Good Enough: A Primer on the Analysis and Interpretation of Noninferiority Trials

Sanjay Kaul, MD, and George A. Diamond, MD

Active-control noninferiority trials are being performed with increasing frequency when standard placebo-controlled trials are considered unethical. Three attributes are optimally required to establish noninferiority: 1) The treatment under consideration exhibits therapeutic noninferiority to the active control; 2) the treatment would exhibit therapeutic efficacy in a placebo-controlled trial if such a trial were to be performed; and 3) the treatment offers ancillary advantages in safety, tolerability, cost, or convenience. Trials designed to show noninferiority require an appropriate reference population, a proven active control and dose, a high level of adherence to treatment, and adequate statistical power. However, the formal analysis of such trials is founded on several assumptions that cannot be validated explicitly. These assumptions are evaluated in the context of 8 recently published noninferiority trials. The analyses in this paper confirm the establishment of noninferiority in only 4 of the 8 trials. The authors conclude that if noninferiority trials are to be applied to clinical and regulatory decisions about the marketing and use of new treatments, these assumptions must be made explicit and their influence on the resultant conclusions assessed rigorously.

Ann Intern Med. 2006;145:62-69. **www.annals.org**
For author affiliations, see end of text.

> A thing moderately good is not so good as it ought to be.
>
> —*Thomas Paine* (1)

The archetypical clinical trial is designed to show that a new treatment is superior to an inactive placebo. In contrast, an active-control noninferiority trial is designed to show that a new treatment is not inferior to standard treatment by a predefined clinically acceptable amount (hence, "good enough"). If noninferiority is established, the utility of the new treatment can be based on ancillary advantages in safety, convenience, or cost (2–10).

Active-control noninferiority trials are being performed with increasing frequency, especially in cardiovascular and oncologic applications when placebo-controlled trials are considered unethical. These trials pose a particular challenge to clinicians because their formal analysis is founded on several assumptions that cannot be validated explicitly (5–8). We enumerate the key assumptions underlying the typical noninferiority trial and quantify the influence of these assumptions on the conclusions derived from the trial. Our explicit goal is to provide the practicing clinician with a minimally technical primer on the interpretation of noninferiority trials. The more technical details underlying the analysis and interpretation of such trials are described in a number of recent reviews (4–15).

Foundations of Noninferiority Analysis

Overview

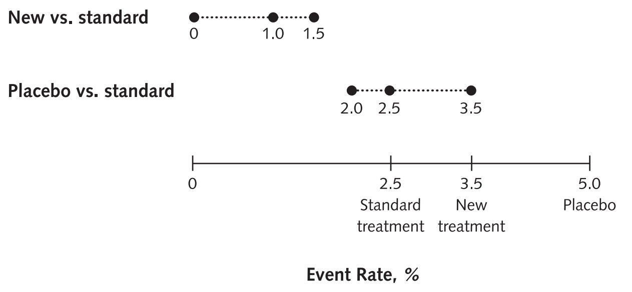
The basic concept of assessing noninferiority is illustrated, with the help of an example, in **Figure 1**. The disease at issue is known to be associated with a serious irreversible clinical outcome, death, and the available standard treatment lowers the mortality rate from 5% (with placebo) to 2.5% (with treatment), an absolute benefit of 2.5 percentage points. A promising new treatment is introduced that has a similar mechanism and predicted benefit but more convenience and fewer side effects. A trial is designed that compares the standard treatment with the new treatment, as a placebo arm is deemed inferior and unethical. What should be the criteria for success of the new treatment?

To be fairly certain that the new treatment is better than placebo, we want to be convinced that its mortality rate is not more than 2.5 percentage points greater than the standard. One such criterion could be that the 95% CI around the trial estimate should not include a mortality rate increase of 2.5 percentage points. This would be fulfilled by observing a mortality rate increase of 1 percentage point (CI, 0 to 2.0 percentage points). The upper limit statistically excludes the mortality rate difference of 2.5 percentage points in persons taking placebo, but there are 2 problems with such a criterion as the basis for a conclusion of noninferiority. First, although the absolute benefit of the new treatment might show that it is likely to be better than placebo, it might be lower than the standard (for example, 2%) and thus still not preferable. Second, in this example, we assume that we know the standard treatment benefit with certainty. In fact, the treatment benefit of 2.5 percentage points would always have a range of uncertainty, for example, a CI of 2 to 3.5 percentage points. For a new treatment to be better than placebo, it would have to be shown that its mortality rate was less than 2.0% (the smallest expected standard treatment effect), not 2.5% (the point estimate of standard treatment effect) and thus no more than 2.0 percentage points (or less) higher than the standard. In summary, the degree of tolerable inferiority, that is, the noninferiority margin, must take into account the uncertainty in the estimated difference over placebo,

See also:

Web-Only
Conversion of figures and tables into slides

Figure 1. **Basic concepts of noninferiority assessment.**



The treatment effects are shown as 95% CIs around the estimates (*dotted line*). Criterion for success is defined as the 95% CI around the estimate of the new versus standard treatment, excluding the smallest expected effect (lower bound of the CI) of the standard treatment over placebo.

and it must be outweighed by the superiority of the new treatment in other respects.

A Hypothetical Trial

Designing a trial based on this conceptual overview, a hypothetical group of investigators reviewed 5 previous trials in which the pooled event rate was 2.3% for standard treatment versus 5% for placebo. The absolute difference in event rate averaged 2.7 percentage points (CI, 2.0 to 3.4 percentage points). The investigators thereby define the lower bound of the CI as their operative noninferiority margin. They then determine that a sample of 1530 patients is required for each treatment group to detect this difference of 2 percentage points with a 1-sided type I error of 2.5% and a type II error of 10%.

On conducting the trial, 46 events (2.9%) are observed among 1600 patients assigned to the new treatment versus 35 events (2.2%) among 1600 patients assigned to the standard treatment, an absolute difference of 0.7 percentage point (CI, −0.4 to 1.8 percentage points). Because the upper bound of 1.8 percentage points lies below the prespecified noninferiority margin of 2.0 percentage points, the investigators formally conclude that the new treatment is noninferior to the standard treatment.

Ideally, a judgment of noninferiority in a trial such as this is founded on 3 prerequisites: 1) The new treatment exhibits therapeutic noninferiority to the standard treatment; 2) the new treatment would exhibit therapeutic efficacy in a placebo-controlled trial, if such a trial were performed; and 3) the new treatment offers ancillary benefits with respect to safety, tolerability, convenience, or cost (**Table 1**).

Determination of Therapeutic Noninferiority

Estimation of the Noninferiority Margin

The critical step in determining therapeutic noninferiority is the selection of the marginal difference (*d*). Statistical reasoning and clinical judgment are commonly used to choose this margin (10–12).

From a statistical perspective, the margin is best determined from a random-effects meta-analysis of historical placebo-controlled studies of the standard treatment (active control). Two key assumptions underlie this determination: 1) the ability to discriminate between effective and ineffective therapies (assay sensitivity or discriminative power) and 2) the applicability of the meta-analysis in the context of the current trial (constancy or representativeness). However, because the noninferiority trial does not have a placebo control, these assumptions are unverifiable. For this reason, the historical trials being examined should exhibit reliable and consistent superiority of the active control over placebo, and the reference population and the experimental protocol in the current active-control trial should be identical to those used in the historical trials (5–10). However, unavoidable inconsistencies in patient characteristics, concomitant medications, intensity of treatment, and temporal improvements in health care can invalidate these key assumptions (5–8, 16), thereby rendering past experience with active control of uncertain relevance to the current study.

Because of this uncertainty, the noninferiority margin is typically defined in terms of some fraction (*f*) of the standard treatment effect to be preserved (9, 11). The choice of *f* is a matter of clinical judgment governed by the maximum loss in efficacy (the magnitude of inferiority) that one is willing to accept in return for potential nonefficacy advantages of the new therapy. Several factors contribute to this judgment, including the seriousness of the clinical outcome (higher values for death or irreversible morbidity), the magnitude of standard treatment effect (smaller values for large effects), and the overall benefit–

Table 1. **Prerequisites of Noninferiority Assessment**

Was therapeutic noninferiority between new and standard treatment (active control) established?
Was the noninferiority margin defined a priori on the basis of statistical reasoning and clinical judgment?
Was the noninferiority assessment adequately powered to minimize statistical uncertainty?
Was the active control effect consistent with that in historical trials?
Were the noninferiority trial design and conduct commensurate with those in historical trials?
Was noninferiority adjudged stable with respect to alternative analytical criteria, including tighter noninferiority margins, relative vs. absolute risk, 1-sided vs. 2-sided confidence intervals, and intention-to-treat vs. per-protocol analysis?

Was therapeutic efficacy of new treatment established?
Was an optimal active control used in the current trial?
Was the new treatment shown to be effective in comparison with putative placebo?
Was a prespecified fraction of active control effect preserved by the new treatment?

Was the new treatment beneficial over standard treatment in nonefficacy terms, such as safety, tolerability, cost, or convenience?

Table 2. **Recent Active-Control Noninferiority Cardiovascular Clinical Trials***

Trial (Reference)	New Treatment	Standard Treatment (Active Control)	Primary Outcome (Trial Duration)	Noninferiority Margin	Criterion for Selection of Margin	New vs. Standard Treatment (95% CI)	Noninferiority Conclusion	Ancillary Benefits of New Treatment
TARGET (17)	Tirofiban	Abciximab	D, MI, uTVR (30 d)	RR, 1.47	0.5*f* of the lower bound of the 95% CI for active control	OR, 1.26 (1.06 to 1.51)†	No	Acceptable safety, lower cost
COBALT (18)	Double-bolus alteplase	Accelerated alteplase	D (30 d)	ARD, 0.4 percentage point	Lower bound of the 95% CI for active control	0.44% (−0.59 to 1.49 percentage points)† 0.44% (−0.79 to 1.68 percentage points)‡	No No	More convenient, unacceptable safety (increased occurrence of hemorrhagic stroke)
GUSTO III (19)	Double-bolus reteplase	Accelerated alteplase	D (30 d)	ARD, 1.0 percentage point	Point estimate of active control	0.23% (−0.51 to 0.98 percentage points)† 0.23% (−0.66 to 1.10 percentage points)‡	Yes No	More convenient, acceptable safety
REPLACE-2 (20)	Bivalirudin	Heparin + abciximab or eptifibatide	D, MI, uTVR, bleeding (30 d)	OR, 1.11	0.5*f* of the lower bound of the 95% CI for active control	OR, 0.92 (0.77 to 1.09)‡	Yes	Superior safety, decreased cost, shorter duration
REPLACE-2§ (20)	Bivalirudin	Heparin + abciximab or eptifibatide	D, MI, uTVR (30 d)			OR, 1.09 (0.92 to 1.32)‡	Yes‖	
VALIANT (21)	Valsartan	Captopril	D (2.1y)	RR, 1.13	0.55*f* of the lower bound for the 95% CI for active control	RR, 1.02 (0.94 to 1.11)‡	Yes	Superior tolerability
A-to-Z (22)	Enoxaparin + tirofiban	Heparin + tirofiban	D, MI, RI (7 d)	OR, 1.14	Lower bound of the 95% CI for active control	OR, 0.88 (0.71 to 1.09)‡	Yes	More convenient, acceptable safety
SYNERGY (23)	Enoxaparin + abciximab, eptifibatide, or tirofiban	Heparin + abciximab, eptifibatide, or tirofiban	D, MI (30 d)	RR, 1.10	Expert consensus (trial committee)	RR, 0.96 (0.87 to 1.06)‡	Yes	More convenient, unacceptable safety (increased occurrence of major bleeding)
SPORTIF V (24)	Ximelagatran	Warfarin	Stroke, systemic embolism (20 mo)	ARD, 2.0 percentage points	Expert consensus (trial committee)	0.45% (−0.13 to 1.03 percentage points)‡	Yes	More convenient, unacceptable safety (increased occurrence of hepatotoxicity)

* A-to-Z = Aggrastat-to-Zocor; ARD = absolute risk difference; COBALT = Continuous Infusion versus Double-Bolus Administration of Alteplase; D = death; *f* = fraction of the standard treatment effect to be preserved; GUSTO III = Global Utilization of Streptokinase and t-PA for Occluded Coronary Arteries III; MI = nonfatal myocardial infarction; OR = odds ratio; REPLACE-2 = Randomized Evaluation in PCI linking Angiomax to Reduced Clinical Events-2; RI = refractory ischemia; RR = relative risk; SPORTIF V = Stroke Prevention Using Oral Thrombin Inhibitor in Atrial Fibrillation V; SYNERGY = Superior Yield of the New Strategy of Enoxaparin, Revascularization, and Glycoprotein IIb/IIIa Inhibitors; TARGET = Do Tirofiban and ReoPro Give Similar Efficacy Trial; uTVR = urgent target vessel revascularization; VALIANT = Valsartan in Acute Myocardial Infarction Trial.

† One-sided 95% CI was used for the primary analysis in TARGET, COBALT, and GUSTO III.

‡ Two-sided 95% CI. Noninferiority was inferred in GUSTO III by using 1-sided, but not 2-sided, 95% CIs (19, 27).

§ Analysis based on efficacy end point without bleeding.

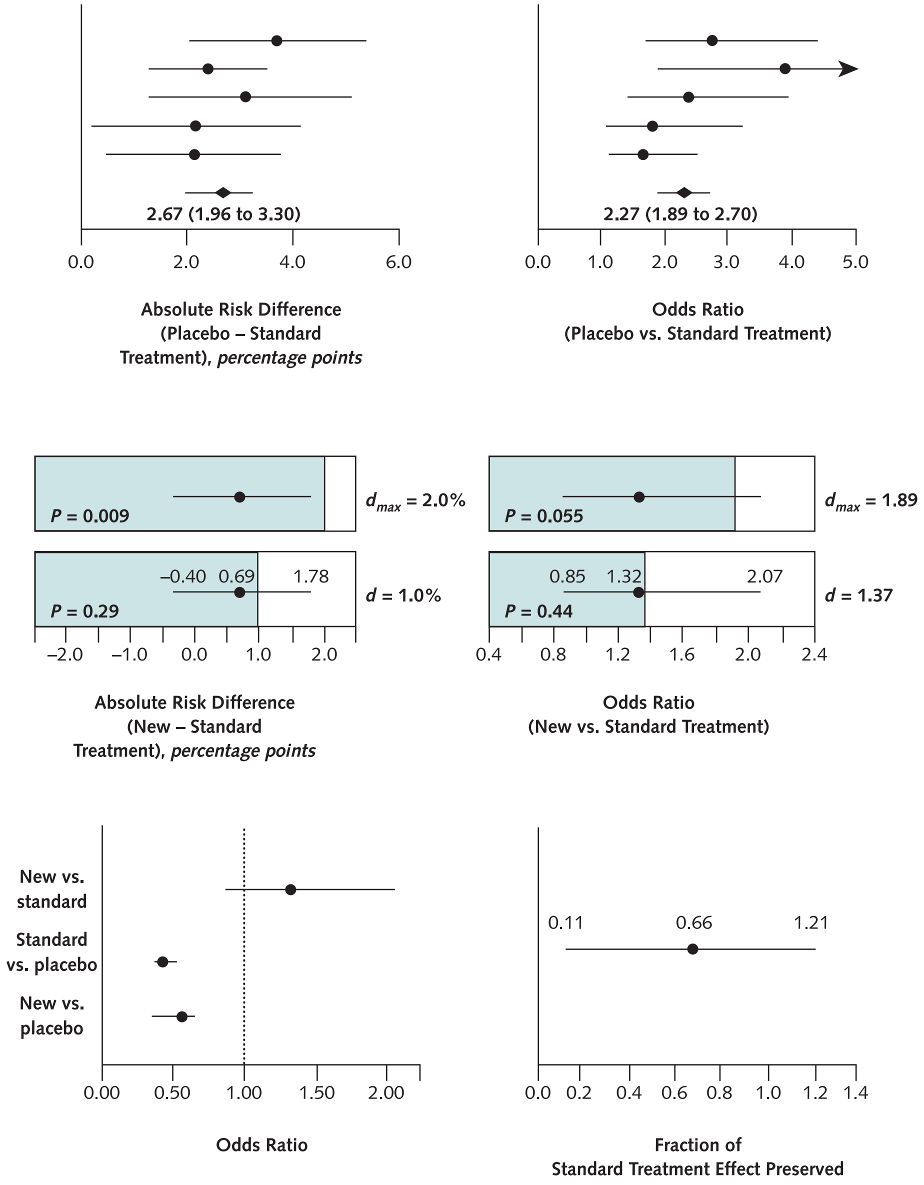
‖ Noninferiority conclusion for efficacy end point in REPLACE-2 was based on 51% preservation of active control (20).

cost and benefit–risk assessment. In the context of oncologic and thrombolytic trials, when mortality is evaluated, the U.S. Food and Drug Administration has suggested an *f* value of 0.5 (9, 11, 16). Typically, the margin is understood to be smaller than the minimally clinically important difference used in conventional superiority trials. Odds or risk ratios of 1.15 to 1.20 (relative differences of 15 to 20 percentage points) have been commonly used as clinically acceptable margins in contemporary noninferiority cardiovascular trials (**Table 2**). Relative scales, such as the odds ratio (OR), for the operative margin are usually preferred over absolute scales, such as the arithmetic risk difference, to "fix" the margin in case of unanticipated dissimilarities in observed and expected event rates (4, 10). The choice of margin has a critical impact on sample size (narrower margins resulting in larger numbers) and on statistical uncertainty (inflation of type I error ["false-positive" or erroneous acceptance of an inferior treatment] with wider margins and of type II error ["false-negative" or erroneous rejection of a truly noninferior treatment] with narrower margins).

In summary, several factors contribute to the selection of the margin. Ultimately, the selection should be justified on statistical, clinical, and regulatory grounds and should be described explicitly in the published report.

In the top section of **Figure 2**, we illustrate the estimation of the noninferiority margin in our hypothetical trial using the criteria described above. The standard treatment effect is derived from a random-effects meta-analysis of the 5 historical placebo-controlled trials. The maximum noninferiority margin ($d_{\max}$) is defined as the lower bound of a 95% CI for the magnitude of that effect—an absolute risk difference of 2.0 percentage points (*left*) or an OR of 1.89 (*right*). Each maximum margin is then weighted by a factor (f) representing the fractional preservation of the standard treatment effect of the new treatment that is considered acceptable ($d = d_{max} \times [1 - f]$ for the absolute risk difference and $d = {d_{max}}^{1-f}$ for the relative OR). Thus, if $f = 0.5$, the operative weighted margin (d) in our hypothetical trial is an absolute difference of 1.0 percentage point (2.0 percentage points × 0.5) and an OR of 1.37 ($1.89^{0.5}$). According to these equations, the greater the value of f, the smaller the margin and the greater the difficulty of judging noninferiority.

Figure 2. **Key steps in noninferiority assessment.**



Top. Estimation of noninferiority margin is illustrated. The standard treatment (active control) effect is derived from analysis of 5 historical trials and is expressed as the absolute risk difference (*left*) or the odds ratio for placebo relative to standard treatment (*right*). Summary effects are estimated by using the DerSimonian–Laird method for random-effects meta-analysis. **Middle.** The CI comparison approach is illustrated, in which noninferiority is established when the upper bound of the 1-sided 97.5% CI (corresponding to a 2-sided 95% CI) lies within the noninferiority zone (d_{max} or d), represented by the shaded area. The analysis is shown for absolute risk difference (*left*) and odds ratio (*right*). Results of hypothesis testing are shown as P values, with a 1-sided P value less than or equal to 0.025 (corresponding to a 1-sided 97.5% CI) as the criterion for noninferiority (13). **Bottom.** The putative placebo approach is shown in the left panel. The effect of the new treatment versus placebo, odds ratio, is derived from the effect of the new versus the standard treatment observed in the current trial and the effect of the standard treatment versus placebo in the historical trials. Superiority over putative placebo is established if the derived odds ratio for the new treatment versus placebo is less than 1.0. The fraction of standard treatment effect preserved by the new treatment using the Hasselblad and Kong method is shown in the right panel (14). Efficacy is established if the lower limit of the CI exceeds a target fractional threshold.

Table 3. **Reanalysis of Recent Active-Control Noninferiority Cardiovascular Clinical Trials***

Trial	Observed OR (New vs. Standard Treatment) (95% CI)	Historical OR (Placebo vs. Standard Treatment) (95% CI)†	Margin OR‡		Noninferiority Conclusion§			Derived OR (New Treatment vs. Placebo) (95% CI)	Efficacy Conclusion: OR (New Treatment vs. Placebo) <1	Fractional Preservation (95% CI)	Fractional Preservation Conclusion, %§	
			d_{max}	d	d_{max}	d	Published Report				Current Analysis	Published Report
TARGET	1.28 (1.02 to 1.61)	2.27 (1.54 to 3.33)	1.54	1.24	No	No	No	0.36 (0.56 to 0.88)	Yes	0.70 (0.39 to 1.01)	39	NA
COBALT	1.06 (0.89 to 1.27)	1.18 (1.06 to 1.28)	1.06	1.03	No	No	No	0.90 (0.74 to 1.10)	No	0.62 (−0.47 to 1.71)	<0‖	NA
GUSTO III	1.03 (0.91 to 1.18)	1.18 (1.06 to 2.28)	1.06	1.03	No	No	Yes	0.88 (0.75 to 1.03)	No	0.79 (−0.02 to 1.60)	<0‖	NA
REPLACE-2	0.92 (0.77 to 1.09)	1.50 (1.23 to 1.82)	1.23	1.11	Yes	Yes	Yes	0.62 (0.48 to 0.80)	Yes	1.21 (0.76 to 1.65)	76	75
REPLACE-2¶	1.09 (0.92 to 1.32)	1.79 (1.42 to 2.27)	1.42	1.19	Yes	No	Yes	0.61 (0.45 to 0.83)	Yes	0.85 (0.51 to 1.20)	51	51
VALIANT	1.03 (0.93 to 1.13)	1.35 (1.20 to 1.52)	1.20	1.10	Yes	No	Yes	0.76 (0.65 to 0.88)	Yes	0.91 (0.58 to 1.24)	58	60
A-to-Z	0.88 (0.71 to 1.09)	1.47 (1.11 to 1.92)	1.11	1.06	Yes	No	Yes	0.60 (0.42 to 0.85)	Yes	1.34 (0.72 to 1.95)	72	NA
SYNERGY	0.96 (0.85 to 1.07)	1.10 (1.02 to 1.18)	1.02	1.01	No	No	Yes	0.87 (0.76 to 0.99)	Yes	1.47 (0.23 to 2.72)	23	NA
SPORTIF V	1.39 (0.91 to 2.12)	2.78 (2.08 to 3.70)	2.08	1.44	No	No	Yes	0.50 (0.30 to 0.83)	Yes	0.68 (0.24 to 1.10)	24	NA

* See text for criteria for establishment of noninferiority, efficacy, and fractional preservation. A-to-Z = Aggrastat-to-Zocor; COBALT = Continuous Infusion versus Double-Bolus Administration of Alteplase; d_{max} = noninferiority margin based on lower limit of the 95% CI for the historical OR; d = noninferiority margin based on 50% of d_{max}; GUSTO III = Global Utilization of Streptokinase and t-PA for Occluded Coronary Arteries III; NA = not available; OR = odds ratio; REPLACE-2: Randomized Evaluation in PCI Linking Angiomax to Reduced Clinical Events-2; SPORTIF V = Stroke Prevention Using Oral Thrombin Inhibitor in Atrial Fibrillation V; SYNERGY = Superior Yield of the New Strategy of Enoxaparin, Revascularization, and Glycoprotein IIb/IIIa Inhibitors; TARGET = Do Tirofiban and ReoPro Give Similar Efficacy Trial; VALIANT = Valsartan in Acute Myocardial Infarction Trial.

† Historical OR for REPLACE-2, VALIANT, SYNERGY, and SPORTIF V are based on random-effects meta-analysis of placebo-controlled trials; TARGET is based on Evaluation of Platelet IIb/IIIa Inhibitor for Stenting (EPISTENT); COBALT and GUSTO III are based on GUSTO I; and A-to-Z is based on Platelet Receptor Inhibition in Ischemic Syndrome Management in Patients Limited by Unstable Signs and Symptoms (PRISM-Plus).

‡ Noninferiority margins are based on d_{max} (lower limit of the 95% CI for the historical OR), and d (50% of d_{max}), which is used as a conservative estimate of the margin.

§ Noninferiority conclusions based on the margins and fractional preservation are compared with published reports.

‖ Fractional preservation < 0% (COBALT and GUSTO III) corresponds to lack of establishment of superiority over the putative placebo.

¶ Analysis based on efficacy end point alone without bleeding.

Statistical Methods Used To Assess Noninferiority

There are 2 basic approaches To the statistical analysis of noninferiority trials. One uses an indirect CI comparison in which noninferiority is inferred if the upper limit of the CI for the difference between the new and the standard treatment is below the lower limit of the 95% CI for the standard treatment effect (the "95-95" approach) (9, 16). The second approach uses a hypothesis-testing framework but switches the roles of the null and alternative hypotheses used in superiority trials. The null hypothesis of inequality (risk difference is greater than or equal to the margin) is rejected in favor of the alternative hypothesis of equality (risk difference is less than the margin) if the 1-sided P value is less than 0.025 (13).

The analysis of our hypothetical trial based on noninferiority margins is summarized in the middle section of **Figure 2**. According to this analysis, the new treatment can be judged noninferior to the standard treatment because the upper bound of 1.8 percentage points for the CI of the absolute risk difference (*left*) is below the investigators' noninferiority margin of 2.0 percentage points ($P =$ 0.009). A judgment of noninferiority cannot be supported for the more restrictive margin of 1.0 percentage point ($P =$ 0.29), nor for similar analyses based on ORs (*right*). Ideally, a judgment of noninferiority would be more reliable if analyses based on absolute and relative difference were concordant, especially if the event rates for the standard treatment in the new trial differ from those in the historical trials.

Determination of Therapeutic Efficacy

As noted, noninferiority trials are conducted on the tacit assumption that the new treatment would exhibit efficacy in a placebo-controlled trial, if such a trial were to be conducted. An inference of efficacy is implied whenever the margin chosen for noninferiority assessment is based on the lower bound of the 95% CI for the standard treatment effect versus placebo. This ensures that the new treatment, although noninferior to the comparator, is also superior to placebo. However, a way to formally, although indirectly, estimate therapeutic efficacy of the new treatment in the absence of a conventional placebo-controlled trial is by the "putative placebo" or "synthesis" approach (6). In terms of ORs, the effect of the new treatment versus placebo is derived from the effect of the new versus the standard treatment in the current trial and the standard treatment versus placebo in the historical trials (5–10): OR (new vs. placebo) ≅ (OR [new vs. standard] × OR [standard vs. placebo]), assuming "constancy" of standard therapy (5–10). Efficacy is thereby established if the OR for the new treatment versus placebo is less than 1.

This approach can also be used to estimate the fraction (f) of the standard treatment effect preserved by the new treatment (5–10, 14, 15). Fractional preservation is established if the lower bound of the 95% CI for this estimate exceeds a prespecified threshold ($0 < f < 1$). As before, the choice of f is a matter of clinical judgment and depends on the critical assumptions of constancy and assay sensitivity.

Analysis of our hypothetical trial in this context is shown in the bottom section of **Figure 2**, in which efficacy of the new treatment is supported by the observation that the derived OR is less than unity (*left*). We also show in the bottom section of **Figure 2** (*right*) that the fraction of the standard treatment effect retained by the new treatment lies within a wide range, 0.11 to 1.21 (14). These data suggest that the new treatment might be only 11% as effective as the standard treatment—much below the 50% threshold suggested by the U.S. Food and Drug Administration for some products (9, 11, 16).

Determination of Ancillary Benefits

As noted, noninferiority should ideally be assessed only if the new treatment offers potential advantages in safety, tolerability, convenience, cost, or some other important facet that justify its use in place of the standard treatment. A formal assessment of superiority, although not currently a regulatory requirement, may nevertheless be desirable and preferably would be specified prospectively within the active-control trial design as a secondary objective (5–7). At the very least, the new treatment should have acceptable safety and tolerability, and the evidence in support of these ancillary benefits should be described explicitly in a published report.

Analyses of Published Trials

We performed a retrospective review of cardiovascular noninferiority trials published over the past 8 years in the *New England Journal of Medicine* and *Journal of the American Medical Association* to determine the degree to which these trials met the standards summarized in **Table 1**. We then compared the published conclusions of those trials based on our re-analysis employing these standards.

In **Table 2**, we summarize the published results of these trials (17–24), one of which (Stroke Prevention Using Oral Thrombin Inhibitor in Atrial Fibrillation V [SPORTIF V]) (24) is quantitatively similar to our hypothetical trial. Statistical reasoning and clinical judgment were specified in the selection of the noninferiority margin in all but 2 trials, Superior Yield of the New strategy of Enoxaparin, Revascularization, and GlYcoprotein IIb/IIIa inhibitors (SYNERGY) and SPORTIF V, in which the operative margins were chosen on the basis of expert consensus among members of the trial committees (23, 24). Six of 8 trials reported a conclusion of noninferiority. Superior or acceptable safety and tolerability were reported in all but 3 trials, Continuous Infusion versus Double-Bolus Administration of Alteplase (COBALT), SYNERGY, and SPORTIF V.

In **Table 3**, we summarize the results of our re-analysis according to therapeutic noninferiority alone, therapeutic efficacy with respect to a putative placebo, and fractional preservation of standard treatment effect. Our noninferiority results correspond with the published conclusions in 5 of 8 trials, 2 in which noninferiority is not established (Do Tirofiban and ReoPro Give Similar Efficacy Trial [TARGET] and COBALT) and 3 in which noninferiority is established (Randomized Evaluation in PCI Linking Angiomax to Reduced Clinical Events [REPLACE]-2, Valsartan in Acute Myocardial Infarction [VALIANT], and Aggrastat-to-Zocor [A-to-Z]) using the lower limit of the 95% CI (d_{max}). When the more restrictive criterion (d) is used, noninferiority can be inferred only for REPLACE-2 (combined efficacy and safety outcome). However, the evidence for noninferiority is robust in these 3 trials because it is established at clinically acceptable marginal (OR $<$1.15) and fractional ($f > 0.5$) thresholds (**Table 3**).

In contrast, although SYNERGY results were consistent with noninferiority according to the margin used in the published report, our re-analysis did not support noninferiority for either margin. Moreover, the synthesis analysis barely established efficacy of the new treatment over a putative placebo—the upper limit of the OR for the new treatment versus placebo was 0.99—and demonstrated a fractional preservation that was less than 25% of the standard. In addition, a statistically significant increase in bleeding was observed with the new treatment (23), thereby calling into question its ancillary benefits.

Absolute versus Relative Difference

In SPORTIF V, the new treatment was not inferior to the standard according to the absolute margin used in the published report (**Table 2**) but was inferior according to the OR margin used in our re-analysis. The reason for this discordance is that the observed warfarin rate of 1.2% was less than half the expected rate of 3.1% (24), thereby inflating the operative margin (absolute difference of 2.0 percentage points) from a relative risk of 1.65 ([3.1 + 2.0]/3.1) to 2.67 ([1.2 + 2.0]/1.2). Hence, a fixed relative difference (rate ratio) should be preferred over a constant absolute difference for the selection of the margin, especially when observed and expected event rates are dissimilar. Thus, taken together with less than 25% fractional preservation (**Table 3**) and the observed increase in liver toxicity with ximelagatran in SPORTIF V (24), a conclusion of noninferiority is not apparent from these data (10).

Unconventional Combined Efficacy and Safety Outcome

The reference of a noninferiority inference to a combined efficacy and safety outcome in REPLACE-2 is highly unusual from a regulatory perspective; separate assessment of efficacy and safety is the norm. In fact, the published noninferiority conclusion in REPLACE-2 depended more on safety than efficacy: a 43% odds reduction in major bleeding compared with a 9% odds increase in efficacy outcome (20). Accordingly, when analysis is restricted to the conventional efficacy outcome alone, noninferiority is supported only for margins greater than an OR of 1.32 (**Table 3**), which is arguably a threshold higher than is clinically acceptable. This observation serves to support the U.S. Food and Drug Administration's conclusion that "sta-

Table 4. **Determinants of Quality for the Assessment of Noninferiority**

The trial should be designed in a manner consistent with the historical placebo-controlled trials on which its analysis will rely.
The active-control comparator should be a well-established, effective standard therapy that has predictable, quantifiable, and consistent treatment effects.
The noninferiority margin should be specified a priori and should be based on statistical reasoning and clinical considerations regarding benefit, risk, and cost.
The trial should be conducted in accordance with appropriately high standards to minimize protocol deviations and optimize therapeutic adherence.
The trial should be analyzed by using stringent rather than liberal criteria with respect to the marginal threshold, confidence interval, and fractional preservation.
The stability of the noninferiority conclusion should be reported with respect to relevant analytic alternatives (per-protocol vs. intention-to-treat and absolute vs. relative outcomes).

tistical noninferiority was not demonstrated for the efficacy end point" (25).

Setting the Value of the Margin

The 2 thrombolytic trials that we re-analyzed, COBALT and Global Use of Strategies to Open Occluded Arteries in Acute Coronary Syndromes (GUSTO) III, illustrate how inconsistencies in the choice and application of the noninferiority margin and the use of 1-sided versus 2-sided CIs can lead to contradictory results. The operative margin for these trials was based on GUSTO I, a mortality trial that showed an absolute difference of 1.0 percentage point (CI, 0.4 to 1.6 percentage points) in favor of the standard treatment (26). However, GUSTO III used the point estimate (1.0 percentage point), whereas COBALT used the more stringent value, the lower bound of the 95% CI (0.4 percentage point). Although these 2 trials reported similar differences in 30-day mortality rates, GUSTO III was interpreted to be consistent with noninferiority (based on 1-sided but not 2-sided 95% CIs) but COBALT was not (**Table 2**). If both studies had used the same criteria, they would have arrived at the same conclusion (27). On the contrary, small differences in the selection of the operative margin had substantial effects on the resultant noninferiority judgment (established with a margin of 1.1 percentage points but not with a margin of 1.0 percentage point in GUSTO III).

Inferring Noninferiority from Superiority Trials

GUSTO III relied on only 1 previous trial to assess the efficacy of the standard treatment (26, 28). In such cases—in which assay sensitivity cannot be supported confidently—an alternative design strategy is to use a placebo arm, if possible, or to make superiority of the new over the standard treatment the principal goal of the trial (16), as was the case in GUSTO III. Because post hoc determinations of noninferiority (as in GUSTO III) are prone to error, especially when the interpretive margin is not fixed a priori, it may be more advisable to incorporate superiority and noninferiority assessments into the prospective trial design and to use the latter to define the requisite sample size (11).

Impact of Margin on Sample Size

Although GUSTO III enrolled more than 15 000 participants, it was not adequately powered for its formal assessment of noninferiority and thus may have suffered from a type II error (failure to identify true benefit). In fact, the narrow margin in COBALT would have necessitated a sample of approximately 50 000 patients in GUSTO III (27). Such sample sizes often render noninferiority trials impractical (27). As a consequence, some have argued for relaxation of these stringent criteria (27). Reconciling these offsetting considerations of trial feasibility and stringency poses a substantial challenge. To cope with these ambiguities, statisticians and clinicians in consultation with the regulatory authorities should determine, during the planning phase of investigation, a consensus noninferiority limit that is clinically relevant and statistically feasible and that balances the contrasting perspectives of the principal stakeholders (patients, investigators, sponsors, regulators, and payers).

Conclusion

Several problems challenge the design, conduct, analysis, reporting, and interpretation of noninferiority trials, and recent meta-analyses confirm that a majority of published trials have substantial methodologic flaws (29, 30). As a result, potentially suboptimal treatments might be introduced into routine clinical practice (29). In this paper, we suggest a variety of ways to identify and mitigate these errors (**Table 4**). Other issues that are crucial to ensuring the validity of noninferiority inference, such as ethical considerations, adequate power, the quality of trial conduct, the choice of analytic strategy (intention-to-treat versus per-protocol), and an alternative Bayesian approach to analysis, are beyond the scope of this paper and have been detailed previously (2–16, 31–34).

In conclusion, if noninferiority trials are to be applied to regulatory and clinical decisions about the marketing and use of new treatments, their assumptions must be made explicit, the criteria on which they are based must be sufficiently justified, and their influence on the resultant conclusions must be assessed rigorously and expressed unambiguously in published reports. Only in this way might we " . . . avoid false claims, inconsistencies, and inappropriate use of suboptimal therapies" (29).

From Cedars-Sinai Medical Center and David Geffen School of Medicine, University of California, Los Angeles, California.

Potential Financial Conflicts of Interest: None disclosed.

Requests for Single Reprints: Sanjay Kaul, MD, Division of Cardiology, Room 5536, South Tower, Cedars-Sinai Medical Center, 8700 Beverly Boulevard, Los Angeles, CA 90048; e-mail, kaul@cshs.org.

Current author addresses are available at www.annals.org.

References

1. **Paine T.** Common Sense, the Rights of Man, and Other Essential Writings; 1792.
2. **Temple R, Ellenberg SS.** Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 1: ethical and scientific issues. Ann Intern Med. 2000;133:455-63. [PMID: 10975964]
3. **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;133:464-70. [PMID: 10975965]
4. **Siegel JP.** Equivalence and noninferiority trials. Am Heart J. 2000;139:S166-70. [PMID: 10740125]
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;22:213-25. [PMID: 12520558]
6. **Hung HM, Wang SJ, O'Neill R.** A regulatory perspective on choice of margin and statistical inference issue in non-inferiority trials. Biom J. 2005;47:28-36. [PMID: 16395994]
7. **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. [PMID: 12520555]
8. **Snapinn SM.** Alternatives for discounting in the analysis of noninferiority trials. J Biopharm Stat. 2004;14:263-73. [PMID: 15206525]
9. **Rothmann M, Li N, Chen G, Chi GY, Temple R, Tsou HH.** Design and analysis of non-inferiority mortality trials in oncology. Stat Med. 2003;22:239-64. [PMID: 12520560]
10. **Kaul S, Diamond GA, Weintraub WS.** Trials and tribulations of non-inferiority: the ximelagatran experience. J Am Coll Cardiol. 2005;46:1986-95. [PMID: 16325029]
11. International Conference on Harmonisation; choice of control group and related issues in clinical trials (ICH E 10); availability—FDA. Notice. Fed Regist. 2001;66:24390-1. [PMID: 12356096]
12. Committee for Proprietary Medicinal Products. Points to Consider on the Choice of Non-Inferiority Margin. The European Agency for the Choice of Medicinal Products. 2004; CPMP/EWP/2158/99 draft. Accessed at http://home.att.ne.jp/red/akihiro/emea/215899en_ptc.pdf on 16 May 2006.
13. **Blackwelder WC.** "Proving the null hypothesis" in clinical trials. Control Clin Trials. 1982;3:345-53. [PMID: 7160191]
14. **Hasselblad V, Kong DF.** Statistical methods for comparison to placebo in active-controlled trials. Drug Info J. 2001;35:435-449.
15. **Holmgren EB.** Establishing equivalence by showing that a specified percentage of the effect of the active control over placebo is maintained. J Biopharm Stat. 1999;9:651-9. [PMID: 10576409]
16. **Temple R.** Policy developments in regulatory approval. Stat Med. 2002;21:2939-48. [PMID: 12325110]
17. **Topol EJ, Moliterno DJ, Herrmann HC, Powers ER, Grines CL, Cohen DJ, et al.** Comparison of two platelet glycoprotein IIb/IIIa inhibitors, tirofiban and abciximab, for the prevention of ischemic events with percutaneous coronary revascularization. N Engl J Med. 2001;344:1888-94. [PMID: 11419425]
18. A comparison of continuous infusion of alteplase with double-bolus administration for acute myocardial infarction. The Continuous Infusion versus Double-Bolus Administration of Alteplase (COBALT) Investigators. N Engl J Med. 1997;337:1124-30. [PMID: 9340504]
19. A comparison of reteplase with alteplase for acute myocardial infarction. The Global Use of Strategies to Open Occluded Coronary Arteries (GUSTO III) Investigators. N Engl J Med. 1997;337:1118-23. [PMID: 9340503]
20. **Lincoff AM, Bittl JA, Harrington RA, Feit F, Kleiman NS, Jackman JD, et al.** Bivalirudin and provisional glycoprotein IIb/IIIa blockade compared with heparin and planned glycoprotein IIb/IIIa blockade during percutaneous coronary intervention: REPLACE-2 randomized trial. JAMA. 2003;289:853-63. [PMID: 12588269]
21. **Pfeffer MA, McMurray JJ, Velazquez EJ, Rouleau JL, Køber L, Maggioni AP, et al.** Valsartan, captopril, or both in myocardial infarction complicated by heart failure, left ventricular dysfunction, or both. N Engl J Med. 2003;349:1893-906. [PMID: 14610160]
22. **Blazing MA, de Lemos JA, White HD, Fox KA, Verheugt FW, Ardissino D, et al.** Safety and efficacy of enoxaparin vs unfractionated heparin in patients with non-ST-segment elevation acute coronary syndromes who receive tirofiban and aspirin: a randomized controlled trial. JAMA. 2004;292:55-64. [PMID: 15238591]
23. **Ferguson JJ, Califf RM, Antman EM, Cohen M, Grines CL, Goodman S, et al.** Enoxaparin vs unfractionated heparin in high-risk patients with non-ST-segment elevation acute coronary syndromes managed with an intended early invasive strategy: primary results of the SYNERGY randomized trial. JAMA. 2004;292:45-54. [PMID: 15238590]
24. **Albers GW, Diener HC, Frison L, Grind M, Nevinson M, Partridge S, et al.** Ximelagatran vs warfarin for stroke prevention in patients with nonvalvular atrial fibrillation: a randomized trial. JAMA. 2005;293:690-8. [PMID: 15701910]
25. **Hughes S.** FDA Approves REPLACE-2 Label for Bivalirudin. HeartWire News. 16 June 2005. Accessed at www.theheart.org on 16 May 2006
26. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. The GUSTO investigators. N Engl J Med. 1993;329:673-82. [PMID: 8204123]
27. **Ware JH, Antman EM.** Equivalence trials [Editorial]. N Engl J Med. 1997;337:1159-61. [PMID: 9329939]
28. **Dundar Y, Hill R, Dickson R, Walley T.** Comparative efficacy of thrombolytics in acute myocardial infarction: a systematic review. QJM. 2003;96:103-13. [PMID: 12589008]
29. **Greene WL, Concato J, Feinstein AR.** Claims of equivalence in medical research: are they supported by the evidence? Ann Intern Med. 2000;132:715-22. [PMID: 10787365]
30. **Le Henanff A, Giraudeau B, Baron G, Ravaud P.** Quality of reporting of noninferiority and equivalence randomized trials. JAMA. 2006;295:1147-51. [PMID: 16522835]
31. **Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ.** Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. JAMA. 2006;295:1152-60. [PMID: 16522836]
32. **Gøtzsche PC.** Lessons from and cautions about noninferiority and equivalence randomized trials [Editorial]. JAMA. 2006;295:1172-4. [PMID: 16522840]
33. **Simon R.** Bayesian design and analysis of active control clinical trials. Biometrics. 1999;55:484-7. [PMID: 11318204]
34. **Diamond GA, Kaul S.** Prior convictions: Bayesian approaches to the analysis and interpretation of clinical megatrials. J Am Coll Cardiol. 2004;43:1929-39. [PMID: 15172393]

Current Author Addresses: Drs. Kaul and Diamond: Division of Cardiology, Room 5536, South Tower, Cedars-Sinai Medical Center, 8700 Beverly Boulevard, Los Angeles, CA 90048.