The Tyranny of Power

Martin Bland

Talk to be presented in New Zealand and Australia, February and March 2008.

I have sat on several grant-giving bodies and refereed proposals for others, seeing many hundreds of power calculations. I have done hundreds more for medical researchers who have put their heads round my door. I have even done a few on my own account. I find myself increasingly dissatisfied with them.

When I began my career in medical statistics, back in 1972, little was heard of power calculations. In major journals, sample size often appeared to be whatever came to hand. For example, in that month, September 1972, the *Lancet* contained 31 research reports which used individual subject data, excluding case reports and animal studies. The median sample size was 33 (quartiles 12 and 85). In the *British Medical Journal* in September 1972, there were 30 reports of the same type, with median sample size 37 (quartiles 12 and 158). None of these publications reported any explanation of 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. One of the few that mentioned them at all (Bottiger and Carlson 1972) merely noted that 'Statistical analyses were performed using methods described by Snedecor (1956)', this being a standard statistical textbook.

Compare the research papers of September 1972 to 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 23351). Power calculations were reported for 4 of the *Lancet* papers and 5 of the *BMJ* papers.

The patterns in the two journals are strikingly similar (Figure 1). For each journal, the differences in sample size, the number of papers reporting power calculations, and the number of individual subject studies published were all statistically significant ($P<0.0001$, $P=0.0003$, $P=0.007$, $P=0.001$, $P=0.02$, $P=0.008$, respectively). The difference in the number of reports is not because of the number of issues; in both years, September was a five issue month.

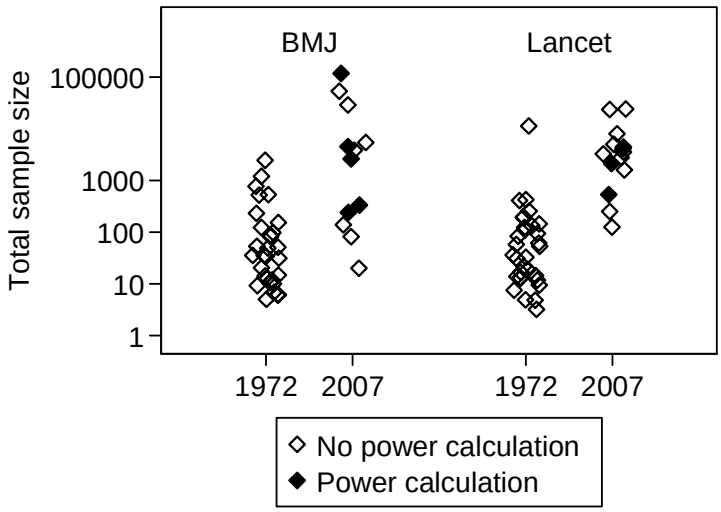

Figure 1. Samples sizes (log scale) for research reports using individual patient data in the Lancet and BMJ for September in 1972 and 2007.

In the past there were problems arising from what might appear to be very small sample sizes. Studies were typically analysed statistically using significance tests, and differences were often not significant. What does "not significant" mean? It means that we have failed to demonstrate that there is evidence against the null hypothesis, for example that there is no evidence for a difference between two types of patient or patients treated with different treatments. This was often misinterpreted as meaning that there is 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 which he had drawn from a systematic review and fledgling meta-analysis of all published randomised trials in these areas.

Acknowledgement of the problems with small samples led to changes. One of these was the precalculation of sample size so as to try to ensure a study which would answer its question. The method which has been almost universally adopted is the power calculation, a method which reflected the significance level approach to analysis.

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 how big a difference we want the study to detect, how big a difference it would be worth knowing about. We then choose a sample size so that, if this were the actual difference in the population, a large proportion of possible samples would produce a statistically significant difference.

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 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 might suggest that they look at their existing records to find some data, or to look at published papers where the same variable has been used. I was once told that no-one had ever made the measurement, in which case, I thought, we are not ready to use it as the outcome measure in a clinical trial. Often we are reduced to saying that we could hope to detect a difference of some specified fraction of a standard deviation. Cohen (1992) 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 effect sought. 'How big a difference do you want to able to detect?' is a question which often provokes from the inexperienced researcher the answer 'Any difference at all'. But this they cannot have, 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 which would be large enough to change treatment policy. In the VenUS III trial of ultrasound aimed to shorten healing time in venous leg ulcers, we said '. . . overall we have estimated that 50% of ulcers in the standard care group will heal within 22 weeks. We estimate that clinicians and patients would, however, value a reduction in healing time of seven weeks (a 32% reduction in healing time, from 22 to 15 weeks) and have based our sample size calculation on this premise. To detect a difference in median healing time of 7 weeks (from 22 weeks to 15 weeks), we require 306 patients in total.' (VenUS III trial protocol). This was based on asking some clinicians and patients what would be sufficient return to justify the extra time involved in ultrasound treatment. This is unusual, however, and more often the difference sought is the researchers' own 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 unlikeliness of treatment changes of reducing mortality by 50% or more. Statisticians might respond to the lack of a soundly based treatment difference to go for by giving a range of sample size and the differences which 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 which can be detected with this sample, then claim that difference as 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. So we need to identify this for the power calculation to design the study. As Chan *et al* (2004, 2004b) found, researchers often change the primary outcome variable after the study has begun, which we might suspect to have been done after they have seen the results of the preliminary analysis, and their original choice may not be reported at all. This would make the P values invalid and over-optimistic.

The next difficulty is the power, the probability that a random sample from the target population will produce a significant difference. This is usually arbitrary, choice

being governed by a trade-off between the chance of a significant difference and the feasibility of getting the required sample. Popular choices in grant applications are 0.90 or 0.80, 90% or 80%. Wallis Simpson famously remarked that no woman could be too rich or too thin, and in health research I would say that no study can be too big or too powerful. I once received the referee's comment that we could design our study for 80% power rather than the 90% we had chosen and so have a smaller sample and cheaper study. I replied that if I could have 99% power I would. I have never liked the idea of staking several years' work on a 4 to 1 chance.

These calculations led to some shocks. I remember a clinician asking me how many patients he would need for a trial aimed at reducing mortality following myocardial infarction by one quarter. I estimated that to reduce mortality from 15% to 11.25% we would need 1715 in each group. Why not round this up to 2000, I suggested, to allow for a few things going wrong? I thought he was going to faint. He thought this was impossible and went off to do a trial which was a tenth of the size, which duly reported a difference in the hoped-for direction, which was not significant.

Other statisticians were more forceful than I was and Peto and Yusuf (1981) led the call for large, simple trials, the first being ISIS-1 (ISIS-1 Collaborative Group, 1986). This was spectacularly successful, as Peto *et al*. (1995) described. It probably explains the hundred-fold increase in sample size reported in Figure 1. No clinical researcher with aspirations to be in the top flight can now be happy unless a trial with a four-figure sample size is in progress.

Not all studies compare two groups to see whether one group has a higher mean or greater proportion than other. Sometimes we want to compare two treatments which we hope will have the same outcome, an equivalence trial. In CAVATAS (CAVATAS investigators 2001), for example, we compared angioplasty with artery grafting for stenosis of the carotid artery. The hope was that in terms of mortality or disabling stroke the two treatments would be similar, in which case angioplasty would be preferable as a less traumatic procedure. Researchers wanted to analyse such trials using the familiar significance tests. However, if the usual test to compare two groups were used, a non-significant result would mean that there was no evidence that a difference existed, not that no difference existed. This led to the of notion of significance being turned round. Instead of testing the null hypothesis that the outcome in the two groups is the same, we test the null hypothesis that the difference does not differ by more than a prespecified amount. In their tables of sample size for clinical trials, Machin and Campbell (1987) give sample sizes for trials to check for differences between two proportions of 0.05, 0.10, 0.15, and 0.20. If the test is significant, we can conclude that there is good evidence that the treatments do not differ by more than this specified amount. If test were not significant, we would conclude that there is insufficient evidence that one or other treatment is not superior by more than the pre-specified amount. A similar procedure works for non-inferiority trials, where we want evidence that our new treatment is no worse than an existing treatment, though we would not mind if it were better. A significant result would mean that we could conclude that our new treatment was not worse than the control treatment by more than the prespecified amount, though we would not conclude that it were better. These analyses seem very awkward and artificial to me and I have never used them, but there are power calculations for them, where we have to determine not only likely properties of the population but what difference we would consider acceptable. For example, Machin and Campbell (1987) give sample sizes for an equivalence trial where we expect one proportion to be 0.20, the second to be 0.15,

and the treatments would be regarded as equivalent if the difference were not more than 0.10. Whether we could regard treatments as equivalent if the death rate on one were 20% and on the other 10% is open to debate.

Another reaction to the problem 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 (Gardner and Altman 1986, Bland 1987). This was motivated by the difficulties of interpreting significance tests, particularly when the result was not significant. Interval estimates for effect sizes were seen as the best way to present the results for most types of study, clinical trials in particular, and significance tests were the alternative only to be used 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 there is no meaningful estimate to be obtained.) This campaign was very successful and 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 standard for the presentation of clinical trials (Begg *et al.* 1996). We insist on interval estimates, confidence or credible, and rightly so.

If we ask researchers to design studies the results of which will be presented as confidence intervals, rather than significance tests, I think that we should base our sample size calculations on confidence intervals, rather than significance tests. It seems quite inconsistent to say that we insist on the analysis using confidence intervals but the sample size should be decided using significance tests.

This is not difficult to do. For example, the International Carotid Stenting Study (ICSS) was designed to compare angioplasty and stenting with surgical vein transplantation for stenosis of carotid arteries, to reduce the risk of stroke. We did not anticipate that angioplasty would be superior to surgery in risk reduction, but that it would be similar in effect. If this were so, we would prefer angioplasty because we would not need to give general anaesthesia to often vulnerable and elderly patients. We might even think some increase in risk would be acceptable in exchange for this advantage. We therefore might think of this as an equivalence or as a non-inferiority trial.

The sample size calculations for ICSS were based on the earlier CAVATAS study (CAVATAS investigators 2001), which had the 3 year rate for ipsilateral stroke lasting more than 7 days = 14%. The one year rate was 11%, so most events were within the first year. There was very little difference between the treatment arms. The width of the confidence interval for the difference between two very 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 this for different sample sizes:

total sample	width of 95% confidence interval in percentage points
500	±6.1
1000	±4.3
1500	±3.5
2000	±3.0

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 observed difference $\pm 1.96\sigma\sqrt{2/n}$, where n is the number in each treatment group and σ is the standard deviation of the economic indicator.

total sample	width of 95% confidence interval in standard deviations
500	±0.18
1000	±0.12
1500	±0.10
2000	±0.09

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 from fully enrolled centres. 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 three-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 three years 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 riskier than surgery or associated with more symptomatic restenosis. The difference detectable with power 80% are 4.7 for 30 day outcome and 5.1 percentage points for survival free of disabling stroke. Similar differences are detectable for secondary outcomes. We expect to achieve this recruitment within 6 years.' (ICSS 2007)

Despite my best attempts, power calculations could not be excluded completely. My collaborators felt the need for the security that a conventional approach offered. However, the main sample size calculation was based on a confidence interval and the study was funded, by research grants from the Stroke Association, Sanofi-Synthélabo and the European Union.

What led me to adopt this approach was that ICSS was to be an equivalence or perhaps a non-inferiority trial. I think that this approach has particular advantages in equivalence trials, as deciding how different treatments could be before being regarded as equivalent strikes me as even more subjective and problematic than deciding what target treatment difference is the smallest that would change practice. There are so many considerations which might affect this and they might be different for patients in different circumstances. However, I can see no reason why we should use such an approach for other kinds of study.

I propose that we estimate the sample size required for a clinical trial or other comparative study by giving estimates of likely confidence interval width for a set of outcome variables. This does 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 with 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 dramatically 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.

References

Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF. (1996) Improving the quality of reporting of randomized controlled trials - The CONSORT statement. *JAMA-Journal of the American Medical Association* **276**, 637-639.

Bland M. (1987) *An Introduction to Medical Statistics*. Oxford University Press, Oxford.

Bottiger LE, Carlson LA. Relation between serum-cholesterol and triglyceride concentration and hemoglobin values in non-anemic healthy persons. *British Medical Journal* 1972; **3**: 731-3.

CAVATAS investigators. (2001) Endovascular versus surgical treatment in patients with carotid stenosis in the Carotid and Vertebral Artery Transluminal Angioplasty study (CAVATAS): a randomised trial. *Lancet* 357, 1729-37.

Chan AW, Hrobjartsson A, Haahr MT, Gøtzsche PC, Altman DG. (2004) Empirical evidence for selective reporting of outcomes in randomized trials - comparison of protocols to published articles. *JAMA-Journal of the American Medical Association* **291**: 2457-2465.

Chan AW, Jeric K, Schmid I, Altman DG. (2004b) Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *Canadian Medical Association Journal* **171**: 735-740.

Cohen J. (1992) A power primer. *Psychological Bulletin* **112**: 155-159.

Gardner MJ and Altman DG. (1986) Confidence intervals rather than P values: estimation rather than hypothesis testing. *British Medical Journal 292*: 746-50.

ICSS (2007) International Carotid Stenting Study Protocol Version 3.2, 22nd November 2007.

ISIS-1 (First International Study of Infarct Survival) Collaborative Group. (1986) Randomized trial of intravenous atenolol among 16,027 cases of suspected acute myocardial infarction. ISIS-l. *Lancet* ii: 57-66.

Machin D and Campbell MJ. (1987) *Statistical Tables for the Design of Clinical Trials* Oxford: Blackwell.

Peto R, Collins R, Gray R. (1995) Large-scale randomized evidence: large, simple trials and overviews of trials. *Journal of Clinical Epidemiology* **48**: 23-40.

Peto R, Yusuf S. (1981) Need for large (but simple) trials. *Thrombosis and Haemostasis* **46**: 325-325.

VenUS III trial prototcol. http://www.venus3.co.uk/ (Accessed 19 December 2007).

Yusuf S, Collins R, Peto R. (1984) Why do we need some large, simple randomized trials? *Statistics in Medicine* **3**: 409-420.