European Heart Journal (2012) **33**, 1318–1324
doi:10.1093/eurheartj/ehs099

**REVIEW**

***Statistical review***

# Non-inferiority study design: lessons to be learned from cardiovascular trials

**Stuart J. Head[1], Sanjay Kaul[2], Ad J.J.C. Bogers[1], and A. Pieter Kappetein[1]***

[1]Department of Cardio-Thoracic Surgery, Erasmus University Medical Center, PO Box 2040, 3000 CA Rotterdam, The Netherlands; and [2]Division of Cardiology, Cedars-Sinai Medical Center, Los Angeles, CA, USA

*Received 5 December 2011; revised 6 March 2012; accepted 28 March 2012; online publish-ahead-of-print 7 May 2012*

The non-inferiority trial design has gained popularity within the last decades to compare a new treatment to the standard active control. In contrast to superiority trials, this design is complex and is based on assumptions that cannot be validated directly. Many readers and even investigators, therefore, have difficulty grasping the full methodological nature of non-inferiority trials. Non-inferiority margins are often arbitrarily chosen such that a favourable margin can bias a trial towards declaring non-inferiority. Pitfalls of non-inferiority trials are not fully appreciated, and without having identified these shortcomings, objective conclusions from non-inferiority trials cannot be made. This methodological review elaborates on what is a non-inferiority trial, why such a trial is performed, what the hazards are, and how conclusions from non-inferiority trials are derived, by providing examples of recent cardiovascular trials.

**Keywords** Non-inferiority • Methodology • Randomized trial • Trial design

## Introduction

Unlike superiority trials that are designed to show that one treatment is better than another, a non-inferiority trial is designed to show that a new treatment is 'not unacceptably worse' than the current standard therapy. Since the introduction of non-inferiority trials in the mid-1990s it has been debated whether such trials should be performed.[1,2] The design of a non-inferiority trial is complicated and is founded on assumptions that are difficult to verify.[3–6] Readers often fail to fully understand the concept, statistical approaches, and conclusions; even some trialists may have difficulties with grasping the sense of a non-inferiority study. Non-inferiority studies often have 'substantial methodological flaws' with the risk of incorrectly claiming non-inferiority.[3] This could potentially expose patients to the possibility of receiving a treatment that is inferior to the 'gold standard'. In addition, the reporting of analyses and conclusions has been shown to be misleading in a review of 116 non-inferiority trials.[3,7]

In the last few years, several cardiovascular trials have been published that compared surgical to catheter-based therapies for the treatment of heart diseases, with a great impact on clinical practice.[8–10] More trials are currently underway and it is crucial that these and future trials are adequately designed, well performed, rigorously analysed, and prudently interpreted.[11]

In this review, we discuss the aspects of non-inferiority trials; when to perform such a study, how to design a non-inferiority trial, and how to derive conclusions from such a trial. To elaborate on these topics, examples of recent cardiovascular trials are provided.

## Superiority, equivalence, non-inferiority

A superiority trial is designed to show that a new treatment is better than an active control or placebo. The null hypothesis states that no difference between treatments exists. The trial is determined to reject this hypothesis and show a statistically significant difference in favour of the new treatment. In equivalence trials, which are rarely performed, the difference between two treatments is predefined as $\Delta$, and the goal of the trial is to demonstrate that treatment with either therapy is equally good and the confidence intervals (CIs) do not exceed a difference of $-\Delta$ and $+\Delta$.

A non-inferiority trial is different as it is designed not to show that treatments are equal, or 'not different', but that the new treatment is not unacceptably worse than, or 'non-inferior' to, an active control. Statistically, such a study differs from an equivalence trial because the $\Delta$ is only one-sided towards $-\Delta$. Non-inferiority is claimed if the lower bound of the CI of the treatment effect

* Corresponding author. Tel: +31 (0)10 70 34375, Fax: +31 (0)10 70 39933, Email: a.kappetein@erasmusmc.nl

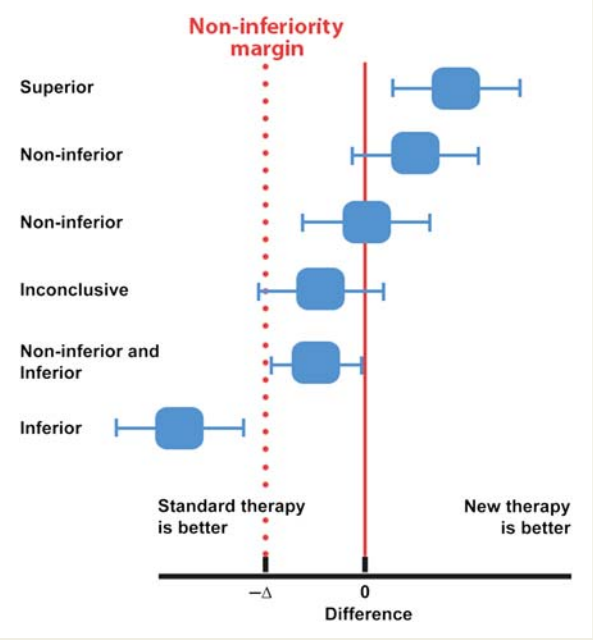


**Figure 1** Possible conclusions from non-inferiority trials.

difference does not exceed $-\Delta$, thus meaning that the risk of it being inferior is within acceptable boundaries (*Figure 1*).

## Why a non-inferiority trial?

Non-inferiority trials have become more popular in the last decades, especially in cancer and cardiovascular studies. A common misunderstanding is that this is caused by safety and efficacy regulations, which would suggest that a new therapy first needs to show non-inferiority before it can be tested in a superiority trial. However, non-inferiority trials were originally designed for studies in which it is unethical to include a placebo arm. For cancer and cardiovascular conditions where a 'gold standard' therapy already exists, it would be unethical to perform a placebo-controlled trial with a newly introduced treatment. For example, elderly patients with symptomatic severe aortic stenosis are generally treated by means of surgical aortic valve replacement (SAVR). Whenever patients are considered to be at too high a risk for surgery, they are managed medically. Transcatheter aortic valve replacement (TAVR) is a new less invasive therapy suited for these extreme high-risk patients, with initially good results from the PARTNER trial.[12,13] However, in lower risk patients TAVR has to compete with the gold standard SAVR, which shows excellent long-term results in these patients. A TAVR vs. medical management trial would therefore be unethical in lower risk patients due to the superiority of SAVR over medical management in patients who are good candidates for surgery. Patients randomized to medical management would then not receive established effective therapy.

Even if a new treatment is shown to be non-inferior to the 'gold standard' therapy with regard to an efficacy endpoint, it would still need to demonstrate an ancillary benefit, i.e. lower procedural risks (safety), favourable costs, or improved convenience for it to be considered the preferred treatment. In the previous example, if TAVR shows non-inferior efficacy (and safety), its preference over SAVR might be potentially justified due to the lower invasiveness (avoidance of sternotomy and cardiopulmonary bypass) and reduced length of stay. An example where a non-inferiority trial would be adequate in a pharmacologic trial is the comparison between warfarin and new anticoagulant drugs. Warfarin has been the standard anticoagulant therapy for over 60 years but has some disadvantages including the requirement for routine monitoring of the international normalized ratio (INR). Several new drugs that are more convenient with regard to drug administration have been shown to demonstrate non-inferiority compared with warfarin.[14,15]

Not only are there clinical indications to perform a non-inferiority trial, but also the costs of a randomized trial are very high, and the stakes for companies are crucial. In a non-inferiority trial investigators can choose unreasonably wide margins and high active control event rates that yield lower sample sizes, and thus improve the trial efficiency, i.e. achieve a positive trial result at a minimized cost. For example, the Stroke Prevention Using Oral Thrombin Inhibitor in Atrial Fibrillation (SPORTIF) V trial used an unreasonably generous non-inferiority margin of a 2% absolute risk difference (ARD) and an expected warfarin event rate of 3.1% per year [equivalent to a relative risk margin of $(3.1 + 2)/3.1 = 1.65$]; with 90% power this produced a sample size of 3156 patients.[16] Using the more accurate expected warfarin event rate of 1.9% per year derived from pooled historical data, the study would have needed 4875 patients for a similar 90% power and a relative risk margin of 1.65 (equivalent to an ARD margin of 1.23%). The sample size would even be 8190 patients if the observed warfarin event rate of 1.2% per year had been used for the sample size calculation.[11] Thus only 39% (3156/8190) of the actually needed sample was included, thereby drastically reducing costs. Although the cost of a trial is merely one of the factors influencing trial design, it should not be the main contributor.

## Methodology of non-inferiority trials

One major issue with a non-inferiority trial is that, unlike a superiority trial, it is biased towards non-inferiority if the trial is poorly designed and sloppily conducted.[17] Part of the basis of a randomized trial is the expected event rate with the corresponding sample size calculation. A non-inferiority trial has the same principle, but an additional non-inferiority margin is included. This margin quantifies when the new therapy is considered to be non-inferior to the standard therapy. Several factors need to be considered during the trial design before a reasonable margin can be adopted. If these factors are not taken into account, it could lead to a phenomenon called 'biocreep' or 'technology creep'; an inferior therapy is granted non-inferiority and becomes the control group in future trials, ultimately leading to an active therapy being no better than a placebo.[4]

### Choice of margin: absolute vs. relative risk difference

A non-inferiority margin can be chosen as an ARD or risk ratio (RR). It is recommended to use a relative difference to account for changes in event rates; fixed RRs provide more conservative

**Table 1 Inflation of the relative risk in the SPORTIF V trial**

| | Expected | Pooled historical | Observed |
|---|---|---|---|
| Standard Rx event rate | 3.1%/year | 1.9%/year | 1.2%/year |
| New Rx event rate acceptable | 5.1%/year | 3.9%/year | 3.2%/year |
| RR | 1.65 | 2.05 | 2.67 |

RR, relative risk difference; Rx, treatment.

margins in trials in which the event rate is unpredictable or the observed rate is lower than expected.[11] In the previously used example of the SPORTIF V trial, non-inferiority was met using an ARD of 2%, even with the observed event rate of 1.2% instead of the expected 3.1%/year. This lower event rate caused inflation of the RR from 1.65 to 2.67 (*Table 1*). Had the investigators fixed the RR at 1.65 (and correspondingly used a more conservative ARD margin of 0.78% [$(1.65 \times 1.2) - 1.2$], non-inferiority would not have been met. However, it is evident that conservative margins result in larger sample sizes.

Margins based on ARD can potentially introduce a bias towards non-inferiority, since it can result in an underpowered trial due to lower than expected event rates.[9,11] For example, in the recent PRECOMBAT trial that compared percutaneous coronary intervention (PCI) with coronary artery bypass grafting (CABG) for left main disease, the expected event rate in the CABG arm was 13%, and the pre-specified margin was an ARD of 7%.[9] In an analysis with a one-sided alpha of 0.025, the upper bound of the difference was 6.3%. Because this was below the predefined margin of 7%, the investigators declared non-inferiority. Had they fixed the margin as an RR [$(13 + 7)/13 = 1.54$], the upper bound of the RR would be 2.12 (1.30, 95% CI: 0.81–2.12), thereby not allowing a claim of non-inferiority. In trials that use an ARD, a judgement of non-inferiority would be more convincing if analyses on the basis of absolute and relative difference were concordant.[3,11]

## Active control event rate

It is crucial that the active control event rate be chosen properly, since an overestimation can result in an underpowered trial. Frequently the event rate is unsubstantiated. For example, the PRECOMBAT trial used a 1-year event rate of 13% based on a previously published meta-analysis, while the actual observed event rate was only 6.7%.[9,18] The investigators could, however, have foreseen differences in the event rate. The meta-analysis was not representative of the current clinical practice as it included four trials that enrolled patients between 1995 and 2000 treated with bare-metal stents, while PRECOMBAT enrolled patients between 2004 and 2009 that were treated with drug-eluting stents. Furthermore, their own clinical practice demonstrated low rates similar to PRECOMBAT, but these data were not taken into account when performing the sample size calculation.[19] An interim analysis during the trial would have demonstrated lower than expected event rates and a sample size adjustment would have been appropriate given the contemporary data.[19] Although the trial extended the primary endpoint to 2 years, this still did not result in an adequate number of events.[20]

In some instances, there are no previous trials to reliably estimate the expected active control event rate. In such cases, investigators have no other option but to extrapolate from their own experiences or use pooled feasibility data for a propensity-matched analysis. An advantage of this technique is that it can provide a ratio of the new treatment vs. the active control. This is, however, often cumbersome due to diverse 'all-comer' patients treated with the control and the highly selected patients treated with the new intervention.

## Nature of events

One must be aware of the fact that the margin should be based on the number and nature of the events that are included in a composite endpoint. The use of composite endpoints that are driven by 'softer' events poses a dilemma in the estimation of the margin. On the one hand, one is willing to accept a greater degree of inferiority (given the ancillary benefits), thereby resulting in a wider margin. On the other hand, 'softer' events occur more frequently and inflate the event rate, which would require more stringent margins. Whether composite endpoints should include both safety and efficacy outcomes remains debatable. For example, in the SYNTAX trial the composite of death, stroke, myocardial infarction, and repeat revascularization was used as the primary endpoint. Some argue that repeat revascularization should not have been included in the endpoint, since this was a 'softer' efficacy event. The primary endpoint of non-inferiority was not met in the analysis that included revascularization, while PCI would have been non-inferior to CABG in the composite analysis without repeat revascularization. However, this composite of death, myocardial infarction, and stroke was not a predefined endpoint. Had it been chosen as the primary endpoint of the trial, sample size adjustments due to a lower event rate would have been required, resulting in a prohibitively large sample size.[21]

The recently published EVEREST II trial randomized patients to percutaneous mitral valve repair or mitral valve surgery.[8] For the primary endpoint, the investigators chose a combination of clinical (death and surgery for mitral valve dysfunction) and echocardiographic endpoints (grade $\geq 3+$ mitral regurgitation), which is unusual for a device vs. surgery trial. Ideally, the regurgitation endpoint should not have been included in the primary endpoint, but this 'softer' and more frequent endpoint drove the event rate. A composite endpoint of death or need for surgery (hard, but less frequent, endpoints) would have required a prohibitively large sample size. In contrast, the PARTNER trial had a clinical primary endpoint, while valve function was considered a secondary endpoint.[12]

## Clinical relevance

A crucial step in determining a margin is to contemplate what difference between therapies is clinically acceptable. An overly conservative margin might result in a high risk of not being able to claim non-inferiority when it actually is non-inferior. Conversely, overly liberal margins could result in a high risk of claiming non-inferiority when it actually is not non-inferior. A reasonable margin would be best derived from a combination of factors: the expected event rate, the duration of follow-up, and the number

and nature of the events. However, arbitrary clinical judgment and the sponsor budget are of a great influence, resulting in a somewhat subjective non-inferiority margin.

A formal approach for choosing the margin is based upon a combination of statistical reasoning and clinical judgment.[4,5,11] The first step is to reliably estimate the efficacy of the active control compared with placebo, often derived from a meta-analysis of historical placebo-controlled superiority trials. The lower 95% CI of this effect is the largest acceptable non-inferiority margin, M1, to provide assurance that the new treatment is at least better than placebo.[4,5,22] The second step in selecting the margin is choosing a reasonable fraction of the control effect (M1) that needs to be preserved, typically set at 50% of M1. This new non-inferiority margin is called M2, and is typically based upon clinical judgment. An example of a trial using this method is the RE-LY trial (*Table 2*).[14] The investigators used a meta-analysis of trials of vitamin K antagonist compared with control therapy in patients with atrial fibrillation. The hazard ratio of 1.46 was used as the margin in RE-LY, which was defined by using half the upper bound of the 95% CI derived from the estimated effect of control therapy over warfarin.

## Follow-up

The duration of follow-up for the primary endpoint is important as well. The shorter the follow-up, the more conservative a margin should be. While after 1 year a certain difference in events might be acceptable, the same difference at 30 days could raise serious concerns regarding the safety of the treatment. This becomes more important whenever a trial is designed with an ARD (Δ) as the non-inferiority margin, as opposed to a trial with a hazard ratio.[23] As shown in *Figure 2*, data from the SYNTAX trial show that the hazard ratio remains constant over time, while the absolute difference may increase.

## Statistical power

The minimal acceptable standard for statistical power in superiority trials generally is 80% with a two-sided alpha of 0.05. Both superiority and non-inferiority trials should ideally be designed with a ≥90% statistical power. In non-inferiority trials this is more crucial, since lower power biases the results towards non-inferiority. In addition, although practice varies, a one-sided alpha of 2.5% is considered to be more robust for non-inferiority assessment; the CI is wider and therefore more likely to cross the non-inferiority margin.

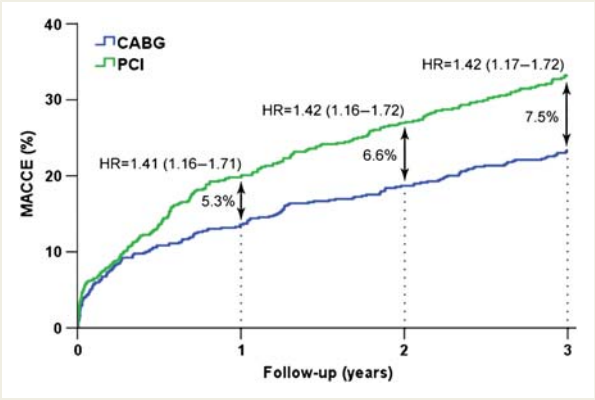


**Figure 2** The influence of the length of follow-up on the non-inferiority margin. Data from the SYTNAX trial demonstrate that the duration of the follow-up is of different influence on an absolute risk difference or risk ratio.

## Assumptions

An adequately powered superiority trial allows one to conclude that a new treatment is superior to placebo. Conclusions from non-inferiority trials, however, are based on assumptions that cannot be verified directly.[11] In contrast to superiority trials, a major issue in non-inferiority trials is that although a new treatment can be non-inferior to the active control, it does not necessarily imply that the active control is more effective, and to what extent, than a placebo. This is referred to as the 'constancy' and 'assay sensitivity' principle. The effect of the active control in relation to the placebo could be different from historical data.[4,5,24,25] For example, in a trial comparing PCI with CABG, if the non-inferiority margin exceeds the treatment difference between CABG and medical treatment, non-inferiority of PCI does not mean it would be superior to medical treatment. To overcome these problems, one can include a third (placebo) arm in a trial, so that a check of the superiority of the active control over the placebo ('assay sensitivity') is available. In case of the example, the PCI vs. CABG trial should include a medical treatment arm, to show that CABG is indeed superior to the placebo. If a third arm is not included, investigators can perform a separate analysis in which the new treatment is compared with historical placebo data, but this relies on the assumption that the observed outcomes are constant over trials ('constancy'). This is frequently not the case as treatment effects can be heterogeneous due to differences in patient populations, outcome definitions, treatment allocation, or other study factors.

# Reporting of non-inferiority trials

## Analysis

Conclusions from non-inferiority trials are highly sensitive to the method of analysis. The intention-to-treat analysis, typically preferred as the more robust analytical framework in a superiority trial, can be biased towards non-inferiority. For example, if a large number of patients 'cross-over'—patients randomized to treatment A receive treatment B or vice versa—groups will be 'blended' and it is likely that outcomes will be similar in an intention-to-treat analysis. In a superiority trial this strengthens the final effect of a difference, because the analysis makes the results of two arms more similar and thus harder to detect a significant difference. Loss to follow-up will also increase the similarity between groups, because of the assumption that none of these patients met the primary endpoint. Other protocol deviations such as non-adherence to the assigned therapy can bias the results towards non-inferiority.[26] Therefore, a non-inferiority trial should always report both the intention-to-treat and per-protocol (or as-treated) analyses, since either analysis has strengths and

**Table 2** Examples of recent non-inferiority trials

| | Device vs. surgery trials | | | | Pharmacologic trials | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Trial, year** | **SYNTAX, 2009** | **PRECOMBAT, 2011** | **PARTNER 1A, 2011** | **EVEREST II, 2011** | **PROTECT AF, 2009** | **RE-LY, 2009** | **RE-LY, 2009** | **ROCKET AF, 2011** | **ARISTOTLE, 2011** |
| New Rx | TAXUS DES | DES | TAVR | Mitraclip | Watchman LAA closure | Dabigatran 150 mg | Dabigatran 110 mg | Rivaroxaban | Apixaban |
| Standard Rx | CABG | CABG | SAVR | MV surgery | Warfarin | Warfarin | | Warfarin | Warfarin |
| Primary endpoint | MACCE | MACCE | All-cause mortality | Freedom from death, MV surgery or MR >2+ | Stroke, cardiovascular death, and systemic embolism | Stroke or systemic embolism | | Stroke or systemic embolism | Stroke or systemic embolism |
| Standard Rx event rate (expected) | 13.2% | 13% | 32% | 90% | 6.15% per 100 patient-years | Not specified | | 2.3% per 100 patient-years | Not specified |
| Standard Rx event rate (observed) | 12.4% | 6.7% | 26.8% | 88% | 4.9% per 100 patient-years | 1.7% per 100 patient-years | | 2.2% per 100 patient-years | 1.6% per 100 patient-years |
| Trial power | 96% | 80% | 85% | 80% | 80% | 84% | | 95% | 90% |
| Alpha | One-sided, 0.05 | One-sided, 0.05 | One-sided, 0.05 | One-sided, 0.05 | One-sided, 0.025 | One-sided, 0.025 | | One-sided, 0.025 | One-sided, 0.025 |
| Sample size | 1800 | 600 | 699 | 279 | 707 | 15000 | | 14000 | 18000 |
| Follow-up duration | 1 year | 1 year | 1 year | 1 year | Mean of 1.5 years | Median 2.0 years | | Median 1.9 years | Median of 1.8 years |
| Standard Rx effect | Not quantified | Not quantified | Not quantified | 90% (84–96%) | 0.36 (0.25–0.53) for stroke and embolism. Not quantified for the endpoint with death included | 0.36 (0.25–0.53) | | 0.36 (0.25–0.53) | 0.36 (0.25–0.53) |
| Non-inferiority margin | ARD = 6.6% | ARD = 7% | ARD = 7.5% | ARD = 31% (PP) | Rate ratio = 2.0 | Relative risk = 1.46 | | Relative risk = 1.46 | Relative risk = 1.44 |
| | RR = 1.51 | RR = 1.54 | RR = 1.23 | | | | | | |
| % preservation of standard Rx effect | . . . | . . . | . . . | 65% of point estimate | . . . | 50% of lower bound of 95% CI of placebo vs. standard | | 50% of lower bound of 95% CI of placebo vs. standard | 50% of lower bound of 95% CI of placebo vs. standard |
| New Rx vs. standard Rx | ARD = 5.5% (2.8–8.3%) | ARD = 2.0% (−1.6–5.6%) | ARD = −2.6% (−9.3–4.1%) | ARD = 15.4% (4.8–26.1%) | Rate ratio = 0.62 (0.35–1.25) | Relative risk = 0.65 (0.52–0.81) | Relative risk = 0.90 (0.74–1.10) | Hazard ratio = 0.79 (0.66–0.96) | Hazard ratio = 0.79 (0.66–0.95) |
| | RR = 1.44 (1.15–1.81) | RR = 1.30 (0.81–2.08) | HR = 0.93 (0.71–1.22) | RR = 2.3 (1.2–4.4) | | | | | |
| Non-inferiority met | No | Yes (ARD margin) No (RR margin) | Yes (ARD margin) Yes (RR margin) | Yes (ARD margin) | Yes (RR margin) | Yes (RR margin) | Yes (RR margin) | Yes (RR margin) | Yes (RR margin) |
| Ancillary advantage | Less invasive, lower stroke | Less invasive, lower stroke | Less invasive | Less invasive, lower bleeding | No lifelong anticoagulation | Lower bleeding, no monitoring | | Lower bleeding, no monitoring | Lower bleeding, no monitoring |

DES, drug-eluting stent; CABG, coronary artery bypass grafting; TAVR, transcatheter aortic valve replacement; MV, mitral valve; LAA, left atrial appendage; MACCE, major adverse cardiac or cerebrovascular events; ARD, absolute risk difference; RR, relative risk; PP, per-protocol; ITT, intention-to-treat; MR, mitral regurgitation; Rx, treatment.
[a]Estimations based on the rates provided in the papers.

limitations. However, the intention-to-treat analysis should be the primary analysis as it preserves the advantages of randomization, while the per-protocol analysis can be used as the supporting sensitivity analysis for non-inferiority assessment.

Patients who cross over or drop out need close examination. If a specific reason for a cross-over or drop-out is found in one treatment group, this shows that the two treatments are not similar by concept, thereby providing evidence of lack of non-inferiority.[26]

## Trial conclusions

Non-inferiority can be concluded when the CI does not exceed $-\Delta$ (the non-inferiority margin). It is, however, often misinterpreted as equivalence. Non-inferiority means that the new treatment is not significantly worse (inferior) than the active control, while equivalence means that the new treatment is not significantly worse (inferior) or better (superior) (*Figure 1*). If non-inferior, the new treatment can be preferred because of an associated ancillary benefit in terms of invasiveness, cost, or convenience.

If the non-inferiority endpoint is not met, the interpretation becomes more difficult. Frequently one concludes that the new treatment is inferior to the active control. It could also mean, however, that the trial result is 'inconclusive'. To conclude which is the case, it depends on the side of the CI being considered (*Figure 1*). An inconclusive result is the case when the mean difference is larger than $-\Delta$ and the lower bound of the CI exceeds $-\Delta$. Inferiority is concluded if the mean difference is smaller than $-\Delta$ and the upper bound of the CI does not exceed the $-\Delta$. From a statistical point of view, a trial can show both non-inferiority and inferiority at the same time (*Figure 1*). This can potentially occur in two ways: (i) if the trial is too large, so that an extremely narrow CI can exclude both 0 and a reasonably conservative margin, or (ii) when the choice of the margin is too generous, providing the opportunity for the CI to fit in between $-\Delta$ and 0. Although rare, it is often the result of a poor trial design and should be avoided. From a clinical standpoint, a treatment can be inferior and non-inferior when non-inferiority is met but the margin might have been chosen too generously. The EVEREST II trial is an example of this, where the MitraClip was non-inferior to surgery but this conclusion was difficult to accept due to unduly wide ARD margins of 31 and 25% for the per-protocol and 'comparison of strategy' analyses, respectively.[8] Even the claim of superior safety of the device was driven by blood transfusions that were more frequent with surgery. Excluding these transfusions, the rate of major adverse events in the MitraClip group was not significantly lower (5 vs. 10% after surgery, $P = 0.23$). Thus, one can reasonably argue that MitraClip is less effective than surgery while not demonstrating a clinically relevant safety advantage. In the EVEREST trial the investigators chose a 65% preservation of the active control (surgery) effect over the placebo. This treatment effect being 90%, the investigators were willing to accept an unreasonably large decline in efficacy. In contrast, the ARISTOTLE trial comparing apixaban with warfarin for atrial fibrillation was designed to maintain at least 50% of the 62% relative reduction in warfarin over the placebo.[27] In general, large standard treatment effects require greater preservation (and correspondingly narrow margins) for non-inferiority assessment.

Even in a non-inferiority trial, a new treatment can show superiority over the active control, a sort of 'bonus' in the trial. This is the case if the lower bound CI exceeds 0 in which there is only a 5% chance (alpha) that the active control is better (*Figure 1*). Sequential testing for superiority is only justified after non-inferiority has been successfully demonstrated. Although somewhat obvious, *post hoc* non-inferiority testing in a negative superiority trial is not appropriate, as the margins are not pre-specified and the trial not adequately powered for non-inferiority.

*Table 1* provides an overview of recent non-inferiority trials. It demonstrates the differences in trial design, conduct, and analysis based on the expected event rate, power, sample size, non-inferiority margin, and preservation of the effect of standard therapy.

# Conclusions

The design and interpretation of non-inferiority trials is more complex than for superiority trials. Therefore, many readers and investigators have difficulties understanding the full concept of these trials. When starting a non-inferiority trial, investigators need to make several assumptions and should be aware of not choosing inaccurate or unreasonably generous active control event rates or non-inferiority margins. For readers, to objectively interpret non-inferiority trial results, one must be conscious of several pitfalls of the methodology. Assay sensitivity and trial inconsistency impede conclusions from non-inferiority trials.

**Conflict of interest:** none declared.

## References

1. Lang J, Cetre JC, Picot N, Lanta M, Briantais P, Vital S, Le Mener V, Lutsch C, Rotivel Y. Immunogenicity and safety in adults of a new chromatographically purified Vero-cell rabies vaccine (CPRV): a randomized, double-blind trial with purified Vero-cell rabies vaccine (PVRV). *Biologicals* 1998;**26**:299–308.
2. Garattini S, Bertele V. Non-inferiority trials are unethical because they disregard patients' interests. *Lancet* 2007;**370**:1875–1877.
3. Kaul S, Diamond GA. Good enough: a primer on the analysis and interpretation of noninferiority trials. *Ann Intern Med* 2006;**145**:62–69.
4. D'Agostino RB Sr, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues—the encounters of academic consultants in statistics. *Stat Med* 2003;**22**:169–186.
5. James Hung HM, Wang SJ, Tsong Y, Lawrence J, O'Neil RT. Some fundamental issues with non-inferiority testing in active controlled trials. *Stat Med* 2003;**22**:213–225.
6. Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ, Group C. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. *JAMA* 2006;**295**:1152–1160.
7. Le Henanff A, Giraudeau B, Baron G, Ravaud P. Quality of reporting of noninferiority and equivalence randomized trials. *JAMA* 2006;**295**:1147–1151.
8. Feldman T, Foster E, Glower DG, Kar S, Rinaldi MJ, Fail PS, Smalling RW, Siegel R, Rose GA, Engeron E, Loghin C, Trento A, Skipper ER, Fudge T, Letsou GV, Massaro JM, Mauri L, EVEREST II Investigators. Percutaneous repair or surgery for mitral regurgitation. *N Engl J Med* 2011;**364**:1395–1406.
9. Park SJ, Kim YH, Park DW, Yun SC, Ahn JM, Song HG, Lee JY, Kim WJ, Kang SJ, Lee SW, Lee CW, Park SW, Chung CH, Lee JW, Lim DS, Rha SW, Lee SG, Gwon HC, Kim HS, Chae IH, Jang Y, Jeong MH, Tahk SJ, Seung KB. Randomized Trial of Stents versus Bypass Surgery for Left Main Coronary Artery Disease. *N Engl J Med* 2011;**364**:1718–1727.
10. Serruys P, Morice M, Kappetein A, Colombo A, Holmes DR, Mack MJ, Ståhle E, Feldman TE, van den Brand MJ, Bass E, van Dyck N, Leadley K, Dawkins KD, Mohr FW, SYNTAX Investigators. Percutaneous coronary intervention versus coronary-artery bypass grafting for severe coronary artery disease. *N Engl J Med* 2009;**360**:961–972.
11. Kaul S, Diamond GA, Weintraub WS. Trials and tribulations of non-inferiority: the ximelagatran experience. *J Am Coll Cardiol* 2005;**46**:1986–1995.

12. Smith CR, Leon MB, Mack MJ, Miller DC, Moses JW, Svensson LG, Tuzcu EM, Webb JG, Fontana GP, Makkar RR, Williams M, Dewey T, Kapadia S, Babaliaros V, Thourani VH, Corso P, Pichard AD, Bavaria JE, Herrmann HC, Akin JJ, Anderson WN, Wang D, Pocock SJ, PARTNER Trial Investigators. Transcatheter versus surgical aortic-valve replacement in high-risk patients. *N Engl J Med* 2011; **364**:2187–2198.
13. Leon MB, Smith CR, Mack M, Miller DC, Moses JW, Svensson LG, Tuzcu EM, Webb JG, Fontana GP, Makkar RR, Brown DL, Block PC, Guyton RA, Pichard AD, Bavaria JE, Herrmann HC, Douglas PS, Petersen JL, Akin JJ, Anderson WN, Wang D, Pocock S, PARTNER Trial Investigators. Transcatheter aortic-valve implantation for aortic stenosis in patients who cannot undergo surgery. *N Engl J Med* 2010;**363**:1597–1607.
14. Connolly SJ, Ezekowitz MD, Yusuf S, Eikelboom J, Oldgren J, Parekh A, Pogue J, Reilly PA, Themeles E, Varrone J, Wang S, Alings M, Xavier D, Zhu J, Diaz R, Lewis BS, Darius H, Diener HC, Joyner CD, Wallentin L, RE-LY Steering Committee and Investigators. Dabigatran versus warfarin in patients with atrial fibrillation. *N Engl J Med* 2009;**361**:1139–1151.
15. Patel MR, Mahaffey KW, Garg J, Pan G, Singer DE, Hacke W, Breithardt G, Halperin JL, Hankey GJ, Piccini JP, Becker RC, Nessel CC, Paolini JF, Berkowitz SD, Fox KA, Califf RM, ROCKET AF Investigators. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. *N Engl J Med* 2011;**365**:883–891.
16. Albers GW, Diener HC, Frison L, Grind M, Nevinson M, Partridge S, Halperin JL, Horrow J, Olsson SB, Petersen P, Vahanian A, SPORTIF Executive Steering Committee for the SPORTIF V Investigators. Ximelagatran vs warfarin for stroke prevention in patients with nonvalvular atrial fibrillation: a randomized trial. *JAMA* 2005;**293**:690–698.
17. Schumi J, Wittes JT. Through the looking glass: understanding non-inferiority. *Trials* 2011;**12**:106.
18. Mercado N, Wijns W, Serruys PW, Sigwart U, Flather MD, Stables RH, O'Neill WW, Rodriguez A, Lemos PA, Hueb WA, Gersh BJ, Booth J, Boersma E. One-year outcomes of coronary artery bypass graft surgery versus percutaneous coronary intervention with multiple stenting for multisystem disease: a meta-analysis of individual patient data from randomized clinical trials. *J Thorac Cardiovasc Surg* 2005;**130**:512–519.
19. Seung KB, Park DW, Kim YH, Lee SW, Lee CW, Hong MK, Park SW, Yun SC, Gwon HC, Jeong MH, Jang Y, Kim HS, Kim PJ, Seong IW, Park HS, Ahn T, Chae IH, Tahk SJ, Chung WS, Park SJ. Stents versus coronary-artery bypass grafting for left main coronary artery disease. *N Engl J Med* 2008;**358**:1781–1792.
20. Correia LC. Stents versus CABG for left main coronary artery disease. *N Engl J Med* 2011;**365**:181; author reply 181–182.
21. Mantovani V, Lepore V, Mira A, Berglin E. Non-inferiority randomized trials, an issue between science and ethics: the case of the SYNTAX study. *Scand Cardiovasc J* 2010;**44**:321–324.
22. Ellenberg SS, Temple R. Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 2: practical issues and specific cases. *Ann Intern Med* 2000;**133**:464–470.
23. Kappetein AP, Feldman TE, Mack MJ, Morice MC, Holmes DR, Stahle E, Dawkins KD, Mohr FW, Serruys PW, Colombo A. Comparison of coronary bypass surgery with drug-eluting stenting for the treatment of left main and/or three-vessel disease: 3-year follow-up of the SYNTAX trial. *Eur Heart J* 2011; **32**:2125–2134.
24. Snapinn SM. Alternatives for discounting in the analysis of noninferiority trials. *J Biopharm Stat* 2004;**14**:263–273.
25. Hasselblad V, Kong D. Statistical methods in for comparison to placebo in active-controlled trials. *Drug Inf J* 2001;**35**:435–449.
26. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. *BMJ* 1996;**313**:36–39.
27. Granger CB, Alexander JH, McMurray JJ, Lopes RD, Hylek EM, Hanna M, Al-Khalidi HR, Ansell J, Atar D, Avezum A, Bahit MC, Diaz R, Easton JD, Ezekowitz JA, Flaker G, Garcia D, Geraldes M, Gersh BJ, Golitsyn S, Goto S, Hermosillo AG, Hohnloser SH, Horowitz J, Mohan P, Jansky P, Lewis BS, Lopez-Sendon JL, Pais P, Parkhomenko A, Verheugt FW, Zhu J, Wallentin L, ARISTOTLE Committees and Investigators. Apixaban versus warfarin in patients with atrial fibrillation. *N Engl J Med* 2011;**365**:981–992.