Doing clinical trials large enough to achieve adequate reductions in uncertainties about treatment effects

Michael J Campbell

Medical Statistics Group, ScHARR, University of Sheffield, Regent Court, Sheffield S1 4DA, UK
Correspondence to: Michael J Campbell. Email: m.j.campbell@sheffield.ac.uk

DECLARATIONS

Competing interest
None declared

Funding
None declared

Ethical approval
Not applicable

Guarantor
MJC

Contributorship
MJC

Acknowledgements
I am indebted to Iain Chalmers and Peter Armitage for helpful comments on earlier drafts of this article.

Recognition of the need for adequate numbers of observations in clinical trials

Some medical researchers and commentators on tests of medical treatments recognized more than two centuries ago that conclusions about the effects of treatments should be based on adequate numbers of people.[1] From the middle of the 19th century onwards, the importance of the 'law of large numbers' became acknowledged increasingly within medical research.

In 1840, for example, Jules Gavarret noted that:

> *Average mortality, as provided by statistics, is never the exact and strict translation of the influence of the test medication but approaches it all the more as the number of observations increases.*
>
> *A therapeutic law can never be absolute; its applications can always oscillate between certain limits which are all the narrower, the more the collected observations are multiplied, and which can be determined with the aid of the numbers constituting the statistics that have provided the law.*
>
> *To be able to decide in favour of one treatment method over another, it is not enough for the method to yield better results; the difference found must also exceed a certain limit, the extent of which is a function of the number of observations.*[2]

In 1854, Thomas Graham Balfour (who later became President of the Royal Statistical Society) was careful not to draw any conclusions about the effects of belladonna in preventing scarlet fever when only two children out of 76 given belladonna and two out of 75 in an untreated comparison group developed the disease. Balfour[3,4] concluded that 'the numbers are too small to justify deductions as to the prophylactic power of belladonna'.

By the end of the 19th century, some medical researchers were exhibiting a very sophisticated understanding of how statistical tests were being used to assess the extent to which different outcomes in treatment comparison groups were likely to reflect the play of chance.[5,6]

Estimating the statistical power of clinical trials

Statistical tests can only be applied once data are available for testing. They are based on the null hypothesis, that is, a hypothesis that there are no differential effects between treatments in a fair (unbiased) comparison. The issue to be addressed in Balfour's experiment could thus be phrased as: 'Can we accept the null hypothesis of no difference between belladonna and no belladonna?' For reliable evaluations, however, we need to think of what alternative hypotheses might be relevant. An alternative hypothesis is one where a difference deemed worthwhile exists. For example, a diet that leads to a loss of half a kilogram in six months may not be deemed worthwhile, whereas one that results in a 10 kg loss may be well worth using. This implies that a limit should be set below which a differential treatment effect is not worth pursuing. Thus the design principle could be stated as 'A trial should be capable of rejecting the null hypothesis, if an alternative hypothesis is true.' The size of the effect specified under the alternative hypothesis is often termed the 'effect size', for example, a difference in means based on continuous data, or a difference in proportions, relative risk or odds

J R Soc Med 2013: **106**: 68–71. DOI 10.1177/0141076813477570

ratio based on binary data. For binary data, it is the number of events in the comparison groups (rather than the number of patients) that is important. It is helpful here to consider a single alternative hypothesis, for example, that the difference between two proportions is a given value.

The first authors to consider alternative hypotheses formally were Jerzy Neyman and Egon Pearson (son of Karl Pearson, who derived the correlation coefficient and the chi-squared test). They specified the concept of statistical power, which is the probability of getting a statistically significant result if the alternative hypothesis is true. They also defined the Type I error, the probability of rejecting the null hypothesis when it is true, and the Type II error, which is the probability of failing to reject the null hypothesis when it is false, which is equal to one minus the power.[7] It is generally accepted that a statistically reliable trial should have a power of at least 80%, that is a Type II error of at most 20%, similar to the requirement of a Type I error of at most 5%. For trials against placebo, the requirement is often 90%.

In his famous book, Austin Bradford Hill[8] discussed the sample size issue and considered what likely outcomes would yield a statistically significant result. Although he stated that one would have to use past experience to decide the kind of difference that a trial might detect, he did not formally discuss statistical power. He concluded that 'we must confess ignorance of the numbers required to give a convincing result' – an example of humility too often rare among researchers today!

Even the most exemplary clinical trials done in the early 1940s did not use statistical arguments to justify their sample sizes (see for example Refs.[9–11]). In the late 1940s, articles began to appear describing how to calculate statistical power,[12,13] and by 1957 a textbook had been published with tables of sample sizes for comparing two groups of equal numbers of patients.[14] One of the earliest clinical trial reports to discuss statistical power, Type I errors, and Type II errors, appears to have been that comparing treatments for solid tumours reported by Zubrod *et al.*[15]

The same year, Armitage[16] reviewed the concept of statistical power in sequential trials, a design derived from the use of sequential testing in industry. This began to be applied in clinical settings in the late 1950s and early 1960s.[17–19] Sequential trials require a statement of statistical power from the outset, so some of the earliest clinical trial reports quoting power used sequential methods. The concept of sequential methods and the Neyman–Pearson approach was subject to an extensive critique by Anscombe,[20] however. He argued that clinical trials were not about making 'accept' or 'reject' decisions, but rather about estimating the range of plausible differential effects between treatments.

Even as late as the 1970s, statistical power was a new concept to many researchers. Freiman *et al.*[21] showed that many authors were still interpreting failure to demonstrate a statistically significant difference as grounds for accepting the null hypothesis and claiming that no difference existed.

Since the 1970s, there have been many papers and books on sample size estimation (see for example Ref.[22]), and increasing numbers of medical journals require authors to justify the sample sizes they have used for the trials they wish to report. Similarly, research funders now often require researchers to estimate the numbers of participants needed to produce statistically robust results. This means that researchers are too often unrealistically optimistic about the sample sizes they will achieve.

Given the uncertainties that surround the values of the parameters required to make sensible guesses about sample sizes, others have warned about the dangers of demanding observance of the 'ritual' of power calculations.[23–26]

There are arguments for and against doing trials that are unlikely to achieve high statistical power. In principle, the choice between continuing to use inadequately evaluated treatments haphazardly outside the context of controlled trials, or offering them within controlled trials, should be easy: one will learn nothing from the former and something from the latter. As it is almost always unrealistic to expect a single study to answer an important question, the results of relatively small trials can contribute to meta-analyses,[24] with emphasis on estimating effect sizes, with associated confidence intervals.

An argument against statistically 'underpowered' trials is that problems may result from 'equivocal' results. There is often an opportune time for a clinical trial to be conducted, when

investigators are willing to accept a 'balance of probabilities' in favour of alternative treatments. Imprecise estimates of treatment effects from an underpowered trial may change this balance of probabilities, and yet still leave considerable doubt as to the relative efficacy of the treatments compared. This may make it difficult to obtain ethics approval and patient consent for an additional trial. In addition, equivocal trials are probably less likely to be submitted and accepted for publication, and so will be less readily available for inclusion in meta-analyses.

Sample size calculations have also been criticized because they depend on the selection of one endpoint when, in practice, trials have several endpoints, and any size of study can be justified by judicious choice of endpoint and power. When investigators have no idea of what should be regarded as a meaningful effect size they will be tempted to focus on significance tests when the purpose of most experiments is estimation. An alternative view has been put by Williamson *et al.*[27] They pointed out that a power calculation forces investigators, *a priori*, to name the main outcome variable, which can then be checked in the analysis, to protect against data dredging. These authors also noted that it makes clear, before the results are known, what the authors considered to be a meaningful effect size. This prevents authors from claiming two treatments are equivalent when there is no statistically significant difference between them, when the observed difference could plausibly have arisen from the hypothesis of a clinically meaningful difference.

Recent developments

To start a clinical trial, or to continue one in the face of accumulating data, is a decision, and so-called Bayesian methods are well suited to making such decisions.[28] These decisions should be informed by systematic reviews of all the evidence on the effects of the treatment in question, but recent surveys have shown that such reviews are rarely done.[29,30] One result is that investigators are often unrealistically optimistic about the likely benefits of new treatments.[31]

Sutton *et al.*[32] describe some methods for calculating the sample size required of a new trial, based on the assumption that its results will contribute to an updated meta-analysis. They suggest that, in some circumstances, new studies, even very large ones, are unlikely to yield any information that would add usefully, let alone overturn, existing evidence. Some systematic approach to assessing existing evidence to inform trial design prior to conducting a new trial would seem an obvious requirement.

Recently, economists have shown interest in trying to estimate the benefits and dis-benefits of doing a trial by converting these estimates into costs. The principle they are exploring is that the expected value of the additional information provided by the proposed trial must exceed the cost of the research being considered (see Ref.,[33] Chapter 7). These approaches remain at an early stage of development.

An approach to trial design which has no formal pretrial sample size calculation is the so-called adaptive design (see, for example Ref.[34]). This is essentially a two-phase trial, with an initial phase used to generate information such as which doses of a drug to use and estimates of the standard deviations of the outcome variables. This information can then inform appropriate changes to the trial protocol, including amended statistical power calculations and target sample size. This is an area of much ongoing research.

Sequential designs, as discussed earlier, require monitoring of the likely effect size, and stopping either because the null hypothesis has been rejected or because the trial has no chance of rejecting the null hypothesis. Neither adaptive nor sequential approaches are conducive to fixed sample sizes. The estimate of the desirable sample size made at the start of the study is actually a random variable in these circumstances. How this can be dealt with in the context of fixed budgets and timescales is a thorny issue that seems likely to remain a topic of active discussion.

Conclusion

Given that some estimate of sample size is required prior to embarking on a trial, proper review of the available evidence, and ways of taking account of this evidence in the design of

J R Soc Med 2013: **106**: 68–71. DOI 10.1177/0141076813477570

trials, are essential. Even so, considerable uncertainty about likely outcomes of a trial may remain. Faced with these dilemmas, the example of Bradford Hill's humility suggests that the best option is honesty about these uncertainties. Proposed studies should not be rejected for funding simply because they fail to meet an arbitrary statistical power threshold.

References

1 Tröhler U. The introduction of numerical methods to assess the effects of medical interventions during the 18th century: a brief history. JLL Bulletin: Commentaries on the History of Treatment Evaluation 2010. See www.jameslindlibrary.org
2 Gavarret LDJ. *Principes généraux de statistique médicale: ou développement des règles qui doivent présider à son emploi*. [*General Principles of Medical Statistics: or the Development of Rules that must Govern their Use*] Paris: Bechet jeune & Labé, 1840
3 Balfour TG. *Quoted in West C. Lectures on the Diseases of Infancy and Childhood*. London: Longman, Brown, Green and Longmans, 1854:600
4 Chalmers I, Toth B. 19th century controlled trials to test whether belladonna prevents scarlet fever. JLL Bulletin: Commentaries on the History of Treatment Evaluation 2009. See www.jameslindlibrary.org
5 Heiberg P. Studier over den statistiske undersøgelsesmetode som hjælpemiddel ved terapeutiske undersøgelser (Studies on the statistical study design as an aid in therapeutic trials). *Bibliotek for Læger* 1897;**89**:1–40
6 Gluud C, Hilden J. Povl Heiberg's 1897 methodological study on the statistical method as an aid in therapeutic trials. JLL Bulletin: Commentaries on the History of Treatment Evaluation 2008. See www.jameslindlibrary.org
7 Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika* 1928;**20A**:175–240
8 Hill AB. *Principles of Medical Statistics*. London: Lancet, 1937
9 Bell JA. Pertussis prophylaxis with two doses of alum-precipitated vaccine. *Public Health Rep* 1941;**56**:1535–46
10 Medical Research Council. Clinical trial of patulin in the common cold. *Lancet* 1944;**2**:373–5
11 Medical Research Council. Streptomycin treatment of pulmonary tuberculosis: a Medical Research Council Investigation. *Br Med J* 1948;**2**:769–82
12 Paulson E, Wallis WA. Planning and analysing experiments for comparing two percentages. In: Eisenhart C, Hastay MW, Wallis WA, eds. *Selected Techniques of Statistical Analysis for Scientific and Industrial Research and Production and Management Engineering*. New York: McGraw-Hill, Chapter 7, 1947:247–66
13 Sillitto GP. Note on approximations to the power function of the 2 × 2 comparative trial. *Biometrika* 1949;**4**:347–52
14 Cochran WG, Cox GM. *Experimental Designs*. 2nd edn. New York: Wiley, 1957
15 Zubrod CG, Schneiderman M, Frei E, *et al.* Appraisal of methods for the study of chemotherapy in man: comparative therapeutic trial of nitrogen mustard and thiophosphoramide. *J Chronic Dis* 1960;**11**:7–33
16 Armitage P. *Sequential Medical Trials*. Springfield, Illinois: Thomas, 1960
17 Kilpatrick GS, Oldham PD. Calcium chloride and adrenaline as bronchial dilators compared by sequential analysis. *Br Med J* 1954;**2**:1388–91
18 Snell ES, Armitage P. Clinical comparison of diamorphine and pholcodine as cough suppressants by a new method of sequential analysis. *Lancet* 1957;**1**:860–2
19 Truelove SC, Watkinson G, Draper G. Comparison of corticosteroid and sulphasalazine therapy in ulcerative colitis. *Br Med J* 1962;**2**:1708–11
20 Anscombe FJ. Review of 'Sequential Medical Trials'. *J Am Stat Assoc* 1963;**58**:365–83
21 Freiman JA, Chalmers TC, Smith H, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: survey of 71 'negative' trials. *N Engl J Med* 1978;**299**:690–4
22 Machin D, Campbell MJ, Tan SB, Tan SH. *Sample Size Tables for Clinical Studies*. 3rd edn. Chichester: Wiley-Blackwell, 2008
23 Detsky AS, Sackett DL. When was a 'negative' clinical trial big enough? How many patients you needed depends on what you found. *Arch Intern Med* 1985;**145**:709–12
24 Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. *Lancet* 2005;**365**:1348–53
25 Bacchetti P, Wolf LE, Segal MR, McCulloch CE. Ethics and sample size. *Am J Epidemiol* 2005;**161**:105–10
26 Bland JM. The tyranny of power: is there a better way to calculate sample size? *Br Med J* 2009;**339**:1133–5 10.1136/bmj.b3985
27 Williamson P, Hutton JL, Bliss J, Blunt J, Campbell MJ, Nicholson R. Statistical review by research ethics committees. *J R Stat Soc A* 2000;**163**:5–13
28 Spiegelhalter DJ, Abrams KR, Myles JR. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Chichester: Wiley, 2004
29 Cooper NJ, Jones DR, Sutton AJ. The use of systematic reviews when designing new studies. *Clin Trials* 2005;**2**:260–4
30 Robinson KA, Goodman SN. A systematic examination of the citation of prior research in reports of randomized, controlled trials. *Ann Intern Med* 2011;**154**:50–5
31 Chalmers I, Matthews R. What are the implications of optimism bias in clinical research? *Lancet* 2006;**367**:449–50
32 Sutton AJ, Cooper NJ, Jones DR, Lambert PC, Thompson JR, Abrams KR. Evidence-based sample size calculations based upon updated meta-analysis. *Stat Med* 2007;**26**:2479–500
33 Briggs A, Claxton K, Sculpher M. Decision modelling for health economic evaluation. Oxford: Oxford University Press, 2006
34 Phillips AJ, Keene ON. Adaptive designs for pivotal trials: discussion points from the PSI adaptive design expert group. *Pharm Stat* 2006;**5**:61–6