Superiority, Equivalence, and Non-Inferiority Trials

Emmanuel Lesaffre, Dr.Sc.

Abstract

When the aim of the randomized controlled trial (RCT) is to show that one treatment is superior to another, a statistical test is employed and the trial (test) is called a superiority trial (test). Often a nonsignificant superiority test is wrongly interpreted as proof of no difference between the two treatments. Proving that two treatments are equal in performance is impossible with statistical tools; at most, one can show that they are equivalent. In an equivalence trial, the statistical test aims at showing that two treatments are not too different in characteristics, where "not too different" is defined in a clinical manner. Finally, in a non-inferiority trial, the aim is to show that an experimental treatment is not (much) worse than a standard treatment. In this report, the three types of trials are compared, but the main focus is on the non-inferiority trial. Special attention is paid to the practical implications when setting up a non-inferiority trial. Illustrations are taken from a clinical trial in osteoarthritis and from thrombolytic research.

When the aim of the study is to show that an experimental (E) treatment is superior to a control (C) treatment, the RCT is called a superiority trial and the associated statistical test is a superiority test. With a significant result, one concludes in a superiority trial that E is different in effect from C, and when the observed result is in favor of E, we conclude that E is statistically, significantly better performing than C. However, in the case of a nonsignificant result, one cannot claim a better performance of E over C, but neither can one claim that E is equally as good as C. In fact, when E and C are not identical treatments, there will be always some small difference in effects, and for each small true difference (Δ), one can establish a sample size such that, with high probability, the null hypothesis (H_0) of equal effect is rejected.[1] Unfortunately, often a nonsignificant result is wrongly interpreted as absence of evidence.

Relative to the development of medical treatments, it is becoming increasingly difficult to develop more powerful drugs, hence the pharmaceutical companies are looking for new treatments that have approximate the same efficacy but demonstrate better quality in other aspects. This trend stimulated the development of RCTs that aim to show that an experimental treatment is not (much) inferior to the control treatment. Such trials are called non-inferiority trials.

In this report, I will endeavor to contrast the basic philosophy of superiority, equivalence, and non-inferiority trials. Further, I will illustrate the practical problems pertaining to setting up and conducting a non-inferiority trial. The concepts will be introduced via a fictive example with thrombolytic drugs, but some illustrations will also be taken from Bingham and colleagues, who performed an osteoarthritis treatment study.[2] Lesaffre[1] used primarily the first part of the two studies comparing etoricoxib 30 mg qd (ET) and celecoxib 200 mg qd (CE) to placebo (PL). In the second part of the trial, two studies were conducted to compare the relative performance of ET and CE with a non-inferiority design. The non-inferiority part of the study will be discussed in greater detail in this paper. For further details on the study, the reader is referred to Lesaffre[1] and the original report.[2]

Superiority Trial

Examples involving the treatment of acute myocardial infarction (MI) patients with thrombolytic drugs are taken predominantly from the area of expertise of the current investigator. In this therapeutic domain, non-inferiority trial

Emmanuel Lesaffre, Dr.Sc., is from the Biostatistical Centre, Catholic University of Leuven, Leuven, Belgium, and Department of Biostatistics, Erasmus MC, Rotterdam, The Netherlands.
Correspondence: Emmanuel Lesaffre, Dr.Sc., Department of Biostatistics, Erasmus MC, Rotterdam, The Netherlands; e.lesaffre@erasmusmc.nl.

Lesaffre E. Superiority, equivalence, and non-inferiority trials. Bull NYU Hosp Jt Dis. 2008;66(2):150-4.

designs were initiated more than a decade ago.

As a first example, the 1993 GUSTO-I study consisted of four arms, but here only the following treatment arms will be considered: streptokinase plus intravenous heparin (C) versus rt-PA (recombinant tissue plasminogen activator) plus intravenous heparin (E).[3] The primary end point is 30-day mortality, a binary outcome equal to 1, when the patient died within 30 days after the onset of an acute myocardial infarction (MI), or 0, otherwise. Those receiving the control treatment numbered 10,370 and the experimental treatment, 10,348. The observed percentage of patients who died within 30 days was 7.4% for the control treatment and 6.3% for the experimental treatment. The chi-square test evaluating H_0: $\Delta = 0$ resulted in a p-value of 0.0028 and a (two-sided) 95% CI for Δ: [0.36%, 1.73%]. The conclusion was that rt-PA has a significantly lower 30-day mortality rate than streptokinase. Suppose that the same result had been obtained with twice 1000 patients, then p = 0.37 and 95% CI for Δ: [-1.21%, 3.21%]. What is the conclusion now? Do we conclude that the two treatments have equal efficacy? Clearly, for a larger sample size, we are able to achieve the difference, but with the total sample size of 2000 patients, it was not possible to see a clear distinction. As explained by Lesaffre,[1] there is an absence of evidence for a different treatment effect, but there is clearly no evidence of absence for a different treatment effect.

Remember that for a (two-sided) superiority trial, the null and alternative hypotheses are H_0: $\Delta = 0$ and H_A: $\Delta \neq 0$, respectively (see Lesaffre[1]). It is useful to compare the statistical aspects of superiority, equivalence, and non-inferiority by making use of the following well known property in statistical testing theory (see Lesaffre[1]): when comparing two means using an unpaired t-test, a significant result at 0.05 ($p < 0.05$) is equivalent to a (two-sided) 95% CI for the difference of means, not including $\Delta = 0$. On the other hand, a nonsignificant result ($p \geq 0.05$) is equivalent to the 95% CI, including $\Delta = 0$. For comparing two proportions, this result holds only approximately, but in practice the equivalence

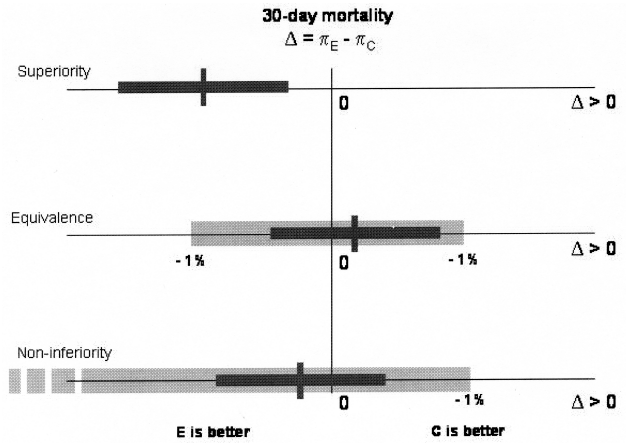

Figure 1 Superiority, equivalence, and non-inferiority concepts.

almost always holds. In Figure 1, the results are shown for a fictive RCT, whereby the 95% CI for the difference in 30-day mortality rates does not embrace zero, implying that E is shown to be superior to C, at $\alpha = 0.05$. In the remainder of the paper, the assumed Δ used in the sample size calculation for a superiority trial will be denoted by Δ_S.

Equivalence Trial

Continuing with the above fictive example, suppose that the aim of the RCT is to show that treatments E and C have equal efficacy. We know from Lesaffre[1] that this is impossible to show with statistical tests; hence, we have to resort to a practical definition of "equally good." Suppose that the clinicians agree upon a (positive) value Δ_E such that two treatments can be considered not to differ (too much) when their true Δ lies in an interval of clinical equivalence $[-\Delta_E, \Delta_E]$. Then the two treatments could be called equivalent if the observed difference and its 95% CI are completely inside the interval of clinical equivalence. In terms of null and alternative hypotheses, proving equivalence boils down to rejecting the H_0: $\Delta > \Delta_E$ or $\Delta < -\Delta_E$ in favor of the alternative hypothesis, H_A: $\Delta_E \leq \Delta \leq \Delta_E$, with an appropriate statistical test. When comparing proportions, the chi-square test needs a simple adaptation. Further, for an equivalence trial, a significant result ($p < 0.05$) means that the two treatments are equivalent, according the definition of equivalence as defined clinically. In Figure 1, a significant result was obtained from a fictive equivalence trial, where Δ_E was defined as 0.01 (1%). In case $p \geq 0.05$, corresponding to a 95% CI that crosses one or both boundary values of the interval of clinical equivalence, the two treatments cannot be called equivalent.

Most equivalence trials are bioequivalence trials that aim to compare a generic drug with the original commercial drug, to show that they have the "same" pharmacokinetic (PK) profile, expressed by the most common PK variables: Cmax, Cmin and AUC (area under the curve). Observe that here Δ_E can be defined as the value for which "the patient will not detect any change in effect when replacing one drug by the other." Further, it is worth mentioning that in bioequivalence trials often a 90% CI is used instead of the 95% CI. Observe, however, that non-inferiority trials (next section) are sometimes (wrongly) referred to as equivalence trials.

Non-Inferiority Trial

Equivalence trials are not appropriate for therapeutic trials. Suppose, for example, that in the second fictive example the 95% CI for Δ crosses the left boundary of the interval of clinical equivalence. This implies that E might be superior to C or, at most, Δ_E worse than C. In a non-inferiority trial, one only defines an upper bound Δ_{NI}, and when the 95% CI for Δ lies left to Δ_{NI}, treatment E will be called non-inferior to C. Observe that in a non-inferiority trial, only one boundary of the classical two-sided 95% CI is looked at. This, in fact, corresponds to a one-sided 97.5% CI that is unbounded

at one side (here, the left side). This one-sided confidence interval corresponds to a one-sided significance level of 0.025. Thus, while in superiority and equivalence tests, one performs the significance testing at a two-sided 0.05 alpha level (which is, in fact, a combination of two one-sided tests) a standard non-inferiority test is performed nowadays at a one-sided 0.025 level. Further, while in superiority trials it is still popular to use p-values for the reporting of the results, in non-inferiority trials, it is customary to use the (two-sided) 95% CI.

Demonstrating non-inferiority necessitates rejecting the null hypothesis (H_0: $\Delta > \Delta_{NI}$) in favor of the alternative hypothesis (H_A: $\Delta < \Delta_{NI}$) with an adapted classical statistical test (as for an equivalence test). A significant result ($p < 0.025$) means, in this case, that the experimental treatment is not (much) worse, i.e., non-inferior, to the control treatment, whereby the conclusion "non-inferior" depends on the chosen value for Δ_{NI}. Figure 1 shows the significant result of a fictive non-inferiority trial, where Δ_{NI} was defined as 0.01 (1%). In case $p \geq 0.025$, i.e., when the right boundary of the one-sided 97.5% CI (two-sided 95% CI) exceeds Δ_{NI}, we have failed to show that the experimental treatment is non-inferior. This does not imply, however, that we have shown treatment C to be superior to E. Observe that the interval of non-inferiority is unbounded on the left side and is, therefore, also referred to as the region of non-inferiority.

Two Examples of a Non-Inferiority Trial

Comparison of Etoricoxib 30 mg and Celecoxib 200 mg in the Treatment of Osteoarthritis

The two randomized, three-arm double-blinded clinical trials described by Bingham and coworkers[2] each contain a non-inferiority assessment of ET versus CE for the treatment of osteoarthritis of the knee and hip, using a time-weighted average (TWA) change from a baseline over 12 weeks of: 1. the WOMAC (Western Ontario and McMaster Universities) Pain Subscale (WOMACPA), 2. the WOMAC Physical Function Subscale (WOMACPH), and 3. the Patient Global Assessment of Disease Status (PGADS) (see Lesaffre[1] or the original publication[2] for more details).

The experimental treatment CE was defined to be non-inferior to ET when the upper bound of the two-sided 95% CIs for the difference between CE and ET was not more than 10 mm for the three primary end points: WOMACPA, WOMACPH, and PGADS. Thus, in order that non-inferiority be shown, all three conditions had to be satisfied. Tables 2A and 2B in Bingham and associates[2] show that these conditions were satisfied for the two studies (95% CIs, entirely below the upper bound). The investigators' conclusion was, therefore, that "etoricoxib 30 mg is comparable with celecoxib 200 mg in osteoarthritis." At first glance, the investigators used a tough criterion for "non-inferiority," only it is not clear how they chose $\Delta_{NI} = 10$ mm.

Finally, in Tables 2A and 2B, not only the 95% CIs are reported but also the p-values. However, all p-values for the comparison of CE with ET are nonsignificant. It is important to realize, though, that these p-values correspond to superiority tests, comparing CE with ET for the three primary end points. Consequently, they do not pertain to the null and alternative non-inferiority hypotheses described above.

ASSENT II Study

One of the first non-inferiority trials in the therapeutic domain of thrombolytics for the treatment of acute MI patients was ASSENT II.[4,5] Specifically, ASSENT II compared the 30-day mortality rates of single-bolus tenecteplase (E) with accelerated infusion of alteplase (C).

At first, the upper bound of the region of non-inferiority Δ_{NI} was taken, 0.01(1%). This implies that E could show a 1% higher 30-day mortality rate than C (with 95% CI inside the region). There was, however, uncertainty about the true 30-day mortality rate (π_C) for the control treatment. Based on historical studies, one concluded that the true value for π_C probably lies between 0.05 (5%) and 0.10 (10%). In relative terms, allowing 1% worse performance for E when the true rate is 5% is 20% compared to 10% when the true control rate is 10%. Therefore, $\Delta_{NI} = 0.01$ for small values of π_C was considered to be unacceptably large and in a second step one changed the non-inferiority region to: when $\pi_C > 0.072$: $\Delta = \pi_E - \pi_C < 0.01$ and when $\pi_C \leq 0.072$: $\pi_E/\pi_C < 1.14$. Observe that the two criteria are equivalent for $\pi_C = 0.072$. For the choice of the cut point, 0.072, we refer the reader to the ASSENT-2 Investigators (1999) study.[4] The values of 0.01 and 1.14 were obtained from the historical study (GUSTO III) and the meta-analysis (see also below). Observe that, at that time, the criterion for non-inferiority was less stringent, since the one-sided 95% CI (two-sided 90% CI) could be used for its definition.

The results for the primary end point were 6.16% for E (tenecteplase) and 6.18% for C (alteplase). The observed 30-day mortality rate was lower than 7.2%. The protocol specified that the relative criterion had to be used in this case. The observed relative risk was 0.997, with 90% CI = [0.904, 1.101], both smaller than 1.14. The conclusion was, therefore, that tenecteplase is non-inferior to alteplase.

A Critical Discussion of the Non-Inferiority Trial

Reasons for Choosing a Non-Inferiority Design

Showing non-inferiority of the experimental treatment E versus the control treatment C can be of interest because of the following:

- *It is not ethically possible (anymore) to do a placebo-controlled trial*. At the time alteplase was introduced (The GUSTO Investigators, 1993), it was not possible any longer to compare its performance against placebo, since streptokinase had been shown to significantly reduce the 30-day mortality in acute MI patients. At that time, it was believed, though, that the experimental treatment had to be better (and proved to be the case, too). This could not be assumed for tenecteplase at the

time of the planning of the ASSENT II study, and hence a non-inferiority design seemed to be a solution.

- *E is not expected to be better than C on a primary efficacy end point, but is better on secondary end points or is safer*. For example, celecoxib is a nonsteroidal antiinflammatory drug (NSAID) that is effective as an analgesic and antiinflammatory agent, but possibly causes serious gastrointestinal side effects. The typical osteoarthritis patient is at higher risk for NSAID gastropathy because of potential interrelated factors and the use of higher doses of NSAIDs for longer periods. However, in clinical studies, COX-2 inhibitors, such as etoricoxib, have a similar efficacy as NSAIDs in the treatment of osteoarthritis pain but with less gastrointestinal side effects. Therefore, a non-inferiority design was found to be appropriate in the Bingham and colleagues study.[2]
- *E is not expected to be better than C on a primary efficacy end point but is cheaper to produce or easier to administer*. For the ASSENT II study, the administration of tenecteplase required the injection of a single bolus, which is definitely easier than the administration of alteplase, a drug that requires a 90 minute (accelerated) infusion.[4]
- *E is not expected to be better than C on a primary efficacy end point in a clinical trial, but compliance will be better outside the clinical trial and hence efficacy will be better outside the trial.* This item is related to the previous one, but it is important to highlight it separately. Suppose that the control treatment requires strict adherence to an oral treatment; for instance, requiring patients to take their medication in regular and narrow time windows. In an RCT, patients are closely followed-up and motivated to comply with the recommended treatment administration, such that adherence to the drug will be relatively high. Outside the RCT, patients will be much less disciplined, and thus the treatment effect will likely be much lower outside the RCT. Consequently, an experimental treatment with about the same efficacy as the control treatment, but with an easier treatment administration, will probably have a higher efficacy outside the trial than the control treatment.

Determining the Non-Inferiority Boundary

Two methods are currently used to choose the non-inferiority margin Δ_{NI}. First, a direct comparison between E and C might be envisaged. In that case, Δ_{NI} is based on purely clinical grounds. It requires clinicians to choose a cut point. Such a cut point is, however, often difficult to choose, especially in life-threatening diseases. For example, how can we motivate the choice of 1% in 30-day mortality for acute MI patients? No clinician will claim that it does not matter that the experimental treatment will save fewer lives than the control treatment. It is also a problem when you are the first in setting up a non-inferiority design in a particular therapeutic area. In such a situation, there is no benchmark value, and you might need to negotiate with the FDA (Food and Drug Administration, Rockville, Maryland, United States) or the EMEA (European Medicines Agency, London, United Kingdom) about an acceptable boundary value. The second reasoning is based on an indirect comparison with the (putative) placebo treatment. In that case, Δ_{NI} is determined based on statistical arguments ensuring that E will likely to be superior to placebo, even when it is allowed to be somewhat worse than C. The second approach could be further refined by combining the clinical and statistical reasoning, resulting in a Δ_{NI} that is clinically acceptable and ensures the superiority of E over placebo.

The second approach requires an estimate of the likely difference of the efficacy between the control treatment and placebo. This could be obtained from, for example, a meta-analysis. Suppose that we find that this difference is 2%, with a 95% CI of [1.7%, 2.3%]. Taking the uncertainty into account on the differential efficacy of the control treatment versus placebo, the experimental treatment can be, at most, 1.7% worse than the control treatment to guarantee (with 95% confidence) that the experimental treatment is better than placebo and thus $\Delta_{NI} < 1.7\%$. It needs to be checked with the clinicians that 1.7% is clinically acceptable or a further decrease is needed. A popular rule is to take half of the minimal estimated difference between control and placebo treatment, e.g., here it would be $1.7\%/2 = 0.85\%$.

The second approach needs a well established, predictable, and quantifiable effect of the control treatment. This requires that multiple placebo-controlled RCTs are available. If this is not the case, then there is always the risk that the experimental treatment cannot be "proven" to be better than placebo. It is also required that the control treatment shows a relatively constant better performance than placebo (constancy assumption). If not, it will be hard to choose the correct value for Δ_{NI}. The problem of an inappropriately estimated effect of control versus placebo treatment is, unfortunately, always lurking because the definition of placebo can change drastically over time.

With respect to the above non-inferiority trials, it seems that Δ_{NI} was chosen on purely clinical grounds for the osteoarthritis trial, while the second approach was used for the determination of the ASSENT II region of non-inferiority.

Intention-to-Treat or Per-Protocol Analysis?

The recommended patient population for a superiority trial is the intention-to-treat (ITT) population. This consists of all patients who were randomized and, thus, all patients who were supposed to be treated. Protocol violators, patients that miss one or more visits, patients that dropout, patients that were randomized into the wrong group, and so forth are analyzed according to the planned treatment. While the ITT approach is not ideal, it is considered to be the most appropriate approach for superiority trials since the ITT

principle implies a conservative effect on the outcome of the trial. That is, when the study is poorly conducted, it will be unlikely that the experimental treatment can be proven to be more efficacious than the control treatment.

For a non-inferiority trial, the ITT analysis does not have a conservative effect. Dropouts and a poor conduct of the study might direct the results of the two arms toward each other. Another possibility is to consider the per-protocol (PP) population, which consists of only the nonprotocol violators. It is, however, not clear whether a PP analysis has the desired conservative effect for a non-inferiority design. Since no good solution to this problem exists, the recommended pragmatic approach for a non-inferiority trial is to perform both analyses, ITT and PP, and hope that the two analyses will confirm each other.

Sample Size Calculations

For a superiority trial, the necessary sample size (N) depends on (among other things) Δ_S, the clinically important difference. For a non-inferiority trial, the necessary sample size depends on (among other things) Δ_{NI}, the upper bound for non-inferiority. When $\Delta_{NI} = \Delta_S$, the necessary sample size for the NI trial is the same under the assumption of $\Delta = 0$ as the sample size for the corresponding superiority trial testing $\Delta = 0$ and if $\Delta = \Delta_S$. On the other hand, Δ_S is typically (much) larger than Δ_{NI}, which causes the sample size for a non-inferiority trial often to be much larger than that of a superiority trial.

Combining Non-Inferiority and Superiority in One Trial

When non-inferiority was planned for and was established on completion of the study, it is natural to ask whether one can go further and also prove superiority with the same data, without a statistical penalty for multiple testing. Or, if one failed to prove superiority, can one then still aim to show non-inferiority with the same data and without penalty.

The following results can be shown (see, e.g., Moyé and coworkers[6]): Applied to the same population (ITT or PP), non-inferiority and superiority both can be tested at 0.05, without a statistical penalty because of the closed testing principle (see also Lesaffre[1]). When non-inferiority is tested first in the PP population and then superiority in the ITT population, there is, again, no penalty to be paid, but the reverse testing order requires a multiplicity adjustment.

Some Final Remarks

The RCT endeavoring to show superiority is the gold standard in clinical trial research. Of the three types of trials, the results of a superiority trial are the simplest regarding interpretation. Due to the difficulty of launching more powerful drugs on the market, the pharmaceutical industry has been forced to look for drugs that may not improve the current, most efficacious medications but are better on other aspects of the treatment. As long as this implies an improvement for the quality of life for the patient, such a development can only be welcome. However, for the clinical researcher, non-inferiority trials are far from easy to conduct and sometimes to interpret. In mathematics, the following rules with A, B, and C as quantitative variables are taught in primary school: 1. if A = B and B = C, then A = C and 2. if A < B and B < C, then A < C. However, what if "≈" means "equivalent" and "ni" means non-inferior? Can we then state the following?

- If A ≈ B and B ≈ C, then A ≈ C?
- If A ni B and B ni C, then A ni C?
- If A ni B and B ni P, then A ni P (called biocreep)?

The answer to these questions is not evident and will depend on the choice of the interval of equivalence and region of non-inferiority, respectively. Different definitions of equivalence and non-inferiority will make the life of a clinical researcher and clinical decision-maker more difficult and harder for them to decide the true message of a non-inferiority trial. Finally, Le Henanff and associates report on a review of non-inferiority and equivalence trials published between January 1, 2003, and December 31, 2004.[7] They concluded the reporting of such trials reveals important deficiencies.

Disclosure Statement

The author has no financial or proprietary interest in the subject matter or materials discussed, including, but not limited to, employment, consultancies, stock ownership, honoraria, and paid expert testimony.

References

1. Lesaffre E. Use and misuse of the p-value. Bull NYU Hosp Jt Dis. 2008;66(2):146-9.
2. Bingham CO III, Sebba AI, Rubin BR, et al. Efficacy and safety of etoricoxib 30 mg and celecoxib 200 mg in the treatment of osteoarthritis in two identically designed, randomized, placebo-controlled, non-inferiority studies. Rheumatology. 2007;46:496-507.
3. The GUSTO Investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. New Eng J Med. 1993;329:673-82.
4. ASSENT-2 Investigators. Single-bonus tenecteplase compared with front-loaded alteplase in acute myocardial infarction: the ASSENT-2 double-blind randomised trial. Lancet, 1999;354:716-22.
5. Lesaffre E, Bluhmki E, Wang-Clow F, et al. The general concepts of an equivalence trial, applied to ASSENT-2, a large-scale mortality study comparing two fibrinolytic agents in acute myocardial infarction. Eur Heart J. 2000;21:1-5.
6. Moyé LA. *Multiple Analyses in Clinical Trials: Fundamentals for Investigators*, New York: Springer-Verlag, 2003.
7. Le Henanff A, Giraudeau B, Baron G, Ravaud P. Quality of reporting of noninferiority and equivalence randomized trials. JAMA. 2006;295(10):1147-51.