ARTICLE

Making sense of equivalence and non-inferiority trials

Ben Ewald
Senior lecturer
Epidemiology and General Practitioner Centre for Clinical Epidemiology and Biostatistics
University of Newcastle
New South Wales

Key words
clinical trials, statistics, treatment efficacy

Aust Prescr 2013;36:170–3

SUMMARY

New drugs are usually compared to a placebo. Sometimes it may be unethical to give patients a placebo, so the new drug is compared with standard treatment.

Trials which compare treatments may not be designed to show that one treatment is superior. These are known as non-inferiority or equivalence trials.

Non-inferiority trials aim to show that the new drug is no worse than standard treatment. Equivalence trials aim to show the new treatment is no better and no worse.

An equivalence boundary should be set before the trial. This is the definition of what would be the minimum important difference between the treatments.

There are several traps in the interpretation of trials of non-inferiority or equivalence. The results can be influenced by many factors including the size of the equivalence boundary and whether an intention-to-treat or 'per protocol' analysis is used.

Introduction

Many clinical trials compare new drugs to placebo. Once there are proven effective treatments for a disease, the clinically important question is whether a new treatment is better than the old one. We would like new treatments to be progressively better than old treatments, but it becomes increasingly difficult to demonstrate the superiority of new treatments if the current treatment achieves most of the possible clinical benefit. New treatments may still be desirable, even when they do not have a superior treatment effect, if they are safer, cheaper, or more convenient. It is also good to have new options for patients who are intolerant of current drugs. This has led to some new drugs coming into use after trials that show they are as good as the old drug, without ever being compared to placebo. These 'active control' trials include equivalence and non-inferiority trials (see Box).

Definition of equivalence and non-inferiority trials

Equivalence trials aim to show that there is no significant difference between treatments

Non-inferiority trials aim to show that one treatment is not significantly worse than another treatment

Confidence intervals

Every experimental trial is subject to the play of random factors that could add a few more successes or failures to the particular group of patients given the experimental or control treatment. While the observed point estimate of effect is the most probable true result, there is a range of values in which we can be confident that the true result will lie. By convention the 95% confidence interval is examined. If this interval does not include the relative risk of 1.0 we accept that there is a difference between treatments. Failing to prove a difference is, however, not the same thing as proving there is no difference.

In Fig. 1 the results of several hypothetical trials are displayed in the same format as a forest plot used in meta-analysis. The line of no difference is when the relative risk is 1.0. A value below 1.0 favours the experimental treatment and a value above 1.0 favours the control. In Fig. 1 all the trials give the same point estimate suggesting that the treatments are equal. In trial 1 there is a wide confidence interval (due to small sample size or poor measurement) so the new treatment could be 10 times better or 10 times worse. This is of no use to a clinician trying to decide whether to use the new treatment. Trial 2 also has a point estimate of no difference, but the 95% confidence interval is smaller due to a larger sample size, and the interval in the larger trial 3 is smaller still. To shrink the 95% confidence interval to zero would take an infinite sample size which is impossible. It is necessary to make a judgement about a boundary that is close enough to 1.0 that we will accept that the result shows equivalence.

Equivalence or non-inferiority?

From a clinician's perspective, if a new drug is not better we at least want to know it is not worse than the old drug. Statisticians use different methods if they are testing only one end of the equivalence boundary. In effect clinicians do not care how far the 'good' end of the 95% confidence interval goes, just as long as the 'bad' end is within an acceptable limit. For this reason most trials will use a non-inferiority analysis, and although it sounds weaker it is just as good as an equivalence analysis.

In an equivalence trial the statisticians look at both ends of the boundary. Is the new drug no better or no worse? (see Fig. 2).

Setting the equivalence boundary

The two main methods for setting the equivalence boundary are clinical and statistical. The statistical method is more widely used.

Clinical method

A group of clinicians give their opinion on the 'minimum clinically important difference' which is the smallest difference that they or their patients would think was important. This might be a change of 5 mmHg of blood pressure or 10 mm on a pain scale. The basis for these arbitrary judgements is rarely explained, such as why it is 10 mm rather than 8 mm or 13 mm. Small differences like this might make all the difference in statistical testing. There is also the problem that although 5 mmHg is a small change in blood pressure it still makes a difference to stroke risk, and 10 mm on a pain scale still affects the patient's comfort.

Statistical method

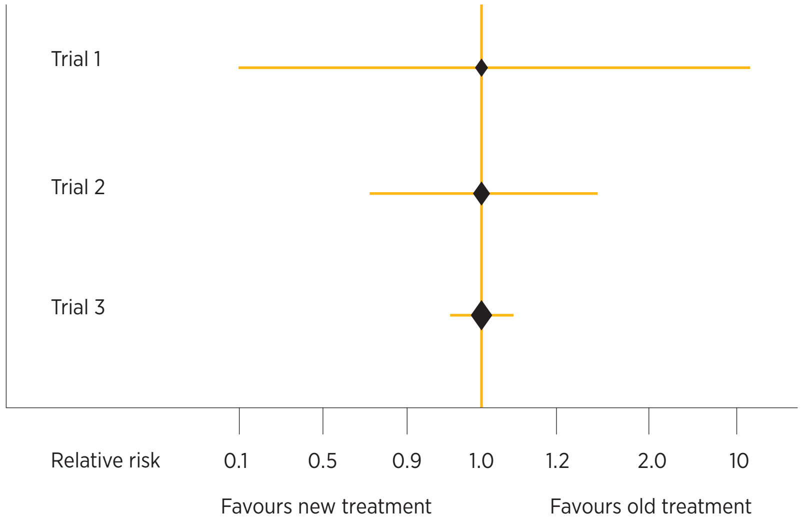
The statistical method relies on examining the difference between the standard treatment and placebo. This is derived from the original placebo-controlled studies of efficacy. The equivalence boundary could be set to prove that the new treatment is no worse than the outcome for placebo in the original trials, although sometimes it is set to be 50% better than placebo.

As the minimum clinically important difference between active treatments is usually smaller than the treatment benefit found in placebo-controlled trials, the sample size required for non-inferiority trials is generally bigger than for superiority trials.

Differences in analysis

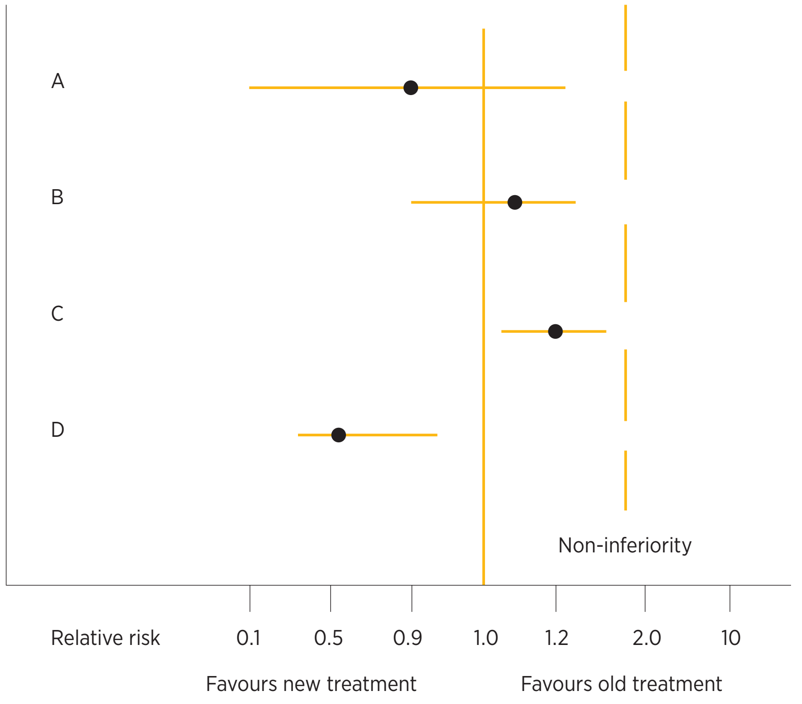
The analysis of a superiority trial should be by 'intention to treat'. This means that the outcomes are measured in patients who were randomised even if they did not take the treatment or complete the trial. This analysis is conservative, because if anything it biases the result towards a null effect. In a non-inferiority trial an intention-to-treat analysis with its bias toward a null effect could be misleading. It is good practice to also perform a 'per protocol' analysis, in which groups are defined by who took the drug rather than just by randomisation.

Fig. 1 **Effect of trial size on the 95% confidence interval**



Confidence intervals reduce with larger trial sizes (represented by the size of the diamonds). A narrow confidence interval increases the chance that the observed result is close to the true value.

Fig. 2 **Spectrum of outcomes included in non-inferiority trials**



All of A B C D are 'non-inferior' to old drug
A + B are not statistically superior or inferior
C is statistically inferior
D is statistically superior

Interpretation

In assessing non-inferiority trials the issues of trial design, such as randomisation, blinding and follow-up, are considered in the same way as they are in trials looking for superiority. However, there are other considerations when assessing non-inferiority trials. It can be difficult to judge if the statistical equivalence boundary has been appropriately set. There is scope for pharmaceutical companies to set the equivalence boundary too wide, making it easy to claim equivalence when it may not exist.

Traps

Proving that two drugs are equivalent could mean that they are both ineffective or even harmful. The evidence for the old drug must be considered when relying on an equivalence trial to show evidence for the efficacy of a new drug. If drug A is superior to placebo and drug B is proved non-inferior to drug A (and becomes the drug of choice because it is cheaper and easier to administer) but later drug C is proved to be non-inferior to drug B, can we be certain drug C is superior to placebo? This problem has been called ‘biocreep’ and could lead to the acceptance of progressively worse treatments if non-inferiority is blindly accepted. It can be avoided by selecting the most effective drug in the class as the control for non-inferiority trials, even if this is not the drug in most common use.

Can the data from a failed superiority trial be used to demonstrate non-inferiority (Fig. 2 – Trial A, B, C)? Can the data from a non-inferiority trial that goes particularly well be used to demonstrate superiority (Fig. 2 – Trial D)? These are controversial questions, however there is a view that if the non-inferiority boundary is selected a priori a failed superiority trial can be taken as evidence of non-inferiority, although the test for statistical significance should be adjusted for multiple comparisons.

Example 1

The RE-LY trial set out to demonstrate non-inferiority of dabigatran versus warfarin for preventing stroke in patients with atrial fibrillation.[1]

Choice of boundary

The non-inferiority boundary was chosen as a relative risk of 1.46 for stroke or systemic embolism. This boundary was derived on statistical grounds from a meta-analysis of trials of warfarin versus placebo and chosen as 50% of the proven benefit of warfarin. Although this may have satisfied the statisticians it is clearly not acceptable to clinicians that a new drug could allow 46% more strokes and still be regarded as non-inferior. As it turned out, dabigatran 110 mg dose reduced the relative risk to 0.91 (95% confidence interval 0.74–1.11). The upper boundary of an 11% increase in strokes is probably acceptable to clinicians and patients.

Analysis

An intention-to-treat analysis was performed. As 99.9% of the patients were followed up, loss to follow-up did not introduce bias. The proportions discontinuing treatment were 14.5% for the low dose and 15.5% for the high dose of dabigatran and 10.2% for warfarin, possibly biasing the relative risk towards 1.0. This could have given a spurious non-inferiority result if the point estimate had been a relative risk greater than 1.0, but would not have had this effect on a point estimate less than 1.0. A per protocol analysis was not done.

The trial set out to demonstrate non-inferiority, but ended up showing superiority of the 150 mg dose over warfarin with a relative risk of 0.66 (95% confidence interval 0.53–0.82) so the intention to treat analysis is appropriate for a claim of superiority (see Fig. 3). If the trial had claimed non-inferiority by showing the relative risk for stroke had a 95% confidence interval extending just short of the boundary (for example to 1.45), it should not have been accepted. To me the possibility that the new drug could lead to a 45% increased risk of stroke would be unacceptable.

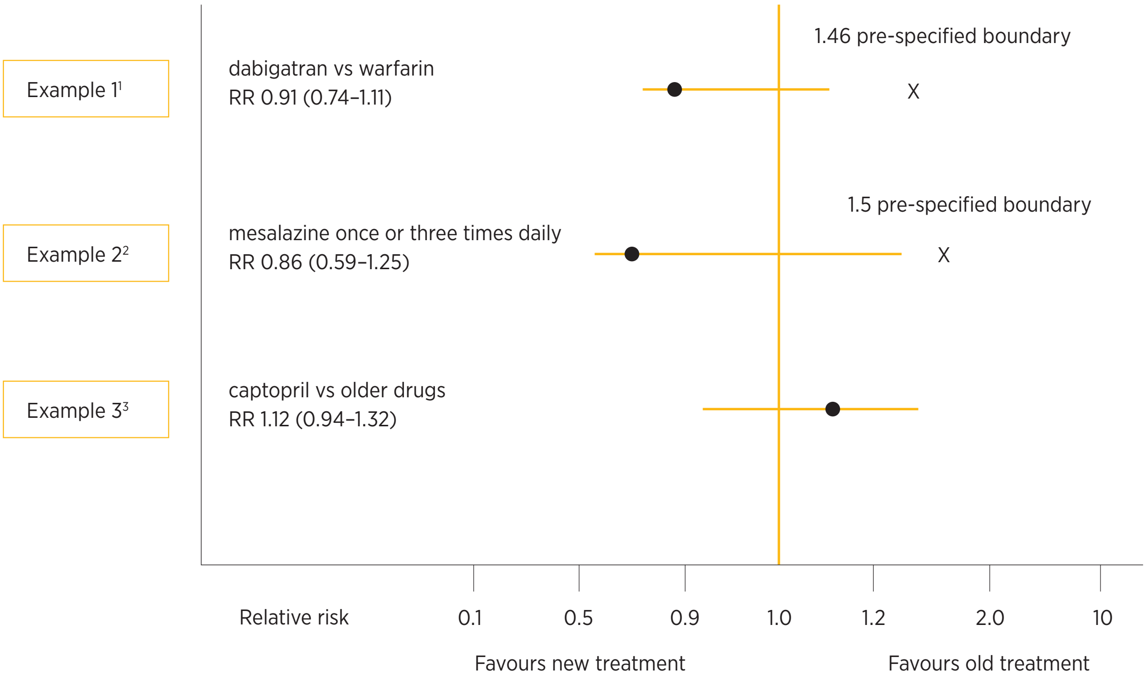
Example 2

A study set out to show that once-daily dosing with mesalazine granules was as good as three-times-daily dosing at inducing remission during first episodes of ulcerative colitis. The rate of non-remission at eight weeks was 24.3% in the three-times-daily group, but only 20.9% in the once-daily group. The relative risk was 0.86 (95% confidence interval 0.59–1.25). The non-inferiority boundary was set at relative risk of 1.6, and as the upper limit of the confidence interval is clear of this, non-inferiority is accepted (see Fig. 3).[2]

Example 3

The Captopril Prevention Project compared the efficacy of the drug to older antihypertensives in the prevention of stroke, myocardial infarction and cardiovascular death. The authors presented both intention to treat and per protocol analyses, showing somewhat worse outcomes for captopril. The adjusted relative risk was 1.12 (95% confidence interval 0.94–1.32). The authors claimed equivalence, but did not pre-specify an equivalence boundary. Patients may not view the possible 32% increase in serious outcomes as equivalent (see Fig. 3).[3]

Fig. 3 **Examples of equivalence and non-inferiority trials**



RR relative risk (confidence interval)

Conclusion

Equivalence and non-inferiority trials are becoming more frequent as use of a placebo control group is no longer justified in many diseases. As well as all the usual issues of trial quality, interpretation of these trials is complicated by the need to establish and justify a minimal clinically important difference. The minimum difference established by statistical means may include values that are not acceptable to clinicians, so this is an issue that warrants close attention.

Conflict of interest: none declared

Acknowledgements: Dr John Attia gave helpful comments on an early draft

REFERENCES

1. Connolly SJ, Ezekowitz MD, Yusuf S, Eikelboom J, Oldgren J, Parekh A, et al; RE-LY Steering Committee and Investigators. Dabigatran versus warfarin in patients with atrial fibrillation. N Engl J Med 2009;361:1139-51.
2. Kruis W, Kiudelis G, Racz I, Gorelov IA, Pokrotnieks J, Horynski M, et al. Once daily versus three times daily mesalazine granules in active ulcerative colitis: a double-blind, double-dummy, randomised, non-inferiority trial. Gut 2009;58:233-40.
3. Hansson L, Lindholm LH, Niskanen L, Lanke J, Hedner T, Niklason A, et al. Effect of angiotensin-converting-enzyme inhibition compared with conventional therapy on cardiovascular morbidity and mortality in hypertension: the Captopril Prevention Project (CAPPP) randomised trial. Lancet 1999;353:611-6.

FURTHER READING

Scott IA. Non-inferiority trials: determining whether alternative treatments are good enough. Med J Aust 2009;190:326-30.