Singapore Med J 2003 Vol 44(4) : 172-174

Randomised Controlled Trials (RCTs) – Sample Size: The Magic Number?

Y H Chan

Clinical Trials and Epidemiology Research Unit
226 Outram Road
Blk A #02-02
Singapore 169039

Y H Chan, PhD
Head of Biostatistics

Correspondence to:
Y H Chan
Tel: (65) 6317 2121
Fax: (65) 6317 2122
Email: chanyh@cteru.gov.sg

INTRODUCTION

A common question posed to a biostatistician from a medical researcher is *"How many subjects do I need to obtain a significant result for my study?"*. **That magic number!** In the manufacturing industry, it is permitted to test thousands of components in order to derive a conclusive result but in medical research, the sample size has to be "just large enough" to provide a reliable answer to the research question. If the sample size is too small, it's a waste of time doing the study as no conclusive results are likely to be obtained and if the sample size is too large, extra subjects may be given a therapy which perhaps could be proven to be non-efficacious with a smaller sample size[1].

Another major reason, besides the scientific justification for doing a study, why a researcher wants an estimate of the sample size is to calculate the cost of the study which will determine the feasibility of conducting the study within budget. This magic number will also help the researcher to estimate the length of his/her study – for example, the calculated sample size may be 50 (a manageable number) but if the yearly accrual of subjects is 10 (assuming all subjects give consent to be in the study), it will take at least five years to complete the study! In that case a multicentre study is encouraged.

STATISTICAL THEORY ON SAMPLE SIZE CALCULATIONS

The **Null Hypothesis** is set up to be rejected. The philosophical argument is: it is easier to prove a statement is false than to prove it's true. For example, we want to prove that "all cats are black", and even if you point to me black cats everywhere, there's still doubt that a white cat could be lying under a table somewhere. But once you bring me a white cat, the hypothesis of 'all cats are black' is disqualified.

Hence if we are interested to compare two therapies, the null hypothesis will be "there is no difference" versus the **Alternative Hypothesis** of "there is a difference". From the above philosophical argument, not being able to reject the null hypothesis does not mean that that it is true (just that we do not have enough evidence to reject).

We want to reject the null hypothesis but could be committing a **Type I Error**: rejecting the null hypothesis when it's true. In a research study, there's no such thing as "my results are correct" but rather "how much error I am committing". For example, if in the population, there are actually no differences between two therapies (but we do not know, that's why we are doing the study) and after conducting the study, a significant difference was found which is given by $p<0.05$.

There are only two reasons for this significant difference (assuming that we have controlled for bias of any kind). One is, there's actually a difference between the two therapies and the other is by chance. The p-value gives us this "amount of chance". If the p-value is 0.03, then the significant difference due to chance is 3%. If the p-value is very small, then this difference happening by chance is "not possible" and thus should be due to the difference in therapies (still with a small possibility of being "wrong").

The other situation is not being able to reject the null hypothesis when it is actually false (**Type II Error**). As mentioned, the main aim of a clinical research is to reject the null hypothesis and we could achieve this by controlling the type II error[2]. This is given by the **Power** of the study (1 – type II error): the probability of rejecting the null hypothesis when it is false. Conventionally, the power is set at 80% or more, the higher the power, the bigger the sample size required.

To be conservative, a **two-sided test** (more sample size required) is usually carried out compared to a **one-sided test** which has the assumption that the test therapy will perform clinically better than the standard or control therapy.

SAMPLE SIZE CALCULATIONS

To estimate a sample size which will ethically answer the research question of an RCT with a reliable conclusion, the following information should be available.

Type of comparison[3]

Superiority trials

To show that a new experimental therapy is superior to a control treatment

Null Hypothesis: The test therapy is not better than the control therapy by a clinically relevant amount.

Alternative Hypothesis: The test therapy is better than the control therapy by a clinically relevant amount.

Equivalence trials

Here the aim is to show that the test and control therapies are equally effective.

Null Hypothesis: The two therapies differ by a clinically relevant amount.

Alternative Hypothesis: The two therapies do not differ by a clinically relevant amount.

Non-inferiority trials

For non-inferiority, the aim is to show that the new therapy is as effective but need not be superior compared to the control therapy. This is when the test therapy could be cheaper in cost or has fewer side effects, for example.

Null Hypothesis: The test therapy is inferior to the control therapy by a clinically relevant amount.

Alternative Hypothesis: The test therapy is not inferior to the control therapy by a clinically relevant amount.

A 1-sided test is performed in this case.

Type of configuration[4]

Parallel design

Most commonly used design. The subjects are randomised to one or more arms of different therapies treated concurrently.

Crossover design

For this design, subjects act as their own control, will be randomised to a sequence of two or more therapies with a washout period in between therapies. Appropriate for chronic conditions which will return to its original level once therapy is discontinued.

Type I error and Power[5]

The type I error is usually set at two-sided 5% and power is at 80% or 90%.

Effect size of therapies

The effect size specifies the accepted clinical difference between two therapies that a researcher wants to observe in a study.

There are three usual ways to get the effect size:

a. from past literature.
b. if no past literature is available, one can do a small pilot study to determine the estimated effect sizes.
c. clinical expectations.

To calculate the sample size, besides knowing the type of design to be used, one has to classify the type of the primary outcome.

Proportion outcomes

The primary outcome of interest is dichotomous (success/failure, yes/no, etc). For example, 25% of the subjects on the standard therapy had a successful outcome and it is of clinical relevance only if we observe a 40% (effect size) absolute improvement for those on the study therapy (i.e. 65% of the subjects will have a successful outcome). How many subjects do we need to observe a significance difference?

For a two-sided test of 5%, a simple formula to calculate the sample size is given by

$$\text{m (size per group)} = \text{c X } \frac{\pi_1(1-\pi_1) + \pi_2(1-\pi_2)}{(\pi_1-\pi_2)^2}$$

where c = 7.9 for 80% power and 10.5 for 90% power, π_1 and π_2 are the proportion estimates.

Thus from the above example, $\pi_1 = 0.25$ and $\pi_2 = 0.65$. For a 80% power, we have

$$\begin{aligned}\text{m (size per group)} &= 7.9 \text{ X } [0.25\ (1-0.25) + \\ &\quad 0.65\ (1-0.65)]/(0.25\text{-}0.65)^2 \\ &= 20.49\end{aligned}$$

Hence 21 X 2 = 42 subjects will be needed.

Table I shows the **required sample size per group** for π_1 & π_2 in steps of 0.1for powers of 80% & 90% at two-sided 5%.

Table I

π	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	199 (266)	62 (82)	32 (42)	20 (26)	14 (17)	10 (12)	7 (9)	5 (6)
0.2	–	294 (392)	82 (109)	39 (52)	23 (30)	15 (19)	10 (13)	7 (9)
0.3		–	356 (477)	93 (125)	42 (56)	24 (31)	15 (19)	10 (12)
0.4			–	388 (519)	97 (130)	42 (56)	23 (30)	14 (17)
0.5				–	388 (519)	93 (125)	39 (52)	20 (26)
0.6					–	356 (477)	82 (109)	32 (42)
0.7						–	294 (392)	62 (82)
0.8							–	199 (266)

Numbers in ( ) are for 90% power

Continuous outcomes

Two independent samples

The primary outcome of interest is the mean difference in an outcome variable between two treatment groups. For example, it is postulated that a good clinical response difference between the active and placebo groups is 0.2 units with an SD of 0.5 units, how many subjects will be required to obtain a statistical significance for this clinical difference?

A simple formula, for a two-sided test of 5%, is

$$\text{m (size per group)} = \frac{2c}{\delta^2} + 1$$

where $\delta = \frac{|\mu_2 - \mu_1|}{\sigma}$ is the standardised effect size and

μ_1 and μ_2 are the means of the two treatment groups
σ is the common standard deviation
c = 7.9 for 80% power and 10.5 for 90% power

From the above example, $\delta = 0.2/0.5 = 0.4$ and for a 80% power, we have m (size per group) = (2 X 7.9)/(0.4 X 0.4) + 1 = 99.75

Hence 100 X 2 = 200 subjects will be needed.

Table II shows the **required sample size per group** for values of δ in steps of 0.1 for powers of 80% & 90% at 2-sided 5%

Paired samples

In this case, we have the pre and post mean difference of the two treatment groups and a simple formula is

$$\text{Total sample size} = \frac{c}{\delta^2} + 2$$

Table III shows the total size required for values of δ in steps of 0.1 for powers of 80% and 90% at two-sided 5%.

SAMPLE SIZE SOFTWARE

There are many sample size calculations software available in the Internet and even on most computers. The main point to note in using a software is to understand the proper instructions of getting the sample size. One could enter some data into a program, and unless an error message is obtained, it is most likely the magic number being generated is accepted by the user. For this number to be "correct", the right formula must be used for the right type of design and primary outcome. It is important to note that nearly all the programs would provide the sample size for one group and not the total (except for paired designs).

A simple-to-use PC-based sample size software, affordable in cost, is Machin's et al[6] Sampsize version 2.1 but it could only be installed for Windows 98 and below. Software with network capabilities are SPSS (www.spss.com), STATA (www.stata.com) and Power & Precision (www.PowerAnalysis.com), just to mention a few. Thomas & Krebs[7] gave a review of the various statistical power analysis software, comparing the pros and cons.

CONCLUSIONS

This article has thus far covered the basic discussions for simple sample size calculations with two aims in mind. Firstly, a researcher could calculate his/her own sample size given the types of design and measures of outcome mentioned above; secondly, it is to provide some knowledge on what information will be needed when coming to see a biostatistician for sample size determination. If one is interested in doing an equivalence/non-inferiority study or with survival outcomes analysis, it is recommended that a biostatistician should be consulted.

REFERENCES

1. Fayers PM & D Machin. Sample size: how many subjects patients are necessary!. British Journal of Cancer 1995; 72:1-9.
2. Muller KE & Benignus VA. Increasing scientific power with statistical power. Neurotoxicology & Teratology, 1992; 14:211-9.
3. Schall R, Luus H & Erasmus T. Type of comparison, introduction to clinical trials, editors Karlberg J & Tsang K, 1998; pp:258-66.
4. Chan YH. Study design considerations — study configurations, introduction to clinical trials, editors Karlberg J & Tsang K. 1998; pp:249-57.
5. Thomas L & Juanes F. The importance of statistical power analysis: an example from animal behaviour. Animal Behaviour, 1996; 52:856-9.
6. D Machin, M Campbell, Fayers P & Pinol A. Sample size tables for clinical studies, 2nd edition. Blackwell Science, 1997.
7. L Thomas & CJ Krebs. A review of statistical power analysis software. Bulletin of the Ecological Society of America 1997, 78(2): 126-39.

Table II

	δ								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
80% power	1,571	394	176	100	64	45	33	26	21
90% power	2,103	527	235	133	86	60	44	34	27

Table III

	δ								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
80% power	792	200	90	52	34	24	19	15	12
90% power	1,052	265	119	68	44	32	24	19	15