REVIEW ARTICLE

Establishing Equivalence or Non-Inferiority in Clinical Trials

Part 20 of a Series on Evaluation of Scientific Publications

Stefan Wellek, Maria Blettner

SUMMARY

Background: An increasing number of clinical trials are being performed to show the absence of relevant differences between the effects of two treatments. The primary care physician makes use of the results of so-called equivalence studies, at least indirectly, practically every day. Equally important are active control clinical trials in which the efficacy of a new treatment has to be proven through demonstrating non-inferiority as compared to a standard treatment.

Methods: Explanation of basic principles and statistical techniques with reference to the original literature; selective searches in the medical literature.

Results: First of all, a suitable distributional parameter must be chosen that can be considered a reasonable measure of dissimilarity of the population effects of the treatments under comparison. The simplest approach to the statistical demonstration of equivalence or non-inferiority is to calculate confidence intervals for that parameter. To keep the required number of subjects for equivalence and non-inferiority studies as low as possible, statistical tests should be used which are optimized with respect to power.

Conclusion: Data from equivalence and non-inferiority studies need to be assessed for statistical significance no less than data that are generated to show that two treatments have different effects. A negative result in a traditional two-sided test does not suffice for statistically proving equivalence.

▶ **Cite this as:**
Wellek S, Blettner M: Establishing equivalence or non-inferiority in clinical trials—part 20 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2012; 109(41): 674–9. DOI: 10.3238/arztebl.2012.0674

Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI) at the University Medical Center of the Johannes Gutenberg University Mainz: Prof. Dr. rer. nat. Wellek, Prof. Dr. rer. nat. Blettner

In a classical randomized controlled trial (RCT) the investigator evaluates the differences between two treatments (or between a treatment and a placebo) (1), with the aim of demonstrating that some new treatment option is superior to the existing standard. In dealing with diseases for which adequate treatment options are already available, it is is often the case that a new drug is developed that costs less than the medication currently being used for the respective indication, or has fewer adverse effects. In this case the aim is to establish the hypothesis that the efficacy of the new drug, compared to existing reference medications, is essentially similar (equivalence) or only marginally lower (non-inferiority). An example of the latter type of study is the CATT trial (Lucentis versus Avastin [2]), which received wide attention in the lay press (3) owing to the high prevalence of the disease concerned (age-related macular degeneration) and the exorbitant costs of the drug demonstrated to be non-inferior (at least € 1 billion annually if used in all eligible patients in Germany alone).

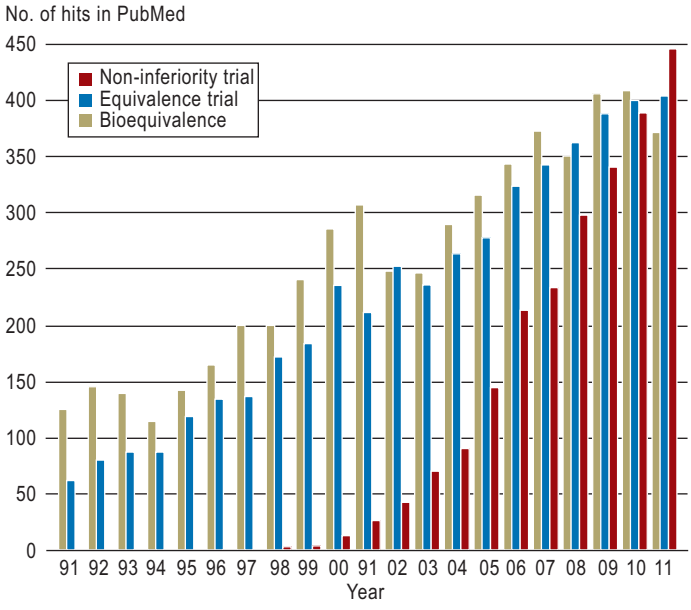
By definition, an equivalence study is conducted to demonstrate that there are no relevant differences in efficacy between two (or more) treatments. Before such studies can be planned and analyzed, the notion of equivalence has to be made precise, i.e., the investigators have to decide what amount of difference between the treatments can be tolerated as clinically irrelevant. The clinically relevant differences must be specified in the study protocol. To this end, a distributional parameter that accounts for these differences is selected. This may be, for example, the difference or the ratio of the expected values of the outcome variable. Furthermore, upper and lower limits are determined for the acceptable deviation from the value this parameter would take on in the case of identical efficacy of the treatments under comparison. The values of these "equivalence margins" are conventionally denoted by the symbols $-\varepsilon_1$ and ε_2, where ε_1 and ε_2 are positive numbers. The clinical research question, the selected clinical endpoint, and the form of the distributions to be compared have to be taken into account when setting ε_1 and ε_2. If, for example, a study is being carried out to demonstrate the equivalence of two antihypertensive agents in reducing diastolic blood pressure after 4

weeks' treatment and the difference $\mu_1-\mu_2$ in mean reduction of the diastolic value attained in the populations is selected as the target parameter, $\varepsilon_1 = \varepsilon_2 = 5$ mm Hg is a sensible choice of equivalence margins.

In a non-inferiority trial, the aim is to show that the new treatment is not relevantly worse than the reference treatment. What constitutes relevant inferiority is defined by a lower limit $-\varepsilon$ (e.g., –5.0 mm Hg for mean reduction in blood pressure) below which the parameter chosen to measure the difference in efficacy between the treatments should not fall.

The importance of equivalence and non-inferiority studies for clinical research has increased steadily over the past 20 years, as can be seen from the number of PubMed hits for the search terms "bioequivalence," "(non)inferiority study (trial)," and "equivalence study (trial)" over the years 1991–2011 *(Figure 1)*. Another indicator for this development is the proportion of drugs approved for the market on the basis of equivalence trials. According to an extrapolation from data in drug reports published by the US Food and Drug Administration (FDA) (4, section 1.4), this proportion was as high as 78% in 2008 *(Figure 1)*.

FIGURE 1



Frequency of equivalence trials: results of a literature search

Inadmissibility of the "naive" approach to testing for equivalence

Testing for equivalence requires statistical procedures other than those used in the classical situation where the aim is to demonstrate superiority. A conventional two-sided test (5) the wrong choice. Actually, it is inadmissible to conclude that the alternative hypothesis of equivalence of the treatments has been proved when such a test yields a negative, i.e. non-significant, result. The type I error in this situation consists of declaring the treatment effects to be similar despite the existence of relevant differences. In the conventional test, the risk of a type I error may amount up to 95%. In other words, a non-significant difference should not be mistaken for significant agreement of treatment effects. A less precise but frequently quoted way of expressing the same fact is: "Absence of evidence is not evidence of absence" (6).

The principle of confidence interval inclusion

Statistically correct confirmatory evaluation of equivalence studies can be done on the basis of confidence intervals. The basic idea behind this approach is remarkably simple and was first proposed in the context of bioequivalence testing (7):

From the data under analysis, one calculates a lower confidence limit C_l and an upper confidence limit C_u for the chosen parameter and compares them with the predefined theoretical equivalence margins $-\varepsilon_1$ and ε_2. If the confidence interval with the limits (C_l, C_u) turns out to be completely included in the theoretical range, one decides in favor of the hypothesis of equivalence. This is the case whenever both the value of C_l is larger than $-\varepsilon_1$ and that of C_u does not exceed ε_2. Otherwise, the null hypothesis of non-equivalence has to accepted. When applying this rule *(Box 1a)*, one must be aware of the following fact: To ensure that the test of equivalence maintains a significance level of 5%, it is not sufficient for the confidence interval to have a two-sided confidence level of 90% (8). Rather, each of the two confidence limits C_l and C_u must have a one-sided confidence level of 95%.

When testing not for equivalence but only for non-inferiority, one needs only the lower confidence limit. Relying on the interval inclusion principle, the testing procedure is as follows: Non-inferiority is declared statistically confirmed if C_l exceeds the lower equivalence margin specified under the hypothesis *(Box 1b, Figure 2)*.

Optimal tests for equivalence and non-inferiority

Tests that are based on the interval inclusion principle control the type I error risk, but are not optimal with regard to power (10) and therefore require larger samples than would ideally be the case.

The statistical literature contains a number of optimal tests for equivalence and non-inferiority hypotheses that cover a large range of situations differing with respect to study design and the nature of the outcome variable (4). The practical implementation of such optimal tests is considerably more complicated than for conventional one- or two-sided tests of significance and requires special algorithms. However, the software is not difficult to use.

A scenario very frequently arising in clinical trials is comparison of two binomial distributions. The optimal procedure for testing for non-inferiority in that setting

BOX 1a

Interval inclusion test for equivalence of two normal distributions with regard to the difference in mean values

Study: Comparison of the efficacy of a new antidepressant (A) and imipramine (B) as reference treatment for major depression
Outcome variable: percentage reduction in Hamilton Depression Scale (HAM-D) score after 6 weeks' treatment.
Distributional assumption: The outcome variable is approximately normally distributed for both treatments, with mean values μ_1 (←group A) and μ_2 (←group B) and unknown common variance σ^2.
Evaluation: Test for equivalence of the means of these distributions, with the maximum tolerated deviation between μ_1 and μ_2 both to the left ($\leftarrow\varepsilon_1$) and the right ($\leftarrow\varepsilon_2$) set at 5.0[%].
The significance level is set at $\alpha = 0.05$, as usual.
Results of the study expressed as sample means and standard deviations:

Group A ($n_1 = 25$): $\bar{X} = 58.9$, $S_X = 5.82$
Group B ($n_2 = 50$): $\bar{Y} = 57.5$, $S_Y = 4.94$

Confidence limits for $\mu_1 - \mu_2$ for a one-sided confidence level of 95%:
From the empirical mean values and standard deviations, the lower and upper confidence limits are calculated by means of the central t distribution using well-known formulas from elementary statistics (9) to be:

$C_l = -1.35$, $C_u = 4.15$

Test decision:
According to the interval inclusion rule, it has to be checked whether both $C_l > -5.0$ and $C_u < 5.0$ are true.
Conclusion: Since the point –1.35 lies to the right of –5.0 and 4.15 to the left of +5.0 on the numerical axis, the null hypothesis of significant differences can be rejected.
Therefore: Decision in favor of equivalence

Alternative formulation of decision rule:
Given the above values for the two standard deviations, the equivalence test leads to a positive decision, provided the two arithmetic means do not deviate from one another by more than 2.25 [%] *(Figure 1)*.
Therefore: The differences in the samples must turn out to be still smaller than the margins specified under the hypothesis.

BOX 1b

What changes when one tests for non-inferiority rather than equivalence (same situation as in *Box 1a*)?

Hypothesis: The working (alternative) hypothesis is now that the true value of μ_1 lies above $\mu_2 - \varepsilon$
(μ_1 [μ_2] = mean percentage reduction in HAM-D under antidepressant A [B] in the population)
Equivalence margin: In the case of non-inferiority, solely the left margin $-\varepsilon$ of the region of clinically irrelevant deviations between μ_1 and μ_2 is of interest.
Diverging from the specifications in *Box 1a*, it is now assumed that the margin ε was set at 2.5 [%] in the study protocol.
Test decision: The decision whether non-inferiority can be regarded as statistically confirmed or not depends exclusively on the lower confidence limit:
The value –1.35 obtained in *Box 1a* lies above the theoretical non-inferiority limit of –2.5.
Therefore: Decision in favor of non-inferiority.

***Note*:** The example shows that the same data have to be evaluated differently when tested for non-inferiority on the one hand and for equivalence on the other. With tolerance reduced to 2.5 the test performed in *Box 1a* would have a negative result, because the right confidence limit lies above +2.5.

is presented in *Box 2,* together with an illustrating example.

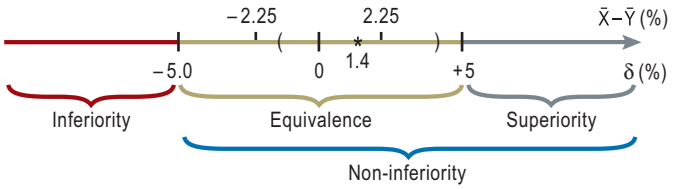
Criteria for assessing publications on equivalence studies

Some basic criteria for the assessment of publications on equivalence and non-inferiority studies are listed in *Box 3*. The *Table* presents the findings of a review of these criteria in articles published in the five leading general medical journals between 2000 and 2011. We found that the error of inferring statistically confirmed equivalence from a non-significant difference is no longer a serious issue in these high-ranking journals. The picture was much less favorable for studies involving tests for two-sided equivalence, where confidence interval inclusion, rather than optimal tests, were used throughout. Moreover, the two-sided confidence level was set at 95%, which means that these confidence interval inclusion tests were carried out in an unnecessarily conservative version *(Box 3, Table)*.

Discussion

Nowadays tests for the confirmatory statistical analysis of equivalence and non-inferiority studies are part of the standard repertoire of medical biometry. In terms of the frequency with which the respective type of study is performed, still the most important field of application for these procedures is the assessment of the bioequivalence of different formulations of the same drug. The methodological peculiarities of bioequivalence studies, which form the basis for market approval of generic drugs, cannot be described in detail in this brief review. (Comprehensive expositions can be found in Chap. 10

FIGURE 2



Visualization of the procedure described in *Box 1a*: values above (below) the numerical axis relate to the treatment difference in the population (in the samples); * = observed mean difference

BOX 2

Test for non-inferiority with regard to the odds ratio in two-armed studies with dichotomous categorization of response

Basic setting, distributional assumption: Parallel group design with binary data (response yes or no); the parameters for statistical analysis are the proportions p_1 (↔ treatment A) and p_2 (↔B) of responders in the underlying populations.

Non-inferiority hypothesis: The true value of the odds ratio $OR = (p_1/(1-p_1))/(p_2/(1-p_2))$ lies above $1-\varepsilon$, with ε as the tolerance specified in the study protocol (e.g., $\varepsilon = 1/3$ or $\varepsilon = 1/2$).

Testing procedure: The test uses as p value (8) the probability $P_{s;\varepsilon}$ that in a situation with the same sample sizes and the same total number s of treatment successes as in the present study, and $1-\varepsilon$ as the true value of the odds ratio, one would obtain at least as many responders in group A as were actually observed.

Example: In the 2010 Lancet study (11) comparing raltegravir (experimental treatment) with lopinavir and ritonavir (positive control) for the treatment of HIV patients with stable viral suppression under previous combination therapy, the following response rates were observed:

Medication	Response +	Response −	Σ
A (raltegravir)	293 (84.4%)	54 (15.6%)	347 (100.0%)
B (lopinavir + ritonavir)	319 (90.6%)	33 (9.4%)	352 (100.0%)
Σ	612	87	699

Setting the non-inferiority margin at 0.5 and using SAS™ software (for details see [4, section 6.6.1]), the p value $P_{s;\varepsilon}$ for this contingency table comes out as 35.04%, far above the usual significance level of 5%. Thus, non-inferiority of raltegravir to the combination therapy with regard to the odds ratio cannot be confirmed on the basis of these data.

BOX 3

Criteria for the evaluation of published trials

(Q1) Testing merely for "absence of evidence," or for equivalence/non-inferiority?

(Q2) Equivalence margin(s) specified *a priori* (without knowledge of the data)?

(Q3) Conclusive justification of the specification of equivalence margin(s)?

(Q4) Optimal test for two-sided equivalence or confidence interval inclusion rule?

(Q5) In the case that interest is in establishing equivalence and the interval inclusion rule is applied: two-sided confidence level 90%, or set unnecessarily conservatively at 95%?

of [4] and in [12–15]). The testing procedure recommended in the guidelines of the regulatory authorities for bioequivalence studies (see [16]), is the one presented in *Box 1a*, for establishing equivalence of two normal distributions with regard to the difference between the non-standardized mean values. This test has to be performed with the (logarithmically transformed) ratios of the measurements from both periods of a crossover trial (17).

Advanced-phase clinical trials are also increasingly being carried out with the aim of demonstrating equivalence or non-inferiority. The majority of these studies are RCTs (1) using an active (positive) control, which means that instead of placebo the participants in the control group receive an established, effective treatment. A major difference from bioequivalence trials is that the endpoint criterion is the response of patients with a real indication for the treatment, not a pharmacokinetic parameter measured in healthy probands. With regard to statistical analysis, the main fact distinguishing studies to demonstrate therapeutic equivalence from bioequivalence trials is that the test of equivalence is often carried out with variables whose distribution is not of the continuous type (and thus not normal) or with data which is subject to censoring. In active control studies, one very often has to perform comparisons between response rates (i.e., binomial proportions) or Kaplan–Meier survival curves. Suitable tests for equivalence and non-inferiority for all these situations can be found in the literature. Currently, most active control trials are planned and evaluated on the basis of non-inferiority tests (18). From the point of view of statistical theory, there are no compelling reasons for this preference. Rather, it is motivated by the fact that given the same lower margin of equivalence and the same desired power, considerably higher sample sizes are needed to establish equivalence in the strict sense, than are required to demonstrate non-inferiority. This difference is easy to explain: a positive result of the test permits a much more precise conclusion in a trial demonstrating equivalence than in one establishing non-inferiority.

Generally, in evaluating a clinical trial from a statistical point of view, it is important to take into account whether it is a study of the classical type or one conducted with the aim of establishing equivalence or non-inferiority. Different types of trials require different procedures for statistical analysis. Tests of equivalence and non-inferiority have been developed to a high level and are widely known, but are not always applied properly with regard to interpreting the results or checking the assumptions made at the outset. The minimum requirements for publications reporting the results of equivalence or non-inferiority trials were published some years ago in an addendum to the CONSORT Statement (18).

TABLE

Distribution of the pros and cons according to the criteria in *Box 3*[*1]

	Q1	Q2	Q3	Q4	Q5
+	180	176	46	0	2
–	10	4	131	23	21
na[*2]	0	10	13	167	167

[*1] Publications in NJEM, Lancet, JAMA, Ann Intern Med, and BMJ in the years 2000–2011 found by a PubMed search using the terms "equivalence" and "non(-)inferiority"

[*2] Not applicable

Conflict of interest statement
Prof. Blettner has received payments for acting as a consultant from Astellas and AstraZeneca. Prof. Wellek declares that no conflict of interest exists.

Manuscript received on 12 January 2012, revised version accepted on 4 July 2012.

Translated from the original German by David Roseveare.

REFERENCES

1. Kabisch M, Ruckes C, Seibert-Grafe M, Blettner M: Randomized controlled trials: part 17 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2011; 108(39): 663–8.
2. The CATT Research Group: Ranibizumab and Bevacizumab for Neovascular Age-Related Macular Degeneration. NEJM 2011; 364: 1897–908.

KEY MESSAGES

- In an equivalence study, it is not admissible to apply a conventional two-sided test and conclude equivalence from a negative result.
- The first step in correct confirmatory analysis of an equivalence or non-inferiority trial is determination of a distributional parameter that can be considered a suitable measure of the difference between the effects of the treatments in the population.
- The simplest approach to the statistical demonstration of equivalence or non-inferiority is then based on the calculation of confidence limits for this parameter.
- The advantage of methods based on confidence limits lies mainly in their simplicity. However, this comes at the price of unnecessarily low power of the tests.
- In the interest of minimizing patient or proband numbers, statistical tests that are optimized with respect to power should also be used in equivalence trials.

3. Kuhrt N: Gleiche Wirkung. Bei Altersblindheit helfen zwei Medikamente. Weiter verbreitet ist das teure. Warum? ZEIT ONLINE 2011. www.zeit.de/2011/20/Pharmaindustrie-Medikamente
4. Wellek S: Testing statistical hypotheses of equivalence and noninferiority. 2nd edition. Boca Raton: Chapman & Hall/CRC 2010.
5. du Prel J, Röhrig B, Hommel G, Blettner M: Choosing statistical tests: part 12 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2010; 107(19): 343–8.
6. Altman DG, Bland JM: Absence of evidence is not evidence of absence. BMJ 1995; 311: 485.
7. Westlake WJ: Use of confidence intervals in analysis of comparative bioavailability trials. J Pharma Sci 1972; 61: 1340–1.
8. du Prel JB, Hommel G, Röhrig B, Blettner M: Confidence interval or p-value? Part 4 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2009; 106(19): 335–9.
9. Hilgers R-D, Bauer P, Schreiber V, Heitmann KU: Einführung in die Medizinische Statistik. 2nd edition. Berlin: Springer-Verlag 2007.
10. Röhrig B, du Prel JB, Wachtlin D, Kwiecien R, Blettner M: Sample size calculation in clinical trials: part 13 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2010; 107(31–32): 552–6.
11. Eron JJ, Young B, Cooper DA, et al.: SWITCHMRK 1 and 2 investigators: Switch to a raltegravir-based regimen versus continuation of a lopinavir-ritonavir-based regimen in stable HIV-infected patients with suppressed viraemia (SWITCHMRK 1 and 2): two multicentre, double-blind, randomised controlled trials. Lancet 2010 30; 375: 396–407. Epub 2010 Jan 12. PubMed PMID: 20074791.
12. Vollmar J (Ed.): Bioäquivalenz sofort freisetzender Arzneiformen. Stuttgart: Gustav Fischer Verlag 1991.
13. Chow SC, Liu JP: Design and Analysis of Bioavailability and Bioequivalence Studies, 3rdEdition. Boca Raton: Chapman & Hall/CRC 2008.
14. Patterson S, Jones B: Bioequivalence and Statistics in Clinical Pharmacology. Boca Raton: Chapman & Hall/CRC Press 2005.
15. Hauschke D, Steinijans VW, Pigeot I: Bioequivalence Studies in Drug Development: Methods and Applications. Chichester: John Wiley & Sons 2007.
16. Food and Drug Administration (FDA): Guidance for industry: Statistical approaches to establishing bioequivalence. Rockville, MD: Center for Drug Evaluation and Research (CDER) 2001.
17. Wellek S, Blettner M: On the proper use of the crossover design in clinical trials: part 18 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2012; 109(15): 276–81.
18. Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ, for the CONSORT Group: Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. JAMA 2006; 295: 1152–60.

Corresponding author
Prof. Dr. rer. nat. Maria Blettner
Institut für Medizinische Biometrie
Epidemiologie u. Informatik der
Johannes Gutenberg-Universität
Obere Zahlbacher Str. 69
55131 Mainz, Germany
blettner-sekretariat@imbei.uni-mainz.de