Was the study big enough? Two "café" rules

WHY IS A SMALL STUDY A PROBLEM?

When reading an article, we often wonder whether the study was large enough. If a study does not find a statistically significant effect (e.g., at $P < 0.05$), it may be because the study was too small or because there actually is no true effect. You should check whether the CIs show that the data are consistent with a clinically important effect, even though the effect was not "statistically significant."

HOW CAN WE TELL WHETHER THE STUDY WAS TOO SMALL?

The CI quantifies the random error and thus the uncertainty associated with the use of study results to draw inferences about wider population effects. The upper and lower limits give us the plausible range of population values. If the CI is very wide, then there is little certainty that the study result is a good estimate, and the study was probably too small. However, if the CI does not cross the value of clinical significance, then the data are not consistent with a clinically important effect no matter how large or small the study, how wide or narrow the CI, and how statistically significant the effect. Because studies sometimes do not report a CI, it is helpful to have an approximate idea of the size requirements of different types of studies.

SAMPLE SIZE AND CIS

The larger a study, the smaller the random error (quantified by the standard error [SE]) and therefore the tighter the CIs (the 95% CI for the true population value is calculated as estimate ± 1.96 SE of the estimate) (Figure 1). The upper and lower limits of the CI give us the plausible range of values for the true, but unknown, population effect. If this is too wide for comfort, then the study is too small, and even large effects may not reach statistical significance.

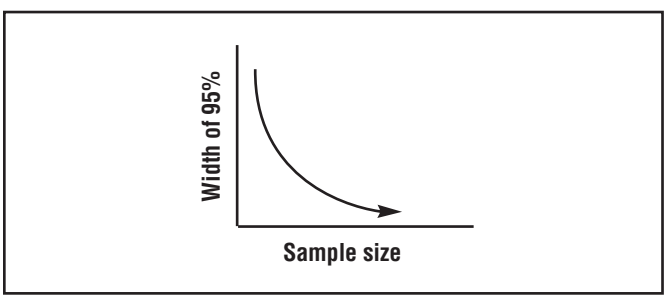

Figure 1. Relation between sample size and CIs.

WHAT IS STUDY "POWER"?

Sometimes studies will report their "power" instead of, or in addition to, CIs. The power is the prestudy probability (given the available knowledge before data collection) that the study will detect (at a certain significance level, such as $P < 0.05$) a minimum effect regarded as clinically significant. Power is calculated before the study to determine the required sample size. After the study is conducted, post hoc power calculations should not be calculated. Once the size of the effect is known, CIs should be used to express the uncertainty of the study estimate.

In this editorial, we will provide you with 2 "café" rules (for when you are discussing studies over an espresso) and then discuss the ideas behind them and some resources for more exact calculations.

HOW DO WE KNOW THE REQUIRED SAMPLE SIZE?

It is helpful to have an approximate idea of the sample size requirements for different types of studies. The first approximate rule is the 50–50 rule for studies looking at such dichotomous (present or absent) outcomes as mortality, hospitalizations, or remissions.

Rule 1. A study with a dichotomous outcome measure needs (approximately) 50 events to occur in the control group to have an 80% power of detecting a 50% relative risk reduction (RRR).

Note that the rule is about the number of persons *with events*, not the number of persons in the study. The events provide the information. For example, if a large study has follow-up that is too short for any deaths to occur, then there is no information about mortality. Either a larger or a longer study is needed.

Events can be increased by choosing higher-risk patients, by increasing the time of follow-up, or by increasing the sample size. The "50" events are approximate, and Table 1 shows how this compares with exact sample size calculations for various control group event risks.

Table 1. Sample size for a dichotomous outcome using the 50–50 rule and "exact" methods

Control group risk	Rule 1 sample size	"Exact" sample size
20%	2 × 250	2 × 219
10%	2 × 500	2 × 474
5%	2 × 1000	2 × 984
1%	2 × 5000	2 × 5066

What happens if we want to detect a smaller difference? The rule here is that to detect a difference one half the size, we need to quadruple the sample size. This is illustrated in Figure 2 for the case of a dichotomous outcome with a 10% control group rate.

For a 50% RRR (RR 0.50), the sample size is 474 persons per group, but to detect a 25% RRR (RR 0.75), we need 2084 persons per group.

(continued on page A-9)

(*continued from page A-8*)

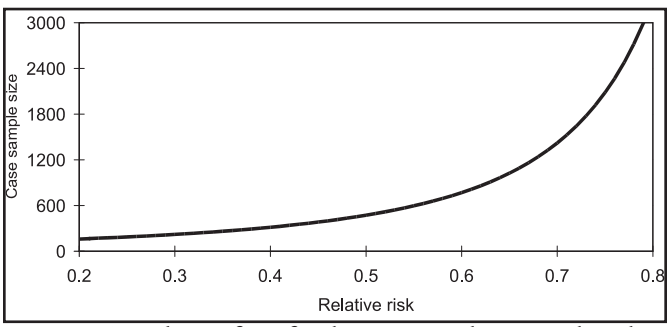

Figure 2. Sample size for a fixed 10% control group risk and a varying relative risk.

Rule 2. A study with a continuous outcome measure needs about 50 persons per group.

With a continuous outcome measure, such as blood pressure or depression score, each person contributes information. So by analogy to the 50–50 rule, we only need 50 persons, because they all have "events" (i.e., outcome measurements). A more exact calculation would require us to specify the number of SDs of difference we wish to reliably detect. Table 2 shows this for a range of SDs.

Table 2. Sample size for different minimum differences specified in SD

Difference	Sample size
1.0 SD	2 × 17
0.7 SD	2 × 33
0.5 SD	2 × 64
0.3 SD	2 × 175
0.1 SD	2 × 1571

We have only outlined some of the issues and some approximate calculations in this editorial. This information should help you to determine whether a study you are reviewing is "big enough" to support a firm conclusion but should not be relied on if you are designing your own study.

WHAT IS BEHIND THESE CALCULATIONS?

The 4 factors that go into a sample size calculation are 1) the minimum difference you think is worth detecting (the "clinically important" difference), 2) the variance (for studies with continuous outcome measures) or the control group risk (for studies of event outcomes), 3) the acceptable level of significance (usually 0.05), and 4) the desired power of the study (the chance that it will detect the minimum difference as statistically significant, usually 80% or 90%). The first 2 can be thought of as the "signal" we are looking for and the "noise" we have to detect it in. The last 2 correspond to the 2 types of error that can result from a hypothesis test discussed previously (1).

Type I errors (α) arise when the null hypothesis (H_0) is rejected when it is true, whereas type II errors (β) arise when H_0 is accepted when it is false. Figure 3 outlines the statistical issues. If you were designing your own study, a number of good books and programs can assist with sample size calculations.

The power of a study is indicated by the area under H_A (hypothesis that there is an effect) to the right of the critical value (i.e., $1 - \beta$). Clearly, the power of the study (the chance that it will detect the minimum difference as statistically significant) will be affected by 1) the minimum difference you think is worth detecting (the distance between the central values of the 2 curves under a hypothesis of H_0 and H_A), 2) the variability of the data (for studies with continuous outcome measures) or the control group risk (for studies of event outcomes), 3) the level of statistical significance (α, the acceptable *P* value, usually 0.05), and 4) the sample size of the study.

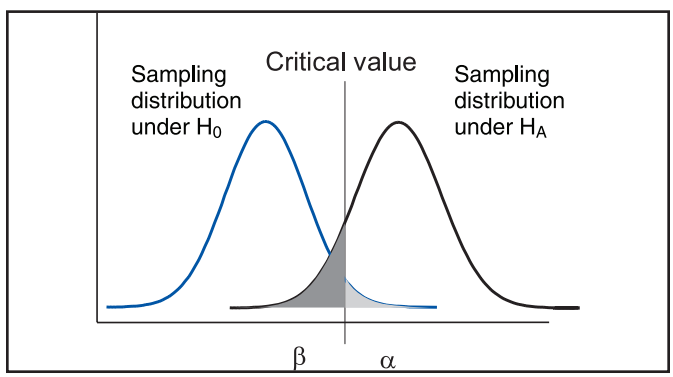

Figure 3. Frequency distributions under H_0 (*blue line*) and H_A (*black line*) showing the probabilities of making a type I (α) or type II (β) error.

RESOURCES

Software: There are many available programs, but a good free one is Power, which is downloadable from http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize.

A good book that covers various situations and has its own software is Machin D, Campbell M, Fayers P, Pinol A. **Sample size tables for clinical studies.** 2nd ed. London, Edinburgh, Malden, and Carlton: Blackwell Science; 1997.

Paul Glasziou, MBBS, PhD
Centre for Evidence-Based Medicine, University of Oxford
Oxford, UK

Helen Doll, MSc, DPhil
Department of Public Health, University of Oxford
Oxford, UK

Reference

1. **Doll H, Carney S.** Statistical approaches to uncertainty: *p* values and confidence intervals unpacked. ACP J Club. 2006 May-Jun;144:A8.