BMJ

BMJ 2012;345:e5278 doi: 10.1136/bmj.e5278 (Published 23 August 2012)

RESEARCH METHODS & REPORTING

Sample size calculations: should the emperor's clothes be off the peg or made to measure?

Ethics committees require estimates of sample size for all trials, but statistical calculations are no more accurate than estimates from historical data. **Geoffrey Norman and colleagues** propose some "one size fits all" numbers for different study designs

Geoffrey Norman *statistician and cognitive psychologist*[1], Sandra Monteiro *graduate student*[2], Suzette Salama *chair of McMaster research ethics board*[3]

[1]Clinical Epidemiology and Biostatistics, McMaster University, MDCL 3519, 1280 Main St W, Hamilton, Ontario L8S 2T1, Canada; [2]Department of Psychology, McMaster University, Hamilton ; [3]Medicine, Hamilton Health Sciences, Hamilton

Conventional wisdom dictates that it is unethical to conduct a study that is so large that excess numbers of patients are exposed or so small that clinically important changes cannot be detected.[1] This implies that there is some optimal sample size that can be calculated using statistical theory and information from previous research. But the choice of sample size is usually a compromise between statistical considerations, which always benefit from increased sample size, and economic or logistical constraints.[2]

Only rarely is sufficient information available to make informed decisions. Moreover, despite the illusion of precision that arises from the application of arcane statistical formulas, in many situations the choice of inputs—the expected treatment effect, the standard deviation, and the power—are subject to considerable uncertainty. As a result, sample size calculations may vary widely.

We argue that, in the absence of good data, it would be better to determine sample size by adopting norms derived from historical data based on large numbers of studies of the same type. We show that for many common situations we can define defensible, evidence based, ranges of sample sizes.

An example

Imagine that you decide to do a study to see if control of primary hypertension is improved by home monitoring. You visit your local statistician for a sample size calculation, as the ethics board insists. The following are some key questions that he or she will ask and some tentative answers.

What is the distribution of blood pressure in the population you intend to study?

One study design might be to randomise people to treatment and control groups, put one group on monitors for a few months, then measure their blood pressure. We would then compare blood pressure in the two groups. To compute sample size, we need to know the standard deviation of systolic (or diastolic) blood pressure in the group we are studying. One recent meta-analysis of interventions to control hypertension gave values of 15-17 mm Hg.[3]

How much do you think your treatment will affect systolic blood pressure?

The most reasonable answer is, "How do I know? That's why I'm doing the study." Regrettably, you have to know to calculate sample size. Fortunately, a recent Cochrane review of 12 randomised trials with over 1200 patients per group provides a guide.[3] The mean difference in systolic pressure was 2.53 mm Hg. Individual study results ranged from a drop of 26.0 mm Hg to a gain of 5.0 mm Hg. If we eliminate the two studies with very small samples of 9 and 18 and use the three largest observed differences, we end up with a mean drop of 6.9 mm Hg from studies with samples of 48, 55, and 76. Conversely, the studies with the three smallest treatment effects (n=123, 326, and 72) showed an average benefit of 1 mm Hg.

What α and β levels do you want?

That's easy. Convention dictates that α (level of statistical significance) is 0.05 and β (the probability of a type II error: rejecting the null hypothesis when the alternate hypothesis is true) is 0.20. (However easy and universal these may be, the choice of a power of 0.80 is logically unsupportable, as shown by Bacchetti.[2] But for the sake of convention, we will proceed.)

We can now do the calculation (box). If we take the extremes, the smallest sample size, based on a reduction of 6.9 mm Hg and an SD of 15, equals 75 per group. The largest, for a 1 mm Hg drop and an SD of 17, equals 4624. The overall average drop of 2.53 corresponds to a sample size of 722. These estimates differ by a factor of 60 even though this was a "best

Correspondence to: G Norman norman@mcmaster.ca

case" situation, in which all studies had reasonable sample size and were viewed as sufficiently homogeneous to be included in a systematic review.

Critics might argue that the choice of the three smallest and largest differences was arbitrary and extreme, but we used it to illustrate the point. We could have used alternative strategies, such as weighting by sample size. But the fact is that all the studies were derived from a Cochrane systematic review, all were examining a single question, all were deemed of sufficient quality to be included in the systematic review, and all were used in the final calculation in the review. On that basis, all are equal candidates for inclusion in a sample size calculation.

In more representative situations where data are lacking, there would be even more "wiggle room." Virtually all statisticians who have been engaged in this activity describe multiple iterations until the computed sample size converges to a desired result.

Other approaches to sample size

Interestingly, inclusion of sample size as an element of research ethics is far more pervasive in health sciences than in other areas such as social sciences. In Canada, the authoritative Tricouncil statement on ethics—produced by the three federal research councils for health, social sciences, and physical science—mentions sample size only in the section on qualitative research. In non-medical disciplines it seems that judgments about adequacy of sample size, if raised at all, are resolved by arguments along the lines of "studies in this area typically use sample sizes about this large." Perhaps this is because biomedical research is more likely to expose participants to real, occasionally life threatening, risks.

Clinical trials also typically include large numbers of people, and the cost per person is high. All these factors increase pressure to arrive at the "right" sample size. However, as Bacchetti argues, any sample size is a compromise and high risks and costs really should be seen as factors to reduce sample size.[2]

Calculation of sample size seems unlikely to disappear, whatever the philosophical flaws in the argument. But in view of the imprecision of the estimates, we need a fundamental rethink of the approach to avoid the calculation being seen (with justification) as simply another hoop to jump through to obtain ethical approval.

We propose a new approach that establishes norms for particular study questions and designs, while not preventing the investigator from producing an individual estimate when the evidence warrants it. The idea stems from a proposal by Bacchetti[2] and from the commonsense idea to use existing data when available to increase precision of estimates.

As we have seen, individual estimates, even when based on previous studies, can vary wildly. Still, it is a large leap of faith to presume that "cultural" norms based on previous research are more defensible. However, in some areas there is good evidence of the magnitude of treatment effects that might be expected. For studies of two groups in which the outcome is either a measured (interval or ratio) dependent variable or a difference in proportions, we argue that there is sufficient evidence from various sources to compute norms for sample sizes that may apply to all such studies. For some regression methods, there are "rules of thumb" that do not require specific information.

Sample size norms for different designs

Differences between groups

Measured outcome variable

The most basic study design on which to base a sample size calculation resembles our introductory example. Participants are assigned to two groups: one group receives a treatment and the other a placebo, the outcome is measured on a continuous scale (such as blood pressure, range of motion, creatinine concentration), and the means of the two groups are compared.

The classic text for sample size and power calculations is Cohen's *Statistical Power Analysis for the Behavioral Sciences*.[5] The basis for the sample size calculation is the effect size—the treatment difference divided by the standard deviation within groups. Based on his experience, Cohen proposed that a small effect size is 0.2, a medium is 0.5, and a large is 0.8. On this basis, the norm for sample sizes would be 400, 64, and 25 respectively.

Although equating of 0.2, 0.5, and 0.8 with small, medium, and large effect sizes has now become almost axiomatic, Cohen did not view them that way.[2] He spends considerable time arguing the reasonableness of these estimates by comparing them with other indices such as overlap of distributions, correlations, and percent of variance, as well as anchoring to commonly accepted scales like intelligence quotient (IQ). Halpern went further and argued that, in the absence of any more specific data, sample size could be based on a medium effect size (0.5), so the default would be n=64.[6]

We are not suggesting that sample size estimates should be based on small, medium, or large effect sizes. Rather we should use norms within research communities—explicitly using archival data to identify representative and expected normative values of effect size. As one example, Lipsey and Wilson examined 302 meta-analyses of 13 000 studies looking at educational and psychological interventions.[7] They found a mean effect across all studies of 0.50, with a standard deviation of 0.29. This large series is quite consistent with Cohen's original estimates and results in a sample size between 26 and 363, again with a best estimate of 64.

Another approach to sample size calculation involves consideration of the minimally important difference (MID). There are several approaches to determining the MID, most commonly by observing change in a cohort of patients who are judged to have had minimal change in their quality of life. One study looked at 38 studies estimating the MID in health related quality of life.[8] The mean MID over 62 estimates was 0.495, with an SD of 0.15. From this survey, the range of sample sizes (±1SD) would be 38 to 134, with a best estimate of 65.

Both these examples estimate that a study with two groups and a continuous outcome might use a sample size of about 60 per group, although anything within the range 25 to 400 is acceptable, with larger samples for treatment-treatment comparisons and smaller samples for comparison with no treatment. They also happen to be consistent with Cohen's medium effect size, although this cannot be seen as justification to adopt 0.5 as a standard since many clinical interventions have much smaller effect sizes.[7]

Binary outcome variable— proportions

Many clinical trials use a binary (death/no death, event/no event) outcome. The sample size formula differs and is dependent on both the base rate of the outcome and the risk reduction. If we can establish a normative range of relative risk reduction to be

Calculating sample size

For a difference between two groups, sample size=$16\times s^2/d^2$

where s is the standard deviation and d is the expected treatment effect.

As Lehr has shown,[4] for α=0.05 and power of 0.80 this is a close approximation to the exact formula. We have deliberately rounded the computed values to avoid the illusion of precision.

expected, then a sample size curve, describing the sample size for a particular base rate, can be easily produced.

Is it plausible that most trials will have risk reductions within a narrow range? Yusuf has argued this case and has produced evidence from 42 cardiovascular trials of chronic interventions (such as aspirin) and 84 trials of acute interventions (such as intravenous nitrates).[9] Relative risk reductions ranged from 8% to 36% (mean 15%) for the chronic interventions and 6% to 24% (mean 19%) for the acute interventions. Averaging all studies, the mean relative risk reduction was 16.5% (SD 8.4%). The table⇓ shows sample sizes for base rates of 2%, 5%, 10%, and 20% using an adapted sample size formula. Though the variation in sample size overall is very large, from 250 to 19 800 per group, for a particular base rate, the range of sample sizes reduces to roughly 10 to one.

Relations between continuous variables

The relation between two continuous variables can be assessed with the correlation coefficient. The standard error of the correlation is roughly $(1-r^2)/\sqrt{(n-2)}$. If we assume that typical correlations are in the range of 0.2 to 0.5, then with $\alpha=0.05$ and a power of 0.80 the estimated sample size $(n=2+8\times(1-r^2)/r^2)$ ranges from 44 to 194. Why 0.2 to 0.5? Pragmatism. A correlation of less than 0.2 accounts for less than 4% of the variance; a correlation of 0.1 accounts for only 1%. It is difficult to imagine why anyone would care about a relation that explains less than 4%. On the other hand, correlations greater than 0.5 are fairly rare and researchers would be unlikely to design a study in the hope of detecting a correlation this large. With these bounds, we might accept any sample size in the range 50-200.

For multivariable analyses such as multiple regression, logistic regression, and factor analysis, everything depends on everything else. No exact predictions are really feasible. Consequently, rules of thumb are often adopted that the sample size should be 5, 10, or 20 times the number of variables. The maths is therefore simple: for five predictors, sample sizes of 25 might be acceptable, and a sample size of 100 would meet the most stringent rule of thumb.

Conclusions

Sample size estimates are like the emperor's clothes; we collectively act in public as if they possess an impressive aura of precision, yet privately we (statisticians) are acutely aware of their shortcomings and extreme imprecision. Clearly, all would benefit from a new approach. In many circumstances, researchers should be encouraged to use the "off the peg" sample sizes we have suggested, although a "made to measure" calculation can be used if sufficient information is available to justify it. More generally, we support the position of Bacchetti[2] that any attempt to determine a precise sample size must necessarily consider more than simple numerical issues and should explicitly deal with broader ethical issues underlying the choice.

Competing interests: All authors have completed the ICMJE unified disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare GN is funded by a Canada research chair; they have no financial relationships with any organisations that might have an interest in the submitted work in the previous three years and no other relationships or activities that could appear to have influenced the submitted work.

Contributors: GN did the calculations and was primarily responsible for writing the paper. SS suggested the topic and reviewed and critiqued multiple drafts. SM conducted literature searches, and reviewed and contributed to drafts of the paper. GN is guarantor.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Altman DG. Why we need confidence intervals. *World J Surg* 2005;29:554-6.
2 Bacchetti P. Current sample size conventions: flaws, harms, and alternatives. *BMC Med* 2010;8:17.
3 Glynn LG, Murphy AW, Smith SM, Schroeder K, Fahey T. Interventions used to improve control of blood pressure in patients with hypertension. *Cochrane Database Syst Rev* 2010;3:CD005182.
4 Lehr R. Sixteen S-squared over D-squared: a relation for crude sample size estimates. *Stat Med* 1992;11:1099-102.
5 Cohen JJ. Statistical power analysis for the behavioral sciences. Erlbaum, 1988.
6 Halpern SD, Karlawish JHT, Berlin JA. The continuing unethical conduct of underpowered clinical trials. *JAMA* 2002;288:358-62.
7 Lipsey MW, Wilson DB. The efficacy of psychological, educational, and behavioral treatment. Confirmation from meta-analysis. *Am Psychol* 1993;48:1181-209.
8 Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care* 2003;41:582-92.
9 Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials? *Stat Med* 1984;3:409-22.

Accepted: 23 May 2012

Cite this as: *BMJ* 2012;345:e5278

Summary points

- Conventional sample size calculations, based on guesses about statistical parameters, are subject to large uncertainties
- There is sufficient evidence to justify establishing expected normative ranges of sample sizes for common research designs
- Normative ranges can be modified if good evidence exists on which to base a sample size calculation

Table

Table 1| Sample sizes for various combinations of relative risk reduction and base rate*

	Relative risk reduction		
Base rate	**0.08**	**0.16**	**0.25**
0.01	247 500	61875	25300
0.02	122500	30625	12550
0.05	47500	11875	4 865
0.1	22500	5625	2 300
0.2	10000	2 500	1025
0.5	2 500	625	255

*Sample size based on the approximate formula $n=16\times(BR(1-BR))/ARR^2$, where BR=base rate and ARR=absolute risk reduction. Sample sizes are rounded to the nearest 5 or 0.