. A Graphical Aid for Determining Sample Size when Comparing Two Independent Proportions. *Biometrics* 1978;34:111–122.
75. Aspirin Myocardial Infarction Study Research Group. A randomized, controlled trial of aspirin in persons recovered from myocardial infarction. *JAMA* 1980;243:661–669.
76. Beta Blocker Heart Attack Trial Research Group. A randomized trial of propranolol in patients with acute myocardial infarction: I. mortality results. *JAMA* 1982;247:1707–1714.

77. CASS Principle Investigators and Their Associates. Coronary artery surgery study (CASS): a randomized trial of coronary artery bypass surgery. Survival data. *Circulation* 1983;68:939–950.
78. The Coronary Drug Project Research Group. The Coronary Drug Project: Design, Methods, and Baseline Results. *Circulation* 1973;47:I–1.
79. Hypertension Detection and Follow-up Program Cooperative Group. Five-year findings of the hypertension detection and follow-up program: I. reduction in mortality of persons with high blood pressure, including mild hypertension. *JAMA* 1979;242:2562–2571.
80. The Intermittent Positive Pressure Breathing Trial Group. Intermittent Positive Pressure Breathing Therapy of Chronic Obstructive Pulmonary Disease A Clinical Trial. *Ann Intern Med* 1983;99:612–620.
81. Multiple Risk Factor Intervention Trial Research Group. Multiple risk factor intervention trial: Risk factor changes and mortality results. *JAMA* 1982;248:1465–1477.
82. Packer M, Carver JR, Rodeheffer RJ, et al. Effect of Oral Milrinone on Mortality in Severe Chronic Heart Failure. *N Engl J Med* 1991;325:1468–1475.
83. Barlow W, Azen S. The effect of therapeutic treatment crossovers on the power of clinical trials. *Control Clin Trials* 1990;11:314–326.
84. Halperin M, Rogot E, Gurian J, Ederer F. Sample sizes for medical trials with special reference to long-term therapy. *J Chronic Dis* 1968;21:13–24.
85. Lakatos E. Sample size determination in clinical trials with time-dependent rates of losses and noncompliance. *Control Clin Trials* 1986;7:189–199.
86. Lavori P. Statistical Issues: Sample Size and Dropout; in Benkert O, Maier W, Rickels K (eds): Methodology of the Evaluation of Psychotropic Drugs. Springer Berlin Heidelberg, 1990, pp 91–104.
87. Newcombe RG. Explanatory and pragmatic estimates of the treatment effect when deviations from allocated treatment occur. *Statist Med* 1988;7:1179–1186.
88. Schork MA, Remington RD. The determination of sample size in treatment-control comparisons for chronic disease studies in which drop-out or non-adherence is a problem. *J Chronic Dis* 1967;20:233–239.
89. Wu MC, Fisher M, DeMets D. Sample sizes for long-term medical trial with time-dependent dropout and event rates. *Control Clin Trials* 1980;1:111–124.
90. Pentico DW. On the Determination and Use of Optimal Sample Sizes for Estimating the Difference in Means. *Am Stat* 1981;35:40–42.
91. Dawson JD, Lagakos SW. Size and Power of Two-Sample Tests of Repeated Measures Data. *Biometrics* 1993;49:1022–1032.
92. Kirby AJ, Galai N, Munoz A. Sample size estimation using repeated measurements on biomarkers as outcomes. *Control Clin Trials* 1994;15:165–172.
93. Laird NM, Wang F. Estimating rates of change in randomized clinical trials. *Control Clin Trials* 1990;11:405–419.
94. Lipsitz SR, Fitzmaurice GM. Sample size for repeated measures studies with binary responses. *Statist Med* 1994;13:1233–1239.
95. Nam J. A Simple Approximation for Calculating Sample Sizes for Detecting Linear Trend in Proportions. *Biometrics* 1987;43:701–705.
96. Overall JE, Doyle SR. Estimating sample sizes for repeated measurement designs. *Control Clin Trials* 1994;15:100–123.
97. Rochon J. Sample Size Calculations for Two-Group Repeated-Measures Experiments. *Biometrics* 1991;47:1383–1398.
98. Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis, ed 2nd. John Wiley & Sons, 2011.
99. Cantor AB. Power estimation for rank tests using censored data: Conditional and unconditional. *Control Clin Trials* 1991;12:462–473.
100. Emrich LJ. Required Duration and Power Determinations for Historically Controlled-Studies of Survival Times. *Statist Med* 1989;8:153–160.

101. Freedman LS. Tables of the number of patients required in clinical trials using the logrank test. *Statist Med* 1982;1:121–129.
102. Gail MH. Applicability of sample size calculations based on a comparison of proportions for use with the logrank test. *Control Clin Trials* 1985;6:112–119.
103. George SL, Desu MM. Planning the size and duration of a clinical trial studying the time to some critical event. *J Chronic Dis* 1974;27:15–24.
104. Halperin M, Johnson NJ. Design and Sensitivity Evaluation of Follow-Up Studies for Risk Factor Assessment. *Biometrics* 1981;37:805–810.
105. Hsieh FY. Sample size tables for logistic regression. *Statist Med* 1989;8:795–802.
106. Lachin JM, Foulkes MA. 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;42:507–519.
107. Lachin JM. Biostatistical Methods: The Assessment of Relative Risks, ed 2nd. John Wiley & Sons, Inc., 2010.
108. Lakatos E. Sample Sizes Based on the Log-Rank Statistic in Complex Clinical Trials. *Biometrics* 1988;44:229–241.
109. Lui KJ. Sample size determination under an exponential model in the presence of a confounder and type I censoring. *Control Clin Trials* 1992;13:446–458.
110. Morgan TM. Nonparametric Estimation of Duration of Accrual and Total Study Length for Clinical Trials. *Biometrics* 1987;43:903–912.
111. Palta M, Amini SB. Consideration of covariates and stratification in sample size determination for survival time studies. *J Chronic Dis* 1985;38:801–809.
112. Pasternack BS, Gilbert HS. Planning the duration of long-term survival time studies designed for accrual by cohorts. *J Chronic Dis* 1971;24:681–700.
113. Rubinstein LV, Gail MH, Santner TJ. Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *J Chronic Dis* 1981;34:469–479.
114. Schoenfeld DA, Richter JR. Nomograms for Calculating the Number of Patients Needed for a Clinical Trial with Survival as an Endpoint. *Biometrics* 1982;38:163–170.
115. Schoenfeld DA. Sample-Size Formula for the Proportional-Hazards Regression Model. *Biometrics* 1983;39:499–503.
116. Taulbee JD, Symons MJ. Sample Size and Duration for Cohort Studies of Survival Time with Covariables. *Biometrics* 1983;39:351–360.
117. 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.
118. Zhen B, Murphy JR. Sample size determination for an exponential survival model with an unrestricted covariate. *Statist Med* 1994;13:391–397.
119. Pasternack BS. Sample sizes for clinical trials designed for patient accrual by cohorts. *J Chronic Dis* 1972;25:673–681.
120. Hjalmarson A, Herlitz J, Malek I, et al. Effect on Mortality of Metoprolol in Acute Myocardial Infarction: A Double-blind Randomised Trial. *The Lancet* 1981;318:823–827.
121. The Norwegian Multicenter Study Group. Timolol-Induced Reduction in Mortality and Reinfarction in Patients Surviving Acute Myocardial Infarction. *N Engl J Med* 1981;304:801–807.
122. Nocturnal Oxygen Therapy Trial Group. Continuous or Nocturnal Oxygen Therapy in Hypoxemic Chronic Obstructive Lung Disease A Clinical Trial. *Ann Intern Med* 1980;93:391–398.
123. Ingle JN, Ahmann DL, Green SJ, et al. Randomized Clinical Trial of Diethylstilbestrol versus Tamoxifen in Postmenopausal Women with Advanced Breast Cancer. *N Engl J Med* 1981;304:16–21.
124. Spriet A, Beiler D. When can 'non significantly different' treatments be considered as 'equivalent'? Br J Clin Pharmacol 1979;7:623–624.
125. Blackwelder WC. "Proving the null hypothesis" in clinical trials. *Control Clin Trials* 1982;3:345–353.
126. Blackwelder WC, Chang MA. Sample size graphs for "proving the null hypothesis". *Control Clin Trials* 1984;5:97–105.

127. Makuch R, Simon R. Sample size requirements for evaluating a conservative therapy. *Cancer Treat Rep* 1978;62:1037–1040.
128. Rothmann MD, Wiens BL, Chan ISF. Design and Analysis of Non-Inferiority Trials. Taylor & Francis, 2011.
129. Donner A, Birkett N, Buck C. Randomization by cluster. Sample size requirements and analysis. *Am J Epidemiol* 1981;114:906–914.
130. Cornfield J. Randomization by group: A formal analysis. *Am J Epidemiol* 1978;108:100–102.
131. Hsieh FY. Sample size formulae for intervention studies with the cluster as unit of randomization. *Statist Med* 1988;7:1195–1201.
132. Lee EW, Dubin N. Estimation and sample size considerations for clustered binary responses. *Statist Med* 1994;13:1241–1252.
133. Murray DM. Design and Analysis of Group-randomized Trials. Oxford University Press, 1998.
134. Urokinase Pulmonary Embolism Trial Study Group: Urokinase-streptokinase embolism trial: Phase 2 results. *JAMA* 1974;229:1606–1613.
135. Roberts R, Croft C, Gold HK, et al. Effect of Propranolol on Myocardial-Infarct Size in a Randomized Blinded Multicenter Trial. *N Engl J Med* 1984;311:218–225.
136. Follmann D. A Simple Multivariate Test for One-Sided Alternatives. *J Am Stat Assoc* 1996;91:854–861.
137. O'Brien PC. Procedures for Comparing Samples with Multiple Endpoints. *Biometrics* 1984;40:1079–1087.
138. Tang DI, Gnecco C, Geller NL. Design of Group Sequential Clinical Trials with Multiple Endpoints. *J Am Stat Assoc* 1989;84:776–779.
139. Tang DI, Geller NL, Pocock SJ. On The Design and Analysis of Randomized Clinical Trials with Multiple Endpoints. *Biometrics* 1993;49:23–30.
140. Hsu J. Multiple Comparisons: Theory and Methods. Taylor & Francis, 1996.
141. Miller RG. Simultaneous Statistical Inference, ed 2nd. Springer New York, 2011.
142. The Coronary Drug Project Research Group. Clofibrate and niacin in coronary heart disease. *JAMA* 1975;231:360–381.
143. Church TR, Ederer F, Mandel JS, et al. Estimating the Duration of Ongoing Prevention Trials. *Am J Epidemiol* 1993;137:797–810.
144. Ederer F, Church TR, Mandel JS. Sample Sizes for Prevention Trials Have Been Too Small. *Am J Epidemiol* 1993;137:787–796.
145. Neaton JD, Bartsch GE. Impact of measurement error and temporal variability on the estimation of event probabilities for risk factor intervention trials. *Statist Med* 1992;11:1719–1729.
146. Patterson BH. The impact of screening and eliminating preexisting cases on sample size requirements for cancer prevention trials. *Control Clin Trials* 1987;8:87–95.
147. Shih WJ. Sample size reestimation in clinical trials; in Peace KE (ed): Biopharmaceutical Sequential Statistical Applications. Taylor & Francis, 1992, pp 285–301.
148. Wittes J, Brittain E. The role of internal pilot studies in increasing the efficiency of clinical trials. *Statist Med* 1990;9:65–72.
149. Ambrosius WT, Polonsky TS, Greenland P, et al. Design of the Value of Imaging in Enhancing the Wellness of Your Heart (VIEW) trial and the impact of uncertainty on power. *Clin Trials* 2012;9:232–246.
150. Steering Committee of the Physicians' Health Study. Final Report on the Aspirin Component of the Ongoing Physicians' Health Study. *N Engl J Med* 1989;321:129–135.
151. The Diabetes Control and Complications Trial Research Group: The Effect of Intensive Treatment of Diabetes on the Development and Progression of Long-Term Complications in Insulin-Dependent Diabetes Mellitus. *N Engl J Med* 1993;329:977–986.
152. Davis BR, Wittes J, Pressel S, et al. Statistical considerations in monitoring the Systolic Hypertension in the Elderly Program (SHEP). *Control Clin Trials* 1993;14:350–361.