Efficacy of experimental treatments compared with standard treatments in non-inferiority trials: a meta-analysis of randomized controlled trials

Darius Soonawala,1,2* Rutger A Middelburg,1 Matthias Egger,3 Jan P Vandenbroucke1 and Olaf M Dekkers1,4

1Department of Clinical Epidemiology, Leiden University Medical Centre, RC Leiden, The Netherlands, 2Department of Infectious Diseases, Leiden University Medical Centre, RC Leiden, The Netherlands, 3Institute of Social and Preventive Medicine, University of Bern, Berne, Switzerland and 4Department of Endocrinology and Metabolic Diseases, Leiden University Medical Centre, RC Leiden, The Netherlands

*Corresponding author. Department of Infectious Diseases C5-P, Leiden University Medical Centre, Albinusdreef 2, 2300 RC Leiden, The Netherlands. E-mail: d.soonawala@lumc.nl

Accepted 6 July 2010

Background There is concern that non-inferiority trials might be deliberately designed to conceal that a new treatment is less effective than a standard treatment. In order to test this hypothesis we performed a meta-analysis of non-inferiority trials to assess the average effect of experimental treatments compared with standard treatments.

Methods One hundred and seventy non-inferiority treatment trials published in 121 core clinical journals were included. The trials were identified through a search of PubMed (1991 to 20 February 2009). Combined relative risk (RR) from meta-analysis comparing experimental with standard treatments was the main outcome measure.

Results The 170 trials contributed a total of 175 independent comparisons of experimental with standard treatments. The combined RR for all 175 comparisons was 0.994 [95% confidence interval (CI) 0.978–1.010] using a random-effects model and 1.002 (95% CI 0.996–1.008) using a fixed-effects model. Of the 175 comparisons, experimental treatment was considered to be non-inferior in 130 (74%). The combined RR for these 130 comparisons was 0.995 (95% CI 0.983–1.006) and the point estimate favoured the experimental treatment in 58% (n = 76) and standard treatment in 42% (n = 54). The median non-inferiority margin (RR) pre-specified by trialists was 1.31 [inter-quartile range (IQR) 1.18–1.59].

Conclusion In this meta-analysis of non-inferiority trials the average RR comparing experimental with standard treatments was close to 1. The experimental treatments that gain a verdict of non-inferiority in published trials do not appear to be systematically less effective than the standard treatments. Importantly, publication bias and bias in the design and reporting of the studies cannot be ruled out and may have skewed the study results in favour of the
Introduction

Non-inferiority trials are increasingly published in the medical literature, increasingly used in drug licensing and have at the same time come under increased scrutiny and criticism, up to the allegation that they are unethical.\(^1\)\(^-\)\(^13\) A verdict of ‘non-inferiority’ leaves readers with the impression that a new experimental treatment is as good as an established standard treatment and that the two can be used interchangeably. However, in such trials, non-inferiority is statistically accepted whenever an experimental treatment is unlikely to be worse than an established treatment by more than a pre-specified amount, the so-called non-inferiority margin. If a relatively wide margin is chosen, new treatments that are actually less beneficial might wrongly be considered as equally effective. This may lead to acceptance and use of new therapies that are actually less effective in a clinically relevant way.\(^10\),\(^12\)

There is concern that non-inferiority trials might be deliberately designed to conceal that a new treatment is somewhat less effective than a standard treatment.\(^10\),\(^12\) Systematic use of too-large non-inferiority margins or systematic biases of design, conduct or reporting of non-inferiority trials may skew results in favour of new treatments.\(^14\)\(^-\)\(^19\) In this meta-analysis we examined one type of systematic bias. If trialists systematically compare slightly less effective new treatments with standard treatments, the combined results from a meta-analysis of many trials in which experimental treatments gain a verdict of non-inferiority, would be expected to favour the standard treatment. In order to test this hypothesis, we performed a meta-analysis of non-inferiority trials published in clinical journals and assessed the average effect of experimental treatments compared with standard treatments. Importantly, the combined estimate from the meta-analysis will not be influenced by the choice of the non-inferiority margins.

Methods

Eligibility and search strategy

We searched for non-inferiority trials on 20 February 2009 using PubMed (National Library of Medicine) with the text words ‘noninferiority’ or ‘non inferiority’ or ‘equivalence’ combined with the text words ‘clinical trial’ or ‘trial’ or ‘trials’ or ‘study’ or ‘studies’, limiting the search to publications from 1991 onwards. The initial search was restricted to six general medicine journals (Annals of Internal Medicine, BMJ, JAMA, Lancet, New England Journal of Medicine and PLoS Medicine). In a second step the search was extended to include the other 115 journals included in PubMed’s selection of core clinical journals (see http://www.nlm.nih.gov/bsd/aim.html for a list of these journals). Although equivalence trials (trials that specify both a lower and an upper equivalence margin)\(^20\),\(^21\) were not eligible for inclusion, we included the term ‘equivalence’ in our search strategy in order not to miss non-inferiority trials that had been described as equivalence trials.

Study selection

All two-arm parallel group non-inferiority trials of an experimental treatment compared with standard therapy were included, independent of the intervention examined in the trial. Articles that were published electronically ahead of print were also considered.

Data extraction

The following information was extracted independently by two investigators (D.S. and R.M.): year of publication, journal, subject area (cardiovascular medicine, infectious diseases, obstetrics and gynaecology, rheumatology, surgery or other), primary endpoint, non-inferiority margin for primary endpoint, expected incidence of the primary endpoint in the standard arm and the point estimate for the comparison of experimental with standard treatment. The primary endpoint was classified into three categories: (i) mortality alone or as part of a combined endpoint; (ii) clinical disease; and (iii) surrogate endpoint (imaging or laboratory test). Disagreements were resolved in consultation with a third investigator (O.D.). If trials presented more than one primary endpoint, the endpoint for which the sample size had been calculated was used. If it was unclear for which endpoint sample size calculations were done, or if no such calculations were reported, one of the primary endpoints was randomly selected. In trials that included several non-inferiority comparisons using the same standard treatment, e.g. when testing two dosages of an experimental therapy, one comparison was selected at random and included in the analysis. If a study reported both intention-to-treat and per-protocol analyses, the result used by the author to determine whether the intervention was non-inferior was extracted. If this was not clear, the per-protocol results were used. The funding source was independently classified by two investigators (D.S. and O.D.) as industry, public or mixed funding. The provision of
study drugs by industry was considered as a source of industry funding.

Data synthesis and analysis
The confidence intervals (CIs) and non-inferiority margins reported by the investigators were used to classify the results as superior, non-inferior, inconclusive or inferior according to the definitions given in the extension of the Consolidated Standards of Reporting Trials (CONSORT) statement to non-inferiority trials.\(^5\) Superiority was assumed if the experimental treatment was significantly (\(P < 0.05\)) more efficacious than the standard treatment. Non-inferiority was assumed if the 95% CI did not include the non-inferiority margin. Results were classified as inconclusive if the 95% CI included the non-inferiority margin. Treatments were assumed to be inferior if the entire 95% CI was significantly worse than the non-inferiority margin.

Results of comparisons were expressed as ratio measures, which we call relative risks (RRs) throughout this article. If the trial reported risk ratios or hazard ratios (HRs) or odds ratios (ORs) from statistical models these were used in the analyses. For trials that reported risk differences, we calculated the risk ratio. For studies reporting continuous endpoints (e.g. blood pressure), results were converted to ORs using the method described by Hasselblad and Hedges,\(^22\) and the OR was then used in further calculations. This method assumes logistic distributions with equal variances in the two treatment groups. Under this assumption the natural logarithm of the OR equals a constant multiplied by the standardized difference between means. If needed, the inverse of this ratio was calculated, so that ratios >1 consistently indicated that the experimental treatment performed better than was expected at the design stage of the trial, and ratios <1 indicated that the standard treatment performed worse than was expected. This ratio could not be calculated for studies that did not report separately the expected, or the observed incidence of the primary endpoint, or studies reporting continuous endpoints.

Statistical analyses were done in Comprehensive Meta-Analysis (version 2.0, Biostat, Englewood, NJ, USA) and Stata (version 10.0, Stata Corporation, College Station, TX, USA).

Results
Literature search and study characteristics
We identified 532 potentially eligible articles and excluded 362 studies for the reasons shown in Figure 1. A total of 170 studies, which were published in 43 different journals, were included (see Appendix Tables 1, 2 and 3 for bibliographic details\(^35\–197\) available as supplementary data at IJE online). Five articles reported the results for two separate comparisons. In total, 175 comparisons were therefore included in the analyses. The oldest non-inferiority study in our selection dates from 1993.\(^23\) Seventy-eight percent of included studies date from 2004 onwards, reflecting an increase in non-inferiority trials in the past 5 years.

The characteristics of the 170 non-inferiority trials are described in Table 1. Of the general medical journals, the New England Journal of Medicine published the largest number of non-inferiority trials; our search found no non-inferiority trials in PLoS Medicine. Most trials were from cardiovascular medicine (\(n = 47\); 28%) or infectious disease (\(n = 43\); 25%). Other fields were obstetrics and gynaecology (7%), oncology (6%), rheumatology (6%), surgery (5%), psychiatry (3%), general medicine (3%), pulmonary medicine (2%), gastroenterology (2%), anaesthesiology (1%), intensive care medicine (1%) and neurology (1%). The majority reported risk differences (\(n = 106\); 61%); 36 studies reported continuous endpoints, which were converted to ORs for the present analysis. Figure 2 shows the type of endpoint, number of participants, point estimate, CI and pre-specified margin
for 33 comparisons that were reported on the RR scale.

For 130 comparisons (74%), we considered the experimental treatment to be non-inferior according to the published criteria. Of note, in 6 of these 130 comparisons the authors deemed the experimental treatment to be clinically inferior based on a secondary endpoint. For 27 comparisons (15%) results were inconclusive, for 15 comparisons (9%) superior and for 3 comparisons (2%) inferior. In 20 instances our assessment differed from the authors’ verdict: in 9 instances we judged the result to be superior where the authors’ verdict was non-inferior, in 6 instances to be inconclusive as opposed to inferior and once to be non-inferior instead of inferior. The authors’ verdict was more favourable to the experimental treatment in four comparisons, each time judging the result to be non-inferior instead of inconclusive.

Meta-analysis
The funnel plot showed a symmetrical distribution of results around RR 1 (Figure 3). Forty-seven percent of comparisons (n = 82) had a point estimate >1 (favouring standard treatment) and 53% (n = 93) <1 (favouring experimental treatment). Of the 130 comparisons judged to be non-inferior, the point estimate favoured experimental treatment in 58% (n = 76) and standard treatment in 42% (n = 54). The combined RR for all 175 comparisons was 0.994 (95% CI 0.978–1.010) using a random-effects model and 1.002 (95% CI 0.996–1.008) using a fixed-effects model. The combined RR for comparisons judged to be non-inferior was 0.995 (95% CI 0.983–1.006). Table 2 shows stratified random-effects meta-analyses according to trial result, measure of effect, type of endpoint, source of funding, by two journal strata and according to whether the judgement of the result was based on an intention-to-treat strata and not. Using a random-effects model, the combined estimate for trials funded by industry was 0.978 (95% CI 0.956–1.000). The combined result for trials funded by public sources was 1.008 (95% CI 0.980–1.038). These two estimates did not differ significantly (P = 0.15 from random-effects meta-regression). All meta-analyses were also performed using a fixed-effects model and are presented in Appendix Table 4 available as supplementary data at IJE online. The main result and the results from the stratified analyses were similar for the random- and fixed-effects meta-analyses except for a difference in the result stratified by funding source.

Non-inferiority margins
The margin was expressed as an RR for 33 comparisons and could be converted from a risk difference to a RR for 91 comparisons. The median pre-specified non-inferiority margin was 1.31 [inter-quartile range (IQR) 1.18–1.59]. Stratified according to trial result, the median margin was 1.42 (range 1.21–4.75) for 3 comparisons judged to be inferior, 1.33 (IQR 1.14–1.51) for 23 comparisons judged to be inconclusive, 1.31 (IQR 1.19–1.59) for 92 comparisons judged to be non-inferior and 1.45 (IQR 1.20–1.75) for 6 comparisons judged to be superior. Stratified by type of endpoint the median margin was 1.34 (IQR 1.19–1.50) for comparisons that had mortality as (part of a combined) endpoint, 1.38 (IQR 1.20–1.70) for comparisons in which clinical disease was the endpoint.
and 1.23 (IQR 1.15–1.35) for comparisons reported as a surrogate endpoint. Stratified according to source of funding the median margin was 1.30 (IQR 1.17–1.50) for 62 trials funded by industry, 1.31 (IQR 1.12–1.63) for 33 trials funded by public sources and 1.35 (IQR 1.23–1.55) for 19 trials with mixed funding.

The ratio of the observed and expected incidence of the primary endpoint in the group that received standard treatment could be calculated for 112 comparisons. Fifty-three percent of comparisons (n = 59) had a ratio <1, indicating that the standard treatment performed better than was expected at the design stage of the trial and 46% (n = 51) >1, indicating that the standard treatment performed worse than was expected. Two ratios were exactly 1. The mean ratio was 0.941 (95% CI 0.859–1.030), meaning that on average the chosen standard treatments performed slightly better than was estimated at the design stage of the trials. Stratified by source of funding, this ratio was 0.974 (95% CI 0.865–1.097) for 55 studies funded by industry and 0.906 (95% CI 0.760–1.080) for 31 studies funded by public sources. These two estimates did not differ significantly (P = 0.5 from t-test for equality of means).

Discussion

In this meta-analysis of trials using a non-inferiority design, experimental treatments were regarded as non-inferior to standard treatments in the majority of studies. The combined RR for these studies comparing experimental with standard treatments was close to 1. For non-inferiority trials published in core clinical journals, this finding contradicts the hypothesis that new treatments that gain a verdict of non-inferiority are systematically less effective than standard treatments.

Our study has several strengths and limitations. We aimed to include all the non-inferiority trials published in these journals, irrespective of the type of endpoints or measures of effect. We restricted the search to the group of core clinical journals, as defined in PubMed, which is the same group of journals as in the Abridged Index Medicus (AIM). This selection covers a wide range of journals from many clinical specialties. Our results may therefore be representative for non-inferiority trials published in other journals. However, external validity may be limited to higher quality journals. If non-AIM journals are of lower quality, the characteristics of non-inferiority trials published in those journals might be different. Furthermore, our search would have missed trials that do not mention the non-inferiority design in the abstract, the title or as a key word. The characteristics of such trials might also differ.

We examined two aspects of non-inferiority trials: first, we combined the results of a large number of non-inferiority trials in a meta-analysis; secondly, we examined the non-inferiority margins chosen by the

<table>
<thead>
<tr>
<th>Study characteristics</th>
<th>n</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>New England Journal of Medicine</td>
<td>33</td>
<td>19</td>
</tr>
<tr>
<td>Lancet</td>
<td>28</td>
<td>16</td>
</tr>
<tr>
<td>Circulation</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>Journal of the American Medical Association</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>British Medical Journal</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Obstetrics and Gynecology</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Paediatrics</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Other (from 36 different journals)</td>
<td>64</td>
<td>38</td>
</tr>
<tr>
<td>Field in medicine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cardiovascular medicine</td>
<td>47</td>
<td>28</td>
</tr>
<tr>
<td>Infectious disease</td>
<td>43</td>
<td>25</td>
</tr>
<tr>
<td>Obstetrics and gynaecology</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>Oncology</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>Rheumatology</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>Surgery</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>Psychiatry</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Other</td>
<td>34</td>
<td>20</td>
</tr>
<tr>
<td>Type of comparison</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drug</td>
<td>133</td>
<td>78</td>
</tr>
<tr>
<td>Procedure</td>
<td>23</td>
<td>14</td>
</tr>
<tr>
<td>Device</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>Source of funding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Industry</td>
<td>94</td>
<td>55</td>
</tr>
<tr>
<td>Public or charity</td>
<td>46</td>
<td>27</td>
</tr>
<tr>
<td>Mixed</td>
<td>19</td>
<td>11</td>
</tr>
<tr>
<td>Not reported</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>Median number of</td>
<td>467</td>
<td></td>
</tr>
<tr>
<td>participants per</td>
<td>(40–20 332)</td>
<td></td>
</tr>
<tr>
<td>comparison (range)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect measure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risk difference</td>
<td>106</td>
<td>61</td>
</tr>
<tr>
<td>Ratio measure</td>
<td>33</td>
<td>19</td>
</tr>
<tr>
<td>HR</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>Risk ratio</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>OR</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Continuous</td>
<td>36</td>
<td>21</td>
</tr>
<tr>
<td>Main analysis used to judge the result</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intention-to-treat analysis</td>
<td>95</td>
<td>54</td>
</tr>
<tr>
<td>Modified intention-to-treat or per protocol analysis</td>
<td>80</td>
<td>46</td>
</tr>
</tbody>
</table>

*Characteristics for 175 comparisons from the 170 trials.
investigators. Importantly, the combined estimate from the meta-analysis will not be influenced by the choice of the non-inferiority margin. The combined estimate will be influenced by the efficacy of standard treatment. If the standard treatment is not effective, the experimental treatment is in fact tested against 'placebo' in a non-inferiority design. Although we did not assess whether the chosen standard treatment represented the best-available comparator, we did assess how standard treatments performed in view of what trialists had expected. On average, the standard treatments performed slightly better than was estimated at the design stage of the trials. Our study did not address other important issues pertaining to non-inferiority trials. For example, we did not assess whether a non-inferiority trial was the appropriate design to use (or whether a superiority design would have been more appropriate) or whether the choice of the non-inferiority margin that was used for the power calculation and statistical testing.

Figure 2: Results for 33 comparisons from 31 trials in which the result was expressed as an RR. Point estimates, CIs and non-inferiority margins (red lines, lighter gray in printed version) are shown.

Figure 3: Funnel plot of the standard error by the log RR for 175 comparisons. Treatment on the X-axis and standard error on the Y-axis. Bias would lead to an asymmetrical appearance of the funnel plot.
made clinical sense. The non-inferiority margin is often criticized as being arbitrary, unacceptably wide or even fraudulent.8,10 The selection of the non-inferiority margin should be based on a combination of statistical reasoning and clinical judgment.9,24 Others have reviewed the rationale for the size of the margins in non-inferiority trials.7,8 They found that the majority of trials did not justify the choice of the margin and that <20% reported a clinical consideration. An in-depth analysis of each trial with subject-matter knowledge on each topic would have been required to judge whether the choice of the margin was adequate. This was beyond the aim of the present analysis.

Does our meta-analysis rebuke some of the criticism aimed at non-inferiority trials? We found that the combined RR for all studies was close to 1. This contradicts the hypothesis that in non-inferiority trials the experimental treatment is generally less effective than the standard treatment. We believe that this is an important, reassuring finding, considering the criticism that has been levelled at non-inferiority trials.1–12 Several issues should nevertheless be considered when interpreting this result. First, current standards for drug approval stipulate that a new treatment should be better than placebo and (at least) non-inferior to the established options. This means that demonstrating non-inferiority can legally suffice for the licensing of a new drug. The underlying assumption is often that a ‘non-inferior’ treatment has added value regarding other properties, such as ease of use, lower costs or fewer adverse effects, which might offset a small loss in efficacy. Sometimes such superior properties, such as costs, are self-evident and do not have to be demonstrated in a trial. Claiming that an agent has less adverse

### Table 2

<table>
<thead>
<tr>
<th>Number of comparisons</th>
<th>RR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall analysis</td>
<td>175</td>
</tr>
<tr>
<td></td>
<td>0.994 (0.978–1.010)</td>
</tr>
<tr>
<td>Stratified analyses</td>
<td></td>
</tr>
<tr>
<td>By result</td>
<td></td>
</tr>
<tr>
<td>Result judged as inferior</td>
<td>3</td>
</tr>
<tr>
<td>Result judged as inconclusive</td>
<td>27</td>
</tr>
<tr>
<td>Result judged as non-inferior</td>
<td>130</td>
</tr>
<tr>
<td>Result judged as superior</td>
<td>15</td>
</tr>
<tr>
<td>By effect measure</td>
<td></td>
</tr>
<tr>
<td>Risk difference</td>
<td>106</td>
</tr>
<tr>
<td></td>
<td>0.996 (0.982–1.010)</td>
</tr>
<tr>
<td>Ratio measure</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>1.012 (0.958–1.069)</td>
</tr>
<tr>
<td>Continuous</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>0.954 (0.826–1.101)</td>
</tr>
<tr>
<td>By type of endpoint</td>
<td></td>
</tr>
<tr>
<td>Mortality as (part of a combined) endpoint</td>
<td>35</td>
</tr>
<tr>
<td>Clinical disease</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>0.998 (0.980–1.017)</td>
</tr>
<tr>
<td>Surrogate endpoint</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>1.000 (0.965–1.037)</td>
</tr>
<tr>
<td>By source of funding</td>
<td></td>
</tr>
<tr>
<td>Industry</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>0.978 (0.956–1.000)</td>
</tr>
<tr>
<td>Public source</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>1.008 (0.980–1.038)</td>
</tr>
<tr>
<td>Mixed</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>1.035 (0.972–1.103)</td>
</tr>
<tr>
<td>Not reported</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>1.018 (0.930–1.113)</td>
</tr>
<tr>
<td>By journal</td>
<td></td>
</tr>
<tr>
<td>N Engl J Med/Lancet/JAMA/BMJ/Ann Intern Med</td>
<td>87</td>
</tr>
<tr>
<td>Other journals</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>0.999 (0.976–1.023)</td>
</tr>
<tr>
<td>By the analysis used to judge the result</td>
<td></td>
</tr>
<tr>
<td>Intention-to-treat analysis</td>
<td>95</td>
</tr>
<tr>
<td>Modified intention-to-treat- or per-protocol analysis</td>
<td>80</td>
</tr>
</tbody>
</table>

All meta-analyses mentioned here were performed using a random-effects model. All meta-analyses were also performed using a fixed-effects model and are shown in Appendix Table 4 available as supplementary data at IJE online. The main result and the results from the stratified analyses were similar for the random- and fixed-effects meta-analyses except for a difference in the result stratified by funding source.
effects should however be based on evidence. A separate analysis of the adverse event data, analyses of combined endpoints or a meta-analysis of several trials might be appropriate and informative to demonstrate superiority in this respect. Of the 175 comparisons in this meta-analysis, we considered the experimental treatment to be non-inferior in 130 (74%) and superior in 15 (9%). Although in 6 of these 145 comparisons the authors deemed the experimental treatment to be clinically inferior based on a secondary endpoint, the majority of published non-inferiority trials can be used to support the registration of a new treatment. The added value and safety of these treatments may not always be self-evident and may not always be demonstrated in the trial. The follow-up time and the sample size of trials are limited, making it improbable that rarer side effects or long-term side effects are detected.

Secondly, for superiority trials, it has repeatedly been described how the outcome may be skewed in favour of the experimental treatment by making convenient choices when designing the study or reporting or publishing the result. This may involve the choice of (the dosage of) the comparator drug, the choice of patients, endpoints or of the type of analysis. It may also involve selective reporting of data or changing the pre-specified endpoint after a study is completed. It is plausible that such mechanisms affect the results of non-inferiority trials. In other words, biased choices in study design and bias due to selective reporting of outcomes may make it more likely that an experimental treatment is considered non-inferior after completion of the trial. We did not have access to the study protocols of the included articles and relied on what was reported as the primary endpoint. Also, we restricted our search to studies that have been published. Unpublished trials are more likely to favour standard treatment. Therefore, if publication bias would be an issue, our results might be skewed in favour of experimental treatments. All these potential sources of bias would remain unnoticed in our meta-analysis. Although the funnel plot showed a symmetrical distribution of results around RR 1, this does not rule out biases. This leaves the possibility that our finding of an overall RR close to unity is skewed in favour of experimental treatment. The finding that studies sponsored by industry were more likely to have results favouring sponsored treatments is in line with other reports. Systematic bias has been suggested as a possible explanation. Our finding could also be due to the play of chance.

Thirdly, the statistical verdict of non-inferiority permits licensing of a drug even if the trial result shows that it is somewhat less effective than standard. Therefore, some treatments that are approved based on non-inferiority testing may be less effective compared with the standard therapy with respect to the primary endpoint. A cascade of non-inferiority trials is possible, in which each next experimental treatment is slightly less effective than the previously established ‘standard’. After several generations of non-inferiority trials, ineffective interventions could be licensed, leading to deteriorating patient care. This outcome has been called ‘bio-creep’. Our results are relevant in this context. Our study showed that of the 130 comparisons judged to be non-inferior, the point estimate favoured the standard treatment in 42% of trials. Biocreep could occur if two or three trials in succession belong to this 42% category and if each next trial adopts the previously demonstrated non-inferior treatment as the new active control treatment. Importantly, our study provided no empirical evidence for or against the existence of biocreep.

In conclusion, the number of non-inferiority trials published in clinical journals has greatly increased. We found that the experimental treatments that gain a verdict of non-inferiority in trials published in core clinical journals are not systematically less effective than the standard treatments. Biases in design, reporting and publication cannot be ruled out and may have skewed the study results in favour of experimental treatments. Continued vigilance is required to assure that non-inferiority trials are used appropriately.

**Supplementary data**

Supplementary data are available at IJE online.

**Acknowledgement**

We acknowledge the contribution of Theo Stijnen, PhD, Professor of Medical Statistics, Department of Medical Statistics and Bioinformatics, Leiden University Medical Centre, for critical discussion of the analyses, and of J.W. Schoones, MA, Walaeus Library, Leiden University Medical Centre, for assistance with the search strategy.

**Conflict of interest:** None declared.

**KEY MESSAGES**

- There is a concern that non-inferiority trials might be deliberately designed to conceal that a new treatment is less effective than a standard treatment. There is little empirical evidence at present to support this notion, however.
The combined RR from 170 randomized trials using a non-inferiority design and published in core clinical journals in recent years was close to 1, favouring neither the experimental nor the standard treatment.

In the majority of trials, the new treatments were considered to be non-inferior. For these trials the combined RR was also close to 1.

The experimental treatments that gain a verdict of non-inferiority do not, therefore, appear to be systematically less effective than the standard treatments.

The evidence from published non-inferiority trials might still be distorted by publication bias, or by a biased choice of standard treatments. Further studies are required to clarify the risk of bias in non-inferiority trials.

References


Willenheimer R, van Veldhuisen DJ, Silke B et al. Effect on survival and hospitalization of initiating treatment for chronic heart failure with bisoprolol followed by enalapril, as compared with the opposite sequence: results of the randomized Cardiac Insufficiency Bisoprolol Study (CIBIS) III. Circulation 2003;112:2426–35.


Efficacy of experimental treatments in non-inferiority trials


Efficacy of experimental treatments in non-inferiority trials


Ringleb PA, Allenberg J, Bruckmann H et al. 30 day results from the SPACE trial of stent-protected angioplasty versus carotid endarterectomy in symptomatic patients: a randomised non-inferiority trial. Lancet 2006;368:1239–47.


Solomon SD, Appelbaum E, Manning WJ et al. Effect of the direct Renin inhibitor aliskiren, the Angiotensin receptor blocker losartan, or both on left ventricular mass in patients with hypertension and left ventricular hypertrophy. Circulation 2009;119:530–37.


Williams RE, Myers E, Derbot J et al. Lumiracoxib 400 mg once daily is comparable to indomethacin 50 mg three times daily for the treatment of acute flares of gout. Rheumatology (Oxford) 2007;46:1126–32.


