Some problems with sample size

Martin Bland

Talk to be presented at the Joint Meeting of the Dutch Pathological Society and the Pathological Society of Great Britain & Ireland, Leeds, July 4th, 2008.

A journey back through time . . .

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).

The patterns in the two journals are strikingly similar. The difference in the number of reports is not because of the number of issues; in both years, September was a five issue month.

Problems with small sample sizes

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.

If there is no difference between two populations, the chance of a significant difference between two samples from them is 0.05, whatever the sample size. If there is a real difference, the chance of a significant difference is small if the samples are

small. We call this probability the power. Hence with small samples differences which are significant are more likely to be spurious than with large samples.

Power calculations

Acknowledgement of the problems with small samples led to changes. One of these was the pre-calculation 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.

For example, consider a case control study. We will have a group of cases of a disease and a group of controls. We have a risk factor, e.g. a gene allele, which is found in about 10% of controls. Is it more common in cases? To estimate the sample size, we say how big a different we want to detect. Suppose the risk factor is twice as common, 20%. In some possible samples the difference will be greater than this, in some it will be less. \the sample size calculation tells us that we need 266 in each group to have power 90% of getting a significant difference at the 5% significance level.

For another example, consider a prospective study. We will have a group of subjects for whom we will determine the presence of allele. We guess that the risk of developing the condition is about 2% in allele negatives and 2.5% in allele positives. (For example, see the paper on diabetes genes, Zeggini *et al*. 2008.) We guess that the risk of developing the condition is about 2% in allele negatives and 2.5% in allele positives. We estimate that about 10% of people will have the allele. The sample size calculation tells us that we need 10,067 subjects in the allele positive group to have power 90% of getting a significant difference at the 5% significance level. We need 100,670 subjects altogether! (Zeggini *et al*., 2008, had 90,000.)

Problems with power calculations: knowledge of the research area

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.

Problems with power calculations: how big a difference do we want to able to detect?

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.

Problems with power calculations: multiple outcomes

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.

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. For example, suppose we have two independent variables. Then the probability that at least one variable will be significant = $1 - (1 - 0.05)^2 = 0.098$. The expected number of significant differences, the average number we would get over many studies when the null hypothesis is true, is $2 \times 0.05 = 0.1$. For 10 independent variables, the probability

that at least one variable will be significant = $1 - (1 - 0.05)^{10} = 0.40$ and the expected number of significant differences = $10 \times 0.05 = 0.5$. For 1,000 independent variables, the probability that at least one variable will be significant = $1 - (1 - 0.05)^{1000} = 1.00$ and the expected number of significant differences = $1000 \times 0.05 = 50$.

We can use the Bonferroni correction. We multiply the P value by the number of tests. E.g. for 10 tests, we demand $P < 0.005$ rather than $P < 0.05$. If any test has $P = 0.005$, the difference overall is significant at the 0.05 level. The disadvantage of this approach is that it is usually conservative, because the tests are not independent. In other words, the corrected P values are too large.

In the case control study example, we wanted to detect a difference between 10% and 20%. If we intend to do 1000 tests, for significance we would demand $P < 0.00005$ rather than $P < 0.05$. We need 266 in each group to have power 90% of getting a significant difference at the 5% significance level. We need 723 in each group to have power 90% of getting a significant difference at the 0.005% significance level.

Is the Bonferroni correction appropriate in a case control study? We know the groups are different, so the composite null hypothesis doesn't apply. However, Bonferroni gives the probability of seeing this difference if all null hypotheses are true, so it is a reasonable approach.

In their study of associations with diabetes, Zeggini *et al.* (2008) say that:

> "We detected at least six previously unknown loci with robust evidence for association, including the JAZF1 ($P < 5.0 \times 10^{-14}$), CDC123-CAMK1D ($P < 1.2 \times 10^{-10}$), TSPAN8-LGR5 ($P < 1.1 \times 10^{-9}$), THADA ($P < 1.1 \times 10^{-9}$), ADAMTS9 ($P < 1.2 \times 10^{-8}$) and NOTCH2 ($P < 4.1 \times 10^{-8}$) gene regions."
> . . .
>
> "We based our further analyses on 2,202,892 SNPs that met imputation and genotyping quality control criteria across all studies."

In fact $0.05/2202892 = 2.270 \times 10^{-8}$, so at least one of these associations does not make it.

Consequences of power calculations

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.

Alternatives to power calculations

Power calculations are not the only way to plan sample sizes. There are confidence interval based methods as well. These may have some advantages, see Bland (2008).

Summing up

- We cannot ignore sample size.
- Small samples increase chance that significant differences are false positive.
- Small samples increase chance that important differences will be missed.
- We must think about sample size when planning a study.
- Sample size calculations are not trivial.

References

Bland JM. (2008) The Tyranny of Power, unpublished, http:\\martinbland.co.uk\talks\tyrpowertalk.htm.

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.

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.

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.

Zeggini E, Scott LJ, Saxena R, Voight BF, for the Diabetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genetics*, 30 March 2008.