Reporting of Noninferiority and Equivalence Randomized Trials
An Extension of the CONSORT Statement

Gilda Piaggio, PhD
Diana R. Elbourne, PhD
Douglas G. Altman, DSc
Stuart J. Pocock, PhD
Stephen J. W. Evans, MSc
for the CONSORT Group

THE CONSOLIDATED STANdards of Reporting Trials (CONSORT) statement was developed to alleviate the problem of inadequate reporting of randomized controlled trials (RCTs),[1-4] which has been associated with biased treatment effects.[5-7] The statement comprises evidence-based recommendations for reporting RCTs, including a flowchart of participants through the trial.

CONSORT's primary focus is on parallel group trials,[1-3] aiming to identify treatment superiority if it really exists. Most CONSORT recommendations apply equally well to other trial designs, but some need adaptation. Herein we extend the CONSORT recommendations to noninferiority and equivalence trials. First, we explain the rationale for and key methodological features of such trials. Second, we consider how commonly noninferiority and equivalence trials are published and provide empirical evidence about their quality. Last, we present an adapted CONSORT checklist for reporting non-

The CONSORT (Consolidated Standards of Reporting Trials) Statement, including a checklist and a flow diagram, was developed to help authors improve their reporting of randomized controlled trials. Its primary focus was on individually randomized trials with 2 parallel groups that assess the possible superiority of one treatment compared with another but is now being extended to other trial designs. Noninferiority and equivalence trials have methodological features that differ from superiority trials and present particular difficulties in design, conduct, analysis, and interpretation. Although the rationale for such trials occurs frequently, those designed and described specifically as noninferiority or equivalence trials appear less commonly in the medical literature. The quality of reporting of those that are published is often inadequate. In this article, we present an adapted CONSORT checklist for reporting noninferiority and equivalence trials and provide illustrative examples and explanations for those items amended from the original CONSORT checklist. The intent is to improve reporting of noninferiority and equivalence trials, enabling readers to assess the validity of their results and conclusions.

JAMA. 2006;295:1152-1160 wwww.jama.com

inferiority and equivalence trials and give illustrative examples (and further elaboration) for those items that have been amended.

For convenience, we will refer to *treatments* and *patients*, although we recognize that not all interventions evaluated in RCTs are technically treatments and the participants in trials are not always patients.

Rationale for Noninferiority or Equivalence Designs

Most RCTs aim to determine whether one intervention is superior to another. By contrast, equivalence trials[8] aim to determine whether one (typically new) intervention is therapeutically similar to another, usually an existing treatment. We use *new* to refer to the treatment under test, and the comparison or reference treatment is often called an *active control*.

A noninferiority trial seeks to determine whether a new treatment is no

Author Affiliations: Statistics and Informatics Services Group, Department of Reproductive Health and Research, World Health Organization, Geneva, Switzerland (Dr Piaggio); Medical Statistics Unit, London School of Hygiene and Tropical Medicine, London (Drs Elbourne and Pocock and Mr Evans) and Centre for Statistics in Medicine, Oxford (Dr Altman), England. **Corresponding Author:** Gilda Piaggio, PhD, Department of Reproductive Health and Research, World Health Organization, 1211 Geneva 27, Switzerland (piaggiog@who.int).

worse than a reference treatment. Because proof of exact equality is impossible, a prestated margin of noninferiority (Δ) for the treatment effect in a primary patient outcome is defined. Equivalence trials are very similar, except that equivalence is defined as the treatment effect being between –Δ and Δ. True (2-sided) equivalence therapeutic or prophylactic trials are rare because most such trials address the question of noninferiority.

In trials that investigate noninferiority, the question of interest is not symmetric.[9] The new treatment will be recommended if it is similar to or better than an existing one, but not if it is worse (by more than Δ). Superiority of the new treatment would be a bonus. This article focuses mainly on noninferiority trials but applies also to 2-sided equivalence trials.

Noninferiority trials are intended to show whether a new treatment has at least as much efficacy as the standard or is worse by an amount less than Δ, often on the premise that it has some other advantage, eg, greater availability, reduced cost, less invasiveness,[10,11] fewer side effects (harms),[12] or greater ease of administration,[13] for instance, one daily dose rather than 2 doses[14] or more than 2 doses.[15] Some noninferiority trials have been criticized for merely studying a new marketable product ("me-too" drugs).[16]

Noninferiority and equivalence trials are not limited to drugs. For example, a new antenatal care model with fewer clinic visits and reduced cost was investigated for its equivalence to the standard model as regards maternal and neonatal outcomes.[17] A noninferiority trial compared 2 interventional strategies for coronary revascularization in diabetic patients.[18]

Methodological Issues in Noninferiority and Equivalence Trials

Noninferiority and equivalence trials present particular difficulties in design, conduct, analysis, and interpretation.[19]

Hypotheses. In a superiority trial the null hypothesis is that treatments are equally effective and the alternative hypothesis is that they differ. A type I error is falsely finding a treatment effect when there is none, and a type II error is failing to detect a treatment effect when truly one exists. In noninferiority trials, the null and alternative hypotheses are reversed; a type I error is the erroneous acceptance of an inferior new treatment, whereas a type II error is the erroneous rejection of a truly noninferior treatment.

Design. A noninferiority or equivalence trial requires that the reference treatment's efficacy is established[20,21] or is in widespread use so that a placebo or untreated control group would be deemed unethical.

Both participants and outcome measures in a noninferiority or equivalence trial should be similar to those in trial(s) that established the efficacy of the reference treatment. Outcome measures should also be similar to those in previous trials.

The required sample size is calculated using the confidence interval (CI) approach, considering where the CI for the treatment effect lies with respect to both the margin of noninferiority Δ and a null effect. Sample size depends on the level of confidence chosen, the risk of type II error (or desired power), and Δ.[22,23] A prestated margin of noninferiority Δ can be specified as a difference in means or proportions or the logarithm of an odds ratio, risk ratio, or hazard ratio. A prestated margin of noninferiority is often chosen as the smallest value that would be a clinically important effect.[24] If relevant, Δ should be smaller than the "clinically relevant" effect chosen to investigate superiority of reference treatment against placebo.[25,26] For example, if mortality with treatment A is better than placebo by 10% (absolute difference), a new treatment B might need to be at least 5% better than placebo (and thus no more than 5% worse than A). The required size of noninferiority trials is therefore usually larger than that for superiority trials.[25] Unfortunately, sample sizes for noninferiority and equivalence trials are often too small.[19,24] Given several previous trials, the effect of the reference treatment can be estimated from a meta-analysis. There are several techniques to determine Δ,[27] its magnitude being influenced by several factors, eg, efficacy, safety, cost, acceptability, and adherence.[28,29]

Conduct. Trial conduct should closely match any trial that demonstrated efficacy of the reference treatment, provided they were of high quality.[20] One should avoid features that might dilute true differences between treatments, thereby enhancing the risk of erroneously concluding noninferiority,[30,31] eg, poor adherence, dropouts, recruitment of patients unlikely to respond, and treatment crossovers.

Analysis. Although a modified hypothesis testing framework exists,[32,33] a more informative CI approach is preferred in the design, analysis, and reporting of noninferiority and equivalence trials.[34]

For superiority trials, intention-to-treat (ITT) analysis (analyzing all patients within their randomized groups, regardless of whether they completed allocated treatment) is recommended.[3] Intention-to-treat analysis often leads to smaller observed treatment effects than if all patients had adhered to treatment. In noninferiority trials, ITT analysis will often increase the risk of falsely claiming noninferiority (type I error),[25] although not always.[35] In practice, ITT analysis is often not possible and one uses a "full analysis set" to describe that patient follow-up, which is "as complete . . . and . . . close as possible" to ITT.[36]

Alternative analyses that exclude patients not taking allocated treatment or otherwise not protocol-adherent could bias the trial in either direction.[37] The terms *on-treatment* or *per-protocol* analysis are often used but may be inadequately defined. Potentially biased non-ITT analysis is less desirable than ITT in superiority trials but may still provide some insight. In noninferiority and equivalence trials, non-ITT

analyses might be desirable as a protection from ITT's increase of type I error risk (falsely concluding noninferiority).[36] There is greater confidence in results when the conclusions are consistent.

Subgroup analysis requires the same caveats in noninferiority trials as it requires in superiority trials. Interim analyses in noninferiority trials have some differences in rationale from superiority trials. If noninferiority is established before the trial is completed, there may be no ethical requirement to stop early because of lack of efficacy.[19] However, other advantages (adverse effects, cost) could justify stopping the trial, to expedite availability of the new treatment. If a treatment is clearly inferior, then stopping the trial (or a particular trial arm) is ethically justi-fied.[38-40] Stopping rules might be asymmetric, a trial being allowed to continue longer if the new treatment appears superior,[41] although this result is unlikely.[19]

Interpretation. Interpreting a non-inferiority trial's results depends on where the CI for the treatment effect lies relative to both the margin of noninfe-riority Δ and a null effect. The observed treatment effect is not by itself sufficiently informative. With 2-sided equivalence the interpretation is analo-gous, but both margins Δ and $-\Delta$ need considering, and claiming equiva-lence requires the CI to lie wholly be-tween $-\Delta$ and Δ.

Many noninferiority trials based their interpretation on the upper limit of a 1-sided 97.5% CI, which is the same as the upper limit of a 2-sided 95% CI. Although both 1-sided and 2-sided CIs allow for inferences about noninferiority, we suggest that 2-sided CIs are appropriate in most noninferi-ority trials.[29] If a 1-sided 5% signifi-cance level is deemed acceptable for the noninferiority hypothesis test[42] (a decision open to question), a 90% 2-sided CI could then be used. The **FIGURE** interprets several possible scenarios with 2-sided CIs for a non-inferiority trial.

Once noninferiority is evident, it is acceptable to then assess whether the new treatment appears superior to the reference treatment, using an appro-priate test or CI (ie, not just the point estimate), preferably defined a priori and with an ITT analysis.[22,28,43]

It is inappropriate to claim noninfe-riority post hoc from a superiority trial unless clearly related to a predefined margin of equivalence. That is, both su-periority and noninferiority hypoth-eses need explicit specification in the trial protocol.[44] It is, however, always reasonable to interpret a CI as exclud-ing an effect of a particular prestated size.[45] Having demonstrated noninfe-riority against reference treatment, some authors then make claims for efficacy of a new treatment relative to placebo by also using evidence from earlier trials of reference treatment vs placebo.[46] Such inferences assume assay con-stancy, ie, current and earlier trials are identical in all relevant aspects,[20] eg, participants, outcomes definition, and use of standard therapy. Regarding pa-tient populations, for example, this im-plies no differences in the effect of treat-ment across subgroups or similar distribution of relevant subgroups. In the absence of assay constancy, an ad-justment method has been pro-posed.[27] Since assay constancy is inevi-tably questionable, any claims regarding efficacy of new treatment relative to pla-cebo require cautious interpretation.

Figure. Possible Scenarios of Observed Treatment Differences for Adverse Outcomes (Harms) in Noninferiority Trials
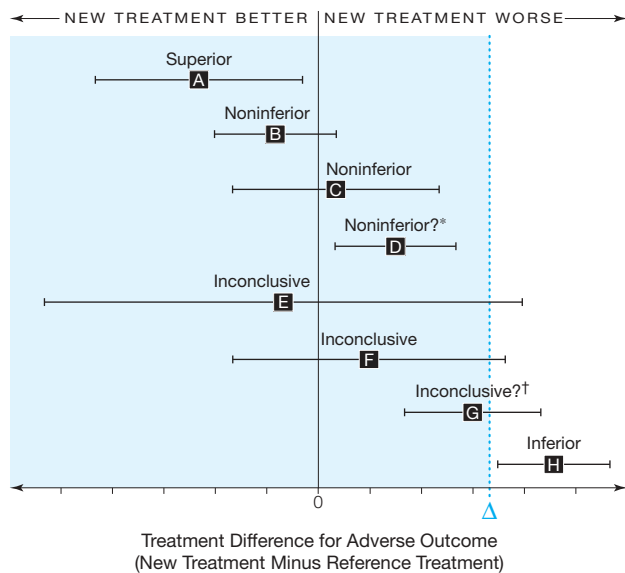
Error bars indicate 2-sided 95% confidence intervals (CIs). Tinted area indicates zone of inferiority. A, If the CI lies wholly to the left of zero, the new treatment is superior. B and C, If the CI lies to the left of Δ and includes zero, the new treatment is noninferior but not shown to be superior. D, If the CI lies wholly to the left of Δ and wholly to the right of zero, the new treatment is noninferior in the sense already defined, but it is also inferior in the sense that a null treatment difference is excluded. This puzzling case is rare, since it requires a very large sample size. It can also result from having too wide a noninferiority margin. E and F, If the CI includes Δ and zero, the difference is nonsignificant but the result regarding noninferiority is inconclusive. G, If the CI includes Δ and is wholly to the right of zero, the difference is statistically significant but the result is inconclusive re-garding possible inferiority of magnitude Δ or worse. H, If the CI is wholly above Δ, the new treatment is in-ferior.[22,43]

*This CI indicates noninferiority in the sense that it does not include Δ, but the new treatment is significantly worse than the standard. Such a result is unlikely because it would require a very large sample size.

†This CI is inconclusive in that it is still plausible that the true treatment difference is less than Δ, but the new treatment is significantly worse than the standard.

How Common Are Noninferiority and Equivalence Trials?

Assessing the frequency of noninferi-ority and equivalence trials is not

straightforward because of inconsistencies in terminology. A search of the Cochrane Controlled Trials Register[47] in October 2004 for the words *equivalence* or *noninferiority* yielded 1021 of 415 918 trials (0.2%), but these figures are likely to be misleading. Not all noninferiority or equivalence trials use these words, and the term *equivalence* is often inappropriately used when reporting negative results of superiority trials; such trials often lack statistical power to rule out important differences.[48,49] Identifying noninferiority trials is difficult because they are often labeled as equivalence trials. A recent study[50] found that only 3 of 188 (1.6%) cancer trials in a PubMed electronic search were designed to evaluate equivalence or noninferiority (Luciano Costa, MD, written communication, April 26, 2005).

From the above findings, it seems that whereas the objective of testing for noninferiority or equivalence is likely to be common, there have been relatively few noninferiority and equivalence trials that are both designed and described as such. However, we perceive that these designs are becoming more widespread.

Quality of Reporting of Noninferiority and Equivalence Trials

We do not know of a study that looked at reporting of a cohort of trials actually designed as noninferiority or equivalence trials. There have been several reviews of quality of trials claiming equivalence, without differentiation between noninferiority and 2-sided equivalence, because many authors use the term *equivalence* to mean either.

Greene et al[49] identified methodological flaws in a systematic review of 88 studies claiming equivalence, published from 1992 to 1996. Equivalence was inappropriately claimed in 67% of them, on the basis of nonsignificant tests for superiority. Fifty-one percent stated equivalence as an aim, but only 23% were designed with a preset margin of equivalence. Only 22%

adopted appropriate practice: a predefined aim of equivalence, a preset Δ, consequent sample size determination, and actually testing equivalence.

Other disease- or field-specific reviews reveal similar findings. Only 2 trials (8%) in a review[51] of 25 RCTs in childhood bacterial meningitis published between 1980 and 2000 that claimed equivalent mortality were designed to test equivalence. A review of 90 RCTs with nonstatistically significant or "negative" results published in 3 surgical journals from 1988 to 1998 found 39% met predefined criteria for establishing equivalence.[52] Only 3 studies in a recent review of 188 cancer trials with negative results used a noninferiority or equivalence analysis.[50] In a review of 20 trials intended to detect equivalence in reproductive health, only 4 stated a margin of equivalence.[53]

McAlister and Sackett[37] evaluated 4 large "negative" RCTs in hypertension as regards methodological requirements for active-control equivalence trials. Only 2 trials published both ITT and per-protocol (or on-treatment) analyses, only 1 trial specified the margin of equivalence in advance, and none was sufficiently large to address the equivalence hypothesis. This illustrates how failure to properly design, conduct, and analyze equivalence trials leads to incorrect conclusions about equivalence.

Extension of Consort Statement

To accommodate noninferiority or equivalence trials, an extension of the CONSORT statement should encompass the following issues: (1) the rationale for adopting a noninferiority or equivalence design; (2) how study hypotheses were incorporated into the design; (3) choice of participants, interventions (especially the reference treatment), and outcomes; (4) statistical methods, including sample size calculation; and (5) how the design affects interpretation and conclusions. Consequences for the CONSORT checklist and flow diagram, including specific changes, are described below.

Checklist

We build on the work of McAlister and Sackett[37] in modifying the CONSORT checklist[2,3] (**TABLE**), especially items 1 to 7, 12, 16, 17, and 20. New text is shown in italics. For each modification, we include 1 or more examples of good reporting (and further elaboration where appropriate). In some examples, we have added text in brackets to explain the context. We mainly concentrate on noninferiority trials but make some reference to equivalence trials which are much less common.

Title and Abstract

Title and Abstract: Item 1. How participants were allocated to interventions (eg, random allocation, randomized, or randomly assigned), *specifying that the trial is a noninferiority or equivalence trial.*

Title. "Oral Pristinamycin versus Standard Penicillin Regimen to Treat Erysipelas in Adults: Randomised, Noninferiority, Open Trial"[54]

Abstract. "Design—Multicentre, parallel group, open labelled, randomised noninferiority trial."[54]

Introduction

Background: Item 2. Scientific background and explanation of rationale, *including the rationale for using a noninferiority or equivalence design.*

Example. "Up to 40 million children worldwide are estimated to suffer from vitamin A deficiency. . . . A dose of 200,000 IU retinyl palmitate to children over 1 year old is most widely used and has generally been regarded as safe and potentially effective. . . . In developing countries, animal products that provide retinyl esters are too expensive. . . . Vegetables and fruit . . . are cheap and good sources of vitamin A in the form of beta carotene. . . . Beta carotene is also considered to be virtually non-toxic. . . . In a preliminary study, . . . after 20 days there was a reversion of the clinical and subclinical signs of vitamin A deficiency in the study group. . . . Since beta carotene is the principal source of vitamin A in developing countries and is

Table. Checklist of Items for Reporting Noninferiority or Equivalence Trials (Additions or Modifications to the CONSORT Checklist are Shown in Italics)

Paper Section and Topic	Item Number	Descriptor (Adapted for Noninferiority or Equivalence Trials)
Title and abstract	1*	How participants were allocated to interventions (eg, "random allocation," "randomized," or "randomly assigned"), *specifying that the trial is a noninferiority or equivalence trial.*
Introduction Background	2*	Scientific background and explanation of rationale, *including the rationale for using a noninferiority or equivalence design*.
Methods Participants	3*	Eligibility criteria for participants *(detailing whether participants in the noninferiority or equivalence trial are similar to those in any trial[s] that established efficacy of the reference treatment)* and the settings and locations where the data were collected.
Interventions	4*	Precise details of the interventions intended for each group, *detailing whether the reference treatment in the noninferiority or equivalence trial is identical (or very similar) to that in any trial(s) that established efficacy,* and how and when they were actually administered.
Objectives	5*	Specific objectives and hypotheses, *including the hypothesis concerning noninferiority or equivalence.*
Outcomes	6*	Clearly defined primary and secondary outcome measures, *detailing whether the outcomes in the noninferiority or equivalence trial are identical (or very similar) to those in any trial(s) that established efficacy of the reference treatment* and, when applicable, any methods used to enhance the quality of measurements (eg, multiple observations, training of assessors).
Sample size	7*	How sample size was determined, *detailing whether it was calculated using a noninferiority or equivalence criterion and specifying the margin of equivalence with the rationale for its choice.* When applicable, explanation of any interim analyses and stopping rules *(and whether related to a noninferiority or equivalence hypothesis)*.
Randomization Sequence generation	8	Method used to generate the random allocation sequence, including details of any restriction (eg, blocking, stratification).
Allocation concealment	9	Method used to implement the random allocation sequence (eg, numbered containers or central telephone), clarifying whether the sequence was concealed until interventions were assigned.
Implementation	10	Who generated the allocation sequence, who enrolled participants, and who assigned participants to their groups.
Blinding (masking)	11	Whether or not participants, those administering the interventions, and those assessing the outcomes were blinded to group assignment. When relevant, how the success of blinding was evaluated.
Statistical methods	12*	Statistical methods used to compare groups for primary outcome(s), *specifying whether a 1- or 2-sided confidence interval approach was used*. Methods for additional analyses, such as subgroup analyses and adjusted analyses.
Results Participant flow	13	Flow of participants through each stage (a diagram is strongly recommended). Specifically, for each group report the numbers of participants randomly assigned, receiving intended treatment, completing the trial protocol, and analyzed for the primary outcome. Describe protocol deviations from trial as planned, together with reasons.
Recruitment	14	Dates defining the periods of recruitment and follow-up.
Baseline data	15	Baseline demographic and clinical characteristics of each group.
Numbers analyzed	16*	Number of participants (denominator) in each group included in each analysis and whether *"intention-to-treat" and/or alternative analyses were conducted*. State the results in absolute numbers when feasible (eg, 10/20, not 50%).
Outcomes and estimation	17*	For each primary and secondary outcome, a summary of results for each group and the estimated effect size and its precision (eg, 95% confidence interval). *For the outcome(s) for which noninferiority or equivalence is hypothesized, a figure showing confidence intervals and margins of equivalence may be useful.*
Ancillary analyses	18	Address multiplicity by reporting any other analyses performed, including subgroup analyses and adjusted analyses, indicating those prespecified and those exploratory.
Adverse events	19	All important adverse events or side effects in each intervention group.
Comment Interpretation	20*	Interpretation of the results, taking into account the *noninferiority or equivalence hypothesis and any other* trial hypotheses, sources of potential bias or imprecision and the dangers associated with multiplicity of analyses and outcomes.
Generalizability	21	Generalizability (external validity) of the trial findings.
Overall evidence	22	General interpretation of the results in the context of current evidence.

*Expansion of corresponding item on CONSORT checklist.[2,3]

non-toxic, we compared retinyl palmitate and beta carotene for treatment of vitamin A deficiency."[55]

Elaboration. The rationale should cite evidence for the efficacy of the reference treatment. If previous trials, or their systematic review, demonstrate the superiority of the reference treatment relative to placebo, they should be cited with effect sizes and CIs. If no such trials exist, other evidence for efficacy of the reference treatment should be given. Evidence for other advantages of the new treatment over the reference treatment, if present, should be given, to justify use of the new treatment, if not inferior. One aim of the current trial might be to provide or support such evidence. In the case of "me-too" drugs, it should be clear whether there are other advantages.

Methods

Participants: Item 3. Eligibility criteria for participants (*detailing whether participants in the noninferiority or equivalence trial are similar to those in any trial[s] that established efficacy of the reference treatment)* and the settings and locations where the data were collected.

Example. "From Sept 1, 1992, to Dec 30, 1994, we enrolled 6628 men and women in 312 health centres in Sweden . . . who had hypertension (blood pressure ≥180 mm Hg systolic, ≥105 mm Hg diastolic, or both), aged 70-84 years. The only difference in inclusion criteria between this trial and the STOP-Hypertension trial was that patients with isolated systolic hypertension could be included in STOP-Hypertension-2, based on previous positive findings in patients with isolated systolic hypertension treated with diuretics and calcium antagonists."[56]

Elaboration. Relevant changes in participants' characteristics compared with previous trial(s) should be reported and explained. Clinical trial participants differ, mainly if time has elapsed between trials; therefore, such description should concentrate in relevant departures (that might affect response to treatments).

Interventions: Item 4. Precise details of the interventions intended for each group, *detailing whether the reference treatment in the noninferiority or equivalence trial is identical (or very similar) to that in any trial(s) that established efficacy,* and how and when they were actually administered.

Example. "[W]e randomly assigned women about to deliver vaginally to receive 600 μg misoprostol orally or 10 IU oxytocin intravenously or intramuscularly, according to practice. . . . The use of uterotonic agents [oxytocin, a type of uterotonic, is the reference treatment] in the management of the third stage of labour reduces the amount of bleeding and the need for blood transfusion . . . "[57]

(The authors reference a Cochrane systematic review, showing that uterotonic agents reduced bleeding and blood transfusions compared with placebo.)

Elaboration. Any differences between the control intervention in the trial and the reference treatment in the previous trial(s) in which efficacy was established should be reported and explained. For example, differences may exist because background treatment and patient management change with time and concomitant therapies may differ.[27] Dose changes may occur: if the dose of the reference treatment is reduced, it might result in reduced efficacy; if it is increased, possibly leading to tolerability problems, the new treatment's advantages could be overestimated.

Objectives: Item 5. Specific objectives and hypotheses, *including the hypothesis concerning noninferiority or equivalence.*

Example. "[A] bodyweight-adjusted single bolus of 0.50-0.55 mg/kg tenecteplase would be equivalent to a 90 min regimen of alteplase for efficacy and safety [the primary endpoint for efficacy was all-cause 30-day mortality from acute myocardial infarction]. In this double-blind, randomised, controlled study, we formally tested this hypothesis."[13]

Elaboration. The authors should specify for which outcomes noninferiority or equivalence hypotheses apply and for which superiority hypotheses apply. Usually the noninferiority or equivalence hypothesis refers to the primary end point, whereas the new treatment is expected to offer other advantages, eg, fewer adverse effects, cost.

Outcomes: Item 6. Clearly defined primary and secondary outcome measures, *detailing whether the outcomes in the noninferiority or equivalence trial are identical (or very similar) to those in any trial(s) that established efficacy of the reference treatment* and, when applicable, any methods used to enhance the quality of measurements (eg, multiple observations, training of assessors).

Example. "Over the past decade seven large, randomised, placebo-controlled trials involving a total of 16,770 patients who underwent percutaneous interventions have established that the overall reduction in the risk of death or nonfatal myocardial infarction 30 days after adjunctive inhibition of platelet glycoprotein IIb/IIIa receptors is 38 percent. Three glycoprotein IIb/IIIa inhibitors were assessed in these trials. The primary end point [in the present trial] was a composite of death, nonfatal myocardial infarction, or urgent target-vessel revascularization within 30 days after the index procedure."[58]

Elaboration. Any differences in outcome measures in the new trial compared with trial(s) that established efficacy of the reference treatment should be noted and justified. In particular, note any changes in timing of evaluation. Ideally, outcomes should remain unchanged, but often insights do lead to change as the understanding, management, and prognosis of a disease improve. For example, early AIDS trials used death outcomes, then deaths became uncommon, so they shifted to AIDS clinical events, then clinical events became uncommon, so they shifted to surrogate markers.

Sample Size: Item 7a. How sample size was determined, *detailing whether it was calculated using a noninferiority or equivalence criterion, and specifying the margin of equivalence with the rationale for its choice.*

Examples. "Considering previous studies, a primary event rate of 3.1% per

year was estimated for patients in both treatment groups. To obtain 90% statistical power with a 1-sided α equal to 0.025, approximately 1600 patient-years of exposure per treatment groups are necessary to establish the noninferiority of ximelagatran compared with dose-adjusted warfarin within 2% per year. . . . Assuming an average follow up of 16 months, approximately 2400 patients are required."[59]

"Sample size was based on . . . [an] 8.0% primary quadruple end point event rate in the control (heparin plus Gp IIb/IIIa blockade) group [reference treatment] and a 12.5% relative reduction in the bivalirudin arm. Using a 2-sided α level of .05 and 3000 patients per group, the trial had a 99% power to detect superiority over the imputed heparin control [historical control] and a 92% power to satisfy noninferiority criteria relative to heparin plus Gp IIb/IIIa."[46]

Elaboration. The margin of noninferiority or equivalence should be specified, and preferably justified on clinical grounds. Its relation to the effect of the reference treatment relative to placebo in any previous trials should be noted (see second example).

Sample size calculations are usually based on the assumption that the point estimate of the difference between treatments will be 0 (as in the first example above). Examples F and G in the Figure would have met the noninferiority criterion had the observed point estimates been 0. That is, the precision of the estimates would have been adequate, had the 2 treatments been equally effective. With a large enough sample, it is possible to demonstrate noninferiority even when the point estimate is between 0 and Δ. If the true effect is assumed to be greater than 0, the sample size will need to be increased, perhaps substantially.

Stopping Rules: Item 7b. When applicable, explanation of any interim analyses and stopping rules (*and whether related to a noninferiority or equivalence hypothesis*).

Example. "Interim safety analyses were planned when 40 and 70 percent of the total number of women had been enrolled. An increased rate of HIV transmission associated with the shorter regimens, as compared with the long-long regimen, would be considered significant if any of the nominal P values for the differences were less than 0.007 in the first interim analysis and less than 0.012 in the second. . . . "[38]

Elaboration. It is customary to base interim stopping criteria on *P* values, and these adjusted *P* values are analogous to widened CIs.

Statistical Methods: Item 12. Statistical methods used to compare groups for primary outcome(s), *specifying whether a 1- or 2-sided confidence interval approach was used.* Methods for additional analyses, such as subgroup analyses and adjusted analyses.

Examples. Binary outcome. "The proportion of the intention-to-treat population experiencing primary events per year for both treatment groups, and the associated 1-sided 97.5% CI for the difference, will be estimated using the time to first event . . . The noninferiority margin (Δ) defined in the primary analysis is based on absolute event rate differences . . . Noninferiority of ximelagatran over warfarin will be accepted [in a 0.025 level test] if the upper bound of the 97.5% CI around the estimated difference in primary event rates lies below Δ. For these studies, an absolute Δ of 2% was adopted. . . . "[59]

Continuous outcome. "Regimens were regarded as equivalent if the difference between treatments in change in FEV1 (using 95% CI) was less than 4% of predicted FEV1 . . . Since we were undertaking an equivalence study, the primary analysis was per protocol but an intention-to treat analysis was also undertaken. The mean difference between treatments and 95% CI for the true difference was obtained from analysis of variance, with adjustment for centre and type of clinic. . . . "[15]

Elaboration. The upper bound of the 1-sided $(1-\alpha) \times 100\%$ CI (or correspondingly, the upper bound of the 2-sided $(1-\alpha/2) \times 100\%$ CI) for the treatment effect has to be below the margin Δ to declare that noninferiority has been shown, with a significance level α.

Both Δ and α should be prespecified in the noninferiority hypothesis.

Results

Numbers Analyzed: Item 16. Number of participants (denominator) in each group included in each analysis and whether "*intention-to-treat" and/or alternative analyses were conducted.* State the results in absolute numbers when feasible (eg, 10/20, not just 50%).

Example. "Efficacy variables were analyzed on an intent-to-treat basis . . . and on an as-treated basis. In the intent-to-treat analysis, patients were considered treatment failures if they made any treatment changes, prematurely discontinued randomized treatment for any reason, or had missing data for 2 consecutive evaluations. In the as-treated analysis, only data from patients continuing randomized treatment were considered for analysis."[60]

Outcomes and Estimation: Item 17. For each primary and secondary outcome, a summary of results for each group and the estimated effect size and its precision (eg, 95% confidence interval). *For the outcome[s] for which noninferiority or equivalence is hypothesized, a figure showing confidence intervals and margins of equivalence may be useful.*

Examples. Inferiority of new treatment, figure legend. "Relative risk of blood loss of 1000 mL or more with misoprostol compared with oxytocin [1.39, 95% CI 1.19 to 1.63]. Vertical dotted lines represent margins of clinical equivalence determined a priori [0.74 and 1.35 on the relative scale]. Solid line represents null effect."[57]

(A figure similar to case G in the Figure was presented on the relative scale.)

Noninferiority of new treatment. "The primary quadruple composite end point of death, myocardial infarction, urgent repeat revascularization, or in-hospital major bleeding by 30 days occurred in 299 (10.0%) of 2991 patients in the heparin plus Gp IIb/IIIa inhibitor group vs 275 (9.2%) of 2975 patients in the bivalirudin group (OR, 0.92; 95% CI, 0.77-1.09; P=0.32). Relative to heparin alone, the imputed

OR was 0.62 (95% CI, 0.47-0.82), satisfying statistical criteria for noninferiority to heparin plus Gp IIb/IIIa blockade and superiority to heparin alone."[46]

(A figure similar to case B in the Figure was presented on the relative scale but without the margin of noninferiority.)

Elaboration. In the first example the new treatment was inferior, but it was uncertain whether the treatment effect was smaller or larger than the margin of equivalence 1.35. The second example demonstrated noninferiority.

Comment

Interpretation: Item 20. Interpretation of the results, taking into account the *noninferiority or equivalence hypothesis and any other* trial hypotheses, sources of potential bias or imprecision and the dangers associated with multiplicity of analyses and outcomes.

Examples. Concluding noninferiority. "According to our definition of equivalence, the efficacy of the . . . long-short regimen (was) statistically equivalent to the efficacy of the long-long regimen . . . The upper limit of the 95 percent confidence interval for the difference between the rates in the two groups was 5.3 percent (close to the boundary of 6.0 percent)."[38]

Concluding inferiority of new drug (or conventional superiority of reference drug). "Although the trial was intended to assess the noninferiority of tirobifan as compared with abciximab, the findings demonstrated that tirobifan offered less protection from major ischemic events than did abciximab. . . . In order to meet the present definition of equivalence, the upper bound of the 95% confidence interval of the hazard ratio for the comparison of tirofiban with abciximab had to be less than 1.47. . . . The primary endpoint occurred more frequently among the 2398 patients in the tirofiban group than among the 2411 patients in the abciximab group (7.6 percent vs. 6.0 percent; hazard ratio, 1.26; . . . two-sided 95 percent

confidence interval of 1.01 to 1.57, demonstrating the superiority of abciximab over tirofiban; P=0.038)."[58]

Concluding noninferiority of new drug from a trial designed to assess superiority. "The SYNERGY protocol prespecified that if enoxaparin was not demonstrated to be superior to unfractionated heparin, a noninferiority analysis was to be performed. . . . Enaxoparin was not superior to unfractionated heparin but was noninferior for the treatment of high-risk patients with non-ST-segment elevation ACS."[44]

Comment

It is not our intent to promote noninferiority or equivalence trials: the design should be appropriate to the question to be answered.[21] Available efficacious reference treatments can make use of placebo controls unethical.[61] But even in cases for which a treatment is efficacious on some measures, eg, depression scales, it may not be for a rarer but more important outcome, eg, suicide.[62] Reports of noninferiority and equivalence trials must be clear enough to allow readers to interpret results reliably. Accordingly, we herein propose extensions to the CONSORT statement to facilitate appropriate reporting of noninferiority and equivalence trials.

We advocate that editors extend support of the original CONSORT statement to include use of this extension to noninferiority and equivalence trials and refer to it in their "Instructions to Authors." Adoption by journals of the original CONSORT statement is associated with improved quality,[63-65] so we hope this proposed extension will result in similar improvements for noninferiority and equivalence trials.

The CONSORT Group continues to update and extend its recommendations. The current recommendations add to recent extensions to cluster randomized trials,[66] and the reporting of harms.[4] Further extensions are in preparation. The current versions of all CONSORT recommendations are available at http://www.consort-statement.org.

REFERENCES

1. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA*. 1996;276:637-639.
2. Moher D, Schulz KF, Altman DG; for the CONSORT Group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet*. 2001; 357:1191-1194.
3. Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med*. 2001;134:663-694.
4. Ioannidis JPA, Evans SJW, Gøtzsche PC, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Ann Intern Med*. 2004;141:781-788.
5. Schulz KF. Randomized trials, human nature, and reporting guidelines. *Lancet*. 1996;348:596-598.
6. Jüni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. *BMJ*. 2001;323:42-46.
7. Egger M, Jüni P, Bartlett C, et al. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? empirical study. *Health Technol Assess*. 2003;7:1-76.
8. Wellek S. *Testing Statistical Hypotheses of Equivalence*. Boca Raton, Fla: Chapman Hall/CRC; 2003.
9. Committee for Proprietary Medicinal Products. *Note for Guidelines on Evaluation of Medicinal Products Indicated for Treatment of Bacterial Infections*. London, England: European Medicines Agency (EMEA); April 22, 2004. Available at: http://www.emea.eu.int/pdfs/human/ewp/055895en.pdf. Accessed July 2004.
10. Durkalski VL, Palesch YY, Pineau BC, et al. The virtual colonoscopy study: a large multicenter clinical trial designed to compare two diagnostic screening procedures. *Control Clin Trials*. 2002;23:570-583.
11. Clinical Outcomes of Surgical Therapy Study Group. A comparison of laparoscopically assisted and

open colectomy for colon cancer. *N Engl J Med*. 2004; 350:2050-2059.

12. Chadwick D; Vigabatrin European Monotherapy Study Group. Safety and efficacy of vigabatrin and carbamazepine in newly diagnosed epilepsy: a multicentre randomised double-blind study. *Lancet*. 1999;354:13-19.

13. Assessment of the Safety and Efficacy of a New Thrombolytic (ASSENT-2) Investigators. Single-bolus tenecteplase compared with front-loaded alteplase in acute myocardial infarction: the ASSENT-2 double-blind randomised trial. *Lancet*. 1999;354:716-722.

14. von Hertzen H, Piaggio G, Ding J, et al. Low dose mifepristone and two regimens of levonorgestrel for emergency contraception: a WHO multicentre randomised trial. *Lancet*. 2002;360:1803-1810.

15. Smyth A, Tan KH-V, Hyman-Taylor P, et al. Once versus three times daily regimens of tobramycin treatment for pulmonary exacerbations of cystic fibrosis– the TOPIC study: a randomised controlled trial. *Lancet*. 2005;365:573-578.

16. Pocock SJ. The pros and cons of non-inferiority (equivalence) trials. In: Guess HA, Kleinman A, Kusek JW, Engel LW, eds. *The Science of Placebo: Towards an Interdisciplinary Research Agenda.* London, England: BMJ Books; 2000:236-248.

17. Villar J, Ba'aqeel H, Piaggio G, et al. WHO antenatal care randomised trial for the evaluation of a new model of routine antenatal care. *Lancet*. 2001;357: 1551-1564.

18. Kapur A, Malik IS, Bagger JP, et al. The Coronary Artery Revascularisation in Diabetes (CARDia) trial: background, aims, and design. *Am Heart J*. 2005;149: 13-19.

19. 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.

20. ICH Steering Committee. *Harmonised Tripartite Guideline: Choice of Control Group and Related Issues in Clinical Trials (E10)*. Geneva, Switzerland: International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use; 2000. Available at: http://www.ich .org/LOB/media/MEDIA486.pdf. Accessed February 9, 2006.

21. Djulbegovic B, Clarke M. Scientific and ethical issues in equivalence trials. *JAMA*. 2001;285:1206-1208.

22. *Points to Consider on Switching Between Superiority and Non-inferiority*. London, England: European Medicines Agency (EMEA); July 27, 2000. Available at: http://www.emea.eu.int/pdfs/human/ewp /048299en.pdf. Accessed February 9, 2006.

23. Tamayo-Sarver JH, Albert JM, Tamayo-Sarver M, et al. Advanced statistics: how to determine whether your intervention is different, at least as effective as, or equivalent: a basic introduction. *Acad Emerg Med*. 2005;12:536-542.

24. Wiens BL. Choosing an equivalence limit for noninferiority or equivalence studies. *Control Clin Trials*. 2002;23:2-14.

25. Jones B, Jarvis P, Lewis JA, et al. Trials to assess equivalence: the importance of rigorous methods. *BMJ*. 1996;313:36-39.

26. Detsky AS, Sackett DL. When was a negative clinical trial big enough? how many patients you needed depends on what you found. *Arch Intern Med*. 1985; 145:709-712.

27. Rothmann M, Li N, Chen G, et al. Design and analysis of non-inferiority mortality trials in oncology. *Stat Med*. 2003;22:239-264.

28. Lewis JA. The European regulatory experience. *Stat Med*. 2002;21:2931-2938.

29. *Points to Consider on the Choice of Non-*

inferiority Margin. London, England: European Medicines Agency (EMEA); February 26, 2004. Available at: http://www.emea.eu.int/pdfs/human/ewp /215899en.pdf. Accessed February 9, 2006.

30. Senn S. *Statistical Issues in Drug Development.* Chichester: John Wiley & Sons; 2002:207-217.

31. Kim MY, Goldberg JD. The effects of outcome misclassification and measurement error on the design and analysis of therapeutic equivalence trials. *Stat Med*. 2001;20:2065-2078.

32. Dunnett CW, Gent M. Significance testing to establish equivalence between treatments with special reference to data in the form of 2 x 2 tables. *Biometrics*. 1977;33:593-602.

33. Dunnett CW, Gent M. An alternative to the use of two-sided tests in clinical trials. *Stat Med*. 1996;15: 1729-1738.

34. Rothman KJ. Significance questing. *Ann Intern Med*. 1986;105:445-447.

35. Brittain E, Lin D. A comparison of intent-to-treat and per-protocol results in antibiotic non-inferiority trials. *Stat Med*. 2005;24:1-10.

36. ICH Steering Committee. *Statistical Principles for Clinical Trials (E9)*. Geneva, Switzerland: International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use; 1998. Available at: http://www.ich .org/LOB/media/MEDIA485.pdf. Accessed February 9, 2006.

37. McAlister FA, Sackett DL. Active-control equivalence trials and antihypertensive agents. *Am J Med*. 2001;111:553-558.

38. Lallemant M, Jourdain G, Le Coeur S, et al. A trial of shortened zidovudine regimens to prevent mother-to-child transmission of human immunodeficiency virus type 1. *N Engl J Med*. 2000;343:982-991.

39. Prandoni P, Bruchi O, Sabbion P, et al. Prolonged thromboprophylaxis with oral anticoagulants after total hip arthroplasty. *Arch Intern Med*. 2002;162: 1966-1971.

40. Chan FKL, Chung S, Suen BY, et al. Preventing recurrent upper gastrointestinal bleeding in patients with *Helicobacter pylori* infection who are taking low-dose aspirin or naproxen. *N Engl J Med*. 2001;344:967-973.

41. Durrleman S, Simon R. Planning and monitoring of equivalence studies. *Biometrics*. 1990;46:329-336.

42. Sackett DL. Superiority trials, non inferiority trials, and prisoners of the 2-sided null hypothesis. *ACP J Club*. 2004;140:A11.

43. Fleiss JL. General design issues in efficacy, equivalency and superiority trials. *J Periodontal Res*. 1992;27: 306-313.

44. SYNERGY Trial Investigators. 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.

45. Elbourne D, Dezateux C, Arthur R, et al; UK Collaborative Hip Trial Group. Ultrasonography in the diagnosis and management of developmental hip dysplasia (UK Hip Trial): clinical and economic results of a multicentre randomised controled trial. *Lancet*. 2002; 360:2009-2017.

46. Lincoff AM, Bittl JA, Harrington RA, et al. Bivalrudin 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: 853-863.

47. *The Cochrane Library*. Issue 3. Chichester, England: Wiley; 2004.

48. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA*. 1994;272:122-124.

49. Greene WL, Concato J, Feinstein AR. Claims of

equivalence in medical research: are they supported by the evidence? *Ann Intern Med*. 2000;132:715-722.

50. Costa LJ, Xavier ACG, del Giglio A. Negative results in cancer clinical trials—equivalence or poor accrual? *Control Clin Trials*. 2004;25:525-533.

51. Krysan DJ, Kemper AR. Claims of equivalence in randomized controlled trials of the treatment of bacterial meningitis in children. *Pediatr Infect Dis J*. 2002; 21:753-758.

52. Dimick JB, Diener-West M, Lipsett PA. Negative results of randomized clinical trials published in the surgical literature: equivalency or error? *Arch Surg*. 2001; 136:796-800.

53. Piaggio G, Pinol AP. Use of the equivalence approach in reproductive health clinical trials. *Stat Med*. 2001;20:3571-3577.

54. Bernard P, Chosidow O, Vaillant L; French Erysipelas Study Group. Oral pristinamycin versus standard penicillin regimen to treat erysipelas in adults: randomised, non-inferiority, open trial. *BMJ*. 2002;325: 864.

55. Carlier C, Coste J, Etchepare M, et al. A randomised controlled trial to test equivalence between retinyl palmate and carotene for vitamin A deficiency. *BMJ*. 1993;307:1106-1110.

56. Hansson L, Lindholm LH, Ekbom T, et al; the STOP-Hypertension-2 study group. Randomised trial of old and new antihypertensive drugs in elderly patients: cardiovascular mortality and morbidity in the Swedish trial in Old Patients with Hypertension-2 study. *Lancet*. 1999;354:1751-1756.

57. Gülmezoglu AM, Villar J, Ngoc NTN, et al; WHO Collaborative Group to Evaluate Misoprostol in the Management of the Third Stage of Labour. WHO multicentre randomised trial of misoprostol in the management of the third stage of labour. *Lancet*. 2001;358: 689-695.

58. Topol EJ, Moliterno DJ, Herrmann HC, et al. Comparison of two platelet glycoprotein IIb/IIIa inhibitors, tirobifan and abciximab, for the prevention of ischemic events with percutaneous coronary revascularization. *N Engl J Med*. 2001;344:1888-1894.

59. Halperin JL; Executive Committee, Sportif III and V Study Investigators. Ximelataragan compared with warfarin for prevention of thromboembolism in patients with nonvalvular atrial fibrillation: rationale, objectives, and design of a pair of clinical studies and baseline patient characteristics (SPORTIF III and V). *Am Heart J*. 2003;146:431-438.

60. Staszewski S, Keiser P, Montaner J, et al. Abacavir-lamivudine-zidovudine vs indinavir-lamivudine-zidovudine in antiretroviral-naïve HIV-infected adults. *JAMA*. 2001;285:1155-1163.

61. Rothman KJ. Placebo mania. *BMJ*. 1996;313:3-4.

62. Gunnell D, Saperia J, Ashby D. Selective serotonin reuptake inhibitors (SSRIs) and suicide in adults: meta-analysis of drug company data from placebo controlled, randomised controlled trials submitted to the MHRA's safety review. *BMJ*. 2005;330:385-388.

63. Moher D, Jones A, Lepage L; CONSORT Group. Use of the CONSORT statement and quality of reports of randomised trials: a comparative before-and-after evaluation. *JAMA*. 2001;285:1992-1995.

64. Egger M, Jüni P, Bartlett C; CONSORT Group. Value of flow diagrams in reports of randomized controlled trials. *JAMA*. 2001;285:1996-1999.

65. Devereaux PJ, Manns BJ, Ghali WA, et al. The reporting of methodological factors in randomized controlled trials and the association with a journal policy to promote adherence to the Consolidated Standards of Reporting Trials (CONSORT) checklist. *Control Clin Trials*. 2002;23:380-388.

66. Campbell MK, Elbourne DR, Altman DG; the CONSORT Group. CONSORT statement: extension to cluster randomised trials. *BMJ*. 2004;328:702-708.

son of results following microvascular decompression. *J Neurosurg*. 2002;96:527-531.

5. Kucerova H, Dostalova T, Prochazkova J, Bartova J, Himmlova L. Influence of galvanic phenomena on the occurrence of algic symptoms in the mouth. *Gen Dent*. 2002;50:62-65.

6. Cheshire WP Jr. The shocking tooth about trigeminal neuralgia. *N Engl J Med*. 2000;342:2003.

CORRECTIONS

Incorrect Formula: In the Special Communication entitled "Reporting of Noninferiority and Equivalence Randomized Trials: An Extension of the CONSORT Statement" published in the March 8, 2006, issue of *JAMA* (2006;295:1152-1160), a formula for a 2-sided confidence interval (CI) used to assess noninferiority of a new treatment compared with a standard one with regard to an outcome was incorrect. On page 1158 in the subsection titled "Elaboration" that reads "the upper bound of the 2-sided $(1-\alpha/2) \times 100\%$ CI for the treatment effect has to be below the margin Δ to declare that noninferiority has been shown . . . " should read "the upper bound of the 2-sided $(1 -2\alpha) \times 100\%$ CI for the treatment effect has to be below the margin Δ to declare that noninferiority has been shown . . . "

Error in Figure: In the Original Contribution entitled "Long-term Renal Outcomes in Patients With Primary Aldosteronism" published in the June 14, 2006, issue of *JAMA* (2006;295:2638-2645), an error occurred in **FIGURE 3**. In the left panel, which reports creatinine clearance values, the scale of the y-axis is not correct, and the data are therefore not plotted correctly. The correct graph appears below.

Creatinine Clearance
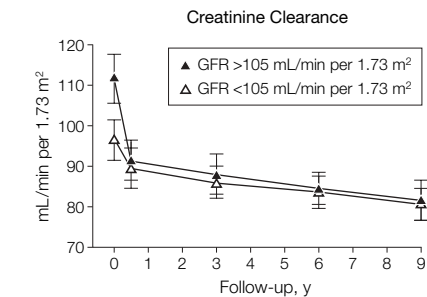
No. of Patients					
GFR <105 mL/min per 1.73 m²	25	25	25	14	8
GFR >105 mL/min per 1.73 m²	25	25	25	14	8