REVIEW ARTICLE

THE CHANGING FACE OF CLINICAL TRIALS

Jeffrey M. Drazen, M.D., David P. Harrington, Ph.D., John J.V. McMurray, M.D., James H. Ware, Ph.D., and Janet Woodcock, M.D., *Editors*

Challenges in the Design and Interpretation of Noninferiority Trials

Laura Mauri, M.D., and Ralph B. D'Agostino, Sr., Ph.D.

From the Division of Cardiovascular Medicine, Brigham and Women's Hospital (L.M.), Harvard Medical School (L.M.), the Department of Mathematics and Statistics, Boston University (R.B.D.), and Baim Institute for Clinical Research (L.M., R.B.D.) — all in Boston. Address reprint requests to Dr. Mauri at Brigham and Women's Hospital, Dept. of Cardiovascular Medicine, 75 Francis St., Boston, MA 02115, or at lmauri@bwh.harvard.edu.

N Engl J Med 2017;377:1357-67.
DOI: 10.1056/NEJMra1510063

Noninferiority clinical trials have become a major tool for the evaluation of drugs, devices, biologics, and other medical treatments. Treatment with placebo or with a no-treatment control in a study is not ethical when an effective treatment has already been established. Effective medical treatments exist for many medical conditions and are the relevant bar to be surpassed by a new treatment. Although some new treatments offer greater efficacy, others may promise greater safety or convenience, or less expense, while providing similar efficacy. The concept of a good substitute was the original rationale for the design of noninferiority trials (i.e., to evaluate a new treatment for efficacy similar to that of an established treatment). Recently, noninferiority trial methods have also been applied in evaluating whether an effective treatment is safe enough. The number of randomized trials assessing noninferiority increased by a factor of 6 in a decade — in 2005, just under 100 trials were listed in MEDLINE under the general rubric of "noninferiority," whereas in 2015, there were almost 600 such trials. These trials span multiple medical and surgical disciplines and diverse treatment strategies.

In this article, we provide a framework for considering the features, including pitfalls, of noninferiority studies. We use cardiovascular treatment trials as examples, although noninferiority trials can be conducted in many fields. These trials include studies designed for regulatory approval of new therapies and trials designed to compare established treatments. In addition, we consider the application of noninferiority concepts and design to emerging areas of clinical investigation. The term "placebo" is used to denote either a true placebo or a no-treatment control in situations in which a true placebo is not available.

A Framework for Noninferiority Studies

Assessing noninferiority in a trial is more complex than assessing superiority, in both the design and analysis phases. Although it is not statistically possible to prove that two treatments are identical, it is possible to determine that a new treatment is not worse than the control treatment by an acceptably small amount, with a given degree of confidence. This is the premise of a randomized, noninferiority trial. The null hypothesis in a noninferiority study states that the primary end point for the experimental treatment is worse than that for the positive control treatment by a prespecified margin, and rejection of the null hypothesis at a prespecified level of statistical significance is used to support a claim that permits a conclusion of noninferiority.[1-3] Figure 1 outlines the statistical evaluation to be used and the range of possible outcomes for a trial designed to demonstrate noninferiority. If the confidence interval for the study results excludes the prespecified

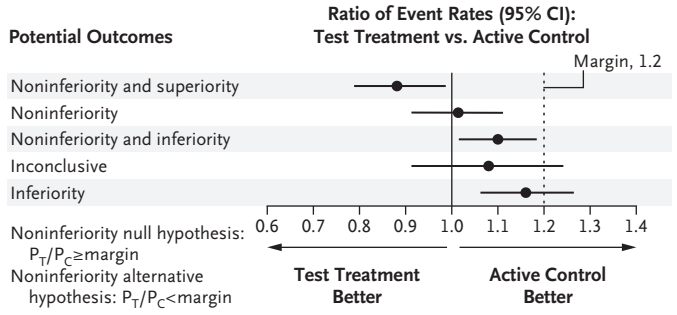


Figure 1. Hypothesis Testing in Noninferiority Trials.

In a noninferiority trial, the null hypothesis states that the primary end point for the new treatment is worse than that of the active control by a prespecified margin, and rejection of the null hypothesis at a prespecified level of statistical significance permits a conclusion of noninferiority. In the example shown, the outcome of interest is a proportion (P) of events that are clinically undesirable (e.g., myocardial infarction). The x axis shows the ratio of proportions (test treatment, or P_T, vs. active control, or P_C). The statistical procedure to test for noninferiority is a one-sided test at an alpha level of significance. Equivalently, one can compute a confidence interval as $100\times(1-2\alpha)$. For this example, if the upper limit of the confidence interval for the relative risk P_T/P_C is less than the margin (shown as a ratio of 1.2), then with 97.5% percent confidence, we can say that the active control is more efficacious than the test treatment by no more than the margin, or that the treatment is noninferior to the active control. There are five potential outcomes of this design (shown with two-sided 95% confidence intervals for simplicity). Noninferiority and superiority of the test treatment are demonstrated if the confidence interval for the ratio between treatments is less than 1, even though a confidence interval that excludes 1 is not necessary to conclude noninferiority. If the confidence interval for the ratio does not exceed the prespecified margin, noninferiority is demonstrated. Paradoxically, both noninferiority and inferiority are demonstrated if the statistical test for noninferiority is met but the confidence interval is above 1. The results are inconclusive if the confidence interval includes the noninferiority margin and does not exclude 1, a finding that suggests an underpowered comparison. Inferiority of the test treatment is shown if the confidence interval excludes 1 and the treatment effect favors the active control.

margin (i.e., the noninferiority margin, also called "delta"), then the conclusion is made that the test treatment is noninferior to the active control. Traditionally, the confidence interval is a 97.5% one-sided or 95% two-sided interval, although sometimes regulatory bodies have agreed to allow 95% one-sided intervals for evaluation of medical devices. For simplicity, the confidence intervals in Figure 1 are all two-sided.

NECESSARY FEATURES OF NONINFERIORITY STUDIES

The following major components of noninferiority study design are listed in Table 1. First, the foundation of the noninferiority trial is one or more prior randomized trials evaluating the superiority of the active control over placebo. Second, an end point is selected, and on the basis of prior experience, the expected performance of the active control is derived.

Third, an acceptable noninferiority margin is defined during the design phase, which preserves a minimum clinically acceptable proportion of the effect of the active treatment as compared with placebo. This margin cannot be greater than the smallest effect size for the active treatment that would be expected in a placebo-controlled trial.[1]

A variety of statistical methods are used to derive the margin. One common approach is to establish a fixed margin[5] based on estimates of the effect of the active comparator in previous studies. The noninferiority study will be successful if the results rule out with a sufficient level of confidence the possibility that the test treatment performs worse than the active control by the specified margin. In the fixed-margin approach, previous studies comparing the active control with placebo are used to derive a single fixed value for the margin. The value recommended in recent guidance from the Food and Drug Administration (FDA)[5] is the lower bound of the 95% confidence interval around the treatment effect of a single placebo-controlled trial or a meta-analysis of such trials, though noninferiority trials are sometimes designed to preserve a specific proportion of the observed treatment effect of the active control.[1] The synthesis method, an alternative to the fixed-margin method, uses the same approach as the fixed method and also accounts for the variability of the treatment effect of active control versus placebo in determining the margin.[5]

Fourth, considerations about the comparator must apply. The study must be designed to adequately distinguish between effective and ineffective therapies, also described as preserving assay sensitivity. More specifically, one would want to be assured that if a placebo had been included, the study design and conduct would have allowed the active control to be shown to be superior to the placebo. This may be difficult to prove within the study, since a placebo group is rarely included, for ethical reasons.

However, this leads to the fifth necessary feature of a noninferiority trial — namely, the design of the new trial preserves the conditions of

Table 1. Features of Noninferiority Studies.

Consideration	Explanation	Challenges
Active control	Select active control on the basis of a previous randomized superiority trial comparing active control with placebo; active control represents current standard of care	Placebo-controlled trials may not have been performed
End-point selection	Is the end point clinically relevant, and are there historical data comparing the active control with placebo for the selected end point?	Composite end points may be difficult to interpret; the relevance of end points may change in the course of follow-up
Choice of noninferiority margin	Is the margin less than the treatment effect of the active control versus placebo? Is there consensus about the margin of reduced effectiveness that is still acceptable in light of potential benefits (e.g., improved safety, lower cost, lower risk of side effects)?	It is important not to accept new therapies that are less effective over time than previous therapies (known as "biocreep"*); historical data are not always available to determine the difference between placebo and control (e.g., in the case of antiinfective agents)
Assay sensitivity	If the active control were compared with placebo, would superiority be evident?	A "positive control" usually cannot be assessed in the study, since placebo is not feasible or ethical
Constancy and metrics	Have the conditions changed between the trial establishing superiority of the active control over placebo and the noninferiority trial? What type of metric (between-group difference in absolute risk or relative risk) is more likely to be constant between studies and therefore a reliable metric for comparison and margin definition?	Characteristics of the study population or concomitant therapies may have changed since the effect of active therapy was established, making a determination of noninferiority unreliable; constancy is not always present for absolute effects; a lower-than-expected event rate may make a risk-difference margin clinically inappropriate if viewed from a relative-risk perspective; a higher-than-expected event rate may result in lower-than-expected power
Execution	Are the assigned treatments administered adequately? Is ascertainment of the end point accurate and complete?	Lack of attention to execution in the control group or misclassification or missing data on the end point may bias the study toward a conclusion of noninferiority
Analysis	If treatment crossover or nonadherence occurs, what is the appropriate analysis (intention-to-treat or per-protocol)?	Treatment crossover may bias an intention-to-treat analysis toward a conclusion of noninferiority, but a per-protocol analysis may also introduce bias, since baseline characteristics are no longer balanced between study groups

* Biocreep was defined in a 1992 "Points to Consider" Food and Drug Administration briefing document.[4]

the trial in which the active control was shown to be effective; this is called the "constancy assumption." An appropriate metric must be used in the noninferiority trial. Because the choice between relative and absolute effects can affect both power and validity, this choice must be carefully considered in the design phase of the study. Figure 1 presents relative risk as the metric for the statistical evaluation. However, there are other ways of evaluating proportions, such as calculation of an odds ratio, hazard ratio (in a time-to-event study), or absolute risk difference. For example, if the proportion of events (an adverse outcome) in the control group is P_C and the proportion of events in the treatment group is P_T, and if the respective values for P_C and P_T are 0.20 and 0.40 in one study and 0.10 and 0.20 in another, the relative risk, P_C/P_T, is 0.5 in both, yet the risk differences are 20 percentage points and 10 percentage points, respectively. In a recent trial,[6] which evaluated the noninferiority of a reduced duration of dual antiplatelet therapy after placement of coronary stents, the difference between absolute and relative differences was pronounced and made it difficult to conclude noninferiority, given a shift from the intended study population to a lower-risk population. The expected rate of the composite primary end point of death, myocardial infarction, stent thrombosis, stroke, or major bleeding was 10%, and the margin of noninferiority for the risk difference was 2 percentage points (equivalent to a 20% relative risk), yet the observed rate of the end point in the control group was only 1.6% because of enrollment of lower-risk participants than anticipated, as well as early termination of the study. Statistically, the noninferiority test excluded the margin of 2 percentage points (upper

limit of the one-sided 95% confidence interval [CI] for the difference between groups, 0.5%; P<0.001), but the noninferiority margin of 2 percentage points represented acceptance of a rate of adverse events that was 3 times as high in the treatment group as in the control group.[6] The investigators were therefore careful to avoid concluding that the experimental treatment was noninferior, despite a significant P value for the statistical test of noninferiority.

The sixth component of noninferiority trials is adequate execution of the trial and ascertainment of outcomes. Incomplete or inaccurate ascertainment of outcomes, as a result of loss to follow-up, treatment crossover or nonadherence, or outcomes that are difficult to measure or subjective, may cause the treatments being compared to falsely appear similar.

Finally, noninferiority designs raise analytic questions that may differ from those in a superiority study. In a superiority study, an intention-to-treat analysis (in which all patients who received the experimental treatment, even if only one dose, are included in the statistical tests for superiority) is used. In a noninferiority study, however, if some patients did not receive the full course of the assigned treatment, an intention-to-treat analysis may produce a bias toward a false positive conclusion of noninferiority by narrowing the difference between the treatments. In some instances, a per-protocol analysis, which excludes patients who did not meet the inclusion criteria or did not receive the randomized, per-protocol assignment, may be preferable in a noninferiority trial. However, a per-protocol analysis may include fewer participants and introduce postrandomization bias. In general, both the intention-to-treat and per-protocol data sets are important. We suggest analyzing both sets and examining the results for consistency. Furthermore, careful consideration and sensitivity analyses may be needed before drawing conclusions about noninferiority.

SPECIAL CHALLENGES WITH NONINFERIORITY DESIGN

A few challenging aspects of noninferiority design deserve mention. Even if there is no placebo group, an implicit superiority comparison between the test treatment and placebo underpins the noninferiority trial. Three-group studies that include a placebo group may allow an explicit comparison, but practical or ethical reasons often preclude randomized assignment to placebo, and instead historical data must be relied on for the placebo comparison. In some cases, historical data for a placebo treatment are not available. In these cases, less effective treatments may stand in for placebo to identify the expected benefit of the active control on which to base the noninferiority margin. In studies of stroke prevention, aspirin has been the comparator for warfarin (the active control), and trials comparing warfarin with aspirin provide estimates of a treatment effect used to set noninferiority margins for novel oral anticoagulants. In the case of coronary stents, bare-metal stents have been used as the reference for the treatment effect of approved drug-eluting stents (the active control) in studies of new drug-eluting stents. Treatment strategies such as percutaneous coronary intervention (PCI) for left main coronary artery disease and transcatheter therapy for valvular heart disease have been compared with surgery, and patients receiving medical therapy have served as a reference group for the treatment effect of surgery. Anti-infective therapies are an example of an area of investigation in which no placebo comparisons are available.[7] Finally, in setting the sample-size goal, the noninferiority margin should not be "back calculated" solely from a feasible sample size. To do so may sufficiently exclude the chosen margin but will not necessarily reflect a conclusion of noninferiority that is clinically meaningful.

Noninferiority cannot be established on the basis of the absence of a significant difference between treatments in a superiority study. A superiority trial may fail to reject the null hypothesis because of lack of power (due to a small sample) and should not be used to support a claim of no difference. As the traditional dictum states, "absence of evidence does not constitute evidence of absence." For example, multiple underpowered trials (studies with <800 participants) showed no significant difference between streptokinase and placebo for the treatment of acute myocardial infarction,[8] yet an adequately powered trial (with >17,000 participants) showed that streptokinase was superior in reducing the outcome of vascular mortality.[9] Although meta-analysis is frequently used to combine data from

underpowered studies, heterogeneity and sources of statistical bias can make the results difficult to interpret[10]; therefore, meta-analysis is a poor substitute for a randomized trial with an adequate sample size.

EXAMPLES OF NONINFERIORITY TRIALS

EVALUATION OF EFFICACY

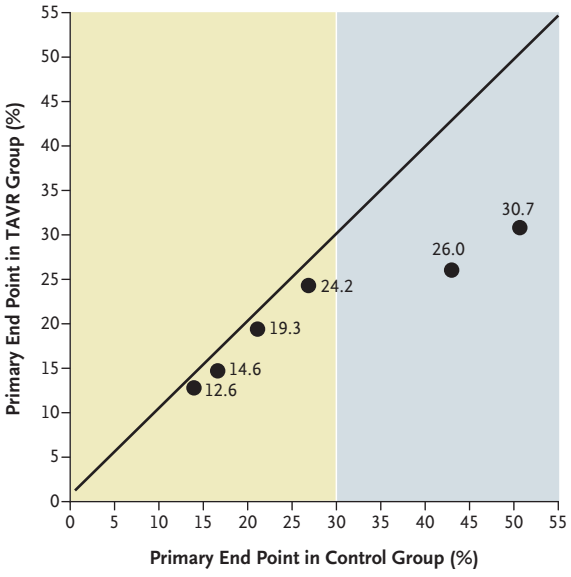
ARISTOTLE, RE-LY, and ROCKET AF Trials

In patients with atrial fibrillation, warfarin reduces the risk of stroke, as compared with placebo or aspirin, but is associated with an increased risk of bleeding and requires frequent blood testing to ensure a therapeutic effect. Several new oral anticoagulant agents are associated with a lower risk of bleeding and offer greater convenience, since they do not require blood testing. These agents have recently been examined and approved by the FDA on the basis of three large noninferiority trials comparing the oral anticoagulants with warfarin for the prevention of stroke or thromboembolism: ARISTOTLE (Apixaban for Reduction in Stroke and Other Thromboembolic Events in Atrial Fibrillation), RE-LY (Randomized Evaluation of Long-Term Anticoagulant Therapy), and ROCKET-AF (Rivaroxaban Once Daily Oral Direct Factor Xa Inhibition Compared with Vitamin K Antagonism for Prevention of Stroke and Embolism Trial in Atrial Fibrillation).[11-13]

Prior randomized trials of warfarin versus aspirin provided the expected rate of stroke or systemic thromboembolism.[14] The noninferiority trials compared new anticoagulants with warfarin in study populations ranging from 14,264 to 18,261 participants randomly assigned to treatment groups, with the relative risk of stroke or thromboembolism as the primary end point and a relative noninferiority margin of less than 1.4. The upper bounds of the one-sided 97.5% confidence interval for the relative risk in each study ranged from 0.95 to 1.11, falling below the prespecified margin and supporting the conclusion of noninferiority in each trial. These studies also showed less frequent intracranial hemorrhage, which, along with greater convenience for patients, has led to the replacement of warfarin with these new anticoagulants as first-line therapy to prevent stroke in many patients with atrial fibrillation.

PARTNER, CoreValve, and SURTAVI Trials

Severe aortic stenosis is associated with heart failure and death if untreated, and surgical aortic-valve replacement (SAVR) is effective in many patients. The availability of valves that can be placed by means of a catheter rather than sternotomy has recently allowed a less invasive approach to treatment. Beginning with the CoreValve and PARTNER 2 (Placement of Aortic Transcatheter Valves 2) studies, which involved patients with severe aortic stenosis who were unlikely to survive surgical repair because of additional medical conditions and advanced age, transcatheter aortic-valve replacement (TAVR) was shown to be superior to balloon aortic valvuloplasty, a palliative procedure, with respect to the reduction in mortality (Fig. 2).[15-18] Studies have subsequently progressed to examine TAVR in patients who are candidates for surgery, as well as in younger patients, when a less invasive procedure with similar efficacy might be preferable. In patients with an intermediate risk of death as a result of surgery (4 to 8% predicted risk), a relative margin of 1.2 was prespecified for the primary composite end point of the relative risk of death or disabling stroke at 2 years in the PARTNER 2A trial. The observed hazard ratio was 0.89, and the upper bound of the 95% confidence interval (1.09) was lower than the margin of 1.2, showing noninferiority. In fact, in a prespecified subgroup of patients in whom femoral access was feasible for TAVR, that procedure was shown to be superior to surgery.[18] Similarly, in the SURTAVI (Surgical Replacement and Transcatheter Aortic Valve Implantation) trial, conducted among patients with a predicted operative mortality of 3 to 15%, placement of a transcatheter valve was shown to be noninferior to surgery, with an absolute risk difference of 7 percentage points (i.e., a margin of 7 percentage points over the expected end-point rate of 14% with surgery).[19] Studies currently under way are examining the noninferiority of TAVR as compared with SAVR in patients at even lower risk for complications (ClinicalTrials.gov numbers, NCT02701283 and NCT02675114). The extended follow-up in these studies will be of particular importance for these healthier cohorts.



Trial	Study and Year	TAVR	Control	Primary End Point	Surgical Risk of Death at 30 Days
		%			
Superiority of TAVR vs. medical therapy	PARTNER[15] 2010	30.7	50.7	Death at 1 yr	Not suitable, ≥50%
	CoreValve[16] 2014	26.0	43.0	Death or stroke at 1 yr	Extreme, ≥50%
Noninferiority of TAVR vs. SAVR	PARTNER[17] 2011	24.2	26.8	Death at 1 yr	High, >10%
	PARTNER 2A[18] 2016	19.3	21.1	Death or disabling stroke at 2 yr	Intermediate, ≥4% and ≤8%
	SURTAVI[19] 2017	12.6	14.0	Death or disabling stroke at 2 yr	Intermediate, 3–15%
	PARTNER 3[20] 2016	14.6	16.6	Death, stroke, or rehospitalization at 1 yr	Low, <4%
	CoreValve[21] 2017	—	—	Death or stroke at 2 yr	Low, <3%

Figure 2. Progression from Superiority Trials to Noninferiority Trials for the Evaluation of Transcatheter Aortic Valves.
A series of trials have been designed to compare transcatheter aortic-valve replacement (TAVR) with the standard of care in patient populations with progressively lower surgical risk. The studies shown in gray are superiority studies that enrolled patients with risk factors prohibiting surgery. In these studies, which compared the outcome after TAVR with the outcome after a palliative procedure, mortality was lower with TAVR. The studies shown in yellow are noninferiority studies involving patients at lower operative risk. These studies show the noninferiority of TAVR as compared with surgical aortic-valve replacement (SAVR). Expected rates are shown for the PARTNER 3 trial,[20] which is still enrolling patients, and are not available for the CoreValve 2017 trial,[21] which is also still enrolling patients. PARTNER denotes Placement of Aortic Transcatheter Valves, and SURTAVI Surgical Replacement and Transcatheter Aortic Valve Implantation.

Trials such as these, which compare very different types of procedures, are susceptible to imbalances in treatment adherence and follow-up because some participants may have a strong preference for one therapy and because blinding is not feasible.[11,12] Although investigators seek to minimize such imbalances with careful informed consent, the problem cannot be prevented altogether. Since incomplete treatment adherence could bias results toward a conclusion of noninferiority, analyses of both the intention-to-treat cohort and the cohort restricted to participants who received the assigned therapy (the as-treated cohort) have been important for these studies. In the PARTNER 2A trial, nonadherence to the randomly assigned treatment differed by a factor of more than 4 between the two studies (7.5% in the SAVR group vs. 1.7% in the TAVR group), but the results in the intention-to-treat and as-treated cohorts were largely similar, with both analyses excluding the prespecified margin of 1.2 for noninferiority (relative risk in the intention-to-treat cohort, 0.92; 95% CI, 0.77 to 1.09; P=0.001 for noninferiority; and relative risk in the as-treated cohort, 0.90; 95% CI, 0.75 to 1.08; P<0.001 for noninferiority). Similarly, in the SURTAVI trial, a modified intention-to-treat analysis, which excluded patients in whom the assigned procedure

was not attempted, was used as the primary analysis, and the results were consistent with the results of the intention-to-treat analysis.[19] Measures to address missing data can also be important in this scenario. Because an analysis excluding patients who did not receive the assigned treatment may introduce imbalances in patient characteristics between the randomized study groups, imputing results to allow a complete intention-to-treat analysis is an additional important method that may be used to avoid bias.[22,23]

EXCEL and NOBLE Trials

Revascularization with coronary-artery bypass grafting (CABG) in patients with left main coronary artery disease was established as superior to medical therapy in randomized trials conducted in the 1970s, with an approximate 7-year increase in median survival with surgery.[24] PCI for treatment of the left main coronary artery has become safer and more common with the use of current coronary stents and procedural techniques. Subgroup analyses in randomized trials comparing coronary stenting with CABG for revascularization of the left main coronary artery showed no significant increase in major adverse cardiac events, a shorter recovery time, and possibly a lower periprocedural risk of stroke with stenting[25,26]; therefore, clinical recommendations have been broadened to accommodate percutaneous treatment.[25-27]

EXCEL (Evaluation of XIENCE versus Coronary Artery Bypass Surgery for Effectiveness of Left Main Revascularization) was a noninferiority trial designed specifically to evaluate PCI versus CABG for left main coronary-artery stenosis.[28] The primary end point was a composite of death, stroke, or myocardial infarction. The study investigators concluded that the results showed the noninferiority of PCI at a median follow-up of 3 years. The Nordic–Baltic–British Left Main Revascularization (NOBLE) trial (NCT01496651) was also a noninferiority trial designed to evaluate PCI versus CABG for left main coronary artery disease, but the composite end point was death, stroke, myocardial infarction, or revascularization, assessed at 5 years. The investigators did not conclude that PCI was noninferior and also concluded that CABG was superior. In EXCEL, at 1 month, patients who underwent PCI had a lower rate of periprocedural myocardial infarction, and therefore a lower rate of the composite primary end point, than patients who underwent CABG. After 3 years, however, the rate of spontaneous myocardial infarction was higher among the patients treated with PCI than among those treated with CABG (Fig. 3). Nonetheless, since the overall rate of the primary end point at 3 years did not differ significantly between the two treatment groups and excluded the prespecified margin, the investigators concluded that the study showed the noninferiority of PCI. The NOBLE trial results, with a 5-year follow-up period, showed a higher primary end-point rate for PCI than for CABG (29% vs. 19%), driven by nonprocedural myocardial infarction and revascularization; the criterion for noninferiority of PCI was not met, and the investigators concluded that the results showed the superiority of CABG.[29] The longer follow-up and more inclusive end point in the NOBLE trial contributed to the difference in the conclusions between this trial and the EXCEL trial. Thus, the components of the composite clinical outcome and the timing of the outcome assessment are important in interpreting the study results and explaining expected treatment results to patients.

EVALUATION OF SAFETY

A noninferiority study design is increasingly being used to evaluate the safety of new therapeutics. A particular challenge in noninferiority design for safety studies is that there are usually no reasonable data to justify the margin for safety. Instead, the study's clinical advisors must decide what level of adverse events is acceptable. That level might vary according to the severity of the events, the absolute risk for the patient population, and the expected benefit of the treatment in question. In the PRECISION (Prospective Randomized Evaluation of Celecoxib Integrated Safety versus Ibuprofen or Naproxen) trial, which evaluated the noninferiority of celecoxib to naproxen for the treatment of arthritis, a relative margin of 1.33 was chosen on the basis of an expected annualized risk of 2% for the primary composite end point of death from cardiovascular causes (including hemorrhage), nonfatal myocardial infarction, or nonfatal stroke.[30] Although this was a three-group trial, the third group did not receive placebo but instead received ibuprofen, as a second noninferiority comparator for celecoxib. During the 10-year study period, the rate of treatment discontinuation was nearly

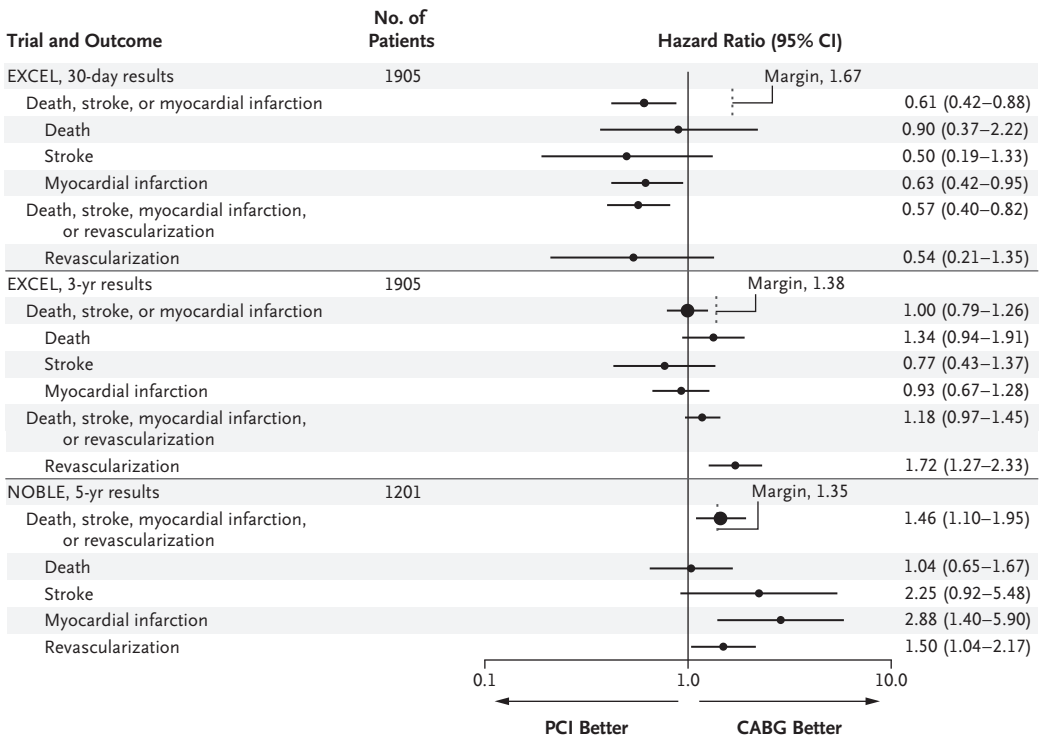


Figure 3. Influence of Timing and End-Point Composition in Noninferiority Trials of Left Main Coronary-Artery Revascularization.

Two trials comparing percutaneous coronary intervention (PCI) with coronary-artery bypass grafting (CABG) have yielded different conclusions. In the EXCEL trial (Evaluation of XIENCE versus Coronary Artery Bypass Surgery for Effectiveness of Left Main Revascularization), the secondary composite end point was death, stroke, or myocardial infarction at 30 days, and the primary composite end point was death, stroke, or myocardial infarction at 3 years. The noninferiority margins for the absolute difference between groups at 30 days and 3 years were 2.0 percentage points and 4.2 percentage points, respectively, which translate roughly to margins of 1.67 and 1.38 for the relative difference between groups. In the NOBLE trial (Nordic–Baltic–British Left Main Revascularization), the composite primary end point was death from any cause, nonprocedural myocardial infarction, any coronary revascularization, or stroke. Although the noninferiority margins were roughly similar in the two trials, the EXCEL results led to the conclusion that PCI was superior to CABG at 30 days and was noninferior at 3 years, and the NOBLE findings, based on longer follow-up and a more inclusive primary end point, led to the conclusion that CABG was superior at 5 years. In the EXCEL trial, the early benefit of PCI was due to avoidance of periprocedural infarction. The late benefit of CABG, on the basis of extended follow-up, was largely due to lower rates of spontaneous myocardial infarction (not shown: 4.3% with PCI vs. 2.7% with CABG; hazard ratio with PCI, 1.60; 95% confidence interval [CI], 0.95 to 2.70; P=0.07 for the superiority of CABG) and revascularization (12.9% vs. 7.6%; hazard ratio, 1.72; 95% CI, 1.27 to 2.33; P<0.001 for the superiority of CABG). Similarly, the NOBLE trial showed that PCI, as compared with CABG, was associated with higher rates of nonprocedural myocardial infarction (7% vs. 2%; hazard ratio, 2.88; 95% CI, 1.40 to 5.90; P=0.004 for the superiority of CABG) and revascularization (16% vs. 10%; hazard ratio, 1.50; 95% CI, 1.04 to 2.17; P=0.03 for the superiority of CABG) at 5 years, leading to the conclusion that CABG was superior (rate of major adverse cardiovascular and cerebrovascular events, 29% with PCI vs. 19% with CABG; hazard ratio, 1.48; 95% CI, 1.11 to 1.96; P=0.007 for the superiority of CABG).

80%, showing that drug trials may also be susceptible to incomplete treatment adherence. Nonetheless, in both the primary intention-to-treat analyses and secondary "on treatment" analyses, celecoxib was noninferior to naproxen and to ibuprofen.

An additional challenge arose when the actual risk of the vascular outcome in the study population was noted to be half the expected risk. Although the use of a relative noninferiority margin would have preserved the validity of a test of noninferiority in this lower-risk population, the

data and safety monitoring board recognized that the study would be underpowered on the basis of an examination of the aggregate event rate (without consideration of the blinded results according to treatment group), and the sample size was therefore augmented from a planned enrollment of approximately 20,000 participants to an eventual enrollment of 24,081 participants. Finally, because a placebo control is not feasible in a study involving patients with chronic pain, the PRECISION trial does not show that there is no increase in cardiovascular risk with any of these medications (i.e., that the medications are noninferior to placebo). Although extending the framework for noninferiority studies of efficacy to evaluate safety can be challenging for a host of reasons, the concept of excluding a prespecified margin remains empirically helpful.

NEW TRIAL METHODS AND THEIR EFFECT ON NONINFERIORITY STUDIES

Simplifying trial conduct (reducing the number of contacts with participants and the number of outcomes assessed) benefits both superiority and noninferiority trials by allowing more reliable ascertainment of a larger sample and may reduce the bias introduced by missing data.[31] However, pragmatic trials that obtain follow-up data from routine clinical care may have imbalances in treatment adherence or imprecise end-point ascertainment, problems that are of particular concern in noninferiority studies. Patient input into trial design may be particularly valuable for noninferiority trials. Given the importance of shared decision making in clinical practice, we believe patients' preferences should be incorporated into both the prespecification of an acceptable margin based on anticipated benefits and risks and the implementation of study results. Finally, noninferiority studies may be used in comparative effectiveness or health services research. Within the framework of value-based health care,[32] evaluating the outcome of treatment in noninferiority designs separately from costs may provide greater reassurance that clinical outcomes remain acceptable or better while efficiencies are provided. Beyond randomized studies, observational data analysis[33-35] and meta-analysis may include testing of noninferiority hypotheses, and prespecification of both the hypothesis and the margin of noninferiority improves the validity of such investigations. Equivalence studies have recently been used in testing biologic agents for similarity to approved agents, with testing for both noninferiority and nonsuperiority as part of the primary analysis.[36-38]

Table 2. Recommendations for the Design, Reporting, and Interpretation of Noninferiority Trials.

CONSORT* recommendations[40]
State hypothesis in terms of noninferiority
Justify choice of noninferiority margins
Describe results with confidence limits for difference or ratio
Food and Drug Administration recommendations[5]
Assess whether active control performed as expected (i.e., determine assay sensitivity)
Be sure noninferiority margin is not larger than the expected difference between active control and placebo
European Medicines Agency recommendations[41]
Make sure the data set for the full analysis, based on the intention-to-treat principle, and the data set for the per-protocol analysis have equal importance, and that their use will lead to similar conclusions for a robust interpretation
Additional recommendations
Compare the noninferiority margin with the expected benefit during design and interpretation
Avoid using composite end points that include discordant components
Perform a sensitivity analysis for missing data (e.g., multiple imputation)[22]

* CONSORT denotes Consolidated Standards of Reporting Trials.

IMPROVING NONINFERIORITY TRIALS

Standards for the design and reporting of superiority trials have been widely disseminated, but adherence to these standards is not universal.[39] Furthermore, unique challenges continue to emerge for noninferiority trials as their uses become both more common and more diverse. The CONSORT (Consolidated Standards of Reporting Trials) group, the FDA, and the European Medicines Agency have promoted specific standards for noninferiority trials (Table 2).[5,40,41] We recommend additional attention to the following items.

First, noninferiority trials should provide an explicit justification of the acceptable margin that is based on a measured or anticipated benefit of the experimental treatment. Some approaches

that may be considered include incorporating decision analysis (population or policy perspective) or patient questionnaires designed to consider how the noninferiority margin and expected benefit are balanced, rather than relying solely on the empiricism of the study investigators or the current expectations of the physician community. Second, we recommend caution when considering composite end points that may include components with discordant benefits and risks. Finally, although avoidance of missing data is an important goal, sensitivity analysis regarding missing data (e.g., with the use of multiple imputation) should be strongly considered in the planning and analysis of noninferiority trials.[22]

CONCLUSIONS

Noninferiority designs are being applied more commonly and more broadly in clinical investigation. Although new challenges emerge with the use of noninferiority studies in diverse settings, the underlying principles for maintaining the validity of such studies should be observed. When appropriately designed and executed, noninferiority trials offer the ability to identify innovative treatment alternatives with clinical value.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

We thank Joanna Suomi, M.S., at the Baim Institute for Clinical Research for her assistance with the preparation of an earlier version of the manuscript.

REFERENCES

1. 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;22:169-86.
2. Blackwelder WC. "Proving the null hypothesis" in clinical trials. Control Clin Trials 1982;3:345-53.
3. Fleming TR, Odem-Davis K, Rothmann MD, Li Shen Y. Some essential considerations in the design and conduct of noninferiority trials. Clin Trials 2011;8:432-9.
4. Center for Drug Evaluation and Research. Points to consider: clinical development and labeling of anti-infective drug products — guidance for industry. Rockville, MD: Food and Drug Administration, 1992.
5. Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research. Non-inferiority clinical trials to establish effectiveness — guidance for industry. Silver Spring, MD: Food and Drug Administration, November 2016.
6. Schulz-Schüpke S, Byrne RA, Ten Berg JM, et al. ISAR-SAFE: a randomized, double-blind, placebo-controlled trial of 6 vs. 12 months of clopidogrel therapy after drug-eluting stenting. Eur Heart J 2015;36:1252-63.
7. Center for Drug Evaluation and Research. Antibacterial drug products: use of noninferiority trials to support approval — guidance for industry. Silver Spring, MD: Food and Drug Administration, November 2010.
8. May GS, Furberg CD, Eberlein KA, Geraci BJ. Secondary prevention after myocardial infarction: a review of short-term acute phase trials. Prog Cardiovasc Dis 1983;25:335-59.
9. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. Lancet 1988;2:349-60.
10. LeLorier J, Grégoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. N Engl J Med 1997;337:536-42.
11. Granger CB, Alexander JH, McMurray JJV, et al. Apixaban versus warfarin in patients with atrial fibrillation. N Engl J Med 2011;365:981-92.
12. Connolly SJ, Ezekowitz MD, Yusuf S, et al. Dabigatran versus warfarin in patients with atrial fibrillation. N Engl J Med 2009;361:1139-51.
13. Patel MR, Mahaffey KW, Garg J, et al. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. N Engl J Med 2011;365:883-91.
14. Hart RG, Pearce LA, Aguilar MI. Meta-analysis: antithrombotic therapy to prevent stroke in patients who have nonvalvular atrial fibrillation. Ann Intern Med 2007;146:857-67.
15. Leon MB, Smith CR, Mack M, et al. Transcatheter aortic-valve implantation for aortic stenosis in patients who cannot undergo surgery. N Engl J Med 2010;363:1597-607.
16. Popma JJ, Adams DH, Reardon MJ, et al. Transcatheter aortic valve replacement using a self-expanding bioprosthesis in patients with severe aortic stenosis at extreme risk for surgery. J Am Coll Cardiol 2014;63:1972-81.
17. Smith CR, Leon MB, Mack MJ, et al. Transcatheter versus surgical aortic-valve replacement in high-risk patients. N Engl J Med 2011;364:2187-98.
18. Leon MB, Smith CR, Mack MJ, et al. Transcatheter or surgical aortic-valve replacement in intermediate-risk patients. N Engl J Med 2016;374:1609-20.
19. Reardon MJ, Van Mieghem NM, Popma JJ, et al. Surgical or transcatheter aortic-valve replacement in intermediate-risk patients. N Engl J Med 2017;376:1321-31.
20. ClinicalTrials.gov. PARTNER 3: the safety and effectiveness of the SAPIEN 3 transcatheter heart valve in low risk patients with aortic stenosis. 2017 (https://clinicaltrials.gov/show/NCT02675114).
21. ClinicalTrials.gov. Medtronic transcatheter aortic valve replacement in low risk patients. 2017 (https://clinicaltrials.gov/ct2/show/NCT02701283).
22. Little RJ, D'Agostino R, Cohen ML, et al. The prevention and treatment of missing data in clinical trials. N Engl J Med 2012;367:1355-60.
23. Freitag G, Lange S, Munk A. Non-parametric assessment of non-inferiority with censored data. Stat Med 2006;25:1201-17.
24. Caracciolo EA, Davis KB, Sopko G, et al. Comparison of surgical and medical group survival in patients with left main equivalent coronary artery disease: long-term CASS experience. Circulation 1995;91:2335-44.
25. Serruys PW, Morice MC, Kappetein AP, et al. Percutaneous coronary intervention versus coronary-artery bypass grafting for severe coronary artery disease. N Engl J Med 2009;360:961-72.
26. Levine GN, Bates ER, Blankenship JC, et al. 2011 ACCF/AHA/SCAI guideline for percutaneous coronary intervention: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines and the Society for Cardiovascular Angiography and Interventions. J Am Coll Cardiol 2011;58(24):e44-e122.
27. Windecker S, Kolh P, Alfonso F, et al. 2014 ESC/EACTS guidelines on myocardial revascularization: the Task Force on Myocardial Revascularization of the European Society of Cardiology (ESC) and the European Association for Cardio-Thoracic Surgery (EACTS) developed with the special contribution of the European Association of Percutaneous Cardiovascular Interventions (EAPCI). Eur Heart J 2014;35:2541-619.

28. Stone GW, Sabik JF, Serruys PW, et al. Everolimus-eluting stents or bypass surgery for left main coronary artery disease. N Engl J Med 2016;375:2223-35.
29. Mäkikallio T, Holm NR, Lindsay M, et al. Percutaneous coronary angioplasty versus coronary artery bypass grafting in treatment of unprotected left main stenosis (NOBLE): a prospective, randomised, open-label, non-inferiority trial. Lancet 2016;388:2743-52.
30. Nissen SE, Yeomans ND, Solomon DH, et al. Cardiovascular safety of celecoxib, naproxen, or ibuprofen for arthritis. N Engl J Med 2016;375:2519-29.
31. Ford I, Norrie J. Pragmatic trials. N Engl J Med 2016;375:454-63.
32. Porter ME. A strategy for health care reform — toward a value-based system. N Engl J Med 2009;361:109-12.
33. Larmore C, Effron MB, Molife C, et al. "Real-world" comparison of prasugrel with ticagrelor in patients with acute coronary syndrome treated with percutaneous coronary intervention in the United States. Catheter Cardiovasc Interv 2016;88:535-44.
34. Kereiakes DJ, Yeh RW, Massaro JM, et al. Stent thrombosis in drug-eluting or bare-metal stents in patients receiving dual antiplatelet therapy. JACC Cardiovasc Interv 2015;8:1552-62.
35. Thourani VH, Kodali S, Makkar RR, et al. Transcatheter aortic valve replacement versus surgical valve replacement in intermediate-risk patients: a propensity score analysis. Lancet 2016;387:2218-25.
36. Barker KB, Menon SM, D'Agostino RB Sr, Xu S, Jin B, eds. Biosimilar clinical development: scientific considerations and new methodologies. Boca Raton, FL: Taylor & Francis Group, 2017.
37. Cohen SB, Genovese MC, Choy EH, et al. Randomized, double-blind, phase 3 study of efficacy and safety of ABP 501 compared with adalimumab in subjects with moderate to severe rheumatoid arthritis. abstract. Arthritis Rheumatol 2015;67: Suppl 10 (http://acrabstracts.org/abstract/randomized-double-blind-phase-3-study-of-efficacy-and-safety-of-abp-501-compared-with-adalimumab-in-subjects-with-moderate-to-severe-rheumatoid-arthritis/).
38. FDA approves Amjevita, a biosimilar to Humira. News release of the Food and Drug Administration, Silver Spring, MD, September 23, 2016 (https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm522243.htm).
39. Rehal S, Morris TP, Fielding K, Carpenter JR, Phillips PP. Non-inferiority trials: are they inferior? A systematic review of reporting in major medical journals. BMJ Open 2016;6(10):e012594.
40. Piaggio G, Elbourne DR, Pocock SJ, Evans SJ, Altman DG. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. JAMA 2012;308:2594-604.
41. Points to consider on switching between superiority and non-inferiority. London: European Medicines Agency, July 2000.