NCBI Bookshelf. A service of the National Library of Medicine, National Institutes of Health.

Smith PG, Morrow RH, Ross DA, editors. Field Trials of Health Interventions: A Toolbox. 3rd edition. Oxford (UK): OUP Oxford; 2015 Jun 1.

Chapter 5 Trial size

1. Introduction to trial size

One of the most important factors to consider in the design of an intervention trial (or indeed in the design of any epidemiological study) is the choice of an appropriate trial size to answer the research question. Trials that are too small may fail to detect important effects of an intervention on the outcomes of interest or may estimate those effects too imprecisely. Trials that are larger than necessary are a waste of resources and may even lead to a loss in accuracy, as it [illegible] maintain data quality and high coverage rates in a large trial than in a smaller one.

The choice of an appropriate trial size may be based on either the precision of outcome measures desired or the power of the trial wanted. In Section 2, there is a discussion of the criteria used to make this choice. In Sections 3 and 4, procedures are given for calculating trial size requirements in the simplest case where two groups of equal size are to be compared. More complex designs are considered in Section 5. Special methods are necessary when the interventions are allocated to groups (for example, communities, schools, or health facilities), rather than individuals, and these are described in Section 6. Following this, in Section 7, two other factors that may influence the choice of trial size are discussed—first, the need to allow for interim analyses of the results (see Section 7.1), and second, the effects of losses to follow-up (see Section 7.2). In Section 8, the consequences of trials that are too small are discussed. Computer programs can be used to carry out sample size calculations, and these are briefly discussed in Section 9.

The procedures described in this chapter should be regarded as providing only a rough estimate of the required trial size, as they are often based on estimates of expected disease rates, subjective decisions about the size of effects that it would be important to detect, and the use of approximate formulae. However, a rough estimate of the necessary size of a trial is generally all that is needed for planning purposes. More comprehensive reviews of methods for the determination of trial size requirements are available (Chow et al., 2008; Machin, 2009), but the methods given in this chapter should be adequate for most purposes.

Readers who are not familiar with methods for the statistical analysis of trial data and, in particular, with the concepts of confidence intervals (CIs) and significance tests may find it helpful to read Chapter 21, Section 2, before embarking on this chapter, which is placed here because of the importance of considering trial size requirements at the design stage of a trial.

A principal objective of most intervention trials is to *estimate the effect* of the intervention on the outcome or outcomes of interest. Any such estimate is subject to error, and this error has two main components: bias and sampling error. Possible sources of *bias* and ways of avoiding them are discussed in Chapters 4, 11, and 21. The second component *sampling error* arises because the trial data come from only a *sample* of the population. This second component of error is the focus of this chapter. Sampling error is reduced when the trial size is increased, whereas bias generally is not.

2. Criteria for determining trial size

2.1. Precision of effect measures

To select the appropriate sample size, it is necessary to decide how much sampling error in the estimate of the effect of the intervention is acceptable and to select the sample size to achieve this precision. When the data are analysed, the amount of sampling error is represented by the width of the *confidence interval* around the estimate of effect. The narrower the CI, the greater the *precision* of the estimate, and the smaller the probable amount of sampling error. When designing a trial, it is necessary therefore to decide the width of an acceptable CI around the chosen intervention effect. Having made this decision, the method to select the required trial size is given in Section 3.

2.2. Power of the trial

An alternative approach is to choose a trial size which gives adequate *power* to detect an effect of a given magnitude. The focus is then on the result of the *significance test* which will be conducted at the end of the trial. The significance test assesses the evidence against the *null hypothesis*, which states that there is no true difference between the interventions under comparison. A *statistically significant* result indicates that the data conflict with the null hypothesis

and that there are grounds for rejecting the hypothesis that there is no difference in the effects of the interventions under study on the outcomes of interest.

Because of the variations resulting from sampling error, it is never possible to be certain of obtaining a significant result at the end of a trial, even if there is a real difference. It is necessary to consider the *probability* of obtaining a statistically significant result in a trial, and this probability is called the *power* of the trial. Thus, a power of 80% to detect a difference of a specified size means that, if the trial were to be conducted repeatedly, a statistically significant result would be obtained four times out of five (80%) if the true difference was really of the specified size. The power of a trial depends on the factors shown in Box 5.1.

The power also depends on whether a one-sided or two-sided significance test is to be performed (see Chapter 21, Section 2.3) and on the underlying variability of the data. How the power may be calculated for given values of these parameters is explained in Section 4.

When designing a trial, the objective is to ensure that the trial size is large enough to give high power *if the true effect of the intervention is large enough to be of public health importance*.

2.3. Choice of criterion

The choice of which criterion (precision or power) should be used in any particular trial depends on the objectives of the trial. If it is known unambiguously that the intervention has some effect (relative to the comparison (control) group), it makes little sense to test the null hypothesis; rather the objective may be to estimate the magnitude of the effect and to do this with some acceptable specified precision.

In trials of new interventions, it is often not known whether there will be any impact at all of the intervention on the outcomes of interest, and what is required is ‘proof of concept’. In these circumstances, it may be sufficient to ensure that there will be a good chance of obtaining a significant result if there is indeed an effect of some specified magnitude. It should be emphasized, however, that, if this course is adopted, the estimates obtained may be very imprecise. To illustrate this, suppose it is planned to compare two groups with respect to the mean of some variable, and suppose the true difference between the group means is D. If the trial size is chosen to give 90% power (of obtaining a significant difference with $p<0.05$ on a two-sided test) if the difference is D, the 95% CI on D is expected to extend roughly from 0.4 D to 1.6 D. This is a wide range and implies that the estimate of the effect of intervention will be imprecise. In many situations, it may be more appropriate to choose the sample size by setting the width of the CI, rather than to rely on power calculations.

2.4. Trials with multiple outcomes

The discussion in Sections 2.1 to 2.3 concerns factors influencing the choice of trial size, with respect to a particular outcome measure. In most trials, several different outcomes are measured. For example, in a trial of the impact of insecticide-treated mosquito-nets on childhood malaria, there may be interest in the effects of the intervention on deaths, deaths attributable to malaria, episodes of clinical malaria, spleen sizes at the end of the malaria season, PCVs at the end of the malaria season, and possibly other measures.

Chapter 12, Section 2 highlights the importance of defining in advance the *primary* outcome and a limited number of *secondary* outcomes of a trial. In order to decide on the trial size, the investigator should first focus attention on the primary outcome, as results for this outcome will be given the most weight when reporting the trial findings, and it is essential that the trial is able to provide adequate results for this outcome. The methods of this chapter can then be used to calculate the required trial size for the primary outcome and each of the secondary outcomes.

Ideally, the outcome that results in the largest trial size would be used to determine the size, as then, for other outcomes, it would be known that better than the required precision or power would be achieved. It is often found, however, that one or more of the outcomes would require a trial too large for the resources that are likely to be available. For example, detecting changes in mortality, or cause-specific mortality, often requires very large trials. In these circumstances, it may be decided to design the trial to be able to detect an impact on morbidity and accept that it is unlikely to be able to generate conclusive findings about the effect on mortality. It is important to point out, however, that, if a trial shows that an intervention has an impact on morbidity, it may be regarded as unethical to undertake a further, larger trial to assess the impact on mortality. For this reason, it is generally advisable to ensure that trials are conducted at an early stage in which the outcome of greatest public health importance is the endpoint around which the trial is planned. This issue is discussed further in Chapter 6.

Sometimes, different trial sizes may be used for different outcomes. For example, it might be possible to design a trial in such a way that a large sample of participants are monitored for mortality, say by annual surveys, and only a proportion of participants are monitored for morbidity, say by weekly visits.

If it is not feasible to design the trial to achieve adequate power or precision for the primary outcome, the trial should either be abandoned or a different primary outcome should be adopted.

2.5. Practical constraints

In practice, statistical considerations are not the only factors that need to be taken into account in planning the size of an investigation. Resources, in terms of staff, vehicles, laboratory capacity, time, or money, may limit the potential size of a trial, and it is often necessary to compromise between the results of the trial size computations and what can be managed with the available resources. Trying to do a trial that is beyond the capacity of the available resources is likely to be unfruitful, as data quality is likely to suffer and the results may be subject to serious bias, or the trial may even collapse completely, wasting the effort and money that have already been expended. If calculations indicate that a trial of manageable size will yield power and/or precision that is unacceptably low, it is probably better not to conduct the trial at all.

A useful approach to examine the trade-off between trial size (and thus cost) and power is to construct *power curves* for one or two of the key outcome variables. Power curves show how power varies with trial size for different values of the effect measure. Figure 5.1 shows power curves for malaria deaths in the mosquito-net trial discussed in Section 2.4, assuming that equal numbers of children are to be allocated to the intervention and control groups and statistical significance is to be based on a two-sided test at the 5% level. *R* represents the rate ratio of malaria deaths in the intervention group, compared to the control group, so that $R = 0.3$ represents a reduction in the death rate of 70%. The assumptions used to construct these curves are described in Section 4. The curves indicate that, if 1000 children were followed for 1 year in each group (making 2000 children in all), there would be about a one in two chance of obtaining a significant result (power = 50%), even if the reduction in the death rate was as high as 70%. A trial five times as large as this would have a good chance (about 80%) of detecting a reduction in the death rate of 50% or more but would be inadequate (about 40%) to detect a 30% reduction in the death rate.

3. Size to give adequate precision

This section describes how the trial size is determined if the aim is to obtain an estimate of the outcome of an intervention with a specified level of precision. The simplest case to consider is where just two groups of about the same size are to be compared (for example, the outcome of an intervention compared with that of a control group, or the comparison of outcomes of two interventions). More complex designs are discussed in Section 5. The methodology varies according to the type of outcome measure; the comparison of proportions, incidence rates, and means are considered in Sections 3.1 to 3.3.

3.1. Comparison of proportions

In this section, outcomes are considered that are *binary* (yes or no) variables. This includes cumulative incidence or *risk*, for example, the proportion of children experiencing at least one episode of clinical malaria during the follow-up period. It also includes examination of the *prevalence* of some characteristic, for example, the presence of a palpable spleen in a survey conducted at the end of the trial.

Suppose the true proportions in groups 1 and 2 are p_1 and p_2, respectively, giving a risk ratio (relative risk) of $R = p_1/p_2$ The approximate 95% CI for *R* extends from R/f to Rf where, in this case, the factor f is given by:

$$\mathrm{f} = \exp\left\{1.96\sqrt{\left[(1-p_1)/\left(np_1\right) + (1-p_2)\,/\,(np_2)\right]}\right\}$$

where *n* is the number of children in each group, and f is commonly called the *error factor*.

The required value of f is chosen, and rough estimates are made of the values of p_2 and *R* to enable the number required in each group *n* to be calculated as:

$$n = (1.96/\log_e f)^2 + \{[(R+1)/(Rp_2)] - 2\}$$

where $\log_e f$ is the *natural logarithm* of *f*.

For example, in the mosquito-net trial, one of the outcomes of interest is the prevalence of splenomegaly (the proportion of children with enlarged spleens) at the end of the trial. Prior data from the trial area suggest that, in the control group, a prevalence of approximately 40% would be expected. Suppose the intervention is expected to roughly halve the prevalence, so that $R = 0.5$ and an estimate of *R* is wanted to within about ± 0.15 This suggests setting *f* to about 1.3 (because then the upper 95% confidence limit on *R* is $Rf = 0.5 \times 1.3 = 0.65$ which is 0.15 above $R(= 0.5)$), and thus $n = (1.96/\log_e 1.3)^2 \{[1.5/(0.5 \times 0.4)] - 2\} = 307$ so that around 300 children would need to be studied in each group.

3.2. Comparison of incidence rates

Suppose a comparison of two groups is required, with respect to the rate of occurrence of some defined event over the trial period. Suppose the true incidence rates are r_1 and r_2 in groups 1 and 2, respectively, where each rate represents the number of events per person-year of observation. The rate ratio *R* (sometimes called incorrectly the relative risk, instead of the relative rate) of the incidence rate in group 1, compared to the incidence rate in group 2, is given by $R = r_1/r_2$ (see Chapter 21, Section 5 for methods of analysis for the comparison of rates). If the total follow-up time for those in each group is *y* years (for example, *y* persons are each followed for 1 year, or $y/2$ are each followed for 2 years), each group is said to experience *y* person-years of observation. The expected numbers of events in the two groups will be $e_1 = yr_1$ and $e_2 = yr_2$ respectively. When the results are analysed, the approximate 95% CI for *R* is expected to extend from R/f to Rf where:

$$f = \exp\{1.96\sqrt{[(1/e_1) + (1/e_2)]}\}.$$

To decide on the necessary size of the trial, make a rough estimate of the likely value of *R*, select the precision that is required by specifying a value for *f*, the error factor, and calculate:

$$e_2 = (1.96/\log_e f)^2 [(R+1)/R].$$

The trial size is then fixed so that the expected number of events in group 2 during the trial period is equal to the calculated value e_2 The expected number of events in group 1 will be Re_2

It should be noted that these methods are only appropriate in the situation where each individual can experience only one event during the trial period or where the number of individuals experiencing multiple events is very small. If most individuals experience at least one event and many experience two or more, it is preferable to define a *quantitative* outcome for each individual, representing the number of events experienced during the trial period, and to use the methods described in Section 3.3.

Example: in the mosquito-net trial, suppose the trial groups are to consist of children aged 0–4 years and that the death rate associated with malaria in the trial area for that age group is estimated to be roughly 10 per 1000 child-years. If group 1 is the intervention group (treated bed-nets) and group 2 is the control group (no protection), *R* represents the ratio of the intervention and control death rates. Suppose *R* is expected to be about 0.4, corresponding to a reduction in the death rate of 60%. Suppose also that *f* is selected to be equal to 1.25, so that the 95% CI for *R* is expected to extend from $(0.4/1.25 = 0.32)$ to $(0.4 \times 1.25 = 0.50)$ In other words, it is desired to estimate the protective efficacy to within about 10% of the true value (i.e. 50–70% around the estimated efficacy of 60%). Then:

$$e_2 = [1.96/\log_e(1.25)]^2 (1.4/0.4) = 270.$$

To expect 270 deaths in the control group, it would be necessary to observe an estimated 27 000 child-years $[= 270/(10/1000)]$. This could be achieved by following 54 000 children for 6 months, or 27 000 children for 1 year, or 13 500 for 2 years, and so on, assuming an expected death rate of ten per 1000 child-years in each of these scenarios. The magnitude of the required trial size (27 000 child-years of observation *in each group*) illustrates that, when rare events are being studied, very large samples are needed to obtain a precise estimate of the impact of an intervention.

3.3. Comparison of means

Quantitative outcomes may be analysed by comparing the means of the relevant variable in the intervention and control groups. This could be the mean of the values recorded at a cross-sectional survey, for example, the mean weight of children in the trial at the end of the trial. Alternatively, it could be the mean of the changes recorded between baseline and follow-up surveys, for example, the mean change in weight (or weight velocity, i.e. the change in weight divided by the time between the two measurements) among the children in the trial.

Suppose the true means in groups 1 and 2 are μ_1 and μ_2. These would generally be compared in terms of the difference in the means, D=$\mu_1 - \mu_2$. The 95% CI for D is given by $D \pm$ f, where:

$$f = 1.96\sqrt{\left[\left(\sigma_1^2 + \sigma_2^2/n\right)\right]}$$

where σ_1 and σ_2 are the standard deviations of the outcome variable in the two groups.

An acceptable value of f is chosen; values of σ_1 and σ_2 are selected, and the required number in each group is calculated as:

$$n = (1.96/f)^2 \left(\sigma_1^2 + \sigma_2^2\right).$$

An estimate of the standard deviation of the outcome variable is often available from other studies. It is usually reasonable to assume that the standard deviation will be roughly similar in the two trial groups. If no other estimate is available, a rough approximation can be obtained by taking one-quarter of the likely range of the variable.

Example: In the mosquito-net trial, another outcome of interest is the PCV, or haematocrit, measured in blood samples taken from the children at the end of the trial. From previous data, the mean PCV in the control group is expected to be about 33.0, with a standard deviation of about 5.0 (the normal range is about 33 ± 10, and it has been assumed that the normal range covers four standard deviations (i.e. ± 2). An increase in mean PCV in the intervention group of between 2.0 and 3.0 is expected, and it is required to estimate the difference D between the two groups to within about 0.5, so that $f = 0.5$. Assuming that the standard deviation is about 5.0 in both groups:

$$n = (1.96/0.5)^2 \left(5.0^2 + 5.0^2\right) = 768.$$

4. Size to give adequate power

The alternative approach to setting trial size is based upon selecting the trial size to achieve a specified *power*. In order to do this, the following must be specified:

1. What size of difference, *D,* between the two groups would be of clinical or public health importance? The trial size will be chosen so it would have a good chance of detecting this size of true difference, i.e. there would be a good chance of obtaining a statistically significant result, thus concluding that there is a real difference between the two trial arms. *D* is the *true* difference between the two groups, not the estimated difference as measured in the trial. Very small differences are generally of no public health importance, and it would not be of concern if they were not detected in the trial. The general principle, in most cases, is to choose *D* to be the *minimum difference* which would be of public health relevance and therefore be important to detect in a trial. Note that ‘detecting’ *D* means that a significant difference is obtained, indicating that there is some difference between the two groups. This does not

mean that the difference is estimated precisely. To ensure a precise estimate is obtained, the approach of Section 3 should be used.

2. Having specified *D*, the investigators must decide how confident they wish to be of obtaining a significant result if this were the true difference between the groups. In other words, the power is set for this value of *D*. Note that, if the true difference between the groups is actually larger than *D*, the power of the trial will be larger than the value set. The required power is specified in the calculations by choosing the corresponding value of z_2, as shown in Table 5.1. Commonly chosen values for the power are 80%, 90%, and 95%, the corresponding values of z_2 being 0.84, 1.28, and 1.64. It would generally be regarded as unsatisfactory to proceed with a trial with a power of less than 70% for the primary outcome, because that means that one would have a more than 30% chance of 'missing' a true difference of *D*.
3. The significance level must also be specified for the comparison of the two groups under study. This is entered into the calculations in terms of the parameterz_1 The commonest choice for the required p-value is 0.05, corresponding to az_1 of 1.96. Alternative values might be 0.01 or 0.001, corresponding toz_1 values of 2.58 or 3.29, respectively. It is assumed throughout this chapter that *two-sided* significance tests are to be used (see Chapter 21, Section 2.3). A significance level of 0.05 is assumed in the numerical examples, unless otherwise stated.
4. In addition, certain additional information must be specified, which varies according to the type of measure being examined. This may be a rough estimate of the rates or proportions that are expected, or an estimate of the standard deviation for a quantitative variable. Note that, if these quantities were known exactly, no trial would be needed! Only rough estimates are required.

Having specified these values, the formulae or tables given in Sections 4.1 to 4.3 can be used to calculate the required trial size.

It is often useful, however, to proceed in the opposite direction, i.e. to explore the power that would be achieved for a range of possible trial sizes and for a range of possible values of the true difference *D*. This enables the construction of *power curves*, as illustrated in Figure 5.1. Formulae for this approach are also given in Sections 4.1 to 4.3.

4.1. Comparison of proportions

The trial size required in each group to detect a specified difference $D = p_1 - p_2$, with power specified by z_2 and significance level specified by z_1, is given by:

$$n = \left[(z_1 + z_2)^2 2p(1-p)\right] / (p_1 - p_2)^2$$

where *p* is the average of p_1 and p_2.

For 90% power and significance at p<0.05, this simplifies to:

$$n = [21p(1-p)] / (p_1 - p_2)^2.$$

Table 5.2 shows the required trial size for a range of values of p_1 and p_2 for 80%, 90%, or 95% power.

To calculate the power of a trial of specified size, calculate as follows, and refer the value ofz_2 to Table 5.1.

$$z_2 = \left(\sqrt{\{n / [2p(1-p)]\}}\right)(|p_1 - p_2|) - z_1.$$

Example: assume that the spleen rate in the control group of the mosquito-net trial is around 40%. To have very high power (say 95%) of detecting a significant effect if the intervention reduces the spleen rate to 30% (so thatp=0.35 the number of children required in each group is given by:

$$n = \left[(1.96 + 1.64)^2 (2 \times 0.35 \times 0.65)\right] / (0.3 - 0.4)^2 = 590.$$

If the true risk ratio is *R* and we wish to power the trial, such that the lower confidence limit on the risk ratio will be greater than or equal to R_L where R_L is the lowest acceptable efficacy (say, for whether or not to implement the intervention in a public health system, i.e. we need to be sure that the efficacy is at least R_L), the required sample size is:

$$n = (z_1 + z_2)^2 [(1 - p_1)/(p_1) + (1 - p_2)/(p_1)]/[\log_e (R/R_L)]^2.$$

4.2. Comparison of incidence rates

For a specified difference $D = r_1 - r_2$ and values of z_1 and z_2 representing the required significance level and power, the required number of person-years in each group is given by:

$$y = \left[(z_1 + z_2)^2 (r_1 + r_2)\right] / (r_1 - r_2)^2$$

where r_1 and r_2 are the expected rates per person-year in the two groups.

A rough estimate of the average of the two rates is therefore required, i.e. $[(r_1 + r_2)/2]$ For 90% power and significance at p<0.05, this formula simplifies to:

$$y = [10.5 (r_1 + r_2)] / (r_1 - r_2)^2.$$

An alternative, but equivalent, formula gives the number of events required in group 2, the control group, in terms of the rate ratio *R*, for which the specified power is required:

$$e_2 = \left[(z_1 + z_2)^2 (1 + R)\right] / (1 - R)^2.$$

This formula was used to construct Table 5.3, which shows the number of events needed in group 2 to detect a rate ratio of *R* with 80%, 90%, or 95% power. The total number of events needed in both groups can be calculated as $e_2(1 + R)$. Since this can be computed without specifying the assumed rates in the two trial groups, this provides a particularly helpful approach when the rates are uncertain. Thus, in an *endpoint-driven trial*, we can specify the number of events that need to be observed to reach the required power, after which recruitment or follow-up may be terminated.

To calculate the power for a given trial size, compute:

$$z_2 = \{\sqrt{[n/(r_1 + r_2)]}\} (|r_1 - r_2|) - z_1$$

where $|r_1 - r_2|$ is the absolute value of the difference between the two rates.

Refer the resulting value of z_2 to Table 5.1 to determine the power of the trial.

Example: Assume, in the mosquito-net trial, that the death rate from malaria in the control group is 10/1000 child-years, so that $r_2 = 0.010$. Eighty per cent power is wanted to detect a significant effect if the true rate in children with bed-nets is reduced by 70% to $r_1 = 0.003$. The number of child-years of observation required in each group is given by:

$$y = \left[(1.96 + 0.84)^2 (0.003 + 0.010)\right] / (-0.007)^2 = 2080.$$

The power curves shown in Figure 5.1 were constructed using the same assumption concerning the death rate in controls. For example, with y = 2000y=2000 and a rate ratio of $R = 0.7$ (corresponding to a death rate of 7 per 1000 child-years in the intervention group), giving a power of 18% (Table 5.1):

$$z_2 = \{\sqrt{[2000 / (0.007 + 0.010)]}\} (|0.007 - 0.010|) - 1.96 = -0.93.$$

These formulae are used to ensure that there is a high probability of rejecting the null hypothesis if the true effect is of the assumed size. However, this may still mean that the lower confidence limit for the effect size is close to the null, and this may provide insufficient evidence to recommend widespread adoption of the intervention. A larger sample size will be needed to ensure that the lower confidence limit exceeds a given value.

Suppose the assumed value of the rate ratio is *R* and that we wish to power the trial so that there is a high probability that the CI excludes a value R_L corresponding to the lower limit of efficacy desired. Then the required sample size is given by the formula:

$$y = (z_1 + z_2)^2 (1/r_1 + 1/r_2) / [\log_e (R/R_L)]^2.$$

Example: In the mosquito-net trial, we found that 2080 child-years were required in each trial group to reject the null hypothesis with 80% power if the true rate ratio *R* was 0.3, corresponding to an efficacy of 70%. Now suppose we wish to ensure that there is an 80% chance that the lower 95% CI for the efficacy exceeds 30%, corresponding to $R_L = 0.7$. Applying the formula, we obtain the following, demonstrating the substantial increase in sample size that this would necessitate:

$$y = (1.96 + 0.84)^2 (1/0.010 + 1/0.003) / [\log_e (0.3/0.7)]^2 = 4732.$$

4.3. Comparison of means

The trial size required in each group to detect a specified difference $D = \mu_1 - \mu_{2,}$ with power specified by z_2 and the significance level specified by $z_{1,}$ is given by:

$$n = \left[(z_1 + z_2)^2 \left(\sigma_1^2 + \sigma_2^2\right)\right] / (\mu_1 - \mu_2)^2$$

where σ_1 and σ_2 are the standard deviations of the outcome variable in groups 1 and 2, respectively.

For 90% power and significance at p<0.05, this simplifies to:

$$n = 10.5 \left(\sigma_1^2 + \sigma_2^2\right) / (\mu_1 - \mu_2)^2.$$

To calculate the power of a trial of specified size, calculate the following, and refer the value of z_2 to Table 5.1:

$$z_2 = \left\{\sqrt{\left[n / \left(\sigma_1^2 + \sigma_2^2\right)\right]}\right\} (|\mu_1 - \mu_2|) - z_1.$$

Estimates of σ_1 and σ_2 may be obtained from previous studies or from a pilot study. If appropriate values cannot be determined, an alternative is to dichotomize the continuous outcome variable and use the sample size formulae for comparison of proportions given in Section 4.1. This will give a conservative estimate of sample size, as it ignores some of the information, but will ensure an adequate sample size in the face of uncertainty regarding the standard deviations.

Example: In the mosquito-net trial, the mean PCV in the control group at the end of the trial is expected to be 33.0, with a standard deviation of 5.0. To have 90% power of detecting a significant effect if the intervention increases the mean PCV by 1.5, the number of children required in each group is given by:

$$n = \left[(1.96 + 1.28)^2 \left(5.0^2 + 5.0^2\right)\right] / (1.5)^2 = 233.$$

Suppose it turns out that only 150 children are available for study in each group. The power in these circumstances is given by the following, corresponding to a power of about 74%:

$$z_2 = \left\{\sqrt{\left[150 / \left(5.0^2 + 5.0^2\right)\right]}\right\} (|1.5|) - 1.96 = 0.64.$$

A summary of the various formulae that have been given for calculating the trial size requirements for the comparison of two groups of equal size is given in Table 5.4.

5. More complex designs

5.1. Two groups of unequal size

Sections 3 and 4 considered the simplest situation where the two groups to be compared are of equal size. Sometimes, there may be reasons for wishing to allocate more individuals to one group than to the other. For example, if an experimental drug is very expensive, it may be desired to minimize the number of patients allocated to the drug, and so the trial may be arranged so that there are two or three patients given the old drug for every patient given the new drug. In order to maintain the same power as in the equal allocation scheme, a larger total trial size will be needed, but the number given the new drug will be smaller. Conversely, in a trial of a new vaccine, it may be decided to allocate twice as many participants to the vaccinated group as are included in the placebo group, in order to increase the size of the safety database for the new vaccine, before it goes into public health programmes.

Let the size of the smaller of the two groups be n_1 and suppose the ratio of the two sample sizes to be k, so that there will be kn_1 individuals in the other group $(k > 1)$ Then, to achieve approximately the same power and precision as in a trial with an equal number n in each group, n_1 should be chosen as:

$$n_1 = n\,(k+1)\,/\,(2k)\,.$$

Examples are shown in Table 5.5 for various values of k. Notice that the number allocated to the smaller group can never be reduced below half the number required with equal groups. Little is gained by increasing k beyond 3 or 4, since, beyond this point, even a substantial increase in n_2 achieves only a small reduction in n_1.

5.2. Comparison of more than two groups

Field trials comparing two groups (for example, intervention and control, or treatment A and treatment B) are by far the commonest. However, in some trials, three or more groups may be compared. For example, in a trial of a new vaccine, there may be four trial groups receiving different doses of the vaccine. It is unusual for field trials to have more than four groups, because of logistical constraints or trial size limitations.

It is suggested that, in designing a trial with three or more groups, the investigator should decide which pair-wise comparisons between groups are of central interest. The methods of Sections 3 and 4 can then be used to decide on the

trial size required in each group. Where there is one control group for comparison with several intervention groups, it is likely that the main pair-wise comparisons will be between each intervention group and the control group. Note, however, that direct comparisons between the intervention groups may then be inadequately powered, since, if each of the interventions has some effect, differences between the intervention groups may be smaller than when each is compared with the control group.

5.3. Factorial designs

As discussed in Chapter 4, Section 3.2, some trials are designed to look simultaneously at the effects of two interventions, using a *factorial design*. In a 2×2 factorial trial of two interventions A and B, for example, participants are randomly allocated between four trial groups receiving A only, B only, both A and B, or a control group receiving neither intervention. If the effects of A and B can be assumed independent, so that the effect of A is the same in the presence or absence of B and vice versa, then this trial design allows us to measure the effects of the two interventions for roughly the price of a single two-group trial measuring the effect of one intervention.

Under these conditions of independence, the main change to the calculation of sample size for a 2×2 factorial trial is that the expected outcome in the intervention and control groups for intervention A has to be adjusted for the expected effect of intervention B. This is explained with an example.

For example, suppose we are interested in the effects of iron supplements (intervention A) and anti-malarial prophylaxis (intervention B) on anaemia during pregnancy. Suppose that the prevalence of anaemia in the control group that receives neither A nor B is expected to be 30%, that each intervention is expected to reduce the prevalence proportionally by 20%, and that these effects are independent. Then the expected prevalences in the four arms of the trial will be: control—30%; A only—24%; B only—24%; A + B—19.2%. In this factorial trial, the effect of intervention A will be estimated by comparing the prevalence between groups A + B and B only, and between group A only and the control group. The overall prevalence in the two groups given intervention A will be $21.6\% [= (24 + 19.2)/2]$ and in the two groups not given A$27\% [= (30 + 24)/2]$ Since the difference in prevalences is slightly smaller than in a simple two-group trial, the total sample size will be somewhat larger for the factorial design.

In some factorial trials, we may wish to look explicitly at whether the effects of the two interventions are independent. This requires a test for *interaction* or *effect modification*, since we are interested in whether the effect of A, for example, differs according to the presence or absence of B. Testing for interaction generally requires a much larger sample size than a simple comparison of two groups. As a rough guideline, the total sample size for a 2×2 factorial trial would need to be multiplied by at least four to detect a substantial interaction (of similar size to the main effects of the interventions) between the effects of two interventions.

5.4. Equivalence and non-inferiority trials

In most field trials, the objective is to determine whether a new intervention is *superior* to a control intervention, for example, an existing intervention. In some cases, however, we may wish to demonstrate that a new intervention is *equivalent*, or at least *not inferior*, to an existing intervention. For example, suppose the current treatment for some condition is known to be highly effective, but it is also expensive and has some unpleasant side effects. Now suppose that a new treatment has been developed which is less costly and has fewer side effects. This would probably be considered for implementation, as long as it is as effective as the old treatment. In this case, we may decide to conduct an *equivalence trial* aimed at determining whether the two treatments have similar efficacy.

For a full discussion of such trials, the reader is referred to Blackwelder (1982) or Wang and Bakhai (2006). However, a simple example is given to illustrate the required sample size calculations.

Example: Suppose that the current treatment for TB has a cure rate of around 90% but requires a prolonged course of treatment. A new shorter-course regimen has been developed which would have advantages, in terms of cost, convenience, and adherence. We wish to carry out a trial to determine whether the cure rate for the short-course regimen is *equivalent* to that of the current regimen. We would usually do this by defining a lower limit for the cure rate, below which we would no longer consider the treatments to be ‘equivalent’. If we set this at 85%, the trial would need to be powered to demonstrate that the difference in cure rates is no more than 5%. The null hypothesis is now that the new treatment is *inferior* to the old treatment, and we power the trial to reject this null hypothesis and declare equivalence of the two treatments if the new treatment has a cure rate that is not inferior to the standard treatment by more than the specified 5%.

Modifying the first equation in Section 4.1 appropriately, we need n patients in each group, where:

$$n = \left[(z_1 + z_2)^2 2p(1-p)\right] / D^2.$$

In this equation, *p* is the expected cure rate of 90% in both groups, assuming equivalence, and *D* is the acceptable margin of inferiority, which is 5% in this example. Thus, for 90% power and a two-sided significance test withp=0.05, we have:

$$n = \left[(1.96 + 1.28)^2 \times 2 \times 0.90 \times 0.10\right] / 0.05^2 = 756.$$

In general, large sample sizes are needed to test equivalence.

6. Interventions allocated to groups

The methods described in Sections 3 to 5 all assume that individuals are to be the units of allocation. In other words, the trial groups will be constructed effectively by making a complete list of the individuals available for the trial and randomly selecting which individuals are to be allocated to each trial group. As explained in Chapter 4, Section 4, however, many field trials are not organized in this way. Instead, groups of individuals are allocated to the interventions under study. These groups are often called *clusters* and may correspond to communities, for example, villages, hamlets, or defined sectors of an urban area; institutions such as schools or workplaces; or patients attending a particular health facility.

Trials in which communities or other types of cluster are randomly allocated to the different arms of the trial are known as *cluster randomized trials*, and sample size calculations for such trials are presented in Section 6.1. *Stepped wedge trials* are a modified form of cluster randomized trial and are discussed in Section 6.2.

6.1. Cluster randomized trials

If clusters are randomly allocated to the different trial arms, the cluster should also be used as the unit of analysis, even though assessments of outcome are made on individuals within clusters (see Chapter 21, Section 8). For example, suppose the mosquito-net trial is to be conducted as follows. A number of villages (say 20) are to be randomly divided into two equal-sized groups. In the ten villages in the first group, the entire population of each village will be given mosquito-nets, while the second group of ten villages will serve as controls. The analysis of the impact of mosquito-nets on the incidence of clinical malaria would be made by calculating the (age-adjusted) incidence rate in each village and comparing the ten rates for the intervention villages with the ten rates for the control villages. This would be achieved by treating the (age-adjusted) rate as the quantitative outcome measured for each village and comparing these, using the unpaired t-test or the non-parametric rank sum test (see Chapter 21, Section 8). If analysing proportions, rather than incidence rates, the principle is the same—the (age-adjusted) proportion would be treated as the quantitative outcome for each cluster.

When allocation is by cluster, the trial size formulae have to be adjusted to allow for intrinsic variation between communities. Suppose first that incidence rates in the two groups are to be compared. The required number of clusters c is given by:

$$c = 1 + (z_1 + z_2)^2 \left[(r_1 + r_2)/y + k^2\left(r_1^2 + r_2^2\right)\right] / (r_1 - r_2)^2.$$

In this formula, *y* is the person-years of observation in each cluster, while r_1 and r_2 are the average rates in the intervention and control clusters, respectively. The intrinsic variation between clusters is measured by *k*, the *coefficient of variation* of the (true) incidence rates among the clusters in each group, and is defined as the standard deviation of the rates divided by the average rate. The value of *k* is assumed similar in the intervention and control groups, so that the *relative variability* remains the same following intervention.

If proportions are to be compared, the required number of clusters is given by:

$$c = 1 + (z_1 + z_2)^2 \left[2p\,(1-p)\,/n + k^2\left(p_1^2 + p_2^2\right)\right]/(p_1 - p_2)^2.$$

In this formula, *n* is the trial size in each community;p_1 and p_2 are the average proportions in the intervention and control groups, respectively; *p* is the average of p_1 and $p_{2,}$ and *k* is the coefficient of variation of the (true) proportions among the clusters in each group.

An estimate of *k* will sometimes be available from previous data on the same clusters or from a pilot study. If no data are available, it may be necessary to make an arbitrary, but plausible, assumption about the value of *k*. For example, $k = 0.25$ implies that the true rates in each group vary roughly between $r_1 \pm 2kr_1$, i.e. between $0.5r$ and $1.5r$. In general, *k* is unlikely to exceed 0.5.

Example: Suppose the mosquito-net trial is to be conducted by allocating the intervention at the village level. The incidence rate of clinical malaria among children before intervention is 10 per 1000 child-weeks of observation, and the trial is to be designed to give 90% power if the intervention reduces the incidence rate by 50%. There are about 50 eligible children per village, and it is intended to continue follow-up for 1 year, so that *y* is approximately 2500 child-weeks. No information is available on between-village variation in incidence rates. Taking $k = 0.25$, the number of villages required per group is given by the following, so that roughly seven villages would be needed in each group:

$$c = 1 + (1.96 + 1.28)^2 \left[(0.01 + 0.005)\,/2500 + 0.25^2\left(0.01^2 + 0.005^2\right)\right]/(0.01 - 0.005)^2 = 6.8.$$

Note that this would give a total of 17 500 child-weeks of observation in each group, compared with 6300 child-weeks if individual children were randomized to receive mosquito-nets. Figure 5.2 shows the number of villages required in each group, depending on the child-weeks of observation per village and the value of *k*.

The effect of group allocation on the total trial size needed will depend on the degree to which individuals within a cluster are more likely to be similar to each other than individuals in a different cluster for the outcome measure in the trial. If there is no heterogeneity between clusters in the outcome of interest, in the sense that the variation between the cluster-specific rates or means is no more than would be expected to occur by chance, due to sampling variations, the total trial size will be approximately the same as if the interventions were allocated to individuals. For most outcomes, however, there will be real differences between clusters, and, in these circumstances, the required trial size will be *greater* than with individual allocation. The ratio of the required trial sizes with cluster and individual allocation is sometimes called the *design effect*. Unfortunately, no single value for the design effect can be assumed, as its value depends on the variability of the outcome of interest between clusters and on the sizes of the clusters, and so it is recommended that the required sample size is estimated explicitly.

Note that, even if the calculations suggest that less than four clusters are required in each group, it is preferable to have at least four in each group. With so few units of observation, the use of non-parametric procedures, such as the rank sum test, is generally preferred for the analysis, and a sample size of at least four in each group is needed to have any chance of obtaining a significant result when this test is used.

It may be possible to reduce the required number of communities by adopting a matched design. For example, this can be done by using the baseline study to arrange the clusters into pairs, in which the rates of the outcome of interest are similar, and randomly selecting one member of each pair to receive the intervention. However, it is difficult to quantify the effect of this approach on the number of clusters required. To do this, information is required on the variability of the treatment effect between communities and on the extent to which the baseline data are predictive of the rates that would be observed during the follow-up period in the absence of intervention, and this information is rarely available. With a paired design, at least six clusters are required in each group in order to be able to obtain a significant difference using a non-parametric statistical test.

Further information on sample size calculations for cluster randomized trials is given in Hayes and Bennett (1999) and Hayes and Moulton (2009).

According to the number of child-weeks of observation in each community and the extent of variation in rates of clinical malaria between communities (*k* is the coefficient of variation of the incidence rates; see text). The average incidence rate of clinical malaria in the absence of the intervention is assumed to be ten per 10 000 weeks of

observation, and the trial is required to have 90% power to detect a 50% reduction in the incidence of malaria at the $p < 0.05$ level of statistical significance.

6.2. Stepped wedge trials

The *stepped wedge* design was introduced in Chapter 4, Section 4.3 and is a modification of the cluster randomized trial, in which all clusters commence the trial in the control group. The intervention is then introduced gradually into the clusters in random order, until, at the end of the trial, all the clusters are in the intervention group.

A consequence of the stepped wedge design is that, at most time points during the trial, there will be unequal numbers of clusters in the intervention and control groups. This means that, when secular trends are accounted for by comparing intervention to control groups at each step, a stepped wedge trial can have lower power and precision than a standard cluster randomized trial of the same size, in which the numbers of intervention and control clusters are equal throughout. When there is zero intra-cluster correlation, the trial will need up to 50% more clusters. To adjust for this, the number of clusters has to be multiplied by a correction factor which depends on the number of 'steps' in the stepped wedge design. If there are five steps, the correction factor is 1.3, rising to approximately 1.4 for numbers of steps between 10 and 20. When intra-cluster correlation is large enough, the gain in efficiency that can be made by taking advantage of the pre–post information on each cluster can overtake this factor, making a stepped wedge trial more efficient than a parallel trial. To be conservative, however, it may be best to inflate the number of clusters.

Example: In the mosquito-net trial discussed earlier, the sample size calculation showed that we needed seven clusters in each arm or a total of 14 clusters. If we now propose to carry out this trial using a stepped wedge trial, a conservative correction would be to multiply this number by 1.4, giving 20 clusters. For example, this might be implemented with ten steps over a 5-year period, providing nets to two randomly chosen clusters each half year.

7. Other factors influencing choice of trial size

7.1. Allowance for interim analyses

It is sometimes desirable to incorporate interim analyses into the trial plan, involving review of the results at (say) 6-monthly or annual intervals. If an interim analysis indicates that there is already strong evidence of the superiority of one of the interventions under study, the trial can be terminated in order that participants are no longer subjected to an intervention which is known to be inferior. The incorporation of interim analyses may be particularly valuable if the trial is planned to continue for several years, with the gradual accumulation of cases of the outcome of interest, or if individuals or communities are entered into the trial sequentially.

There are also disadvantages in carrying out interim analyses, however. If the trial is terminated early, because the intervention appears to be beneficial, there may be no opportunity of detecting any long-term effects of the intervention, including how efficacy changes with time or long-term adverse consequences of the intervention. Also, although a significant effect of the intervention may be demonstrated, the precision of the estimate of effect may be too low to be of much value.

If, after careful consideration, it is decided that interim analyses are to be conducted, these need to be planned in the trial design. It is necessary to employ a more stringent significance level for each analysis (interim and final) to maintain the same overall level of significance.

Details of the implications of interim analyses are given by Geller and Pocock (1987). As a rough guide, the following approach is suggested. It is rarely advantageous to plan for more than three or four interim analyses. It is recommended therefore that, for trials planned to continue for 2–4 years, the trial plan should include no more than two interim analyses (plus the final analysis). To compensate for this, the maximum trial size (i.e. the maximum person-years of observation if the trial proceeds to completion) should be increased by about 15%. A stringent significance level of $p=0.01$ should be used at each interim analysis to decide whether or not the trial should be terminated. This means that, if the trial proceeds to completion, an *unadjusted* $p<0.04$ would correspond to an *adjusted* $p<0.05$ if the interim analyses are taken into account, i.e little power has been lost in performing the interim analyses.

7.2. Allowance for losses

Losses to follow-up occur in most longitudinal studies. Individuals may be lost, because they move away from the trial area, they die from some cause unrelated to the outcome of interest, they refuse to continue with the trial, they are away from home at the time of a follow-up survey, or for some other reason.

Losses like these are of concern for two reasons. First, they are a possible source of *bias*, as the individuals who are lost often differ in important respects from those who remain in the trial. Second, they reduce the size of the sample available for analysis, and this decreases the power or precision of the trial.

For these reasons, it is important to make every attempt to reduce the number of losses to a minimum. However, it is rarely possible to avoid losses completely. The extent of the problem will vary, according to circumstances, but, as a rough guide, in a longitudinal trial of a rural community with 2 years of follow-up, losses of around 20% would not be unusual.

The reduced power or precision resulting from losses may be avoided by increasing the initial sample size, in order to compensate for the expected number of losses. For example, if sample size calculations suggest that 240 subjects are required and a 20% loss rate is expected, the sample size should be increased to 300 (because 80% of 300 gives 240). It is important to stress that sample size inflation only deals with the problem due to the reduction in the size of the sample available for analysis; it does not solve any potential problems due to bias. So, even if the sample size has been inflated to allow for losses to follow-up, it is still necessary to strive to minimize losses, in order to avoid bias.

8. The consequences of trials that are too small

The methods outlined in this chapter for selecting an adequate sample size have been available for many years, but it is probably not an exaggeration to state that the majority of intervention trials are much too small. Although there is an increasing awareness of the need to enrol a large enough sample, this chapter is concluded by discussing the consequences of choosing a sample size that is too small.

First, suppose that the intervention under study has little or no effect on the outcome of interest. The difference observed in a trial is likely therefore to be non-significant. However, the width of the CI for the effect measure (for example, the relative risk) will depend on the sample size. If the sample is small, the CI will be very wide, and so, even though it will probably include the null value (a zero difference between the groups, or a relative risk of 1), it will extend to include large values of the effect measure. In other words, the trial will have failed to establish that the intervention is unlikely to have an effect of public health or clinical importance. For example, in the mosquito-net trial, suppose only 50 children were included in each group, and suppose the observed spleen rates in the two groups were identical at 40%, giving an estimated relative risk of $R = 1$. The approximate 95% CI for R would extend from 0.62 to 1.62 (see Section 3.1). A relative risk of 0.62 would imply a very substantial effect, i.e. a reduction in spleen rate from 40% to 25%, and this small trial would be unable to exclude such an effect as being very unlikely. If the sample size in each group were increased to 500, the 95% CI would extend only from 0.86 to 1.16, a much narrower interval.

Suppose that the intervention does have an appreciable effect. A trial that is too small will have low power, i.e. it will have little chance of giving a statistically significant difference. In other words, there is little chance to demonstrate that the intervention has an effect. In the example, if the true effect of the intervention is to reduce the spleen rate from 40% to 25%, a sample size of 50 in each group would give a power of only 36%. A total of 205 children would be needed in each group to give 90% power (Table 5.2). Even if a significant difference is found, the CI on the effect will still be very wide, so there will be uncertainty at the end of the trial whether the effect of the intervention is small and unimportant, or very large and of major importance.

The conduct of trials that are too small has consequences extending beyond the results of the specific trial. There is considerable evidence that trials showing large effects are more likely to be published than those showing little or no effect. Suppose a number of small trials of a specific intervention are conducted. Because of the large sampling error implied by small sample sizes, a few of these trials will produce estimates of the effect of the intervention that are much larger than the true effect. These trials are more likely to be published, and the result is that the findings in the literature are likely to overestimate considerably the true effects of interventions. This publication bias is much smaller for larger trials, because a large trial showing little or no effect is more likely to be published than a small trial with a similar difference.

9. Computer software for sample size calculations

Most of the formulae given in this chapter are simple enough to do by hand, with the aid of a simple calculator. However, computer software is also available to carry out some of these calculations. This can be particularly helpful when a large number of calculations need to be carried out, for example, to explore sample size requirements for different outcomes or under different assumptions, or to produce power curves. Most statistical packages have some

provision for sample size calculations. Here we mention three packages which readers may find helpful when planning field trials.

The *sampsi* command in Stata allows the user to obtain the required sample size for the comparison of means or proportions. Alternatively, if the chosen sample sizes are entered, the user can determine the power that these will provide. The command allows for different sample sizes in the two trial arms. The sample size formulae used by this package differ slightly from those presented in this book, but the results should be quite similar in most cases.

The POWER and GLMPOWER procedures in the statistical analysis program package SAS can handle sample size calculations for a range of situations, including survival analysis, as can the PASS module (a trial version of which is available at <http://www.ncss.com>).

A variety of free sample size calculators may be found on the Internet. These include the program PS which is described by Dupont and Plummer (1990) and available at <http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize>); and Open Epi which can be downloaded from <http://www.openepi.com/Downloads/Downloads.htm>.

Table 5.6 gives a spreadsheet which facilitates the calculation of the required size (number of clusters) for a cluster randomized trial, using the formulae given in Section 6 (as in Hayes and Moulton, 2009).

It is fairly straightforward to set up a spreadsheet, for example, in Excel, to apply any of the formulae given in this chapter. The freeware computer package Epi-Info has a useful component, called StatCalc, for calculating sample sizes for simple trials.

References

Blackwelder, W. C. 1982. 'Proving the null hypothesis' in clinical trials. *Control Clinical Trials*, **3**, 345–53.10.1016/0197-2456(82)90024-1 [PubMed: 7160191] [CrossRef]

Chow, S.-C., Shao, J., and Wang, H. 2008. *Sample size calculations in clinical research*, 2nd ed. New York: Chapman & Hall/CRC Press, Taylor & Francis.

Dupont, W. D. and Plummer, W. D., Jr. 1990. Power and sample size calculations. A review and computer program. *Control Clinical Trials*, **11**, 116–28. [PubMed: 2161310]

Geller, N. L. and Pocock, S. J. 1987. Interim analyses in randomized clinical trials: ramifications and guidelines for practitioners. *Biometrics*, **43**, 213–23.10.2307/2531962 [PubMed: 3567306] [CrossRef]

Hayes, R. J. and Bennett, S. 1999. Simple sample size calculation for cluster-randomized trials. *International Journal of Epidemiology*, **28**, 319–26.10.1093/ije/28.2.319 [PubMed: 10342698] [CrossRef]

Hayes, R. J. and Moulton, L. H. 2009. *Cluster randomized trials*. Boca Raton: Chapman & Hall/CRC.10.1201/9781584888178 [CrossRef]

Machin, D. 2009. *Sample size tables for clinical studies*. Oxford: Wiley-Blackwell.

Wang, D. and Bakhai, A. 2006. *Clinical trials : a practical guide to design, analysis, and reporting*. London: Remedica.

Figures

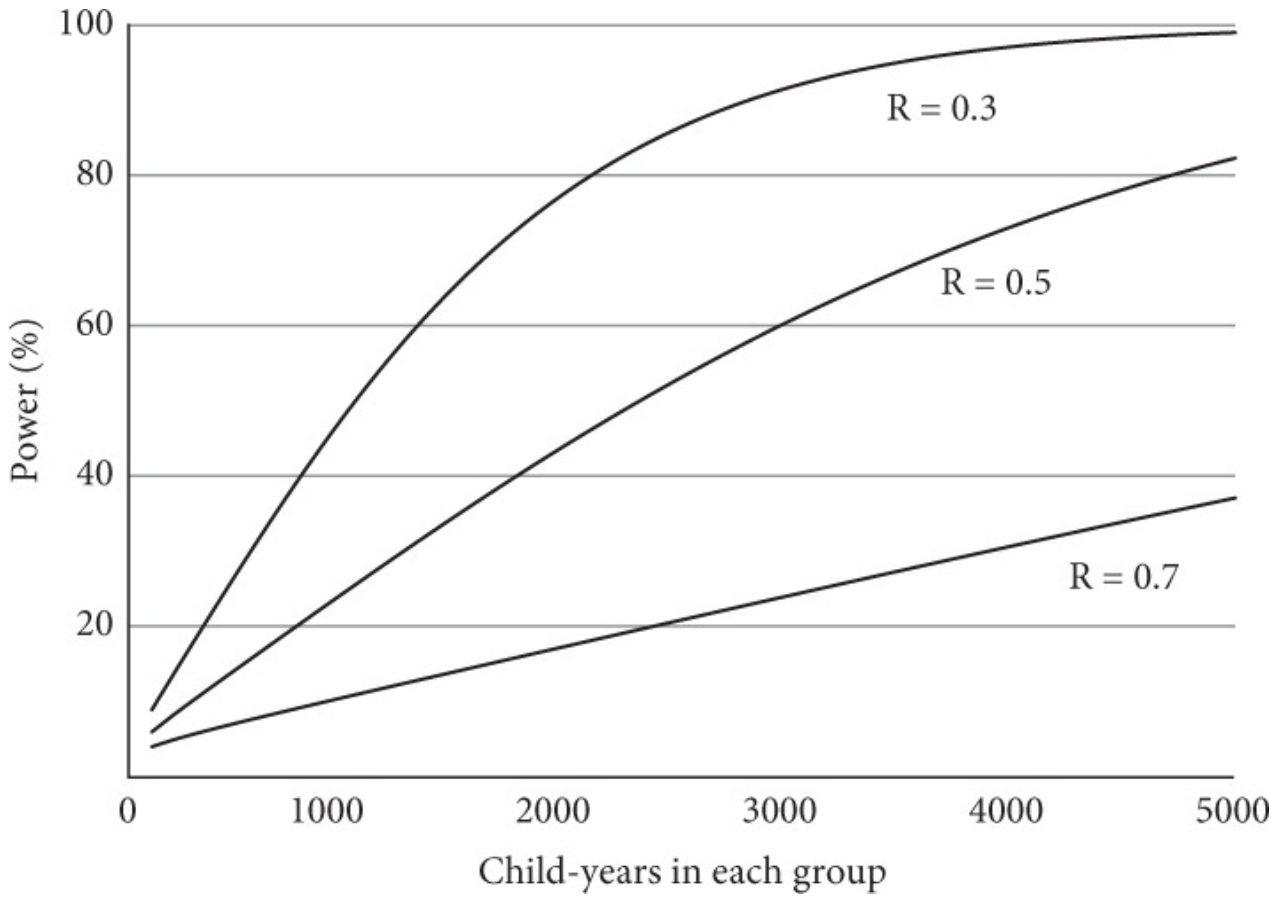

Figure 5.1

Power curves for a trial of the effect of mosquito-nets on malaria deaths.

Malaria death rate in the control group assumed to be 10/1000/year. R, relative rate in the intervention group. Assumes equal-sized groups, two-sided test, and significance $p < 0.05$.

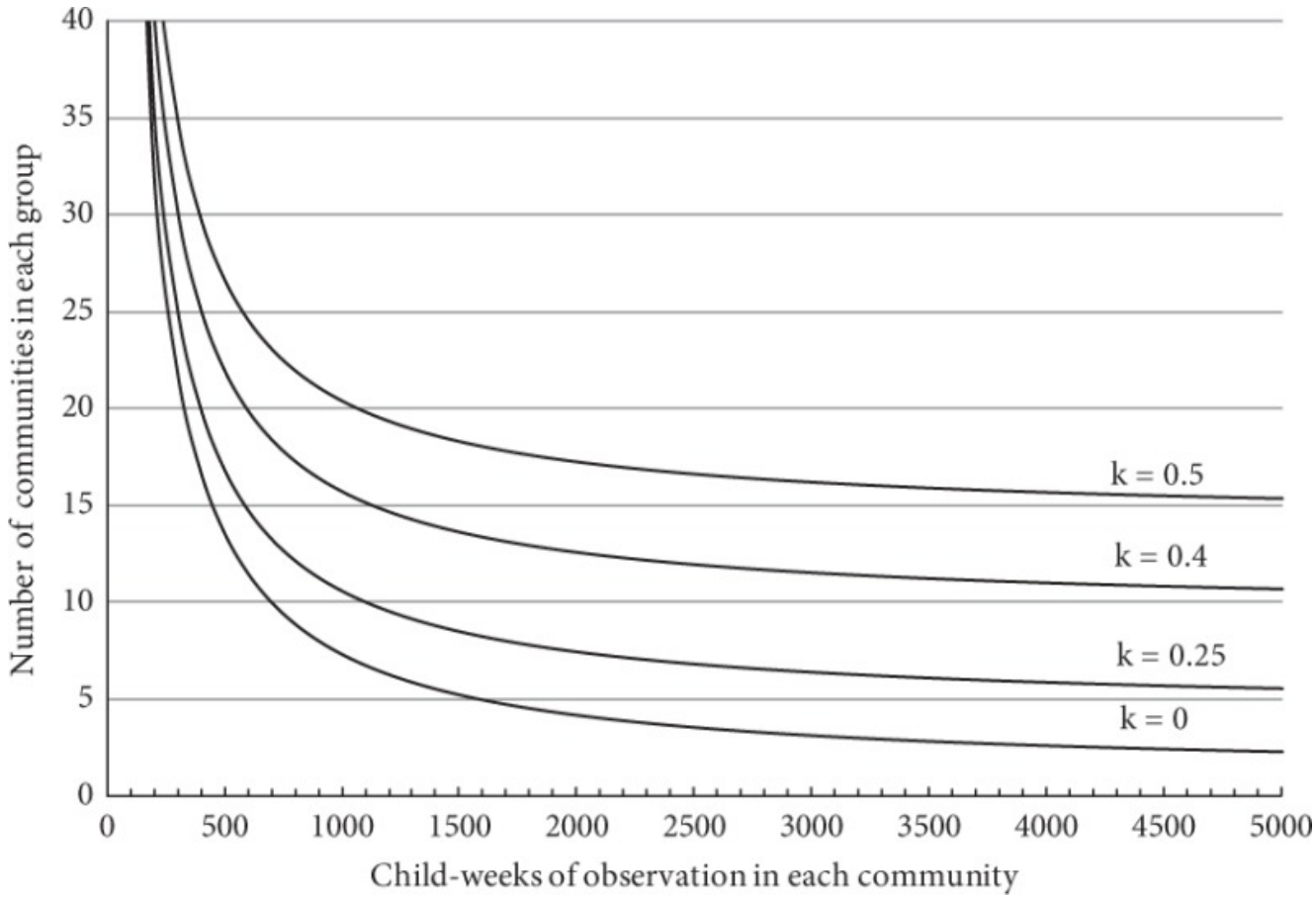

Figure 5.2

Number of communities required in each group in a trial of the effect of mosquito-nets against clinical malaria.

Tables

Table 5.1 Relationship between z_2 and % power (numbers in the body of the table show power corresponding to each value of z_2

First decimal place of z_2										
z_2	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
−3.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
−2.0	2.3	1.8	1.4	1.1	0.8	0.6	0.5	0.3	0.3	0.2
−1.0	15.9	13.6	11.5	9.7	8.1	6.7	5.5	4.5	3.6	2.9
−0.0	50.0	46.0	42.1	38.2	34.5	30.9	27.4	24.2	21.2	18.4
+0.0	50.0	54.0	57.9	61.8	65.5	69.1	72.6	75.8	78.8	81.6
+1.0	84.1	86.4	88.5	90.3	91.9	93.3	94.5	95.5	96.4	97.1
+2.0	97.7	98.2	98.6	98.9	99.2	99.4	99.5	99.7	99.7	99.8
+3.0	99.9	99.9	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Note: for example, $z_2 = -0.7$ corresponds to a power of 24.2%.

Table 5.2 Sample size requirements for comparison of proportions

Smaller prop. p_1	Difference $D = p_2 - p_1$											
	0.05	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**	**0.55**	**0.60**
0.05	435	141	76	50	36	28	22	18	15	13	11	10
	583	189	102	67	48	37	30	25	21	18	15	13
	719	233	126	83	60	46	37	30	26	22	19	16
0.10	686	200	101	63	44	33	26	21	17	14	12	10
	919	268	135	84	59	44	34	28	23	19	16	14
	1134	330	166	104	72	54	42	34	28	24	20	17
0.15	906	251	122	74	50	37	28	22	18	15	13	10
	1212	336	163	98	67	49	38	30	24	20	17	14
	1497	415	201	122	83	60	46	37	30	25	21	18
0.20	1094	294	139	82	55	40	30	24	19	16	13	11
	1464	394	186	110	74	53	40	31	25	21	17	15
	1808	486	230	136	91	66	50	39	31	26	21	18
0.25	1250	329	153	89	59	42	31	24	19	16	13	11
	1674	441	205	119	79	56	42	32	26	21	17	14
	2067	544	253	147	97	69	52	40	32	26	21	18
0.30	1376	357	163	94	61	43	32	24	19	16	13	10
	1842	478	219	126	82	58	43	33	26	21	17	14
	2274	590	270	156	101	71	53	40	32	26	21	17
0.35	1470	376	170	97	63	44	32	24	19	15	12	10
	1968	504	228	130	84	58	43	32	25	20	16	13
	2430	622	282	160	103	72	53	40	31	25	20	16
0.40	1533	388	174	98	63	43	31	24	18	14	11	
	2052	520	233	131	84	58	42	31	24	19	15	
	2534	642	287	162	103	71	52	39	30	24	19	
0.45	1564	392	174	97	61	42	30	22	17	13		
	2094	525	233	130	82	56	40	30	23	18		
	2586	648	287	160	101	69	50	37	28	22		
0.50	1564	388	170	94	59	40	28	21	15			
	2094	520	228	126	79	53	38	28	21			
	2586	642	282	156	97	66	46	34	26			
0.55	1533	376	163	89	55	37	26	18				
	2052	504	219	119	74	49	34	25				
	2534	622	270	147	91	60	42	30				
0.60	1470	357	153	82	50	33	22					
	1968	478	205	110	67	44	30					
	2430	590	253	136	83	54	37					
0.65	1376	329	139	73	44	28						
	1842	441	186	98	59	37						

Smaller prop. p_1	**Difference** $D = p_2 - p_1$											
	0.05	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**	**0.55**	**0.60**
	2274	544	230	121	72	46						
0.70	1250	294	122	63	36							
	1674	394	163	84	48							
	2067	486	201	104	60							
0.75	1094	251	101	50								
	1464	336	135	67								
	1808	415	166	83								
0.80	906	200	76									
	1212	268	102									
	1497	330	126									
0.85	686	141										
	919	189										
	1134	233										
0.90	435											
	583											
	719											

Shown in the body of the table are the sample sizes required in each group to give the specified power.*

* Upper figure: power, 80%; middle figure: power, 90%; lower figure: power, 95%. Using a two-sided significance test with $p < 0.05$. The two groups are assumed to be of equal size.

Table 5.3 Sample size requirements for comparison of rates

Relative rate R *	Expected events in group 2 to give+		
	80% power	90% power	95% power
0.1	10.6	14.3	17.6
0.2	14.7	19.7	24.3
0.3	20.8	27.9	34.4
0.4	30.5	40.8	50.4
0.5	47.0	63.0	77.8
0.6	78.4	105.0	129.6
0.7	148.1	198.3	244.8
0.8	352.8	472.4	583.2
0.9	1489.6	1994.5	2462.4
1.1	1646.4	2204.5	2721.6
1.2	431.2	577.4	712.8
1.4	117.6	157.5	194.4
1.6	56.6	75.8	93.6
1.8	34.3	45.9	56.7
2.0	23.5	31.5	38.9
2.5	12.2	16.3	20.2
3.0	7.8	10.5	13.0
5.0	2.9	3.9	4.9
10.0	1.1	1.4	1.8

Numbers in the body of the table are expected number of events required in group 2 to give specified power if relative rate in group 1 is *R*.

* *R*, ratio of incidence rate in group 1 to incidence rate in group 2.

+ Using a two-sided significance test with $p < 0.05$. The two groups are assumed to be of equal size.

Table 5.4 Summary of formulae for calculating trial size requirements for comparison of two groups of equal size

Type of outcome	Formula	Notation	Section in text
A: Choosing trial size to achieve adequate precision			
Proportions:	$n = (1.96/\log_e f)^2 \{[(R+1)/(Rp_2)] - 2\}$	n = number in each group R = prop. in group 1/prop. in group 2 Gives 95% CI from R/f to Rf	3.1
Rates:	$e_2 = (1.96/\log_e f)^2 [(R+1)/R]$	e^2 = expected events in group 2 R = rate in group 1/rate in group 2 Gives 95% CI from R/f to Rf	3.2
Means:	$n = (1.96/f)^2 (\sigma_1^2 + \sigma_2^2)$	n = number in each group σ_i = SD in group i D = mean in group 1 − mean in group 2 Gives 95% CI of $D \pm f$	3.3
B: Choosing trial size to achieve adequate power			
Proportions:	$n = \left[(z_1 + z_2)^2 2p(1-p)\right]/(p_1 - p_2)^2$	n = number in each group p^i = proportion. in group i p = average of p_1 and p_2	4.1
Rates:	$y = \left[(z_1 + z_2)^2 (r_1 + r_2)\right]/(r_1 - r_2)^2$	y = person-years in each group r^i = rate in group i	4.2
Means:	$n = \left[(z_1 + z_2)^2 (\sigma_1^2 + \sigma_2^2)\right]/(\mu_1 - \mu_2)^2$	n = number in each group σ_i = SD in group i μ_i = mean in group i	4.3

$z_1 = 1.96$ for significance at p<0.05

Power 80%, 90%, 95%

$z_2 = 0.84,$
$1.28,\ 1.64$

Table 5.5 Trial size necessary to achieve approximately the same power in a trial with two groups, one of which contains *k* times as many individuals as the other

k	n_1	n_2	$n_1 + n_2$
1	n	n	$2n$
2	$0.75n$	$1.5n$	$2.25n$
3	$0.67n$	$2.0n$	$2.67n$
4	$0.62n$	$2.5n$	$3.12n$
5	$0.60n$	$3.0n$	$3.60n$
10	$0.55n$	$5.5n$	$6.05n$
100	$0.50n$	$50.0n$	$50.50n$

Table 5.6 Spreadsheet calculation of the number of clusters required in an unmatched cluster randomized trial. Table 5.6 shows the calculations, for some example situations, for (a) comparison of proportions and (b) comparison of rates. Formulae are given which allow the calculation of required trial size for any (unmatched) cluster randomized trial in an Excel spreadsheet

(a) Comparison of proportions

Significance level	Power	z_1	z_2	$(z_1 + z_2)^2$	p_1	% reduction	p_2	Person-years per cluster	*k*	# clusters per arm	Rounded up
A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**
0.95	0.80	1.96	0.84	7.85	2.0%	50%	1.0%	500	0.25	**8.08**	**9**
0.95	0.80	1.96	0.84	7.85	3.0%	50%	1.5%	500	0.25	**6.51**	**7**
0.95	0.80	1.96	0.84	7.85	4.0%	50%	2.0%	500	0.25	**5.73**	**6**
0.95	0.90	1.96	1.28	10.51	2.0%	50%	1.0%	500	0.25	**10.48**	**11**
0.95	0.90	1.96	1.28	10.51	3.0%	50%	1.5%	500	0.25	**8.38**	**9**
0.95	0.90	1.96	1.28	10.51	4.0%	50%	2.0%	500	0.25	**7.33**	**8**
0.95	0.80	1.96	0.84	7.85	2.0%	50%	1.0%	250	0.25	**12.71**	**13**
0.95	0.80	1.96	0.84	7.85	3.0%	50%	1.5%	250	0.25	**9.57**	**10**
0.95	0.80	1.96	0.84	7.85	4.0%	50%	2.0%	250	0.25	**8.01**	**9**
0.95	0.90	1.96	1.28	10.51	2.0%	50%	1.0%	250	0.25	**16.68**	**17**
0.95	0.90	1.96	1.28	10.51	3.0%	50%	1.5%	250	0.25	**12.48**	**13**
0.95	0.90	1.96	1.28	10.51	4.0%	50%	2.0%	250	0.25	**10.38**	**11**

(b) Comparison of rates

Significance level	Power	z_1	z_2	$(z_1 + z_2)^2$	r_1	% reduction	r_2	Person-years per cluster	*k*	# clusters per arm	Rounded up
A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**
0.95	0.8	1.96	0.84	7.85	0.050	50%	0.025	300	0.25	6.46	**7**
0.95	0.8	1.96	0.84	7.85	0.050	45%	0.028	300	0.25	7.99	**9**
0.95	0.8	1.96	0.84	7.85	0.050	40%	0.030	300	0.25	10.18	**11**
0.95	0.8	1.96	0.84	7.85	0.050	35%	0.033	300	0.25	13.44	**14**
0.95	0.8	1.96	0.84	7.85	0.050	30%	0.035	300	0.25	18.57	**19**
0.95	0.8	1.96	0.84	7.85	0.050	50%	0.025	300	0.20	5.58	**6**
0.95	0.8	1.96	0.84	7.85	0.050	45%	0.028	300	0.20	6.86	**7**
0.95	0.8	1.96	0.84	7.85	0.050	40%	0.030	300	0.20	8.68	**9**
0.95	0.8	1.96	0.84	7.85	0.050	35%	0.033	300	0.20	11.39	**12**
0.95	0.8	1.96	0.84	7.85	0.050	30%	0.035	300	0.20	15.65	**16**

Excel expressions:

C = NORMSINV(1− (0.5*(1 − A))); D = NORMSINV(B)

K = 1 + E*((F*(1 − F)/I) + (H*(1 − H)/I) + (J*J)*((F*F) + (H*H)))/((H − F)^2)

L = INT(K) + 1

Boxes

Box 5.1 The power of the trial depends on:

1. The value of the true difference between the study groups, in other words, the true effect of the intervention. The greater the effect, the higher the power to detect the effect as statistically significant for a trial of a given size.
2. The trial size. The larger the trial size, the higher the power.
3. The probability level (p-value) at which a difference will be regarded as ‘statistically significant’.

Bookshelf ID: NBK305517