Commentary
Noninferiority trials
Steven M Snapinn

Merck Research Laboratories, West Point, Pennsylvania, USA

Abstract

Noninferiority trials are intended to show that the effect of a new treatment is not worse than that of an active control by more than a specified margin. These trials have a number of inherent weaknesses that superiority trials do not: no internal demonstration of assay sensitivity, no single conservative analysis approach, lack of protection from bias by blinding, and difficulty in specifying the noninferiority margin. Noninferiority trials may sometimes be necessary when a placebo group can not be ethically included, but it should be recognized that the results of such trials are not as credible as those from a superiority trial.

Keywords: assay sensitivity, blinding, clinical trials, equivalence, intention-to-treat

Introduction

In one of the biggest dilemmas facing cardiovascular clinical research, clinical trials are increasingly being required to show benefits on clinical end-points rather than surrogate end-points, while at the same time the incremental benefits of newer treatments are getting smaller. These two factors have a huge impact on sample size, which has led some investigators to design trials to show that the new treatment has an effect similar to that of the standard, rather than outright superiority. Recent examples of fibrinolytic trials that have demonstrated similar effects of two drugs are ASSENT (Assessment of the Safety and Efficacy of a New Thrombolytic)-2, GUSTO (Global Use of Strategies to Open Occluded Coronary Arteries)-III, and COBALT (Continuous Infusion Versus Double-Bolus Administration of Alteplase) [1–4]. However, as discussed by several authors [5–8], there are issues with trials of this type that make them considerably less credible than superiority trials.

Terminology

'Noninferiority' is a relatively new term that has not been universally adopted, and in the past noninferiority and equivalence trials, which have an important distinction, have both been referred to as 'equivalence trials'. To make the confusion even worse, both of these terms are somewhat misleading.

It is fundamentally impossible to prove that two treatments have exactly equivalent effects. Equivalence trials, therefore, aim to show that the effects differ by no more than a specific amount. This tolerance is known as the equivalence margin, and is often denoted by the symbol δ. In an equivalence trial, if the effects of the two treatments differ by more than the equivalence margin in either direction, then equivalence does not hold. Noninferiority trials, on the other hand, aim to show that an experimental treatment is not worse than an active control by more than the equivalence

ITT = intention-to-treat.

margin. An improvement of any size fits within the definition of noninferiority. Bioequivalence trials are true equivalence trials, but it is difficult to imagine any trial comparing the clinical effects of an experimental treatment and active control that would not more appropriately be termed a noninferiority trial.

Assay sensitivity

Probably the greatest difficulty with noninferiority trials relates to the issue of assay sensitivity, or the ability of a specific clinical trial to demonstrate a difference between treatments if such a difference truly exists. A trial that successfully demonstrates superiority has simultaneously demonstrated assay sensitivity. However, a noninferiority trial that successfully finds the effects of the treatments to be similar has demonstrated no such thing. A well-executed clinical trial that correctly demonstrates the treatments to be similar can not be distinguished, on the basis of the data alone, from a poorly executed trial that fails to find a true difference. Therefore, a noninferiority trial must rely on an assumption of assay sensitivity on the basis of information external to the trial, such as the quality control procedures or the reputation of the investigator.

The International Conference on Harmonization guidelines [9] list a number of factors that can reduce assay sensitivity. These include poor compliance with the study medication, poor diagnostic criteria, excessive variability of measurements, and biased end-point assessment. In order to be credible, therefore, noninferiority trials must attempt to avoid these factors to every possible extent, and even then might not be able to escape suspicion. For example, a successful superiority trial can be very credible despite a moderately large rate of discontinuation from study drug, but a successful noninferiority trial would be less so, because discontinuations can obscure a true treatment effect and thus reduce assay sensitivity.

Analysis of noninferiority trials

Intention-to-treat (ITT) is widely recognized as the most valid analytic approach for superiority trials that involve long-term end-point follow up, because it adheres to the randomization procedure and is generally conservative [10]. Although some might argue that the ITT analysis is overly conservative, most would agree that a positive ITT analysis of a superiority trial is convincing.

Unfortunately, no such conservative analysis exists for noninferiority trials. For example, including data after study drug discontinuation in the analysis, as ITT does, tends to bias the results toward equivalence, which could make a truly inferior treatment appear to be noninferior. The per-protocol analysis, on the other hand, excludes data from patients with major protocol violations. However, excluding these data can substantially bias the results in either direction. For example, patients in a survival trial might discontinue study medication due to the development of heart failure, which is a strong risk factor for mortality.

Therefore, noninferiority trials are often analyzed using ITT and per-protocol approaches, and only if both approaches support noninferiority is the trial considered positive. Even in this case, however, the possibility of bias can not be ruled out, and it can be awkward to have different analytic strategies for superiority and noninferiority trials.

Blinding

Blinding is one of the most important bias-avoiding techniques available to clinical trialists. It is not always feasible to blind the investigator or patient to the treatment regimen, but blinded end-point determination is nearly always possible and should be done, particularly when the end-point has a subjective component. However, blinding does not protect against bias nearly as well in a noninferiority trial as it does in a superiority trial. In a superiority trial, a blinded investigator can not consciously or subconsciously influence the results to support a preconceived belief in superiority, but in a noninferiority trial there is no protection against a blinded investigator biasing the results toward a preconceived belief in equivalence by assigning similar ratings to the treatment responses of all patients.

Specifying the noninferiority margin

It can be quite difficult to specify an appropriate noninferiority margin. There are two basic approaches, both of which have serious drawbacks. One approach is to specify the equivalence margin on the basis of a clinical notion of a minimally important effect. However, this is clearly subjective, and it is possible with this approach to set the equivalence margin to be greater than the effect of the active control, which could lead to harmful treatments fitting within the definition of noninferiority.

To avoid this, the equivalence margin is often chosen with reference to the effect of the active control in historical placebo-controlled trials. When the equivalence margin is chosen in this way, there is some basis on which to claim that a positive noninferiority trial implies that the new treatment is superior to placebo. However, this claim requires an assumption that the effect of the active control in the current trial is similar to its effect in the historical trials. That assumption can be undermined by differences with respect to design features (eg the patient population, dosage regimen of the active control, end-point definition or concomitant therapies), or by an inconsistency in the effect of the active controls among the historical placebo-controlled trials (beyond that expected by random chance). For this reason, the equivalence margin usually includes some type of buffer. Rather than basing it on the full predicted effect of the active control, it is often based on the lower bound of a confidence interval for that effect

(accounting for within-trial and trial-to-trial variability), or on preservation of a specific fraction (eg 50%) of the effect of the active control.

Sample size

Although noninferiority trials typically have smaller sample sizes than active-controlled superiority trials, they can have considerably larger sample sizes than placebo-controlled trials. This is because the equivalence margin is often much smaller than the treatment difference for which a placebo-controlled trial is powered. In addition, the sample size of a noninferiority trial is very sensitive to the assumed effect of the new treatment relative to the active control; the sample size can be considerably larger if the two treatments are assumed to be equivalent than if the new treatment is assumed to be slightly more effective than the active control.

For example, COBALT used a noninferiority margin of 0.4%, and the sample size of 7169 was based on the assumption that the experimental treatment was superior to the standard by 0.9% [3]. Had the investigators assumed that the treatments were equivalent, a sample size of approximately 50 000 would have been required [4]. By contrast, a placebo-controlled trial intended to demonstrate a 2.5% reduction (from 10 to 7.5%) with 80% power would require approximately 4000 patients.

Assessing superiority

Both noninferiority and superiority can be assessed in the same clinical trial without statistical penalty. One consequence of this is that most, if not all, active-controlled clinical trial protocols should define a noninferiority margin and include a noninferiority hypothesis. If a protocol has a noninferiority hypothesis but no superiority hypothesis, it is still valid to do both tests. However, if a protocol has a superiority hypothesis but no noninferiority hypothesis, adding a noninferiority hypothesis after the trial is complete would be problematic because of the subjective component in the definition of the noninferiority margin.

Conclusion

There are inherent problems with noninferiority trials that make their results clearly less credible than those of a placebo-controlled trial. However, it is not always possible to include a placebo treatment for ethical reasons, and there will always be a need for clinical trials to test new treatments that are either no more effective than the standard (but which may offer some other advantage, such as better safety or more convenient dosing) or offer such a small increase in efficacy that the size of a superiority trial would be prohibitive. Trialists will continue to do noninferiority trials because there is often no alternative. Anyone performing such a trial or evaluating its results, however, must be aware of the issues and account for them appropriately.

References

1. Assessment of the Safety and Efficacy of a New Thrombolytic (ASSENT-2) Investigators: **Single-bolus tenecteplase compared with front-loaded alteplase in acute myocardial infarction: the ASSENT-2 double-blind randomised trial.** *Lancet* 1999, **354**:716–722.
2. The Global Use of Strategies to Open Occluded Coronary Arteries (GUSTO III) Investigators: **A comparison of reteplase with alteplase for acute myocardial infarction.** *N Engl J Med* 1997, **337**:1118–1123.
3. The Continuous Infusion Versus Double-Bolus Administration of Alteplase (COBALT) Investigators: **A comparison of continuous infusion of alteplase with double-bolus administration for acute myocardial infarction.** *N Engl J Med* 1997, **337**:1124–1130.
4. Ware JH, Antman EM: **Equivalence trials.** *N Engl J Med* 1997, **337**: 1159–1161.
5. Garbe E, Röhmel J, Gundert-Remy U: **Clinical and statistical issues in therapeutic equivalence trials.** *Eur J Clin Pharmacol* 1993, **45**:1–7.
6. Jones B, Jarvis P, Lewis JA, Ebbutt AF: **Trials to assess equivalence: the importance of rigorous methods.** *Br Med J* 1996, **313**:36–39.
7. Hauck WW, Anderson S: **Some issues in the design and analysis of equivalence trials.** *Drug Information J* 1999, **33**:109–118.
8. Hwang IK, Morikawa T: **Design issues in noninferiority/equivalence trials.** *Drug Information J* 1999, **33**:1205–1218.
9. International Conference on Harmonization: **Choice of control group in clinical trials.** *Federal Register* 1999, **64**:51767–51780.
10. Lewis JA, Machin D: **Intention to treat: who should use ITT?** *Br J Cancer* 1993, **68**:647–650.

Author's affiliation: Merck Research Laboratories, West Point, Pennsylvania, USA

Correspondence: Steven M Snapinn, Merck Research Laboratories, BL3-2, West Point, PA 19486, USA. Tel: +1 610 397 2935; fax: +1 610 397 2931; e-mail: snapinns@merck.com

commentary

review

reports

primary research