The ethics of two- and one-sided hypothesis tests for clinical trials

A Owen

Department of Cardiology, Kent and Canterbury Hospital, Ethelbert Road, Canterbury, Kent CT1 3NG, UK

Abstract

In medical research, hypothesis tests are used to determine if a novel treatment is efficacious in comparison to a control treatment. Traditionally two-sided tests are recommended and one-sided tests are not. In this review the arguments for two- and one-sided tests are presented. It is argued that in many applications one-sided tests are appropriate and that two-sided tests may expose research subjects to unnecessary risk, and are therefore unethical.

Introduction

Randomized controlled clinical trials are conducted to determine if a novel treatment is superior to a standard treatment (or placebo if no standard treatment exists). Traditionally, the data generated from such trials are used to determine the efficacy of the novel treatment through the use of the concept of hypothesis testing. Null and alternative hypotheses are specified *a priori* which are mutually exclusive and exhaustive. The alternative hypothesis is specified so that it supports the research question. The data from the trial are used to see if there is evidence that the null hypothesis can be rejected, in which case we accept the alternative hypothesis and conclude that the research question is answered in the affirmative. If the null hypothesis is not rejected we cannot conclude that it is true, for lack of evidence against the null hypothesis is not evidence for it; the null hypothesis can never be proven.[1]

Hypothesis tests

Hypothesis tests are either two-sided or one-sided. A two-sided test has a null hypothesis of the form 'novel and standard treatment are the same', and an alternative hypothesis of the form 'novel and standard treatment are not the same'. A one-sided test has a null hypothesis of the form 'novel treatment is the same as, or worse than, the standard treatment', with an alternative hypothesis of the form 'novel treatment is better than the standard treatment'.

For any hypothesis test it is necessary to specify the type I error that we are prepared to accept. This is the probability of rejecting the null hypothesis (and thus accepting the alternative hypothesis) when, in fact, it is true (false positive). Arbitrarily, this is traditionally set at 0.05. The *P* value obtained from testing the hypothesis with the data obtained from a trial is the probability of obtaining the observed result (or a more extreme one) when the null hypothesis is true. If the *P* value is less than the specified type I error we accept the alternative hypothesis.

Rejection of the null hypothesis of a one-sided test allows us to conclude that the novel treatment is better than the standard treatment (i.e. it gives us the direction of the treatment effect). If the null hypothesis is not rejected, however, we cannot draw any conclusions as to whether the treatments are the same or if the novel treatment is worse than the standard treatment. For a two-sided test, rejection of the null hypothesis tells us that the treatments are different but not whether the novel treatment is efficacious or harmful. In fact, since no two treatments can ever be identical the null hypothesis of equality will always be rejected, given a sufficiently large sample size. In practice, however, for two treatments that are fairly similar, trials with practical sample sizes will not be able to reject the null hypothesis. To obtain useful information from a two-sided test, it is therefore necessary to consider it as a combination of two one-sided tests:[2]

(1) The null hypothesis of 'novel treatment is as good as or worse than standard treatment' and an alternative hypothesis of 'novel treatment is better than standard treatment'; and
(2) The null hypothesis of 'novel treatment is as good as or better than standard treatment' and an alternative hypothesis of 'novel treatment is worse than standard treatment'.

Test (1) tests for novel treatment efficacy and test (2) tests for novel treatment harm. The sum of the specified type I errors for each test gives the type I error for the two-sided test. In practice this decomposition is rarely explicitly made and it has to be assumed that the type I error has been divided equally between tests. Another way of looking at this is to note that all two-sided tests with a type I error set at 0.05 are really one-sided tests with the type I error set at 0.025 (for both harm and efficacy).

Professor Owen is Consultant Cardiologist at Kent and Canterbury Hospital. He undertakes research in collaboration with Canterbury Christchurch University into the prevention of cardiovascular disease. He is a past chairman of the local NHS research ethics committee and is currently a member of the University research ethics committee. He has an interest in the conduct and interpretation of clinical trials and has recently completed an MSc in Applied Statistics.

The appropriateness of two- or one-sided tests has been the subject of controversy for over half a century.[3–10] One-sided tests have greater power for a given sample size or alternatively require a smaller sample size[10] to obtain the same power (the reduction in sample size is always less than 50% and is typically around 20%). This has led some investigators to use one-sided tests inappropriately to make a result of borderline statistical significance appear significant – for example, converting a *P* value of 0.06 (two-sided) to a *P* value of 0.03 (one-sided). Thus, one-sided tests are treated with suspicion and are generally discouraged.[11] Regulatory authorities on both sides of the Atlantic, and many journal editors, stipulate that two-sided tests should be used to evaluate clinical trial data. Consequently, investigators are compelled to use two-sided tests irrespective of their appropriateness.

This compulsion to use two-sided tests when a one-sided test would be appropriate has a number of important consequences. Additional patients need to be recruited and exposed unnecessarily to an experimental treatment that may be harmful. In addition, these patients are then not available to participate in other studies. The need to recruit additional patients increases the cost of the trial, potentially diverting resources away from the development of other treatments. For rare conditions, where the viability of a trial may be constrained by the ability to recruit patients, the extra patients needed for a two-sided test may make the trial impractical. In such situations new treatments would not be developed, to the detriment of patients.

What are the arguments for two- and one-sided tests? The proponents of two-sided tests assert that, because the treatment effect may go in the direction of efficacy or harm, it is necessary to test for both these possibilities. The proponents of one-sided tests assert that the type of test used should reflect the research question, which for the majority of phase III trials that seek to establish the efficacy of a novel treatment, means using a one-sided test. The principle is that there is no interest in establishing that a treatment is harmful (although this possibility is acknowledged), as if it has not been established as beneficial it will not be used in clinical practice and not be submitted for regulatory approval. Thus, a one-sided test is appropriate. When a novel treatment is tested against a standard treatment and the treatment effect goes in favour of the latter, the former may be either harmful (i.e. worse than placebo) or merely less good than the standard treatment. In either case, the novel treatment would not supersede the standard treatment, so again a one-sided test is appropriate. Some proponents of two-sided tests reject one-sided tests as a consequence of misunderstanding their rationale.[10] They mistakenly believe that the rationale for a one-sided test is that if the investigator is confident his treatment is not harmful then there is no need to test for this possibility. Clearly an investigator cannot *know* this until the trial has been undertaken. Such prior beliefs have no place in classical hypothesis testing and are not the basis for using a one-sided test. There will be some circumstances when a two-sided test is clearly appropriate. For example, interim analyses should be two-sided as interest lies in identifying both efficacy and harm. The occurrence of either will result in early termination of the trial. When a trial has two or more novel treatments, two-sided tests are necessary to compare them. Two-sided tests may well be appropriate for hypothesis-generating pilot studies and dose finding trials.

These principles carry over in a straightforward manner to equivalence trials. The objective of such trials is to test whether a novel treatment is equivalent to the standard treatment. In this context equivalent means nearly the same as, which is established when the treatment effect differs by no more than a pre-specified quantity – the minimum clinically relevant difference. Such trials are appropriate when the novel treatment has advantages on cost, side effects, etc., but is not thought likely to be more efficacious. Thus, interest lies in establishing that the novel treatment is neither greatly worse nor greatly better than the standard treatment. This is therefore a two-sided problem requiring a two-sided test. In some circumstances investigators may not be interested in establishing that the novel treatment is no better than the standard treatment, but merely that it is no worse. Such trials are known as non-inferiority trials and the pre-specified minimum clinically relevant difference is known as the non-inferiority margin. Such trials present a one-sided problem requiring a one-sided test.

To clarify the practical differences of using two- or one-sided tests it is helpful to consider an example. The VEST study[12] examined the use of the experimental drug Vesnarinone in patients with heart failure. Theoretical considerations and pilot studies had suggested that this agent might be efficacious in patients with heart failure. There were three arms to the study (placebo and doses of 30 and 60 mg). Over 1000 patients were recruited to each arm, based on sample size calculations for two-sided hypothesis tests. Interim analyses were undertaken and the trial was not stopped early. The trial was terminated as planned when the specified number of deaths had occurred in the placebo group. The two-sided hypothesis test (placebo against 60 mg dose) revealed a treatment effect in the direction of harm with a *P* value of 0.02. This is not highly statistically significant, but was sufficient for the investigators to conclude that Vesnarinone was unlikely to have a place in the treatment of patients with heart failure. Vesnarinone has not been submitted for regulatory approval. Now suppose that this trial had been conducted using a one-sided hypothesis test to answer the research question: is Vesnarinone efficacious in patients with heart failure? First we note that fewer patients would have been needed; and secondly, as with the two-sided case the trial would not have been stopped early. The final one-sided analysis would not have rejected the null hypothesis and we could conclude that there was no evidence that the treatment was efficacious (this is not the same as concluding that there was evidence of harm). Vesnarinone would therefore not have been submitted for regulatory approval, the same outcome as with the two-sided test, but achieved with fewer patients. Importantly, fewer patients would have been exposed to Vesnarinone and fewer patients would have died as a consequence. Had Vesnarinone turned out to be efficacious this would have been established by both two- and one-sided tests, but the latter would have achieved it with fewer patients.

The CAST[13] trial has been cited[10] as an example of the importance of using a two-sided test to ensure that when the novel treatment turns out to be harmful, this is

identified and patients are not put at avoidable risk. In fact, on the contrary, this trial is an example of the appropriate use of a one-sided test. The trial tested the hypothesis that the use of the antidysrhythmic agent Flecainide to suppress asymptomatic ventricular rhythm disturbances following a myocardial infarction would improve prognosis. The trial used a one-sided test to determine if Flecainide was efficacious: it was not designed to test if Flecainide was harmful, as there was no interest in this. Sample size calculations were undertaken on this basis. It was determined that 4,400 patients were required. The study was discontinued (after 2,309 patients had been recruited) at an interim analysis because the predetermined stopping conditions for harm had been met. The important point to note here is that interim analyses should be two-sided (as above) but the final analysis should be one-sided, as was intended in this case. Flecainide is not used for the purpose tested in the trial, as the trial has not demonstrated it to be efficacious. The fact that harm happened to be demonstrated at an interim analysis is not relevant, other than to stop the trial. The fact that the trial was stopped early in no way undermines the case for a one-sided test.

Investigators are likely to be interested in the possibility of harm (although this is not the main research question). This can be assessed by a prespecified secondary outcome, as is usual for safety considerations. As such it would not carry the same weight as that of a primary analysis and would be considered as hypothesis generating.

Summary

To enable investigators to test for the possibility of harm with the same rigour as that for efficacy it is necessary to recruit additional patients to undertake a two-sided test over and above those required for a one-sided test. When the novel treatment turns out to be efficacious the two-sided approach means that additional patients will have been recruited unnecessarily. When, however, the novel treatment turns out to be harmful the two-sided approach means that these additional patients will have been recruited for the sole purpose of demonstrating that the novel agent is harmful. This necessarily involves observing an increase in morbidity and/or mortality in the treatment group. How can this be ethically justified? When patients consent to participate in clinical trials they are led to believe that the purpose of the trial is to determine if a novel treatment is efficacious (although clearly potential risks are explained). They are not told that the purpose of the trial is also to determine whether the novel treatment is harmful – if they were they may be less likely to consent to participate.

These points should be considered carefully before undertaking trials with two-sided tests and patients should be fully informed of the purpose of the trial in this respect.

References

1 Fisher RA. *The Design of Experiments*. London: Oliver and Boyd, 1935
2 Dannet CW, Gent M. An alternative to the use of two-sided tests in clinical trials. *Stat Med* 1996;**15**:1729–38
3 Jones LV. Tests of hypothesis: One-sided vs two-sided alternatives. *Psychol Bull* 1949;**46**:43–6
4 Burke CJ. A brief note on one-sided tests. *Psychol Bull* 1953;**50**: 384–7
5 Overall JE. Tests of one-sided versus two-sided hypotheses in placebo controlled trials. *Neuropsychopharmacology* 1990;**3**:233–5
6 Peace KE. One-sided or two-sided p values: Which most appropriately address the question of drug efficacy? *J Biopharm Stats* 1991;**1**:133–8
7 Dubey SD. Some thoughts on one-sided and two-sided tests. *J Biopharm Stats* 1991;**1**:139–59
8 Overall JE. A comment concerning one-sided tests of significance in new drug applications. *J Biopharm Stats* 1991;**1**:157–60
9 Bland JM, Altman DG. One and two-sided tests of significance. *BMJ* 1994;**309**:248
10 Moye LA, Tita ATN. Defending the rationale for the two-sided test in clinical research. *Circulation* 2002;**105**:3062–5
11 Altman DG. *Practical Statistics for Medical Research*. London: Chapman and Hall, 1991
12 Cohn J, Goldstein SC, Feenheed S, *et al*. A dose dependent increase in mortality seen with Vesnarinone among patients with severe heart failure. Vesnarinone study group. *N Engl J Med* 1998;**339**:1810–16
13 The CAST Investigators. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *N Engl J Med* 1989;**321**:406–12