Quality of Reporting of Noninferiority and Equivalence Randomized Trials

Anne Le Henanff, MSc
Bruno Giraudeau, PhD
Gabriel Baron, MSc
Philippe Ravaud, MD, PhD

Context Noninferiority and equivalence trials aim to show that the experimental treatment is not clinically worse than (noninferior) or clinically similar to (equivalent) a control active treatment. These study objectives imply particular planning and analysis.

Objective To assess the methodologic quality of reports of randomized controlled trials of noninferiority and equivalence.

Design We searched MEDLINE and the Cochrane Central Register of Controlled Trials for reports of randomized controlled trials of noninferiority and equivalence hypotheses published between January 1, 2003, and December 31, 2004.

Main Outcome Measures Data extracted by use of a standardized form involved assessment of choice of noninferiority or equivalence margins, sample size calculation, sets of patients analyzed, method of statistical testing and reporting results, and conclusions.

Results A total of 162 reports were included in the analysis (116 reports of noninferiority and 46 of equivalence). The margin defining noninferiority or equivalence was described in most reports (156 [96.3%]), with justification of the margin in only 33 (20.4%). Almost one quarter of the reports (35 [21.6%]) did not describe a sample size calculation, and an additional 11 (6.8%) did not take into account a prespecified noninferiority or equivalence margin. Less than half of the reports (69 [42.6%]) described both an intent-to-treat (ITT; all randomized patients are included in the analysis) or modified ITT (patients who never received treatment are excluded) and per-protocol (patients who did not complete the treatment are excluded) analysis, and only about half of those (39 [56.5%]) described both types of results. Results were displayed with confidence intervals in 136 reports (84.0%). Only 33 articles (20.3%) fulfilled reporting requirements specific to noninferiority and equivalence trials, 4 of them (12.1%) with misleading conclusions.

Conclusions Reporting of noninferiority and equivalence trials has important deficiencies: absence of noninferiority or equivalence margin, only an ITT (or a per-protocol) analysis performed, and results not adequately reported. Moreover, even for articles fulfilling these requirements, conclusions are sometimes misleading.

JAMA. 2006;295:1147-1151 wwww.jama.com

A NONINFERIORITY OR EQUIVALence trial aims to demonstrate that the experimental treatment is not clinically worse than the comparator (an active control treatment) by more than a prespecified small amount (Δ), known as the noninferiority or equivalence margin.[1] According to the International Conference on Harmonization guidelines,[2] the term *noninferiority* is used when referring to a 1-sided trial (difference in response lower than Δ); *equivalence*, when referring to 2-sided trials (difference in response between −Δ and +Δ). The new treatment is expected to have noninferior or equivalent efficacy compared with the standard treatment but could have advantages in safety, convenience (eg, administration once a day instead of 3 times a day), or cost. The new treatment may also present an alternative or second-line therapy.[3]

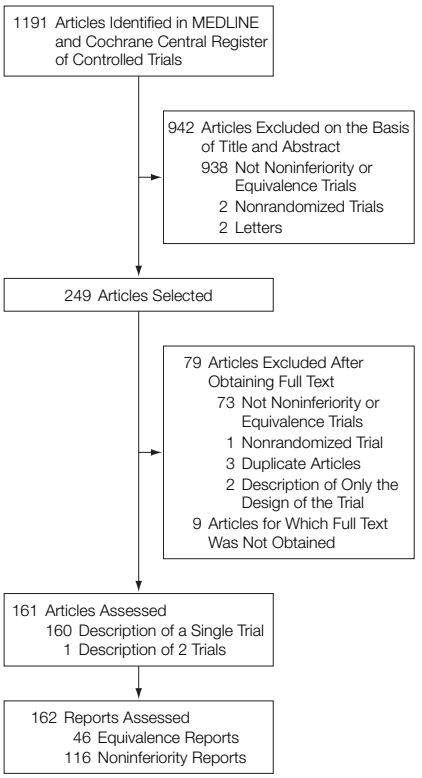
Noninferiority or equivalence trials imply particular planning and analysis.[4-9] The noninferiority margin is taken into account in the formulation of the sample size calculation.[10,11] The margin must be smaller than or equal to "the smallest value that would represent a clinically meaningful difference, or the largest value that would represent a clinically meaningless difference."[12] The determination of this margin must be based on both statistical reasoning and clinical judgment.[13] For analysis, intent-to-treat (ITT) and per-protocol analyses should be performed.[2,5,14] Results are usually displayed as confidence intervals (CIs) around the observed differences in responses, and conclusions are then drawn by comparing these CIs with the prespecified noninferiority margin.[3]

An important review of reports claiming equivalence published between 1992 and 1996,[15] assessing whether these claims were consistent with methods and results, concluded that about half were not aimed at demonstrating equivalence. We sought a new study with a different mode of

See also pp 1152 and 1172.

Author Affiliations: Institut National de la Santé et de la Recherche Médicale (INSERM) U738, Paris, and Département d'Epidémiologie Biostatistique et Recherche Clinique, Groupe Hospitalier Bichat-Claude Bernard (Assistance Publique, Hôpitaux de Paris), Faculté Xavier Bichat (Université Paris 7), Paris, France (Ms Le Henanff, Mr Baron, and Dr Ravaud); INSERM Centre d'Investigation Clinique 202, Tours, INSERM U 717, Paris, Université François-Rabelais, Tours, and Centre Hospitalier Règional Universitaire de Tours (Dr Giraudeau), France.
Corresponding Author: Bruno Giraudeau, PhD, INSERM CIC 202, 10 Bd Tonnellé, 37032 Tours cedex, France (giraudeau@med.univ-tours.fr).

Figure. Study Screening Process



selecting trials (trials aiming to investigate either noninferiority or equivalence instead of trials claiming equivalence) from a sample of recent trials conducted after the publication of guidance from regulatory authorities. Therefore, we systematically reviewed noninferiority and equivalence randomized controlled trials published in 2003 and 2004, focusing on methodologic elements specific to such trials. We aimed to highlight the areas about which misconceptions and problems in the reporting of noninferiority or equivalence trials prevail.

METHODS

Search Strategy

We performed a computerized search of the MEDLINE databases and the Cochrane Central Register of Controlled Trials using the search terms *equivalence OR equivalent OR noninferiority OR noninferior*. We identified all reports of clinical trials published in the English language from January 1, 2003, to December 31, 2004. One of us (A.L.H.) then screened the titles and abstracts to identify potentially relevant studies, and the final selection was made from reading the full text. Articles were included only if the study was identified as a randomized controlled trial assessing noninferiority or equivalence (excluding bioequivalence studies). Letters and articles for which only the abstract was available or reports describing only the design of the trial were excluded. Articles were screened for duplicate publication (ie, the same trial described in several articles), and only the article presenting the main results was selected.

Evaluation of Methodologic Quality

Two independent reviewers (A.L.H. and G.B.) tested a data extraction form with a distinct set of 10 articles during a training session. Then, during a meeting, they discussed the interpretation of the different items. To assess interobserver reliability, a subsample of 30 articles was randomly selected from the sample of included articles, and the 2 reviewers independently extracted information. For the remaining articles, a single reviewer (A.L.H.) extracted information. Reviewers were not blinded to the journal name and authors.

The following data were extracted from all reports:

1. Characteristics of the report, including year of publication, number of patients, number of groups, use of a placebo group (and if so, justification of the use of such a group), interventions studied, whether the choice of noninferiority or equivalence was clearly detailed, and kind of study (noninferiority or equivalence).

2. Choice of noninferiority or equivalence margins. We determined the main outcome measure (continuous, binary, or time to event). We noted whether the margin was defined, whether it was crude or relative, and the justification for the choice of margin.

3. Calculation of sample size. We determined which trials included such a calculation and whether the calculation took into account a noninferiority or an equivalence trial. We described whether the sample size was increased because of the proportion of foreseen dropouts. We noted also whether reports detailed all the elements of the sample size calculation (type I and type II error rates, margin defining noninferiority or equivalence, common standard deviation, difference between the groups).

4. Choice of primary and secondary analysis: per-protocol, ITT, or modified ITT. The ITT analysis includes all randomized patients in the groups to which they were randomly assigned, regardless of their compliance with the entry criteria, the treatment they actually received, and subsequent withdrawal from treatment or deviation from the protocol. The modified ITT analysis excludes patients who never received treatment or who were never evaluated while receiving treatment.[16] The per-protocol analysis includes only patients who satisfied the entry criteria of the trial and who completed the treatment as defined in the protocol.

5. Method of statistical testing. We noted whether a noninferiority or an equivalence analysis and a superiority one were a priori planned and, if so, which was the main analysis. We also determined whether the results were presented with CIs or *P* values. We noted the size of the CI and whether it was 1- or 2-sided. For statistical tests described, we noted whether they took into account the noninferiority or equivalence margin and whether they were 1- or 2-sided.

6. We identified a subsample of reports satisfying the 4 following specific requirements for a noninferiority or equivalence trial: noninferiority or equivalence margin defined, sample size calculation taking into account this margin, both ITT (or modified ITT) and per-protocol analyses, and use of CIs to report results.

7. For the reports that satisfied these requirements, we determined whether the authors' conclusions matched the results, taking into account the prespecified margin.

Statistical Analysis

Categorical variables were described with frequencies and percentages. The degree of agreement between the 2 reviewers was determined with use of the κ coefficient. All data analyses involved use of SAS version 9.1 (SAS Institute Inc, Cary, NC).

RESULTS

Selected Reports

A flowchart of the selection of reports is presented in the FIGURE. The electronic search yielded 1191 citations. From this list, 249 potentially relevant articles were selected after screening titles and abstracts, and, finally, 162 reports were selected after full-text reading. We thus reviewed 116 (71.6%) noninferiority trials and 46 (28.4%) equivalence trials.

Interobserver Reproducibility

The κ coefficients varied between 0.42 and 1.00. The lowest κ was observed for the item that assessed whether the conclusions of the article agreed with the results, but it was associated with a 74% concordance rate.

Description of the Trials

Twenty-one reports (13.0%) were published in general journals. Thirty reports (18.5%) concerned trials with more than 2 groups (8 of them having a placebo group), and the median number of randomized patients was 333 (interquartile range, 201, 656) (TABLE 1). A total of 139 reports (85.8%) compared different treatments (102 [87.9%] of the noninferiority and 37 [80.4%] of the equivalence reports). Twenty-three reports (14.2%) described comparisons of the same pharmacological treatment. The term *noninferiority* or *equivalence* trial was present in the title of 21 reports (13.0%) and in the abstract of 161 (99.4%).

Choice of the Noninferiority or Equivalence Margin

The noninferiority or equivalence margin was specified in 156 reports (96.3%) (TABLE 2). This margin was relative (eg, relative risk) in 22 reports (13.6%).

Table 1. Characteristics of Reports of Randomized Controlled Trials of Noninferiority and Equivalence

Characteristics of the Trial	No. (%)		
	Noninferiority Reports (n = 116)	Equivalence Reports (n = 46)	All Reports (N = 162)
Patients randomized, median (IQR), No.	378 (209-723)	277 (156-524)	333 (201-656)
Study described as double blind	65 (56.0)	29 (63.0)	94 (58.0)
Use of a placebo arm	8 (6.9)	3 (6.5)	11 (6.8)
To show the noninferiority or equivalence between 1 or 2 experimental treatments vs placebo	1 (0.9)	3 (6.5)	4 (2.5)
To show the superiority of 1 or 2 of the treatments compared by a noninferiority or equivalence analysis	7 (6.0)	0	7 (6.3)
Comparison of different treatments	102 (87.9)	37 (80.4)	139 (85.8)
Comparison of same treatments	14 (12.1)	9 (19.6)	23 (14.2)
2 Strategies	9 (7.8)	7 (15.2)	16 (9.9)
2 Doses	2 (1.7)	1 (2.2)	3 (1.9)
2 Durations	3 (9.6)	1 (2.2)	4 (2.5)

Abbreviation: IQR, interquartile range.

Table 2. Methodological Quality of Reports of Randomized Controlled Trials of Noninferiority or Equivalence

Characteristics of the Trial	No. (%)		
	Noninferiority Reports (n = 116)	Equivalence Reports (n = 46)	All Reports (N = 162)
Noninferiority or equivalence margin reported	112 (96.6)	44 (95.7)	156 (96.3)
Justification of the choice of the margin	24 (20.7)	9 (19.6)	33 (20.4)
Statistical considerations	3 (2.6)	2 (4.3)	5 (3.1)
Clinical considerations or results of a previous study	13 (11.2)	5 (10.9)	18 (11.1)
Both statistical considerations and clinical considerations or results of a previous study	8 (6.9)	2 (4.3)	10 (6.2)
Sample size calculated	92 (79.3)	35 (76.1)	127 (78.4)
Sample size taking into account the noninferiority or equivalence margin	81 (69.8)	35 (76.1)	116 (71.6)
Presence of all the elements needed for recalculation	63 (54.3)	24 (52.2)	87 (53.7)
Analysis			
ITT or modified ITT	90 (77.6)	27 (58.7)	117 (72.2)
Per-protocol	74 (63.8)	35 (76.1)	109 (67.3)
Both per-protocol and ITT, or per-protocol and modified ITT	51 (44.0)	18 (39.1)	69 (42.6)
Statistical methods			
Results with a CI	95 (81.9)	41 (89.1)	136 (84.0)
1-Sided	13 (11.2)	0	13 (8.0)
2-Sided	82 (70.7)	41 (89.1)	123 (75.9)
97.5% CI	3 (2.6)		3 (1.9)
95% CI	68 (58.6)	29 (63.0)	97 (59.9)
Results with a *P* value	58 (50.0)	20 (43.5)	78 (48.1)
Statistical test taking into account the noninferiority or equivalence margin	33 (28.4)	6 (13.0)	39 (24.1)
1-Sided test	31 (26.7)	1 (2.2)	32 (19.8)
2-Sided test	2 (1.7)	5 (10.9)	7 (4.3)
Results with either a CI or a *P* value associated with a statistical test taking into account the noninferiority or equivalence margin	106 (91.4)	43 (93.5)	149 (92.0)

Abbreviations: CI, confidence interval; ITT, intent to treat.

Justification for the margin chosen was given in 33 reports (20.4%) and was based on clinical considerations or results of a previous study in 18 reports (11.1%), statistical considerations in 5 (3.1%), and both in 10 (6.2%).

Calculation of Sample Size

Of 127 reports (78.4%) including a sample size calculation, only 116 (91.3%) described a calculation taking into account the noninferiority or equivalence margin (Table 2). Of the 127 reports, 40 were missing some elements necessary to reproduce the calculation. In 49 reports (30.2%; 39 [33.6%] noninferiority and 10 [21.7%] equivalence), the sample size was reported to be increased because of the foreseen proportion of dropouts.

Analysis Sets

We could not determine the population analyzed in 14 reports (8.6%). A per-protocol analysis was described in 109 reports (67.3%) and ITT or modified ITT in 117 (72.2%) (Table 2). Sixty-nine articles (42.6%) reported that both per-protocol and ITT or modified ITT analyses were performed (51 [44.0%] noninferiority and 18 [39.1%] equivalence). The results of both analyses were detailed in 39 (56.5%) of those 69 reports, whereas other reports indicated only whether results were the same.

Statistical Methods

A total of 136 reports (84.0%) displayed results with a CI (95 [81.9%] noninferiority and 41 [89.1%] equivalence) (Table 2). The CI was mostly 2-sided, even for noninferiority trials (82 [70.7%] of the reports). For the 13 noninferiority trials reporting a 1-sided CI, only 3 used a 2.5% type I error rate. Of the 136 reports (84.0%) describing analysis by CI, 101 (74.3%) indicated the type I error used in the sample size calculation. Thirty-four (21.0%) described a CI whose size was not in accordance with the type I error rate used in the sample calculation (27 [23.3%] noninferiority).

In 78 reports (48.1%), results were displayed using *P* values, but only 39 (24.1%) of the statistical tests performed took into account the noninferiority or equivalence margin, meaning that half of these tests were for superiority purposes. In the end, 149 articles (92.0%) described results with either CIs or a statistical test taking into account the noninferiority or equivalence margin.

Exhaustive Fulfilled Requirements

A subsample of 33 reports (20.4%; 24 [20.7%] noninferiority and 9 [19.6%] equivalence) satisfied the 4 requirements specific to noninferiority and equivalence trials we described previously. Moreover, if we also include justification of the margin as an equivalent requirement, this figure decreases to 7 reports (4.3%).

Conclusions of the Study

Of the 33 reports satisfying our quality requirements, conclusions of 4 (12.1%) were highly misleading with their claim of equivalence or noninferiority, although their results were inconclusive.

Table 3. Practical Recommendations for Planning and Reporting the Analysis of Data From Noninferiority or Equivalence Trials

Points to Consider	Recommendations for Planning	Recommendations for Reporting
Noninferiority or equivalence margin	The choice of the margin must be based on statistical and clinical considerations. If available, use practical recommendations from regulatory authorities.	Report the margin and the justification for its choice.
Sample size calculation	Take into account the noninferiority or equivalence margin.	Report all the elements necessary to reproduce the calculation. Report the proportion of dropouts foreseen.
Choice of the sets of patients to analyze	Study both per-protocol and intent-to-treat sets of patients.	Report precisely the sets of patients analyzed and detail the results of both the intent-to-treat and per-protocol analyses.
Statistical method	Study confidence intervals in agreement with the type I error rate used in the sample size calculation.	Report confidence intervals of treatment difference (or treatment ratio) and whether they are 1- or 2-sided.
Conclusion		Conclude noninferiority or equivalence only if both ITT and per-protocol analyses permit that. Restate the prespecified noninferiority or equivalence margin used. Use a standard vocabulary in coherence with the aim of the trial (ie, treatment A is "noninferior to" or "equivalent to" treatment B).

Abbreviation: ITT, intent-to-treat.

COMMENT

We assessed the methodologic quality of reports of randomized controlled trials of noninferiority and equivalence. Our results highlight important deficiencies in the reporting of these trials, which could be improved on in the planning stages. TABLE 3 points out what should be planned for and reported from noninferiority and equivalence trials. These recommendations focus on the methodologic elements specific to noninferiority and equivalence trials (ie, the definition of the noninferiority or equivalence margin, the sets of patients to analyze, and the statistical methods to use), provided the usual requirements for all randomized trials are fulfilled (eg, randomization, concealment, blinding).

Prespecifying the noninferiority or equivalence margin is necessary and has to be clinically justified.[12] This prespecification is all the more important because some reported margins were so large that they were clearly unconvincing (eg, in one report, the 95% CI of an

odds ratio was required to be included in the interval from –0.33 to +3.0) and because an unjustified a priori margin will allow external reviewers to a posteriori consider it clinically questionable.

For noninferiority or equivalence trials, ITT analysis may lead to biased conclusions because of protocol violators and withdrawals. Also, dropouts and nonadherent participants from the 2 groups are potentially different, which may also bias a per-protocol analysis. Thus, both analyses are required and considered to have equal importance in drawing a conclusion.[5,14] Reporting the results of only 1 of the analyses may reflect either ignorance about noninferiority and equivalence trials or a deliberate intention to mask some of the results, potentially modifying the interpretation of the results and preventing readers from drawing definitive conclusions.

The adequacy of a noninferiority or equivalence claim could be judged only in the subsample of 33 well-reported trials in our study. However, even in this selective subsample, we identified 4 (12.1%) reports with misleading conclusions, which misidentified noninferiority or equivalence. These reports are in addition to those confusing noninferiority and equivalence, which is acknowledged to be of lesser importance. In addition, the wording used in conclusions often lacks precision in describing noninferiority and equivalence. It is surely preferable to use standard vocabulary such as treatment A "is not inferior to" (or "is equivalent to") treatment B "with regard to the margin prespecified at Δ" rather than stating a treatment is "not substantially lower than" or "similar to" a control treatment.

Our study of noninferiority and equivalence trial reports has some limitations. First, the sample of studied reports may not be exhaustive: some trials planned with a noninferiority or equivalence aim may have been missed even though we used standard key words. Second, because we studied reports, we were unable to detect trials planned as superiority trials and secondarily reported as noninferiority or equivalence trials after failure to demonstrate superiority. We suspect that some of the articles correspond to this scenario. Discrepancies may exist between the reports and the actual protocol used,[17-19] and some deficiencies may be due only to poor reporting (eg, the absence of elements needed for sample size calculation).

In conclusion, our findings highlight the need for improved planning, analysis, and reporting of randomized controlled trials of noninferiority and equivalence. To improve the planning and reporting of such trials, we present practical recommendations to help researchers enhance the methodologic quality and the reporting of noninferiority and equivalence trials (Table 3). We moreover stress the importance of drawing a conclusion by comparing the obtained results with the prespecified margin, which is mainly a guideline, and to use a standard vocabulary, thus avoiding potentially misleading conclusions.

Author Contributions: Dr Giraudeau had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Study concept and design: Le Henanff, Giraudeau, Baron, Ravaud.
Acquisition of data: Le Henanff, Baron.
Analysis and interpretation of data: Le Henanff, Giraudeau, Baron, Ravaud.
Drafting of the manuscript: Le Henanff, Giraudeau.
Critical revision of the manuscript for important intellectual content: Giraudeau, Baron, Ravaud.
Statistical analysis: Le Henanff, Giraudeau, Baron, Ravaud.
Administrative, technical, or material support: Ravaud.
Study supervision: Giraudeau, Baron, Ravaud.
Financial Disclosures: None reported.

REFERENCES

1. European Medicines Agency. Guideline on the choice of the noninferiority margin. European Medicines Agency Web site. Available at: http://www.emea.eu.int/pdfs/human/ewp/215899en.pdf. Accessed October 24, 2005.
2. ICH topic E9: statistical principles for clinical trials. US Food and Drug Administration. Available at: http://www.fda.gov/cber/gdlns/ICHclinical.pdf. Accessed October 24, 2005.
3. Jones B, Jarvis P, Lewis JA, et al. Trials to assess equivalence: the importance of rigorous methods. *BMJ*. 1996;313:36-39.
4. Durrleman S, Simon R. Planning and monitoring of equivalence studies. *Biometrics*. 1990;46:329-336.
5. Garrett AD. Therapeutic equivalence: fallacies and falsification. *Stat Med*. 2003;22:741-762.
6. Piaggio G, Pinol APY. Use of the equivalence approach in reproductive health clinical trials. *Stat Med*. 2001;20:3571-3587.
7. D'Agostino RB Sr, Massaro J, Sullivan L. Noninferiority trials: design concepts and issues: the encounters of academic consultants in statistics. *Stat Med*. 2003;22:169-186.
8. Ebbutt AF, Frith L. Practical issues in equivalence trials. *Stat Med*. 1998;17:1691-1701.
9. Röhmel J. Therapeutic equivalence investigations: statistical considerations. *Stat Med*. 1998;17:1703-1714.
10. Blackwelder WC. "Proving the null hypothesis" in clinical trials. *Control Clin Trials*. 1982;3:345-353.
11. Blackwelder WC, Chang MA. Sample size graphs for "proving the null hypothesis." *Control Clin Trials*. 1984;5:97-105.
12. Wiens BL. Choosing an equivalence limit for noninferiority or equivalence studies. *Control Clin Trials*. 2002;23:2-14.
13. International Conference on Harmonization. *Guidance on Choice of Control Group and Related Design and Conduct Issues in Clinical Trials (ICH E 10)*. Bethesda, Md: US Food and Drug Administration; 2000.
14. Brittain E, Lin D. A comparison of intent-to-treat and per-protocol results in antibiotic noninferiority trials. *Stat Med*. 2005;24:1-10.
15. 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.
16. Hill CL, LaValley MP, Felson DT. Secular changes in the quality of published randomized clinical trials in rheumatology. *Arthritis Rheum*. 2002;46:779-784.
17. Hill CL, La Valley MP, Felson DT. Discrepancy between published report and actual conduct of randomized clinical trials. *J Clin Epidemiol*. 2002;55:783-786.
18. Chan AW, Hrobjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA*. 2004;291:2457-2465.
19. Pildal J, Chan AW, Hrobjartsson A, Forfang E, Altman DG, Gøtzsche PC. Comparison of descriptions of allocation concealment in trial protocols and the published reports: cohort study. *BMJ*. 2005;330:1049-1052.