The tyranny of power: is there a better way to calculate sample size?

John Martin Bland

Martin Bland's extensive experience in reviewing and using power calculations has led him to believe that it is time to replace them

Department of Health Sciences, University of York, Heslington, York YO10 5DD
mb55@york.ac.uk
Accepted: 12 June 2009

Cite this as: *BMJ* 2009;339:b3985
doi: 10.1136/bmj.b3985

When I began my career in medical statistics, back in 1972, little was heard of power calculations. In major journals, sample size often seemed to be whatever came to hand. For example, in September 1972, the *Lancet* contained 31 research reports that used individual subject data, excluding case reports and animal studies. The median sample size was 33 (quartiles 12 and 85). In the same month the *BMJ* had 30 reports of the same type, with median sample size 37 (quartiles 12 and 158). None of these publications explained the choice of sample size, other than it being what was available. Indeed, statistical considerations were almost entirely lacking from the methods sections of these papers.

Compare the research papers of September 1972 with those in the same journals in September 2007, 35 years later. In the *Lancet*, there were 14 such research reports, with median sample size 3116 (quartiles 1246 and 5584), two orders of magnitude greater than in 1972. In September 2007, the *BMJ* carried 12 such research reports, with median sample size 3104 (quartiles 236 and 23 351). Power calculations were reported for four of the *Lancet* papers and five of the *BMJ* papers.

The patterns in the two journals are strikingly similar. For each journal, sample sizes increased almost a 100-fold, the proportion of papers reporting power calculations increased from none to one third, and the number of studies of individual participants was less than half that in 1972. The difference in the number of reports is not because of the number of issues; in both years, September was a five issue month. I suggest that the changes in sample size result from the adoption of power calculations.

SUMMARY POINTS

- Most medical research studies have sample sizes justified by power calculations
- Power calculations are based on significance tests
- Many journals require results to be presented with confidence intervals
- Sample size calculations should be based on the width of a confidence interval, not power

Power calculations

In the past there were problems arising from what now seem to be very small sample sizes. Studies were typically analysed using significance tests, and differences were often not significant. What does "not significant" mean? It means that we have not found evidence against the null hypothesis—for example, that there is no evidence for a difference between two types of treatments. This was often misinterpreted as meaning that there was no difference. Potentially valuable treatments were being rejected and potentially harmful ones were not being replaced. I recall Richard Peto presenting a (never published) study of expert opinion on three approaches to the treatment of myocardial infarction, as expressed in leading articles in the *New England Journal of Medicine* and the *Lancet*, and contrasting this with the exactly opposite conclusions that he had drawn from a systematic review and meta-analysis of all published randomised trials in these areas.

Acknowledgment of the problems with small samples led to changes. One of these was the advance calculation of sample size to try to ensure that a study would answer its question. The method that has been almost universally adopted reflects the significance level approach to analysis, the so called power calculation. (In practice, power is seldom calculated, though it is used. It is chosen by the researchers in advance, usually to be 0.90 or 0.80.)

The idea of statistical power is deceptively simple. We are going to do a study where we will evaluate the evidence using a significance test. We decide what the outcome variable is going to be and what the comparison is going to be. For example, the outcome variable might be systolic blood pressure and the comparison would be between mean blood pressure in two groups. We then decide what the test of significance would be, such as a two sample t test comparing mean systolic pressure. We decide how big a difference we want the study to detect—that is, how big a difference it would

be worth knowing about. For a two sample *t* test of mean systolic pressure, this could be the difference in mean pressure that would lead us to adopt the new treatment. We then choose a sample size so that if this difference were the actual difference in the population, a large proportion of possible samples would produce a statistically significant difference. This proportion is the power.

Statistical formulas to determine power for different significance tests are incorporated in many computer programs, both specialist sample size software and some general statistical packages. For many of these calculations we need to supply some other information about the outcome variable. For mean blood pressure, we would also require the standard deviation of blood pressure measurements in the population we wish to study. To compare two proportions, we would need to supply the expected proportion in one of the groups in addition to the difference between them.

There are problems with power calculations, however, even for simple studies. To do them, we require some knowledge of the research area. For example, if we wish to compare two means, we need an idea of the variability of the quantity being measured, such as its standard deviation; if we wish to compare two proportions, we need an estimate of the proportion in the control group. We might reasonably expect researchers to have this knowledge, but it is surprising how often they do not. We are then reduced to saying that we could hope to detect a difference of some specified fraction of a standard deviation. Cohen[1] has dignified this by the name "effect size," but the name is often a cloak for ignorance.

If we know enough about our research area to quote expected standard deviations, proportions, or median survival times, we then come to a more intractable problem: the guesswork as to the effect sought. Inexperienced researchers often answer the question, "How big a difference do you want to able to detect?" with, "Any difference at all." But no sample is so large that it has a good chance of detecting the smallest conceivable difference.

One recommended approach is to choose a difference that would be clinically important—one large enough to change treatment policy. In the Venus II trial of the effect of larval therapy on healing of venous leg ulcers, researchers determined the clinically important difference in healing time by asking patients what mattered to them.[2] This is unusual, however, and more often the difference sought is the researchers' idea. An alternative is to say how big a difference the researchers think that the treatment will produce. Researchers are often wildly optimistic, and funding committees often shake their heads over the implausibility of treatment changes reducing mortality by 50% or more.

Statisticians consulted for power calculations might respond to the lack of a soundly based target difference by giving a range of sample sizes and the differences that each might detect for the researchers to ponder at leisure, but this only puts off the decision. Researchers might use this to follow an even less satisfactory path, which is to decide how many participants they can recruit, find the difference that can be detected with this sample, then claim that difference to be the one they want to find. Researchers who do this seldom describe the process in their grant applications.

In a clinical trial, we usually have more than one outcome variable of interest. If we analyse the trial using significance tests, we may carry out a large number of tests comparing the treatment groups for all these variables. Should we do a power calculation for each of them? If we test several variables, even if the treatments are identical the chance that at least one test will be significant is much higher than the nominal 0.05. To avoid this multiple testing problem, we usually identify a primary outcome variable. We need to identify this for the power calculation to design the study. Researchers often don't seem to appreciate the importance of the primary outcome variable. They change it after the study has begun, perhaps after they have seen the results of the preliminary analysis, and in many cases the original choice is not reported at all.[3 4] This makes the reported P values invalid, overoptimistic, and potentially misleading.

Power calculations led to the call for large, simple trials,[5 6] the first being ISIS-1.[7] This was spectacularly successful.[8] It probably explains the 100-fold increase in sample size from 1972 to 2007.

Confidence intervals

Another reaction to the problems of small samples and of significance tests producing non-significant differences was the movement to present results in the form of confidence intervals, or bayesian credible intervals, rather than P values.[9 10] This was motivated by the difficulties of interpreting significance tests, particularly when the result was not significant. Interval estimates for differences were seen as the best way to present the results for most types of study, particularly clinical trials, and significance tests were to be used only when an estimate was difficult or impossible. (In some situations, of course, a significance test is the better approach—when the question is primarily, "Is there any evidence?" and no meaningful estimate can be obtained.)

Many major medical journals changed their instructions to authors to say that confidence intervals would be the preferred or even required method of presentation. This was later endorsed by the wide acceptance of the CONSORT statement on the presentation of clinical trials.[11 12] We insist on interval estimates and rightly so.

If we ask researchers to present their results as confidence intervals rather than significance tests, I think we should also ask them to base sample size calculations on confidence intervals. It is inconsistent to say that we insist on the analysis using confidence intervals but that the sample size should be decided using significance tests.

This is not difficult to do. For example, the International Carotid Stenting Study (ICSS)[13] compared the risk of stroke after angioplasty and stenting with that after surgical resection of the atheromatous plaque

Sample size calculations for International Carotid Stenting Study

Total sample size	Width of 95% confidence interval: For estimating difference between two proportions having an event that is expected to occur in 14% of participants (% points)	Width of 95% confidence interval: For estimating difference between two means of a quantitative variable (SD)
500	±6.1	±0.18
1000	±4.3	±0.12
1500	±3.5	±0.10
2000	±3.0	±0.09

causing stenosis of carotid arteries. We expected that angioplasty would have a similar effect to surgery on risk reduction. The primary outcome variable was to be long term survival free of disabling stroke. There was to be an additional safety outcome of death, stroke, or myocardial infarction within 30 days and a comparison of cost. I calculated sample size based on an earlier study that reported a three year rate for ipsilateral stroke lasting more than seven days of 14%.[14] The one year rate was 11%, so most events were within the first year. There was little difference between the treatment arms. The width of the confidence interval for the difference between two similar percentages is given by observed difference $\pm 1.96\sqrt{(2p(100-p)/n)}$, where n is the number in each group and p is the percentage expected to experience the event. If we put p=14%, we can calculate the confidence interval for different sample sizes (table). Similar calculations were done for other dichotomous outcomes. For health economic measures, the difference is best measured in terms of standard deviations. The width of the confidence interval is expected to be the observed difference $\pm 1.96\sigma\sqrt{(2p/n)}$, where n is the number in each treatment group and σ is the standard deviation of the economic indicator (table).

These calculations were subsequently amended slightly as outcome definitions were modified. This is the sample size account in the protocol:

> The planned sample size is 1500. We do not anticipate any large difference in the principal outcome between surgery and stenting. We propose to estimate this difference and present a confidence interval for difference in 30-day death, stroke or myocardial infarction and for 3-year survival free of disabling stroke or death. For 1500 patients, the 95% confidence interval will be the observed difference ±3.0 percentage points for the outcome measure of 30-day stroke, myocardial infarction and death rate and ±3.3 percentage points for the outcome measure of death or disabling stroke over 3 years of follow-up. However, the trial will have the power to detect major differences in the risks of the two procedures, for example if stenting proves to be much more risky than surgery or associated with more symptomatic restenosis. The differences detectable with a power of 80% are 4.7 percentage points for 30-day outcome and 5.1 percentage points for survival free of disabling stroke. Similar differences are detectable for secondary outcomes.[13]

Despite my best attempts, we could not exclude power calculations completely. However, the main sample size calculation was based on a confidence interval, and the study was funded.

Base sample sizes on estimation

I propose that we estimate the sample size required for clinical trials or other comparative studies by giving estimates of the likely width of the confidence interval for a set of outcome variables. This does not mean that we would not need to think about sample size; we would still have to decide whether the confidence interval was narrow enough to be worth obtaining. It does mean that we would no longer have to choose a primary outcome variable, a practice which, as noted above, is widely abused. It would have real advantages in large trials that include both clinical and economic assessment.

Power calculations have been useful. They have forced researchers to think about sample size and the likely outcome of the planned study. They have been instrumental in increasing sample sizes to levels where studies can provide much more useful information. But they have many problems, and I think it is time to leave them behind in favour of something better.

I thank Doug Altman, Martin Brown, Nicky Cullum, James Raftery, and David Torgerson for comments on an earlier draft.

Competing interests: None declared.

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

1 Cohen J. A power primer. *Psychol Bull* 1992;112:155-9.
2 Petherick ES, O'Meara S, Spilsbury K, Iglesias CP, Nelson EA, Torgerson DJ. Patient acceptability of larval therapy for leg ulcer treatment: a randomised survey to inform the sample size calculation of a randomised trial. *BMC Med Res Methodol* 2006;6:43.
3 Chan AW, Hrobjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials—comparison of protocols to published articles. *JAMA* 2004;291:2457-65.
4 Chan AW, Jeric K, Schmid I, Altman DG. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *CMAJ* 2004;171:735-40.
5 Peto R, Yusuf S. Need for large (but simple) trials. *Thromb Haemos* 1981;46:325.
6 Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials? *Stat Med* 1984;3:409-20.
7 ISIS-1 (First International Study of Infarct Survival) Collaborative Group. Randomized trial of intravenous atenolol among 16 027 cases of suspected acute myocardial infarction. ISIS-l. *Lancet* 1986;ii:57-66.
8 Peto R, Collins R, Gray R. Large-scale randomized evidence: large, simple trials and overviews of trials. *J Clin Epidemiol* 1995;48:23-40.
9 Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *BMJ* 1986;292:746-50.
10 Bland M. *An introduction to medical statistics*. Oxford: Oxford University Press, 1987.
11 Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA* 1996;276:637-9.
12 Moher D, Schulz KF, Altman DG, CONSORT Group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* 2001;357:1191-4.
13 Featherstone RL, Brown MM, Coward LJ. International Carotid Stenting Study: protocol for a randomised clinical trial comparing carotid stenting with endarterectomy in symptomatic carotid artery stenosis. *Cerebrovasc Dis* 2004;18:69-74.
14 CAVATAS Investigators. Endovascular versus surgical treatment in patients with carotid stenosis in the Carotid and Vertebral Artery Transluminal Angioplasty Study (CAVATAS): a randomised trial. *Lancet* 2001;357:1729-37.