Research methods

Series editors
J M Stephenson,
A Babiker

Study size

Angie Wade

Research is only worth doing if it provides useful information. Medical research usually consists of studying groups of individuals with the aim of answering a predefined research question. Most commonly in the field of sexually transmitted infections (STI), the prevalence of a virus or abnormality is to be estimated or prevalences compared, either over time or between different groups of people. Alternatively, several therapies may be compared within a randomised controlled trial. One question that arises at the start of any study is "How many individuals should be included in this study?" There are several ways of answering this question. The number included may be based on practical issues—for instance, the length of time available to the researcher together with the time taken to recruit, treat, and test each individual and the expected patient accrual rate. These factors will vary from researcher to researcher and between different sources of patients—for example, accrual rates will differ between different hospitals. To use such variable quantities to determine the number needed to effectively answer a given research question is clearly flawed. Ethically it is wrong to either underrecruit or overrecruit. On the one hand we may be left with insufficient numbers to conclusively answer the question. On the other hand, if we overrecruit, then the best scenario is not only do we waste time, but also we subject more individuals than necessary to any inconvenience associated with being studied. In the worst scenario, we may be allowing individuals to receive inferior treatment after sufficient numbers have been recruited to ensure that the best treatment is known.

Many researchers associate sample size calculation purely with randomised controlled trials. Most of the studies presented in this journal do not fall into this category. However sample size estimation before study commencement is important for all types of study, including prevalence studies and observational comparisons. This article highlights the need for consideration of study size over and above issues of feasibility and practicality. Information is presented on how to determine an appropriate sample size for the most commonly used study designs within the field of STI.

Key messages

- Sample size is an important issue for all contributors of studies to this journal.
- The interpretation of study results depends on the sample size included that study.
- Sample size formulas are given for the most common scenarios encountered in the field of STI.

Why size matters

A prevalence study finds that 25% of women in the Outer Hebrides have HSV-2 antibodies.

We cannot interpret this information without knowing the numbers this figure was based on. For example, one out of four (25%) is a much less precise estimate than 100 out of 400 (25%). A proper presentation of the results would include a confidence interval. The 95% confidence interval for the first scenario is (0.6, 80.6%) and for the second (21.0, 29.5%). For each of the examples, these are the ranges within which we are 95% confident the population prevalence lies. As we would expect, with a much larger sample size, in the latter scenario we can make a more precise statement about the likely population prevalence.

20% of pregnant Outer Hebrideans have HSV-2 antibodies compared with 20% of Inner Hebrideans.

If the above statement were true and we randomly sampled and tested 30 individuals in each group, then we would expect to see about six antibody positive individuals in each group. However, we would not be unduly surprised to find four individuals with antibodies in one of the groups and eight in the other.

However, we can quantify how likely, in the absence of a difference, we are to falsely conclude from our study that there is one. The *statistical significance* of a study is the probability that we will falsely identify a difference when none exists. The larger the study, the less likely this is to happen.

Question: How do we know that our study will not indicate there are differences when none exist?
Answer: We don't.

20% of pregnant Outer Hebrideans have HSV-2 antibodies compared with 10% of mainland Scots.

If the above statement were true and we randomly sampled and tested 30 individuals in each group, then we would expect to see about six antibody positive individuals in the first group and three in the latter. However, we would not be unduly surprised to find four individuals with antibodies in each of the groups.

Department of Paediatric Epidemiology and Biostatistics, Institute of Child Health, 30 Guilford Street, London WC1N 1EH, UK
A Wade
awade@ich.ucl.ac.uk

Accepted for publication
7 June 2001

Question: How do we know that our study will not fail to identify a difference of clinical importance that truly exists?
Answer: We don't.

However, we can quantify how likely we are to find a difference of a given size if it exists. The *power* of a study, usually represented as a percentage, is the ability of a study of a given size to detect a difference of a given magnitude. The larger the difference the smaller the number needed.

Simple sums for sample sizing

Choosing the best sample size is not a precise art. Equations exist for calculating the sample sizes needed to obtain a specified precision or to identify differences of a given size. The latter of these are called *power calculations*. Sample size calculation relies on "guesstimates" of unknown quantities and hence obtained sizes are by definition unlikely to be correct. The extent to which sample size determination is influenced by erroneous estimates can be investigated by trying out different guesstimates in the formulas. Below are details of sample size calculation for the most common scenarios in STI studies.

(i) The approximate number of individuals required to estimate a prevalence within ± e% is given by the formula:

$$N = 4 \times \frac{\text{prevalence } (100 - \text{prevalence})}{e^2}$$

The number required depends on the prevalence the study is designed to estimate!

Note that if the prevalence is x% then the sample size required is the same as if estimating a prevalence of (100 − x)%.

For example, if the true prevalence is 25% and we want to estimate this prevalence to within ±5% then

$$4 \times \frac{25(100 - 25)}{5^2} = 300$$

individuals are required.

Of course we do not know this prevalence before we undertake the study. Our guesstimate of 25% may be inaccurate. If the true prevalence is actually 35%, then

$$4 \times \frac{35(100 - 35)}{5^2} = 364$$

individuals would be required to give the same level of precision. If only 300 individuals are included (based on the guesstimate of 25%) then the prevalence will be estimated less precisely than anticipated.

Figure 1 shows the numbers required to detect various prevalences to within ±1%. For example, to estimate a prevalence of 25% with this precision will require approximately 7500 individuals. (Note that the same number would be required to estimate 75% —that is, 100 − 25%, with the same precision.)

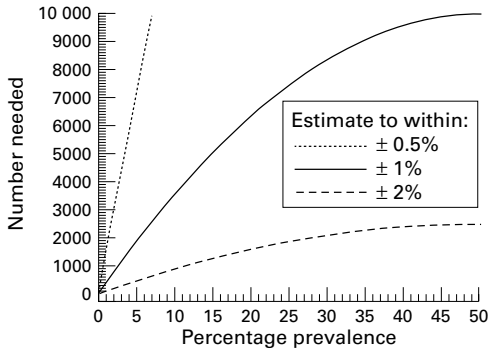

Figure 1 Number need to estimate a single prevalence to within plus or minus 1%, 2%, and 0.5% with 95% confidence.

If a prevalence needs to be estimated with greater precision then the sample size must be increased, for less precision smaller sample sizes are required.

To estimate prevalences more or less precisely, the estimates given for ±1% can be divided by the required precision squared. For example:

(a) To estimate to within ±2% the numbers given for 1% need to be divided by 2^2 (=4), this curve is shown on the figure. To estimate 25% to within ±2% requires approximately 7500/4 or 1875 individuals.

(b) To estimate to within ±0.5% the numbers given for 1% need to be divided by 0.5^2 (= 0.25). Part of this curve is shown on the figure. To estimate 25% to within ±0.5% requires approximately 7500/0.25 or 30 000 individuals.

(ii) If the prevalences within different groups are p1% and p2%, the approximate numbers of individuals required to detect this difference with 80% power at the 5% significance level are:

$$N \text{ (per group)} = \frac{8(p1\%(100 - p1\%) + p2\%(100 - p2\%))}{(p1\% - p2\%)^2}$$

Table 1 shows N for selected p1% and p2%.

For greater power the sample sizes need to be increased. Similarly, larger samples are required to detect smaller differences between prevalences.

For 90% power, the samples need to be increased by about one third.

For example, if 10% of Hebrideans and 20% of mainland Scots have HSV-2 infection, then 200 Hebridean and 200 mainland Scots need

Table 1 Number required in each group to detect differences with 80% power at the 5% significance level given that group prevalences are p1% and p2%

Prevalence (p1%):	*Prevalence (p2%):*						
	10	*20*	*30*	*40*	*50*	*60*	*70*
5	440	74	33	19	12	8	5
10		200	60	30	17	11	7
20			296	80	37	20	12
30				360	92	40	21
40					392	96	40
50						392	92
60							360

to be tested to detect this difference with 80% power at the 5% significance level. For 90% power a total of 266 (= 200 × 1.33) will need to be tested in each group.

As before, the sample size estimate is based on guesstimates of the study outcome. For the purposes of sample size determination, the magnitude of the difference in prevalences between groups may be chosen on the basis of what is a clinically important difference that we want to detect. For example, the best available evidence may suggest that the Hebrideans only have 5% fewer HSV-2 infections but we decide to undertake a study to detect a difference of 10% since this could represent a clinically important difference (in terms of allocating resources, etc), whereas 5% would not.

Often there is more than one factor to consider. For example, Outer Hebrideans may be older than the average Scottish person and it would be of interest to know whether the difference in prevalence of HSV-2 can be explained by the difference in ages of the groups. Sample size estimation in these scenarios is more complex and will depend on the extent to which the factors are related.[1]

MORE OF ONE THAN THE OTHER

When comparisons are made between two or more groups, power will be maximised (for a given overall total number enrolled) if each group is of the same size. To accommodate an imbalance in numbers while retaining the same power, the total sample size needs to be increased accordingly.[2]

ONES THAT GOT AWAY

Calculations give the estimated numbers required for statistical analysis. Sometimes it will be necessary to recruit many more to ensure that sufficient are obtained for that analysis. On the basis of the expected refusal and compliance rates, the study should be designed so that sufficient are contacted/approached/recruited to account for these losses.

ONES THAT NEVER HAPPEN

In some studies, individuals are recruited and monitored for variable lengths of time with the aim of comparing times to some defined event, such as death or infection, between subgroups of individuals. The nature of the outcome is that not all individuals will have the event and a straightforward comparison of percentages dying or becoming infected is invalidated because of the variable follow up times. Similarly, comparing average times to death/infection would also be invalid. A special type of analysis is required[3] and the sample size will be based on the number of individuals for whom the event occurs.[4] So for a rare event larger numbers will be required. Extending the length of follow up may be used to increase the numbers of events occurring.

ONES THAT FLOCK TOGETHER

If treatments are allocated on a group basis, the effective sample size will not equal the total number of individuals. Adjustment must be made according to the extent to which individual outcomes are influenced by the fact that they form part of a homogeneous group.[5 6] For example, the introduction of trained specialists within randomly selected clinics[7] to identify whether these specialists have an effect on adverse outcomes. Individuals from the same clinic may be more alike, perhaps because of social and ethnic similarities, than those from different clinics in terms of their tendency to be recorded as having a problem even before any intervention. A study that randomises by clinic will effectively be putting groups of similar individuals into the same arm of the trial en bloc. A similar situation occurs when individuals within the same family, ward/hospital, etc, are jointly allocated to treatments.

Summary

In conclusion, this short article has addressed the issues surrounding the determination of appropriate study size. Only the sample size equations relating to the most common study types in the field of STI are presented. For more complex but common situations, the important issues to consider are highlighted. Many other sample size equations exist—for example, when the outcome is continuous, such as CD4 count, or more than two groups are compared. Several useful references are given in the further reading section. This article raises awareness of the need to perform studies of the correct size for the given purpose and informs the researcher in sexually transmitted infections of the particular points that they may need to consider.

1 Cohen J. *Statistical power analysis for the behavioural sciences.* 2nd ed. New Jersey: Lawrence Erlbaum Associates, 1988.
2 Altman DG. *Practical statistics for medical research.* London: Chapman and Hall, 1991:459–60.
3 Collett D. *Modelling survival data in medical research.* London: Chapman and Hall, 1994.
4 Machin D, Campbell MJ. *Statistical tables for the design of clinical trials.* Oxford: Blackwell Scientific, 1987.
5 Hauck WW, Gilliss CL, Donner A, *et al.* Randomization by cluster. *Nursing Res* 1991;**40**:356–8.
6 Reading R, Harvey I, McLean M. Cluster randomised trials in maternal and child health: implications for power and sample size. *Arch Dis Child* 2000;**82**:79–83.
7 Osman NB, Challis K, Folgosa E, *et al.* An intervention study to reduce adverse pregancy outcomes as a result of syphilis in Mozambique. *Sex Transm Inf* 2000;**76**:203–27.

Further reading

Du V Florey C. Sample size for beginners. *BMJ* 1993;**306**:1181–4.
Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical and continuous outcomes in two group comparisons. *BMJ* 1995;**311**:1145–8.
Day SJ, Graham DF. Sample size estimation for comparing two or more treatment groups in clinical trials. *Stats Med* 1991;**10**:33–43.
Hsieh FY. Sample size tables for logistic regression. *Stats Med* 1989;**8**:795–802.
Julious SA, Campbell MJ. Sample size calculations for paired or matched ordinal data. *Stats Med* 1998;**17**:1635–42.
Kirkwood BR. *Essentials of medical statistics.* Chapter 26. Oxford: Blackwell Scientific, 1988.