Kidney Disease and Population Health

nephron Clinical Practice

Nephron Clin Pract 2011;118:c319–c323
DOI: 10.1159/000322830

Published online: February 3, 2011

Sample Size Calculations

Marlies Noordzij[a] Friedo W. Dekker[b] Carmine Zoccali[c] Kitty J. Jager[a]

[a]ERA-EDTA Registry, Department of Medical Informatics, Academic Medical Center, University of Amsterdam, Amsterdam, and [b]Department of Clinical Epidemiology, Leiden University Medical Centre, Leiden, The Netherlands; [c]CNR-IBIM, Clinical Epidemiology and Pathophysiology of Renal Diseases and Hypertension, Renal and Transplantation Unit, Ospedali Riuniti, Reggio Calabria, Italy

Key Words
Sample size · Power · Study design · Epidemiology · Statistics · Nephrology

Abstract
The sample size is the number of patients or other experimental units that need to be included in a study to answer the research question. Pre-study calculation of the sample size is important; if a sample size is too small, one will not be able to detect an effect, while a sample that is too large may be a waste of time and money. Methods to calculate the sample size are explained in statistical textbooks, but because there are many different formulas available, it can be difficult for investigators to decide which method to use. Moreover, these calculations are prone to errors, because small changes in the selected parameters can lead to large differences in the sample size. This paper explains the basic principles of sample size calculations and demonstrates how to perform such a calculation for a simple study design.

Introduction

The sample size is the number of patients or other experimental units that should be included in a study to be able to answer the research question. The main aim of sample size calculations is to determine the number of participants required to detect a clinically relevant treatment effect. Optimizing the sample size is extremely important. If the sample size is too small, one may not be able to detect an important effect, while a sample that is too large may be a waste of time and money. Determining the sample size is one of the first steps in the design of a trial, and methods to calculate the sample size are explained in several conventional statistical textbooks [1, 2]. However, it is difficult for investigators to decide which method to use, because there are many different formulas available, depending on the study design and the type of outcome studied. Furthermore, these calculations are sensitive to errors, because small differences in selected parameters can lead to large differences in sample size. In this paper, we explain the basic principles of sample size calculations based on an example describing a hypothetical randomized controlled trial (RCT) on the effect of erythropoietin (EPO) treatment on anaemia in dialysis patients.

The Basic Principles of Clinical Studies: An Example

Suppose one wishes to study the effect of EPO treatment on haemoglobin levels in anaemic dialysis patients (haemoglobin <13 g/dl in men and <12 g/dl in women) [3]. These patients are randomized to receive either EPO

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

1660–2110/11/1184–0319$38.00/0

Accessible online at:
www.karger.com/nec

Marlies Noordzij, PhD
ERA-EDTA Registry, Department of Medical Informatics
Academic Medical Center, University of Amsterdam, PO Box 22700
NL–1100 DE Amsterdam (The Netherlands)
Tel. +31 20 566 78 73, Fax +31 20 691 98 40, E-Mail m.noordzij@amc.uva.nl

Table 1. Overview of the components required for sample size calculations

Component	Synonyms	Definition	Conventional values
Alpha	type I error p value significance level	the chance of a false-positive result	0.05 or 0.01
Beta	type II error	the chance of a false-negative result	0.20 or 0.10
Power (1 – beta)		the chance of finding a statistically significant difference between the groups when this difference exists in reality	0.80 or 0.90
Minimal clinically relevant difference	MCRD	the minimal difference between the groups that a researcher considers clinically relevant and biologically plausible	–
Variance	standard deviation[1]	the variability of the outcome measure	–

[1] In the case of a continuous outcome.

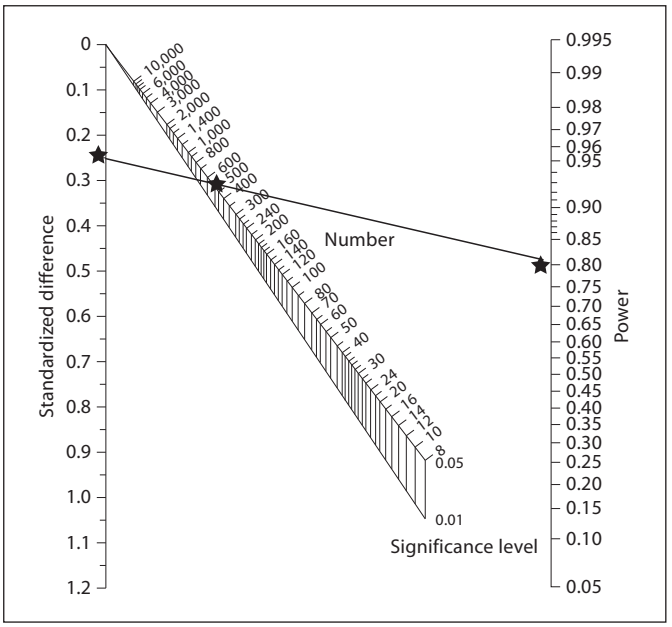

Fig. 1. Nomogram for the calculation of sample size or power (adapted from Altman [4], with permission).

or placebo treatment. The primary outcome of this study is a continuous one, namely haemoglobin level. After the intervention period, haemoglobin levels in the treated and placebo groups are compared. Of course, we hope to find a statistically significant difference in haemoglobin level between the group treated with EPO and the placebo group. Intuitively, we expect that the more patients we include in our study, the more significant our difference will be. To determine how many patients we actually need to include in our RCT to detect a clinically relevant effect of EPO, we need to perform a sample size calculation or estimation.

In the case of a simple study design, such as our RCT on EPO treatment, a graphical method can be used to estimate the sample size required for the study. Figure 1 shows an example of a nomogram for sample size estimation as published by Altman [4]. From this nomogram, we can read that we need a few parameters to estimate the required sample size, i.e. the standardized difference in a study, the power and the significance level.

To be able to use such a nomogram or another method for sample size calculation, it is helpful to have some understanding of the basic principles of clinical studies. When performing a clinical study, an investigator usually tries to determine whether the outcomes in two groups are different from each other. In most cases, individuals treated with a certain drug or other health intervention are compared with untreated individuals. In general, the 'true effect' of a treatment is the difference in a specific outcome variable, in our example haemoglobin level, between treated and untreated individuals in the population. However, in clinical research, effects are usually studied in a study sample instead of in the whole population and as a result two fundamental errors can occur, which are called type I and type II errors. The values of these type I and type II errors are important components in sample size calculations. In addition, it is necessary to have some idea of the results expected in a study to be able to calculate the sample size. These components of sample size calculations are described below and are summarized in table 1.

Components of Sample Size Calculations

Type I Error (Alpha)

The type I error, also called alpha, the significance level or the p value, represents the chance that a researcher concludes that two groups differ when in reality they do not or, in other words, the chance of a false-positive conclusion. Most commonly, alpha is fixed at 0.05, meaning that a researcher desires a less than 5% chance of drawing a false-positive conclusion.

Power

Investigators can also draw a false-negative instead of a false-positive conclusion. They then conclude that there is no difference between two groups when in fact there is. The chance of a false-negative conclusion is called a type II error (beta). Beta is conventionally set at a level of 0.20, which means that a researcher desires a less than 20% chance of a false-negative conclusion.

For the calculation of the sample size, one needs to know the beta or the power of a study. The power is the complement of beta, i.e. 1 – beta. This means that the power is 0.80 or 80% when beta is 0.20. The power represents the chance of avoiding a false-negative conclusion or, in other words, the chance of detecting a specified effect if it really exists.

Minimal Clinically Relevant Difference

The minimal clinically relevant difference is the smallest effect between the studied groups that the investigator wants to be able to detect. It is the difference that the investigator believes to be clinically relevant and biologically plausible. In the case of a continuous outcome variable, the minimal clinically relevant difference is a numerical difference. For instance, if systolic blood pressure were the outcome of a trial, an investigator could choose a difference of 10 mm Hg as the minimal clinically relevant difference. If a trial had a binary outcome, such as the development of catheter-related bacteraemia (yes/no), a relevant difference between the event rates in both treatment groups should be estimated. For example, the investigator could choose a difference of 10% between the percentage of infections in the treatment group and that in the control group as the minimal clinically relevant difference.

Variability

Finally, the sample size calculation is based on the population variance of the outcome variable. In general, the greater the variability of the outcome variable, the larger the sample size required to assess whether an observed effect is a true effect. In the case of a continuous outcome variable, the variability is estimated by means of the standard deviation (SD). The variance is usually unknown, and therefore investigators often use an estimate obtained from a pilot study or a previously performed study.

Estimating Sample Size Using Graphical Methods

Now that we understand the separate components of sample size calculations, we can use the nomogram as published by Altman [4] (fig. 1) to estimate the sample size required for our RCT on EPO treatment in dialysis patients. Suppose we consider a difference in haemoglobin level of 0.50 g/dl between the group treated with EPO and the placebo group as clinically relevant and we specified such an effect to be detected with 80% power (0.80) and a significance level alpha of 0.05. The last value we need for the calculation is the population variance. Previously published reports on similar experiments using similar measuring methods in similar patients suggest that our data will be approximately normally distributed, and we estimate that the SD will be around 1.90 g/dl.

To use this nomogram, one needs the standardized difference, which can simply be calculated by dividing the minimal clinically relevant difference (0.50 g/dl) by the SD in the population (1.90 g/dl). For our example, this yields 0.50/1.90 = 0.26. We can now use the nomogram to estimate the sample size by drawing a straight line between the value of 0.26 on the scale for the standardized difference and the value of 0.80 on the scale for power and reading off the value on the line corresponding to alpha = 0.05, which gives a total sample size of 450, i.e. 225 per group. However, although this nomogram seem to work well for our example, one should keep in mind that these graphical methods often make assumptions about the type of data and statistical tests to be used. In many cases, it is therefore more appropriate to apply statistical formulas to calculate the required sample size.

Estimating Sample Size Using a Formula

Based on our trial example, we will now demonstrate how sample size can be calculated. We will use the simplest formula for a continuous outcome variable, such as haemoglobin level, and equal sample sizes in the treated (EPO) and control (placebo) groups [5]:

$$N = 2[(a + b)^2\sigma^2]/(\mu_1 - \mu_2)^2$$

Table 2. Multipliers for conventional values of alpha and beta

	Alpha		Beta			
	0.05	0.01	0.20	0.10	0.05	0.01
Multiplier	1.96	2.58	0.842	1.28	1.64	2.33

where N is the sample size in each of the groups, μ_1 is the population mean in treatment group 1, μ_2 is the population mean in treatment group 2, $\mu_1 - \mu_2$ is the minimal clinically relevant difference, σ^2 is the population variance (SD), a is the conventional multiplier for alpha and b is the conventional multiplier for power.

Again, we chose a power of 0.80, an alpha of 0.05 and a minimal clinically relevant difference in haemoglobin level between the two groups of 0.50 g/dl ($\mu_1 - \mu_2$). Because we chose the significance level alpha to be 0.05, we should enter the value 1.96 for a in the formula. Similarly, because we chose beta to be 0.20, the value 0.842 should be filled in for b in the formula. These multipliers for conventional values of alpha and beta can be found in table 2.

The final value we need for the calculation is the population variance (SD) of 1.90 g/dl. Entering all values in the formula yields:

$$2 \times [(1.96 + 0.842)^2 \times 1.90^2]/0.50^2 = 226.7.$$

This means that a sample size of 227 subjects per group is needed to answer the research question. This sample size is in line with the number of 225 subjects per group which we estimated from the nomogram.

Different Study Designs and Situations

In our example, the outcome variable is a continuous one. However, in many trials the outcome variable may be, for example, binary (e.g. yes/no) or survival (e.g. time to event). If this is the case, one still needs the four basic components, but different formulas should be used and other assumptions may be required.

Also, for different types of study designs, different methods for sample size calculation should be used. First of all, it is important to realize that sample size calculations are not required in all types of studies. These calculations are especially of interest in the context of hypothesis testing, as in trials aiming to show a difference between groups. If one just wants to know the occurrence of a certain disease (incidence or prevalence), as is the case in registry studies, sample size calculation is probably not necessary or even not possible. Also, for observational studies aimed at the discovery or exploration of effects, sample size is not of major importance.

So, sample size calculations are especially of interest in the design of an RCT. Because a lot of money is invested in this type of study, it is important to be sure that a sufficient number of patients are included in the study to find a relevant effect if it exists. However, sample size calculations are also sometimes needed in studies with other designs, such as case-control or cohort studies, and different formulas for sample size calculation are required in these cases [6, 7]. In the case of a clinical trial testing the equivalence of two treatments rather than the superiority of one over the other, another approach for sample size calculation is necessary. These equivalence or non-inferiority trials usually demand greater sample sizes [8].

Several software programs such as nQuery Advisor and PASS can assist in sample size calculations for different types of data and study designs. In addition, there are some websites that allow free sample size calculations, but not all of these programmes are reliable. However, because many methods are not straightforward, we recommend consulting a statistician in all but the most basic studies.

Difficulties in Sample Size Calculations

Although sample size calculations are useful, especially because they force investigators to think about the planning and likely outcomes of their study, they have some important drawbacks. Firstly, some knowledge of the research area is needed before one can perform a sample size calculation, and lack of this knowledge is often a problem. Secondly, it is necessary to choose a primary outcome in order to calculate the required sample size, while many clinical trials aim to study several outcomes. Researchers often change the planned outcome(s) after their study has begun, making the reported p values invalid and potentially misleading [9]. Furthermore, the required sample size is very sensitive to the values the investigator chooses for the basic components in the calculation. Based on our example, namely an RCT on EPO treatment, we show how selection of alpha, beta and the minimal clinically relevant difference can influence the results of sample size calculations. Choosing a higher power leads to a larger sample size. Since beta is the complement of the power, a higher power automatically

means a lower beta, indicating a lower chance of drawing a false-negative conclusion. If we were to choose a power of 0.90 instead of 0.80, the conventional multiplier for beta in the formula would be 1.28 instead of 0.842 (table 1), and this would yield a larger sample size:

$$2 \times [(1.96 + 1.28)^2 \times 1.90^2]/0.50^2 = 303.2.$$

Similarly, choosing a lower significance level alpha, indicating a lower chance of drawing a false-positive conclusion, leads to a larger sample size. So, if we were to choose a lower alpha of 0.01 instead of 0.05, we would have to use 2.58 as the conventional multiplier for alpha instead of 1.96, resulting in a larger sample size:

$$2 \times [(2.58 + 0.842)^2 \times 1.90^2]/0.50^2 = 338.2.$$

These calculations with different values for alpha and beta clearly show that using a sample size that is too small leads to a higher risk of drawing a false-positive or false-negative conclusion. Finally, the choice of the minimal clinically relevant difference has the largest influence. The smaller the difference one wants to be able to detect, the larger the required sample size. If we aimed to detect a difference of 0.3 g/dl instead of 0.5 g/dl, the calculation would yield:

$$2 \times [(1.96 + 0.842)^2 \times 1.90^2]/0.30^2 = 629.8.$$

These examples show the most important drawback of sample size calculations; investigators can easily influence the result of their sample size calculations by changing the components in such a way that they need fewer patients, as that is usually what is most convenient to the researchers. For this reason, sample size calculations are sometimes of limited value.

Furthermore, more and more experts are expressing criticism of the current methods used. They suggest introducing new ways to determine sample size, for example estimating the sample size based on the likely width of the confidence interval for a set of outcomes [9]. However, consensus about these alternative methods has not yet been reached.

Conclusions

Because there are many different methods available to calculate the sample size required to answer a particular research question and because the calculations are sensitive to errors, performing a sample size calculation can be complicated. We therefore recommend caution when performing the calculations or asking for statistical advice during the designing phase of the study.

Acknowledgements

The research leading to the findings reported herein has received funding from the European Community's Seventh Framework Programme under grant agreement No. HEALTH-F2-2009-241544.

References

1 Altman DG: Practical Statistics for Medical Research. London, Chapman & Hall, 1991.
2 Bland M: An Introduction to Medical Statistics, ed 3. Oxford, Oxford University Press, 2000.
3 World Health Organization: Nutritional Anemia. Report of a WHO Scientific Group. Geneva, World Health Organization, 1968.
4 Altman DG: Statistics and ethics in medical research. III. How large a sample? Br Med J 1980;281:1336–1338.
5 Florey CD: Sample size for beginners. BMJ 1993;306:1181–1184.
6 Machin D, Campbell M, Fayers P, Pinol A: Sample Size Tables for Clinical Studies, ed 2. London, Blackwell Science, 1997.
7 Lemeshow S, Levy PS: Sampling of Populations: Methods and Applications, ed 3. New York, John Wiley & Sons, 1999.
8 Christensen E: Methodology of superiority vs. equivalence trials and non-inferiority trials. J Hepatol 2007;46:947–954.
9 Bland JM: The tyranny of power: is there a better way to calculate sample size? BMJ 2009;339:1133–1135.