Sample size determination in clinical research: 2

Laofe Ogundipe

In the second of this two-part series, comparative study design is considered and, using examples, power calculation is explained. The information required for calculating sample size for comparative studies is highlighted.

SAMPLE SIZE FOR COMPARATIVE STUDIES

Comparative studies are common clinical studies and include case control studies, cohort studies and clinical trials. We may wish to compare two groups on a binary outcome variable, i.e. the outcome of interest has one of two categories, e.g. improved/not improved or relapsed/not relapsed. We may also be interested in comparing the values of a continuous variable in two independent groups, e.g. the level of thiamine in subjects with chronic alcohol dependence *vs* non-alcohol dependent subjects or the brain weight of schizophrenics *vs* controls.

Comparing groups on a binary outcome measure

Example 1: A new drug was discovered recently and it was reported that it will prolong abstinence in people with alcohol dependence syndrome who completed detoxification. We want to do a randomized controlled trial of this drug compared with placebo. How many subjects do we need for the trial?

Step 1:

We need the following information:

1. What proportion of subjects in the placebo group do we expect to relapse? This is usually obtainable from the literature. Let us assume that our literature search shows that 68% of people with alcohol dependence syndrome who recently completed detoxification relapse within 12 months. Therefore, relapse rate in the placebo group = P1 = 0.68.
2. What size of difference in relapse rates between the two groups (clinically important difference or relevant effect size) do we consider clinically significant? At 1 year follow-up for example, what difference in the proportions of the two groups who have not relapsed do we consider clinically significant? Because of the high failure rate in alcoholism, if the new drug can achieve a reduction in relapse rate of 28% over placebo we will take this as clinically significant. Therefore, the proportion we expect to relapse in the intervention group is 40%. Relapse rate in intervention group = P2 = 0.40.
3. What is the chance of a false positive error (a type 1 error) that we are prepared to accept? A false positive result occurs when we mistakenly conclude that there is a difference between the two groups when in reality there is no difference. If we do the same experiment 100 times, we accept that such an error may be made on five occasions, hence the significance level (probability of a type 1 error) will be 5% ($P = 0.05$).
4. What is the chance of a false negative error (a type 2 error) that we are prepared to accept? A false negative result occurs when we mistakenly conclude that there is no difference between the two groups when in reality a difference exists. Again, we are prepared to accept such an error in no more than one in 20 occasions so the probability (B), of making this type of error will be 5%. The power of the study is the probability of not making a type 2 error. Power = 1–B = 95%. Usually a power of 80% or 90% is used in clinical research.

Step 2: Calculate the 'standardized difference'. This is the effect size (the magnitude of the difference between the two groups) divided by the standard deviation. For a difference between proportions, the formula for standardized difference is calculated as in *Figure 1.*

Step 3: We now know the standardized difference (0.56), we have decided on a significance level of 0.05 and a power of 95%. Calculating sample size then becomes very straightforward. We simply read the sample size on Altman's nomogram (*Figure 2*) by drawing a line between the standardized difference of 0.56 and the power required of 0.95 (Altman, 1980). This gives a sample size of 160 in each group. We may not be able to recruit 320 people for the study, therefore in our sensitivity analysis, we may decide on a power of 80%. Using Altman's nomogram again, this gives a sample size of about 90 in each group.

Comparing two independent groups on a continuous outcome measure

Example 2: Suppose that we are planning a randomized controlled trial of Pabrinex in alcohol-dependent people to see if Pabrinex supplement will increase thiamine levels in the experimental compared with control groups after 2 weeks of treatment. How many people do we need to recruit for the study?

$$\text{Standardized difference} = \frac{P1-P2}{\sqrt{\bar{p}(1-\bar{p})}} = \frac{0.68-0.40}{0.54 \times 0.46} = \frac{0.28}{0.498} = = 0.56225$$

where P1=proportion who relapse in the first group, P2=proportion who relapse in the second group, $\bar{p}$= the average of the two proportions (0.68 + 0.4)/2

Figure 1. Calculating standardized difference for categorical variables as in example 1. From Altman (1982).

Dr Laofe Ogundipe is Specialist Registrar in Psychiatry, Lyme Brook Mental Health Centre, Newcastle

Correspondence to: Dr L Ogundipe, Deramore, Ercall Heath, Nr Tibberson, Shropshire TF10 8NQ

Step 1:

1. What is the standard deviation of thiamine in chronic alcoholics? Instead of the proportions required for binary outcome variables as in example 1, we need the standard deviation for a continuous outcome variable. This is also usually obtainable from the literature. Herve et al (1995) reported the mean thiamine level in alcohol dependent people to be 90.8 nmol/litre with a standard deviation of 25.7 nmol/litre.
2. What effect size do we consider to be clinically significant? We will take an increase of 25% of the mean level in the experimental group over the placebo group to be clinically significant. Therefore effect size = 0.25 x 90.8 = 22.7.
3. We will assume a probability of false positive error of 5% and of false negative error of 5% (giving a power of 95%).

Step 2: Calculate the standardized difference from the formula in *Figure 3*.

Step 3: Using Altman's nomogram, by drawing a line between the standardized difference of 0.88 and the required power of 0.95, the sample size appropriate for the study will be 68 subjects in each group. If a power of 80% is required, draw a line between 0.88 and 0.8 on Altman's nomogram and the sample size will be 40 in each group.

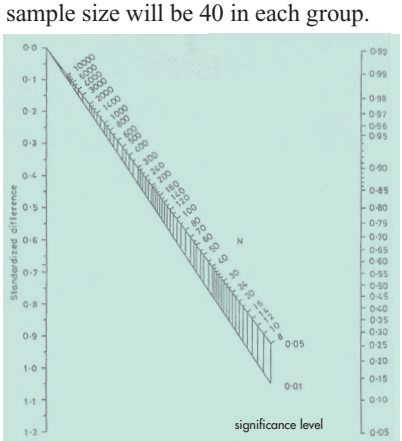

Figure 2. Nomogram for calculating sample size or power. From Altman (1980).

$$\text{Standardized difference} = \frac{\text{effect size}}{\text{sd}} = \frac{22.7}{25.7} = 0.88$$

Figure 3. Calculating standardized difference for continuous variable as in example 2. sd = standard deviation.

Once we know the standardized difference, we can always calculate the sample size required using Altman's nomogram. There are various formulae for calculating the sample size for comparative studies (Machin and Campbell, 1987). For simplicity, these are omitted here but once you use Altman's nomogram, sample size calculation can be very simple and graphical.

BASIC ASSUMPTIONS

For the above calculations it is assumed that the continuous variable has a normal distribution in the population, the categorical variable has normal approximation to the binomial distribution, and the two groups being compared are of equal sizes. If unequal sizes are desired, the calculated sample size should be modified according to the ratio of the two sample sizes.

DOES SIZE MATTER?

It is not always correct to reject a study, as is often done in journal clubs, because 'it is not powerful, it has too small a sample.' The power of a study is not an all-or-none phenomenon in which the study is of no use if the sample size is small. If a sample size is limited, the relevant questions to ask are:

1. Given the limited sample size, what effect size could be detected?
2. Given the limited sample size, what is the power of the study?

It is a good approach to consider the three items together: the power, the effect size and the sample size. It is important to know what power could be achieved with a given sample and using a statistical package this is easy. For example, a power of 80% requires only 60% of the sample size required by a power of 95% (*Figure 4*). A power of 50% requires only 30% of the sample size required by a power of 95% (MacRae, 1992). Similarly, a large effect size, for example 40% response rate, requires a smaller sample size than a small effect size, a response rate of 15%.

Power of study	Sample size required in comparison with that required for a power of 95%
95%	100%
90%	81%
80%	60%
50%	30%

Figure 4. Different sample sizes for different powers of study.

CONCLUSION

In the era of evidence-based medicine, sample size calculation is a skill that many doctors would find useful. A priori determination of sample size is important because with too small a sample, a clinically significant difference may not be detected and it will be impossible to make precise generalization to the parent population. With too large a sample, any difference, however small and clinically insignificant, will become statistically significant.

'Samples which are too small can prove nothing; samples which are large enough can prove anything' (Sackett et al, 1991).

We need to strike a balance between the cost and the usefulness of the sample. Sample size matters, but up to a certain point. Above a sample size of 200 for example, the sample size will need to increase considerably to make an appreciable difference to its usefulness. This is because in order to double the usefulness of a study, we need to increase the size fourfold. The usefulness of a study relates to the square of the sample size, not to the absolute sample size.

Whatever sample size is used, however, readers will want to understand the conjecture and the rationale behind the results of the study. HM

The author would like to thank Peter Jones, Professor of Mathematical Statistics, Dr Richard Hodgson, Consultant Psychiatrist, City General Hospital, Stoke-on-Trent and Dr Sayeed Haque, Medical Statistician, South Birmingham NHS Trust, for their comments on earlier drafts of this paper. Any remaining error is solely the authors responsibility. Figure 2 *is reproduced by kind permission of BMJ Publishing Group.*

Altman DG (1980) Statistics and ethics in medical research. III How large a sample? *Br Med J* **281:** 1336–8

Altman DG (1982) *How Large a Sample?* British Medical Association, London

Herve C, Beyne P, Letteron P, Delacoux E (1995) Comparison of erythrocyte transketolase activity with thiamine and thiamine phosphate ester levels in chronic alcoholic patients. *Clin Chim Acta* **234**: 91–100

Machin D, Campbell M (1987) *Statistical Tables for the Design of Clinical Trials.* Blackwell, Oxford

MacRae K (1992) *Statistics in Psychiatric Research.* WB Saunders, London

Sackett DHB, Haynes RB, Tugwell P, Guyatt G (1991) *Clinical Epidemiology: A Basic Science for Clinical Medicine.* Little Brown and Company, London