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

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. 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 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

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

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.

It is generally accepted that 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.

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; 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. 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.), 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.

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, 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. 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. 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. One result is that investigators are often unrealistically optimistic about the likely benefits of new treatments.

Sutton et al. 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., 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.). 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
Doing clinical trials to achieve adequate reductions in uncertainties about treatment effects

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

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 & Labe, 1840
5 Heiberg P. Studier over den statistiske undersøgelsesmetode som hjælpmiddel ved terapeutiske undersøgelser (Studies on the statistical study design as an aid in therapeutic trials). Bibliotek for Læger 1897:89:1–40
9 Bell JA. Pertussis prophylaxis with two doses of alum-precipitated vaccine. Public Health Rep 1941;56:1353–46
13 Sillitto GP. Note on approximations to the power function of the $2 \times 2$ comparative trial. Biometrika 1949;4:347–52
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
28 Spiegelhalter DJ, Abrams KR, Myles JR. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Chichester: Wiley, 2004
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
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