BMJ Open

Non-inferiority trials: inferior? A systematic review of selected journals

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2016-012594
Article Type:	Research
Date Submitted by the Author:	11-May-2016
Complete List of Authors:	Rehal, Sunita; MRC Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology; MRC Clinical Trials Unit at UCL, London Hub for Trials Methodology and Research Morris, Tim; MRC Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology; MRC Clinical Trials Unit at UCL, London Hub for Trials Methodology Research Fielding, Katherine; MRC Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology; London School of Hygiene & Tropical Medicine, MRC Tropical Epidemiology Group, Department of Infectious Disease Epidemiology Carpenter, James; MRC Clinical Trials Unit at UCL, London Hub for Trials Methodology Research; London School of Hygiene & Tropical Medicine, Department of Medical Statistics Phillips, Patrick; MRC Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology
Primary Subject Heading:	Medical publishing and peer review
Secondary Subject Heading:	Research methods
Keywords:	non-inferiority, systematic review, randomised controlled clinical trials, clinical trial

Non-inferiority trials: inferior?

A systematic review of selected journals

*Sunita Rehal, statistician[1,2], Tim P. Morris, statistician[1,2], Katherine Fielding, reader in medical statistics and epidemiology[1,3], James R. Carpenter, professor of medical statistics[1, 2,4], Patrick P.J. Phillips, senior statistician[1]

[1]MRC Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology, London, UK

[2]London Hub for Trials Methodology Research, MRC Clinical Trials Unit at UCL, London, UK

[3] MRC Tropical Epidemiology Group, Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, UK

[4] Department of Medical Statistics, London School of Hygiene & Tropical Medicine, LondonWC1E 7HT, UK.

Correspondence to Sunita Rehal:*

MRC Clinical Trials Unit at UCL

Institute of Clinical Trials & Methodology

Aviation House

125 Kingsway

London WC2B 6NH

e-mail: s.rehal@ucl.ac.uk

Telephone number: 02076704702

Key words: non-inferiority, systematic review, clinical trial, randomised controlled clinical trial

Word count:

4081

ABSTRACT

Objective

To assess the adequacy of reporting of non-inferiority trials alongside the consistency and utility of current recommended analyses and guidelines.

Design

Review of randomised clinical trials that used a non-inferiority design published between January 2010 and May 2015 in medical journals that had an impact factor greater than 10 (*JAMA Internal Medicine, Archives Internal Medicine, PLOS medicine, Annals of Internal Medicine, BMJ, JAMA, Lancet and New England Journal of Medicine*).

Data sources

Ovid (MEDLINE).

Methods

We reviewed articles for non-inferiority and assessed the following: choice of non-inferiority margin and justification of margin; power and significance level for sample size; patient population used and how this was defined; any missing data methods used and assumptions declared; and any sensitivity analyses used.

Results

A total of 168 trial publications were included. Most trials declared non-inferiority (132; 79%). The non-inferiority margin was reported for 98% (164) but less than half reported any justification for the margin (77; 46%). While most chose two different analyses (91; 54%) the most common being intention-to-treat or modified intention-to-treat and per-protocol, a large number of articles only chose to conduct and report one analysis (65; 39%), most commonly the intention-to-treat analysis. There was lack of clarity or inconsistency between the type I error rate and corresponding confidence intervals for 73 (43%) articles. Missing data were rarely considered with (99; 59%) not declaring whether imputation techniques were used.

Conclusion

Reporting and conduct of non-inferiority trials is inconsistent and does not follow the recommendations in available statistical guidelines, which are not wholly consistent themselves. Authors should clearly describe the methods used, and provide clear descriptions

of and justifications for their design and primary analysis. Failure to do this risks misleading conclusions being drawn, with consequent effects on clinical practice.

Strengths and limitations of this study

- This research clearly demonstrates the inconsistency in recommendations for non-inferiority trials provided by guidelines for researchers and this is reflected within this review
- Highlights missing data and sensitivity analyses in the context of non-inferiority trials
- Provide recommendations using examples for researchers using the non-inferiority design
- Justification of the choice of the margin was recorded as such if any attempt was made to do so. And so one could argue that inadequate attempts were counted as a 'justification', however there was good agreement between reviewers when independently assessed.
- Only one reviewer extracted information from all articles and therefore assessments may be subjective. However, there was good agreement when a random 5% of papers were independently assessed.

INTRODUCTION

Non-inferiority trials are designed to assess if a new intervention is "acceptably worse"(1) when compared to a standard treatment or care. Non-inferiority and equivalence are sometimes, mistakenly, used interchangeably. Equivalence trials are designed to show that a new intervention performs not much worse and not much better than a standard intervention. Both trial designs are different to superiority trials, which aim to show that a new intervention performs better when compared to a control. Trials that use a non-inferiority design are only appropriate if the intervention has some other benefit, such as less intensive treatment, lower cost or fewer side effects(1).

Poor trial quality can bias trial results towards achieving no difference between treatments(2). This creates more challenges in non-inferiority trials than superiority trials as such bias can produce false positive results for non-inferiority(3-5). The increasing use of this design(6-8) means it is even more important for trialists to understand the issues around quality in the design and analysis of non-inferiority trials.

There are several guidelines available to aid researchers using a non-inferiority design, where various considerations of the design are explained and discussed (table 1). The CONSORT extension statements(1, 9) focus on the reporting of non-inferiority trials, with the most recent 2012 statement being an elaboration of the 2006 statement. The draft FDA 2010(2) document focuses on all aspects and issues relative to non-inferiority trials and gives general guidance. The EMEA 2000 guideline(10) discusses switching between non-inferiority and superiority designs and the EMEA 2006(11) guideline discusses the choice of the non-inferiority margin, taking into account two- and three-arm trials. The ICH E9 and E10 guidelines(12, 13) are general statistical guidance documents addressing issues for all clinical trials and designs. SPIRIT(14) is a guidance document for protocols for all trial designs and includes discussions of recently developed methodology.

Table 1: Summary of guidelines

	Justification of margin	Who is included in analysis	Confidence interval	Missing data	Sensitivity analyses
CONSORT 2006(1)	"Margin should be specified and preferably justified on clinical grounds"	"Non-ITT analyses might be desirable as a protection from ITTs increase in type I error. There is greater confidence in results when the conclusions are consistent." Intent-to-treat: "Analysing all patients within their randomized groups, regardless of whether they completed allocated treatment is recommended" Per-protocol: "Alternative analyses that exclude patients not taking allocated treatment or otherwise not protocol-adherent could bias the trial in either direction. The terms on-treatment or per-protocol analysis are often used but may be inadequately defined."	"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 noninferiority trials. If a 1-sided 5% significance level is deemed acceptable for the noninferiority hypothesis test (a decision open to question), a 90% 2-sided CI could then be used."		
CONSORT 2012(9)		"Should be indicated if conclusions are related to PP analysis, ITT analysis or both and if the conclusions are stable between them."	"The two-sided CI provides additional information, in particular for the situation in which the new treatment is superior to the reference treatment"		"Sensitivity analysis is discussed through an example: Study endpoints were analysed primarily for the per protocol population and repeated, for sensitivity reasons, for the intention-to-treat (ITT) population."
Draft FDA 2010(2)	"Whether M1 (the effect of the active control arm relative to placebo) is based on a single study or multiple studies, the observed (if there were multiple studies) or anticipated (if there is only one study) statistical variation of the treatment effect size should contribute to the ultimate choice of M1, as should any concerns about constancy. The selection of M2 (the largest clinically acceptable difference of the test treatment compared to the active control) is then based on clinical judgment regarding how much of the M1 active comparator treatment effect can be lost. The exercise of clinical judgment for the determination of M2 should be applied after the determination of M1 has been made based on the historical data and subsequent analysis"	"It is therefore important to conduct both ITT and 'as-treated' analyses in non-inferiority studies." Intent-to-treat: "preserve the principle that all patients are analyzed according to the treatment to which they have been randomized even if they do not receive it"	"Typically, the one-sided Type I error is set at 0.025, by asking that the upper bound of the 95% CI for control-treat be less than the NI margin. If multiple studies provide very homogeneous results for one or more important endpoints it may be possible to use the 90% lower bound rather than the 95% lower bound of the CI to determine the active control effect size"	"Poor quality can reduce the drug's effect size and undermine the assumption of the effect size of the control agent, giving the study a 'bias towards the null'."	

ICH E9(12)	"This margin is the largest difference that can be judged as being clinically acceptable"	"In confirmatory trials it is usually appropriate to plan to conduct both an analysis of the full analysis set and a per protocol analysis… In an equivalence or non-inferiority trial use of the full analysis set is generally not conservative and its role should be considered very carefully." Intent-to-treat: "subjects allocated to a treatment group should be followed up, assessed and analysed as members of that group irrespective of their compliance to the planned course of treatment" Full analysis set: "The set of subjects that is as close as possible to the ideal implied by the intention-to-treat principle. It is derived from the set of all randomised subjects by minimal and justified elimination of subjects." Per-protocol: "The set of data generated by the subset of subjects who complied with the protocol sufficiently to ensure that these data would be likely to exhibit the effects of treatment, according to the underlying scientific model. Compliance covers such considerations as exposure to treatment, availability of measurements and absence of major protocol violations."	"For non-inferiority trials a one-sided interval should be used. The choice of type I error should be a consideration separate from the use of a one-sided or two-sided procedure."	"Imputation techniques, ranging from LOCF to the use of complex mathematical models may be used to compensate for missing data"	"An investigation should be made concerning the sensitivity of the results of analysis to the method of handling missing values, especially if the number of missing values is substantial."
ICH E10(13)	"The determination of the margin in a non-inferiority trial is based on both statistical reasoning and clinical judgment"				
SPIRIT(14)		Use an example where "non-inferiority would be claimed if both ITT and PP analysis show conclusions of NI." Intent-to-treat: "In order to preserve the unique benefit of randomisation as a mechanism to avoid election bias, an "as randomised" analysis retains participants in the group to which they were originally allocated. To prevent attrition bias, out-come data obtained from all participants are included in the data analysis, regardless of protocol adherence." Per-protocol and modified intention-to-treat: "Some trialists use other types of data analyses (commonly labelled as "modified intention to treat" or "per protocol") that exclude data from certain participants—such as those who are found to be ineligible after randomisation or who deviate from the intervention or follow-up protocols. This exclusion of data from protocol non-adherers can introduce bias, particularly if the frequency of and the reasons for non-adherence vary between the study		"Multiple imputation can be used to handle missing data although relies on untestable assumptions"	"Sensitivity analyses are highly recommended to assess the robustness of trial results under different methods of handling missing data"

		groups."			
EMEA 2006(11)	"The choice of delta must always be justified on both clinical and statistical grounds"		"A two-sided 95% CI (or one-sided 97.5% CI) is constructed. The interval should lie entirely on the positive side of the margin. Statistical significance is generally assessed using the two-sided 0.05 level of significance (or one-sided 0.025)"		
EMEA 2000(10)		"ITT and PP analyses have equal importance and their use should lead to similar conclusions for robust interpretation"	"A two-sided confidence interval should lie entirely to the right of delta. If one-sided confidence is used then 97.5% should be used"		"It will be necessary to pay particular attention to demonstrating the sensitivity of the trial by showing similar results for the full analysis set and PP analysis set"

There is some inconsistency between these guidelines regarding the conduct of non-inferiority trials (table 1) which may adversely affect the overall quality and reporting of non-inferiority trials. Non-inferiority trials require more care around certain issues, and so clear guidance on how to design and analyse these trials are necessary. Some of these issues that can influence inferences made about non-inferiority are outlined below.

First, the non-inferiority margin – the value that allows for a new treatment to be "acceptably worse"(1) – is used as a reference for conclusions about non-inferiority. It is recommended that this margin is chosen on a clinical basis, meaning the maximum clinically acceptable extent to which a new drug can be less effective than the standard of care and still show evidence of an effect(15). However, it is unclear whether statistical considerations should impact on the choice of an appropriate margin (table 1).

Second, it is important to choose who is included in analyses for non-inferiority trials. The intention-to-treat analysis (includes all randomised patients irrespective of post-randomisation occurrences) is preferred for superiority trials as it is likely to lead to a treatment effect closer to having no effect, and so is conservative(16). For non-inferiority trials, the intention-to-treat (ITT) analysis can bias towards the null, which may lead to false claims of non-inferiority(17). The alternative per-protocol (PP) analysis is often considered instead. But as the PP analysis allows for the exclusion of patients, it fails to preserve a balance of patient numbers between treatment arms (i.e. randomisation) that ITT analysis does and can cause bias in either direction, depending on who the analysis excludes(18). Guidelines often recommend performing both the ITT and PP analyses, although definitions are inconsistent (table 1). Other frequently used classifications such as modified intention-to-treat (mITT), which aims to contain 'justifiable' exclusions (e.g. patients who never had the disease of interest) from the ITT analysis, are also defined inconsistently(19).

Third, while two-sided 95% confidence intervals are widely used for superiority trials, there is some inconsistent advice as whether to calculate 90% or 95% confidence intervals for non-inferiority trials and whether these should be presented as one-sided or two-sided intervals.

Fourth, the handling of missing data is generally discussed for all trials but rarely in the specific context of non-inferiority trials. Methods recommended to handle missing data vary between guidelines (table 1). Methods to handle missing data often contain untestable assumptions and so sensitivity analyses are essential to test the robustness of conclusions under different assumptions(12). However, it is unclear what sensitivity analyses are appropriate for non-inferiority trials.

Given the inconsistency between guidelines, we hypothesised that poor conduct and reporting would be associated with demonstrating non-inferiority. This review investigates the quality of conduct and reporting for non-inferiority trials in a selection of high-impact journals over a five-year period. We also provide recommendations to aid trialists who may consider a non-inferiority design.

METHODS

Medical journals with an impact factor greater than 10 according to the ISI web of knowledge(20) were included in the review (correct at time of search on 31st May 2015), the rationale being that articles published in these journals are likely to have the highest influence on clinical practice and be the most rigorously conducted and reported due to the thorough editorial process. We searched Ovid (Medline) using the search terms "noninferior", "non-inferior", "noninferiority" and "non-inferiority" in titles and abstracts between 1st January 2010 and 31st May 2015 in *New England Journal of Medicine*, *Lancet*, *JAMA*, *British Medical Journal*, *Annals of Internal Medicine*, *PLOS Medicine* and *Archives of Internal Medicine* (descending impact order). From 2013, Archives of Internal Medicine was renamed JAMA Internal Medicine, and therefore both journals have been included in this review. Eligibility of articles was assessed via abstracts by two reviewers (SR and TM). Articles included were non-inferiority randomised controlled clinical trials. Articles were excluded if the primary analysis was not for non-inferiority. Systematic reviews, meta-analyses and commentaries were also excluded. Few trials were designed and analysed using Bayesian methods, and were therefore excluded for consistent comparability in frequentist methods.

Before performing the review, a data extraction form was developed to extract information from articles. Information extracted was with regards to the primary outcome. The form was standardised to collect information on year of publication, non-inferiority margin (and how the margin was justified), randomisation, type of intervention, disease area, sample size, analyses performed (how this was defined and what was classed as primary/secondary), primary outcome, p-values (and whether this was for a superiority hypothesis), significance level of confidence intervals (and whether both bounds were reported), imputation techniques for missing data, sensitivity analyses, conclusions of non-inferiority and whether a test for superiority was pre-specified. See supplement for further details on methods.

A quality grading system was developed based on whether the margin was justified (yes vs. no/poor), how many analyses were performed on the primary outcome (<2 vs. ≥2) and whether the type I error rate was consistent with the significance level of the confidence interval (yes vs. no/unclear). Articles were classed as "excellent" if all these criteria were fulfilled and were classed as "poor" if none were fulfilled. Articles which satisfied one criterion were classed as "fair" and articles that provided two of the three criteria were classed as "good". The results of this grading were compared to inferences on non-inferiority to assess if the quality of reporting was associated with concluding non-inferiority.

Additional published supplementary content was only accessed if it specifically referred to the information we were extracting within articles. As a sub-study, all statistical methods, outcomes and sample sizes from protocols and/or supplementary content were reviewed from New England Journal of Medicine as the journal is known to specifically request and publish protocols and statistical analysis plans alongside accepted publications.

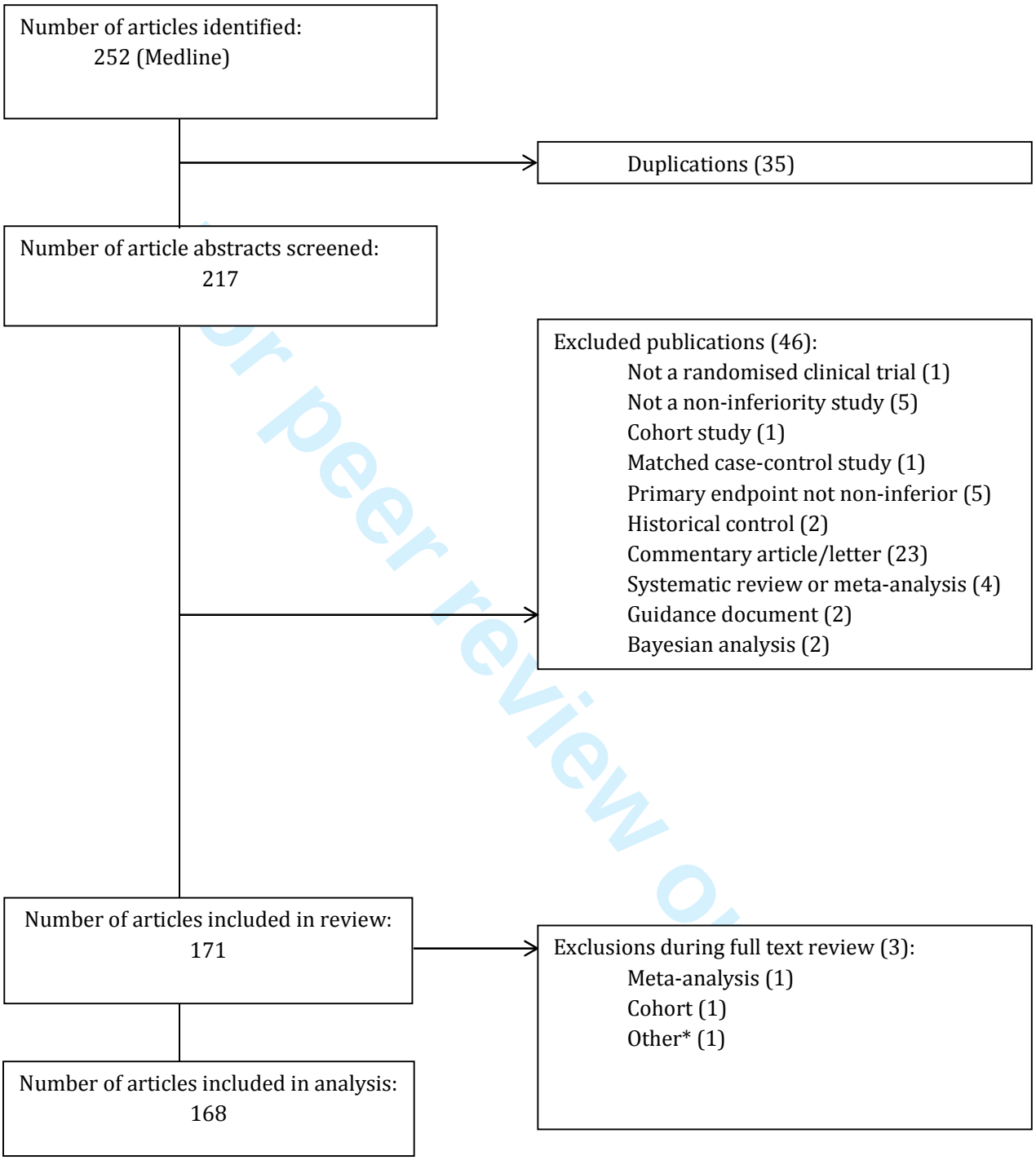
Assessments were carried out by one reviewer (SR), with a random selection of 5% independently reviewed (PP). Any assessments that required a second opinion were independently reviewed (TM). Any discrepancies were resolved by discussion between reviewers.

All analyses were conducted using Stata version 14.

RESULTS

Our search found 252 articles. After duplicate publications were removed, 217 were screened for eligibility using their titles and abstracts. A total of 46 articles were excluded leaving 171 articles to be reviewed. A further three articles were excluded during the full-text review leaving 168 articles (figure 1).

Figure 1: flow chart of eligibility of articles



* Secondary analyses. Primary analyses for the same study was included in the review

General characteristics of the included studies are shown in table 2.

Table 2: General characteristics

	All articles (n=168)	Including NEJM protocols (n=61)
Characteristics	n (%)	n(%)
Journal		
NEJM	61 (36%)	61
Lancet	64 (38%)	
JAMA	19 (11%)	
BMJ	8 (5%)	
Annals of Internal Medicine	5 (2%)	
PLOS Medicine	7 (4%)	
Archives of Internal Medicine	2 (1%)	
JAMA of Internal Medicine	2 (1%)	
Year of publication		
2010	26 (15%)	9 (15%)
2011	27 (16%)	9 (15%)
2012	29 (17%)	8 (13%)
2013	39 (23%)	19 (31%)
2014	27 (16%)	10 (16%)
2015	20 (12%)	6 (10%)
Type of intervention		
Drug	112 (67%)	44 (72%)
Surgery	22 (13%)	7 (11%)
Other	34 (20%)	10 (16%)
Randomisation		
Patient	163 (97%)	59 (97%)
Cluster	5 (3%)	2 (3%)
Power		
80%	6 (36%)	19 (31%)
85%	11 (7%)	5 (8%)
90%	65 (39%)	26 (43%)

71 to 99% (Excluding the above)	21 (12%)	11 (18%)
Not reported/unclear	10 (6%)	0
Composite outcome		
Yes	78 (46%)	37 (61%)
No	90 (54%)	24 (39%)
Disease		
Heart disease	30 (18%)	13 (21%)
Blood disorder	19 (11%)	6 (10%)
HIV	18 (11%)	2 (3%)
Non-infectious disease	18 (11%)	7 (11%)
Cancer	16 (10%)	8 (13%)
Diabetes	11 (7%)	2 (3%)
Infectious disease	8 (5%)	1 (2%)
Thromboembolism	6 (4%)	6 (10%)
Tuberculosis	6 (4%)	4 (7%)
Skin infection	5 (4%)	2 (3%)
Malaria	4 (2%)	1 (2%)
Urinary tract infection	3 (2%)	0
Arthritis	3 (2%)	1 (2%)
Opthomology	3 (2%)	1 (2%)
Pneumonia	3 (2%)	1 (2%)
Complications in pregnancy	3 (2%)	0
Stroke	3 (2%)	2 (3%)
Testing method	3 (2%)	1 (2%)
Appendicitis	2 (1%)	1 (2%)
Depression	2 (1%)	0
Hepatitis C	2 (1%)	2 (3%)

Margin

The non-inferiority margin was specified in 164(98%) articles and was justified in less than half of articles 76(45%). The most common justification was on a clinical basis (29 (17%)) which was often worded ambiguously and with little detail. A total of 14(8%) used previous findings from past trials or statistical reviews to justify the choice of the margin (table 3).

Table 3: Justification of choice of margin, total number of patient populations considered for analyses and patient population included in analysis

	All articles (N=168)	Including NEJM protocols (N=61)
	n (%)	n (%)
Justification of NI margin		
Made no attempt for justification	90 (54%)	22 (36%)
Clinical basis. No evidence for consultation with external expert group, and no reference to previous trials of the control arm	32 (19%)	11 (18%)
Preservation of treatment effect based on estimates of control arm effect from previous trials	13 (8%)	14 (23%)
Expert group external to the authors. No reference to previous trials of the control arm	6 (4%)	3 (5%)
The same margin as was used in other similar trials	5 (3%)	2 (3%)
10-12% recommended by FDA guidelines	4 (2%)	1 (2%)
General comment that margin was decided according to FDA/regulatory guidance.	4 (2%)	0
Clinical basis and based on previous similar trial. No evidence for consultation with external expert group, and no reference to previous trials of the control arm	3 (2%)	0
Based on registry/development program	0	2 (3%)
Other*	11 (7%)	6 (10%)
Number of analyses		
One	65 (39%)	15 (25%)
Two	91 (54%)	38 (62%)
Three	10 (6%)	7 (11%)
Not defined	2 (1%)	1 (2%)
Analysis		
ITT	129 (77%)	44 (72%)
PP	90 (54%)	35 (57%)
mITT	34 (20%)	17 (28%)
As-treated	4 (2%)	6 (10%)
Other	20 (12%)	10 (16%)
Unclear	2 (1%)	2 (3%)

*See supplement

Patients included in analysis

Over a third of articles 65(39%) declared only one analysis (table 3 and 1a of supplement). The majority of trials classed ITT analysis as primary and PP analyses as secondary (figure 1a online supplement). PP analyses were performed in 90(54%) trials of which 11(12%) did not define what was meant by "per-protocol" (table 3 and table 1b online supplement). Definitions of the

PP population contained various exclusions, mostly regarding errors in randomised treatment or treatment received.

Type I error rate

Consistency between the type I error rate and confidence intervals reported was moderate at 95(57%) (table 4). Most articles, 69(41%), used a one-sided 2.5% or (numerically equivalent) two-sided 5% significance level (table 5) and some used a one-sided 5% significance level 46(27%). The majority of articles presented two-sided confidence intervals (147; 88%) and 19(11%) articles presented one-sided confidence intervals. Most two-sided confidence intervals were at the 95% significance level: 125(74%).

Table 4: Consistency of type I error rate with significance levels of confidence intervals over year of publication

	Year of publication						
	2010	**2011**	**2012**	**2013**	**2014**	**2015**	Total
All articles (N=168)							
Yes	11 (42%)	15 (56%)	15 (52%)	24 (62%)	19 (70%)	11 (55%)	95 (57%)
No	5 (19%)	4 (15%)	4 (14%)	5 (13%)	5 (19%)	3 (15%)	26 (15%)
Not reported	10 (38%)	8 (30%)	10 (34%)	10 (26%)	3 (11%)	6 (30%)	47 (28%)
NEJM subgroup (N=61)							
Yes	7 (78%)	6 (67%)	5 (63%)	14 (74%)	8 (80%)	4 (67%)	44 (72%)
No	1 (11%)	2 (22%)	2 (25%)	3 (16%)	2 (20%)	1 (17%)	11 (18%)
Not reported	1 (11%)	1 (11%)	1 (13%)	2 (11%)	0	1 (17%)	6 (10%)

Table 5: Significance level of a) type I error rate and b) confidence intervals for all articles by whether confidence interval was one or two-sided

a) Type I error rate (%)			
	One sided	Two sided	Not reported
0.8	0	1 (1%)	0
1.25	3 (2%)	0	0
2.45	1 (1%)	0	0
2.5	40 (24%)	2 (1%)	2 (1%)
5	46 (27%)	29 (17%)	15 (9%)
10	1 (1%)	2 (1%)	0
Not reported	3 (2%)	0	23 (14%)
b) **Significance level of confidence interval (%)**			
	One sided	Two sided	Not reported
90	1 (1%)	14 (8%)	1 (1%)
95	14 (8%)	125 (74%)	0
97.5	4 (2%)	7 (4%)	0
Other	0	1 (1%)	0
Not reported	0	0	1 (1%)

Missing data and sensitivity analyses

99(59%) trials did not report whether or not any imputation was done and only 12(7%) explicitly declared that no imputation was used. Assuming a worst-case scenario or multiple imputation were the most common methods used (table 6). The number of imputations used for multiple imputation was specified in 8/11 articles and 4/11 stated at least one of the assumptions from Rubin's rules(21). 64(38%) trials reported using sensitivity analyses to test robustness of conclusions of the primary outcome; of these 27/64 (42%) were related to assumptions about the missing data (table 6).

Table 6: Reporting of a) missing data and b) sensitivity analyses

	n (%)
a) Imputation performed	
Yes	**56 (33%)**
Worst case scenario	19 (34%)
Multiple imputation	11 (20%)
Last observation carried forward	8 (14%)
Complete case analysis	6 (11%)
Best case scenario	2 (4%)
Last observation carried forward and worst case scenario	2 (4%)
Best case/worst case scenario	3 (5%)
Mean imputation	1 (2%)
Complete case analysis, multiple imputation using propensity scores and multiple imputation using regression modelling	1 (2%)

Other and worst case scenario	1 (2%)
Other	1 (2%)
No	**12 (7%)**
Not reported	**99 (59%)**
Unclear	**1 (1%)**
Including NEJM protocols (N=61)	
Yes	22 (36%)
No	7 (11%)
Not reported	31 (51%)
Unclear	1 (2%)
b) Sensitivity analyses performed	
Yes	**64 (38%)**
Patient population	13 (20%)
Competing risks	2 (3%)
Statistical modelling	2 (3%)
Adjusted for baseline variables	1 (2%)
Excluded protocol violations	1 (2%)
On-treatment	1 (2%)
Patient population/other	1 (2%)
Unclear	2 (3%)
Other	15 (23%)
Missing data	**27 (42%)**
Best case/worst case scenario	5
Complete case analysis	3
Imputation of missing values	3
Multiple imputation	3
Worst case scenario	3
Baseline observation carried forward	1
Baseline observation carried forward and complete case analysis	1
Complete case analysis, multiple imputation using propensity scores and multiple imputation using regression modelling	1
Complete case analysis and missing not at random	1
Complete case analysis and best case scenario	1
Different methods	1
Last observation carried forward	1
Modelling	1
Observed-failure	1
Worst case scenario and last observation carried forward	1
No	**103 (61%)**
Unclear	**1 (1%)**

Including NEJM protocols	
Yes	38 (62%)
No	23 (38%)

Study conclusions

There were 7(4%) articles that could not make definitive conclusions. For example, if all analyses conducted had to demonstrate non-inferiority to conclude a treatment was non-inferior, and only one of the analyses did, then non-inferiority could not be concluded and could not be rejected. Non-inferiority was declared in 132(79%) articles. 10 of these had made some reference with equivalence studies within the article (See supplement for details),

Superiority analyses were performed in 37(22%) trials after declaring non-inferiority, of which 27/37 (73%) had explicitly pre-planned for superiority analyses. P-values were reported in 98(58%) articles, of which 29/98 (30%) were testing a superiority hypothesis.

Subgroup of trials with published protocols

Additional information from protocols published by NEJM was extracted for 57 of 61 articles. Including this additional information provided by NEJM improved reporting of results across all criteria: 39(64%) articles justified the choice of the non-inferiority margin compared to 19(31%); most planned two or more analyses 45(74%) compared to 37(61%) (there were a couple of cases where two analyses were planned in the protocol but only one was stated in the published article); consistency between type I error rates and confidence intervals was 44(72%) compared with 36(59%); imputation techniques were considered in 29(48%) compared with 17(28%) articles and sensitivity analyses were considered in 38(62%) articles compared with 25(41%). The majority of articles concluded non-inferiority with 8(13%) not determining non-inferiority. Most articles that concluded superiority 14(23%) pre-planned for it 9/14 (64%). Few articles 8/40 (20%) presented superiority p-values.

Association between quality of reporting and conclusions

Overall, there was a suggestive trend between the quality of reporting and concluding non-inferiority: χ_1^2=3.76; p=0.05 (Cochran-Armitage test; table 7). Trials that were poorly reported were less likely to conclude non-inferiority than those that satisfied two or all criteria from justifying the choice of the margin, reporting two or more analyses or reporting a confidence interval consistent with the type I error rate.

Table 7: Quality of reporting of trials associated with conclusions of non-inferiority

Grade	Concluded non-inferiority			
	Yes (N=132)	No (N=29)	Other (N=7)	**Total (N=168)**
	n (%)	n (%)	n (%)	**n (%)**
Excellent[1]	11 (73%)	2 (13%)	2 (13%)	**15**
Good[2]	55 (86%)	9 (14%)	0 (0%)	**64**
Fair[3]	48 (80%)	8 (13%)	4 (7%)	**60**
Poor[4]	18 (62%)	10 (34%)	1 (3%)	**29**

*Excluding trials that concluded 'other': χ_1^2=3.76; p=0.05 (Cochran-Armitage test)

[1] Excellent if margin justified, ≥2 analyses on patient population performed, type I error rate consistent with significance level of confidence interval

[2] Good if fulfilled two of the following: margin justified, ≥2 analyses on patient population performed, type I error rate consistent with significance level of confidence interval

[3] Fair if fulfilled one of the following: margin justified, ≥2 analyses on patient population performed, type I error rate consistent with significance level of confidence interval

[4] Poor if margin not justified, <2 analyses on patient population performed, type I error rate not consistent with significance level of confidence interval

DISCUSSION

Reporting of non-inferiority trials is poor and is perhaps partly due to disagreement between guidelines on vital issues. There are some aspects that guidelines agree on, such as a requirement for the non-inferiority margin to be justified, but we find that this recommendation is neglected by the majority of authors. It is remarkable that several authors performed only one analysis for the primary outcome and the lack of consistency between the significance level chosen in sample size calculations and the confidence interval reported further highlights confusion of non-inferiority trials. Not knowing how to deal with missing data nor appropriate sensitivity analyses, also adds to the confusion. The combination of these recent findings assessed from high impact journals and the inconsistency in guidelines indicate: 1)the non-inferiority design is not well understood by those using the design and 2)optimum methods to compliment the non-inferiority design do not exist.

There was some suggestion that poorly reported trials were less likely to demonstrate non-inferiority. It is therefore essential to ensure that what is reported at the end of a trial was pre-specified before the start of a trial: scientific credibility and regulatory acceptability of a non-inferiority trial rely on the trial being well-designed and conducted according to the design(22). Almost 80% of studies concluded non-inferiority, although it is unclear whether this is due to the reporting in articles or publication bias. It appears that positive results (i.e. alternative hypotheses) are published more often, regardless of trial design, as this number is consistent with other studies that found that more than 70% articles that had positive results are published for superiority trials (23, 24).

More than half of articles reported p-values, of which approximately a third reported p-values for a two-sided test for superiority. P-values, if reported, should be calculated for one-sided tests corresponding to the non-inferiority hypothesis; that is, with H_0: δ = margin. P-values for superiority should not be presented unless following demonstration of non-inferiority, where a pre-planned superiority hypothesis is tested(25).

Comparison with other studies

The value of the non-inferiority margin was almost always reported but more than half of articles made no attempt to explain how the choice was justified. While justification of the margin is low, this is actually an improvement from Schiller et al who reported 23% articles made a justification(26), although this difference could be because only high impact journals were included in this review. There were equally as many articles that planned and reported an ITT analysis compared with articles that performed ITT and PP analyses. This is surprising given that CONSORT 2006 state that an ITT analysis can bias non-inferiority trials towards showing non-inferiority(1). These results were lower than found by Wangge et al(27) who reported 55% used either an ITT or PP and 42% used both ITT and PP. Most articles presented two-sided 95% confidence intervals which is consistent with results from Le Henanff et al(28).

Clinical considerations(1, 2, 9, 11-13) to justify the choice of the margin often had poor justifications, such as “deemed appropriate” or “consensus among a group of clinical experts”.

Non-inferiority is only meaningful if it has strong justification in the clinical context and so should be reported. If the justification includes a measurable reduction in adverse events, these should be measured and the benefit should be demonstrated. Guidelines recommend that the choice of margin should be justified primarily on clinical grounds, however, previous trials and historical data should also be considered if available. As an example, Gallagher et al(29) justify the choice of the margin providing as much information as possible by including references to all published reports and providing data from the institution where the senior author is based. If the choice of the margin is based on a group of clinical experts, authors should provide information on how many experts were involved and how many considered the choice of the margin being acceptable: a consensus among a group of 3 clinicians from one institution is different from a consensus of 20 clinicians representing several institutions.

Definitions provided by authors were inconsistent under what they classed as ITT, PP, mITT and as-treated, for example "all patients randomised who received at least one dose of treatment" was defined at least once in each classification. According to the guidelines, the PP definition excludes patients from the analysis but it is unclear what those exclusions are. The ambiguity of how per-protocol is defined was evident in this review as definitions provided by authors could not be succinctly categorised.

Many articles presented only one analysis, despite most guidelines recommending at least two analyses(1, 2, 9, 10, 12). Unfortunately, guidelines differ in their advice on which of the two analyses should be chosen to base conclusions on. This regrettable, state of affairs was clearly reflected in our review.

Both the ITT and PP analyses have their biases and so neither can be taken as a "gold standard" for non-inferiority trials. The analysis of the primary outcome is the most important result for any clinical trial. It should be pre-defined in the protocol what patients should adhere to and should be considered at the design stage what can be done to maximize adherence. It should be made clear exactly who is included in analyses given the variety of definitions provided by various authors, particularly for per-protocol analyses where definitions are subjective. Such differences in definitions may be may be superficially small but could in fact make critical differences to the results of a trial.

Poor reporting of whether the hypothesis test was one-sided or two-sided or absence of the type I error rate in the sample size calculation meant over a quarter of articles were not clearly consistent with regards to the type I error rate and corresponding confidence interval.

Most guidelines advise presenting two-sided 95% confidence intervals and this is what most articles presented. However, this recommendation may cause some confusion between equivalence and non-inferiority trials. A 5% significance level is maintained using 95% confidence intervals in equivalence trials for two-sided hypotheses whereas non-inferiority takes a one-sided hypothesis and so a 90% confidence interval should be calculated. If a one-sided type I error rate of 2.5% is used in the sample size calculation then this corresponds to the stricter two-sided 95% confidence intervals, not a one-sided 95% confidence interval(30).

The power and type I error rate should be clearly reported within sample size calculations and whether the type I error rate is for a one-sided or two-sided test. For example, the CAP-START trial used a one-sided significance test of 0.05 with two-sided 90% confidence intervals and the authors provide exact details of the sample size calculation in the supplementary appendix(31). If

presenting one bound of the confidence interval throughout an article, this must be done clearly and consistently as described by Schulz-Schupke et al, Lucas et al, Gulmezoglu et al(32-34). Recently, JAMA have introduced a policy to present the lower bound of the confidence interval with the upper bound tending towards infinity(35) and this has been put into practice in recent non-inferiority trials(36-39).

It is unclear whether the potential issues surrounding missing data is well recognised for non-inferiority studies given that the majority of articles did not explicitly state whether or not methods to handle missing outcome data would be considered. Most trials that used multiple imputation stated the number of imputations used but few discussed the assumptions made, which are particularly critical in this context. Some missing data are inevitable, but naïve assumptions and/or analysis threaten trial validity for both ITT and per-protocol analyses(14), particularly in the non-inferiority context where more missing data can bias towards demonstrating non-inferiority(40).

It is recommended for trials to clearly report whether imputation methods to handle missing data was or was not performed. If imputation was used it should be clearly stated what method was used along with any assumptions made, following the guidelines of Sterne et al(41).

Only about a third of articles reviewed reported using sensitivity analyses. There was some confusion between sensitivity analyses for missing data, and secondary analyses. Sensitivity analyses for missing data should keep the primary analysis model, but vary the assumptions about the distribution of the missing data, to establish the robustness of inference for the primary analysis to the inevitably untestable assumptions about the missing data. By contrast, secondary analysis with regards to excluding patients for the primary outcome is attempting to answer a separate, secondary question(42). Thus, while EMEA 2000 and CONSORT 2012 describe this as sensitivity analysis (and many papers we reviewed followed this), in general this will not be the case, and conflating the two inevitably leads to further confusion.

The focus of the analysis for non-inferiority trials should be on patients who behaved as they were supposed to within a trial, i.e. the per-protocol population. But rather than excluding patients from the per-protocol analyses, an alternative approach would be to make an assumption about the missing data for patients who do not adhere to the pre-defined per-protocol definition and then impute missing outcomes for these patients as if they had continued in the trial without deviating. Sensitivity analyses should then be used to check robustness of these results. However, currently, it is unclear what methods are appropriate to achieve this goal.

Subgroup of trials with published protocols

The mandatory publication of protocols taken from NEJM publications improved results for all criteria assessed. This reiterates the findings from Vale et al who evaluated the risk of bias assessments in systematic reviews assessed from published reports, but had also accessed

protocols directly from the trial investigators and found that deficiencies in the medical journal reports of trials does not necessarily reflect deficiencies in trial quality(43). Given this, it is clear that a major improvement in the reporting of non-inferiority trials would result if all journals followed the practice. Since publication of e-supplements is very cheap, there appears to be no reason not to do this.

CONCLUSION

Our findings suggest clear violations of available guidelines, including the CONSORT 2006 statement (published four years before the first paper in our review) which concentrates on improving how non-inferiority trials are reported and is widely endorsed across medical journals.

There is some indication that the quality of reporting for non-inferiority studies can affect the conclusions made and therefore the results of trials that fail to clearly report the items discussed above should be interpreted cautiously. It is essential that justification for the choice of the non-inferiority margin becomes standard practice, providing the information early on when planning a study including as much detail as possible. If journals enforced a policy where authors must justify the choice of the non-inferiority margin prior to accepting publication, this would encourage authors to provide robust justifications for something so critical given that clinical practice may be expected to change if the margin of non-inferiority is met.

Sample size calculations include consideration of the type I error rate, which should be consistent with the confidence intervals as these provide inferences made for non-inferiority when compared against the margin. Inconsistency between the two may distort inferences made, and stricter confidence intervals may lack power to detect true differences for the original sample size calculation. If any imputation was performed then this should be detailed along with its underlying assumptions, supplemented with sensitivity analyses under different assumptions about the missing data. There is an urgent need for research into appropriate ways of handling missing data in the per-protocol analysis for non-inferiority trials; once resolved, this analysis should be the primary analysis.

Information that is partially pre-specified before the conduct of a trial may inadvertently provide opportunities to modify decisions that were not pre-specified at the time of reporting without providing any justification. It is therefore crucial for editors to be satisfied that criteria are defined *a priori*. A compulsory requirement from journals to publish protocols as e-supplements and even statistical analysis plans along with the main article would avoid this ambiguity.

References

1. Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ, Group C. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. Jama. 2006;295(10):1152-60.
2. Food DA, H.H.S. Draft Guidance for Industry Non-Inferiority Clinical Trials. 2010.
3. Schumi J, Wittes JT. Through the looking glass: understanding non-inferiority. Trials. 2011;12:106.
4. Treadwell JR, Uhl S, Tipton K, Shamliyan T, Viswanathan M, Berkman ND, et al. Assessing equivalence and noninferiority. Journal of clinical epidemiology. 2012;65(11):1144-9.
5. Snapinn SM. Noninferiority trials. Curr Control Trials Cardiovasc Med. 2000;1(1):19-21.
6. Murthy VL, Desai NR, Vora A, Bhatt DL. Increasing proportion of clinical trials using noninferiority end points. Clinical cardiology. 2012;35(9):522-3.
7. Suda KJ, Hurley AM, McKibbin T, Motl Moroney SE. Publication of noninferiority clinical trials: changes over a 20-year interval. Pharmacotherapy. 2011;31(9):833-9.
8. Vermeulen L. Gain in Popularity of Noninferiority Trial Design: Caveat Lector. Pharmacotherapy. 2011;31(9):2.
9. Piaggio G, Elbourne DR, Pocock SJ, Evans SJ, Altman DG, Group C. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. Jama. 2012;308(24):2594-604.
10. Committee for Medicinal Products for Human Use (CHMP) CfPMP. Points to Consider on Switching Between Superiority and Non-inferiority London, England2000 [cited 2015 November 3rd]. Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003658.pdf.
11. Committee for Medicinal Products for Human U, Efficacy Working P, Committee for Release for C. Committee for Medicinal Products for Human Use (CHMP) guideline on the choice of the non-inferiority margin. Statistics in medicine. 2006;25(10):1628-38.
12. International conference on harmonisation; guidance on statistical principles for clinical trials; availability--FDA. Notice. Federal register. 1998;63(179):49583-98.
13. Food, Drug Administration HHS. International Conference on Harmonisation; choice of control group and related issues in clinical trials; availability. Notice. Federal register. 2001;66(93):24390-1.
14. Chan AW, Tetzlaff JM, Gotzsche PC, Altman DG, Mann H, Berlin JA, et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. Bmj. 2013;346:e7586.
15. GAO. New drug approval. FDA's Consideration of Evidence from Certain Clinical Trials 2010 [cited 2016 11th April]. Available from: http://www.gao.gov/assets/310/308301.pdf.
16. Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. Bmj. 2010;340:c869.
17. D'Agostino RB, Sr., Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues - the encounters of academic consultants in statistics. Statistics in medicine. 2003;22(2):169-86.
18. Matsuyama Y. A comparison of the results of intent-to-treat, per-protocol, and g-estimation in the presence of non-random treatment changes in a time-to-event non-inferiority trial. Statistics in medicine. 2010;29(20):2107-16.
19. Abraha I, Montedori A. Modified intention to treat reporting in randomised controlled trials: systematic review. Bmj. 2010;340:c2697.
20. ISI Web of Knowledge [cited 2015 May 31st]. Available from: http://admin-apps.webofknowledge.com/JCR/JCR.

21. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons; 1987.
22. Hwang IKM, T. Design Issues in Noninferiority Equivalence Trials. Drug Information Journal. 1999;33:1205-18.
23. Vickers A, Goyal N, Harland R, Rees R. Do certain countries produce only positive results? A systematic review of controlled trials. Control Clin Trials. 1998;19(2):159-66.
24. Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. The Cochrane database of systematic reviews. 2009(1):MR000006.
25. Tunes da Silva G, Logan BR, Klein JP. Methods for equivalence and noninferiority testing. Biol Blood Marrow Transplant. 2009;15(1 Suppl):120-7.
26. Schiller P, Burchardi N, Niestroj M, Kieser M. Quality of reporting of clinical non-inferiority and equivalence randomised trials--update and extension. Trials. 2012;13:214.
27. Wangge G, Klungel OH, Roes KC, de Boer A, Hoes AW, Knol MJ. Room for improvement in conducting and reporting non-inferiority randomized controlled trials on drugs: a systematic review. PloS one. 2010;5(10):e13550.
28. Le Henanff A, Giraudeau B, Baron G, Ravaud P. Quality of reporting of noninferiority and equivalence randomized trials. Jama. 2006;295(10):1147-51.
29. Gallagher TQ, Hill C, Ojha S, Ference E, Keamy DG, Williams M, et al. Perioperative dexamethasone administration and risk of bleeding following tonsillectomy in children: a randomized controlled trial. Jama. 2012;308(12):1221-6.
30. Lesaffre E. Superiority, equivalence, and non-inferiority trials. Bulletin of the NYU hospital for joint diseases. 2008;66(2):150-4.
31. Postma DF, van Werkhoven CH, van Elden LJ, Thijsen SF, Hoepelman AI, Kluytmans JA, et al. Antibiotic treatment strategies for community-acquired pneumonia in adults. The New England journal of medicine. 2015;372(14):1312-23.
32. Schulz-Schupke S, Helde S, Gewalt S, Ibrahim T, Linhardt M, Haas K, et al. Comparison of vascular closure devices vs manual compression after femoral artery puncture: the ISAR-CLOSURE randomized clinical trial. Jama. 2014;312(19):1981-7.
33. Lucas BP, Trick WE, Evans AT, Mba B, Smith J, Das K, et al. Effects of 2- vs 4-week attending physician inpatient rotations on unplanned patient revisits, evaluations by trainees, and attending physician burnout: a randomized trial. Jama. 2012;308(21):2199-207.
34. Gulmezoglu AM, Lumbiganon P, Landoulsi S, Widmer M, Abdel-Aleem H, Festin M, et al. Active management of the third stage of labour with and without controlled cord traction: a randomised, controlled, non-inferiority trial. Lancet. 2012;379(9827):1721-7.
35. Kaji AH, Lewis RJ. Noninferiority Trials: Is a New Treatment Almost as Effective as Another? Jama. 2015;313(23):2371-2.
36. Lee CH, Steiner T, Petrof EO, Smieja M, Roscoe D, Nematallah A, et al. Frozen vs Fresh Fecal Microbiota Transplantation and Clinical Resolution of Diarrhea in Patients With Recurrent Clostridium difficile Infection: A Randomized Clinical Trial. Jama. 2016;315(2):142-9.
37. Rahman NM, Pepperell J, Rehal S, Saba T, Tang A, Ali N, et al. Effect of Opioids vs NSAIDs and Larger vs Smaller Chest Tube Size on Pain Control and Pleurodesis Efficacy Among Patients With Malignant Pleural Effusion: The TIME1 Randomized Clinical Trial. Jama. 2015;314(24):2641-53.
38. Stevenson AR, Solomon MJ, Lumley JW, Hewett P, Clouston AD, Gebski VJ, et al. Effect of Laparoscopic-Assisted Resection vs Open Resection on Pathological Outcomes in Rectal Cancer: The ALaCaRT Randomized Clinical Trial. Jama. 2015;314(13):1356-63.
39. Fleshman J, Branda M, Sargent DJ, Boller AM, George V, Abbas M, et al. Effect of Laparoscopic-Assisted Resection vs Open Resection of Stage II or III Rectal Cancer on Pathologic Outcomes: The ACOSOG Z6051 Randomized Clinical Trial. Jama. 2015;314(13):1346-55.
40. Gotzsche PC. Lessons from and cautions about noninferiority and equivalence randomized trials. Jama. 2006;295(10):1172-4.

41. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj. 2009;338:b2393.
42. Morris TP, Kahan BC, White IR. Choosing sensitivity analyses for randomised trials: principles. BMC medical research methodology. 2014;14:11.
43. Vale CL, Tierney JF, Burdett S. Can trial quality be reliably assessed from published reports of cancer trials: evaluation of risk of bias assessments in systematic reviews. Bmj. 2013;346:f1798.

Competing interests

The authors declare that they have no competing interests

Contributions

SR: Conception, data extraction, analysis, wrote the manuscript.

TM: Data extraction, critical revision of the manuscript

KF: Critical revision of the manuscript

JC: Critical revision of the manuscript

PP: Data extraction, analysis, critical revision of the manuscript

Acknowlegements

Sunita Rehal is supported by a Medical Research Council studentship. Other authors received no specific funding for this work

Data Sharing

No additional data available

Supplement

Methods

Data extraction form

The form was tested by two reviewers (SR & TM) on articles, included in this review, until agreement was achieved between both reviewers. Justifications for the choice of the non-inferiority margin were reviewed by two reviewers due to its complexity. The power from the planned sample size calculation was recorded from the methods section. We recorded what analyses was used for the primary outcome and we noted how this was defined according to authors. This was either extracted from the main text or from the CONSORT flow chart. Definitions that were provided but not classed as ITT, PP, mITT or as-treated were categorised accordingly.

Definition of patient population

If definitions were provided on what patient population was included in analyses but were not classed by authors, then the definitions were categorised as follows:

- All patients randomised into the study were analysed was classed as an *intention-to-treat* analysis
- Patients who were excluded after administration of treatment (e.g. withdrawals, loss to follow up, compliance) was classed as a *per-protocol* analysis
- Patients who were excluded after administration of treatment, but the exclusion was not treatment related (e.g. patients who did not have the disease of interest) was classed as a *modified intention-to-treat* analysis
- Analysis based on what treatment patients actually received as opposed to the treatment that was allocated at the time of randomisation was classed as an *as-treated* analysis

Determining whether the analysis of the patient population was primary or secondary

Information on whether a patient population was considered as a primary analysis or secondary analysis (for the same primary outcome) was collected. The population was assumed primary if only one analysis was reported. If more than one analysis was performed but it was not clearly described which was to be taken as the primary and/or secondary analysis, the primary analysis was assumed to be whatever was presented in the results section of the abstract and secondary if not presented in the abstract but stated elsewhere within the article. If all results were presented for all populations in the abstract, then both were assumed as primary unless non-inferiority was concluded on only one patient population. Analysis was assumed secondary if the patient population was stated but not defined or if the results of the analysis were not presented in the article.

Results

Reasons for "Other" justification of non-inferiority margin

For all articles

There were 12(7%) justifications classed as "other":

- Based on previous trial. No evidence for consultation with external expert group, and no reference to previous trial of the control arm
- Based on unpublished data. No evidence for consultation with external expert group, and no reference to previous trials of the control arm
- Clinical basis and based on previous trials and guidelines. No evidence for consultation with external expert group, and no reference to previous trials of the control arm
- Clinical basis. Attempted to justify based on preservation of treatment effect, but were unable to do so due to paucity of previous trials.
- Expert group external to the authors and previous trial. No reference to previous trial of the control arm
- Justified based on treatment effect of control, but margin actually bigger than control arm treatment effect
- Placebo controlled study. Clinical basis, previous trials and literature review
- Preservation of treatment effect. Reference to separate paper justifying margin
- Regulatory guidelines (WHO), but recommendation is for superiority. No evidence for consultation with external expert group, and no reference to previous trials of the control arm
- Synthesis approach
- Unclear

NEJM protocols

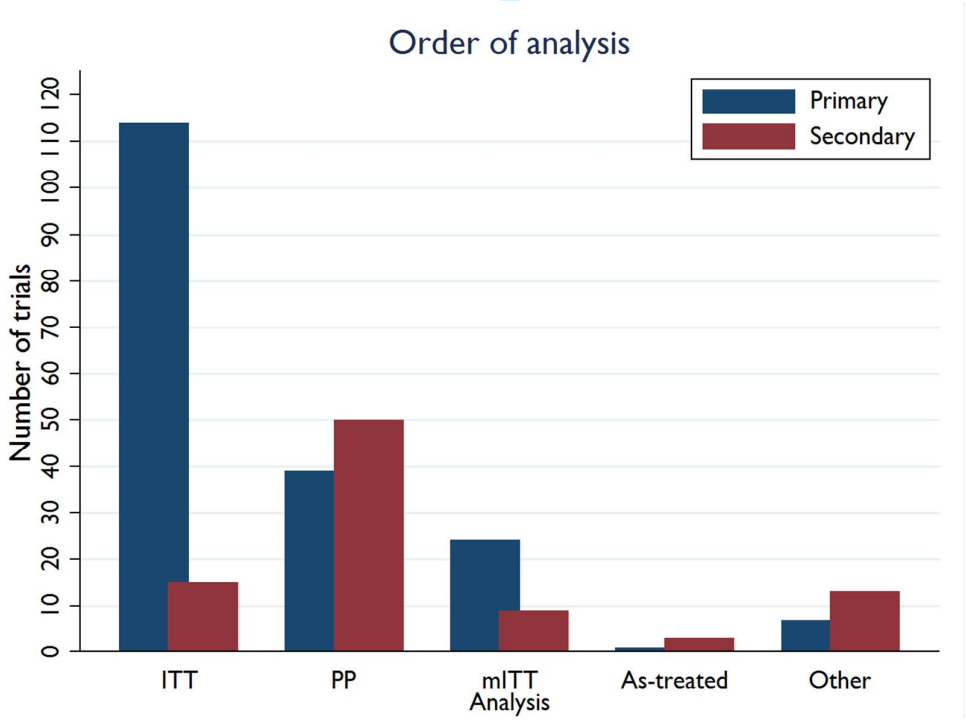
There were 6 (10%) justifications classed as "other":

- Based on previous trial. No evidence for consultation with external expert group, and no reference to previous trial of the control arm
- General comment that margin was decided according to FDA request
- Justified based on treatment effect of control, but margin actually bigger than control arm treatment effect
- Preservation of treatment effect based on estimates of control arm effect from previous trials and clinical basis
- Preservation of treatment effect based on estimates of control arm effect from previous trials, clinical basis and according to FDA guidelines
- Preservation of treatment effect. Reference to separate paper justifying margin

Table 1a: Type of analysis chosen

	All articles	NEJM protocols
Analysis	**n (%)**	**n (%)**
ITT only	54 (32%)	12 (20%)
PP only	3 (2%)	0
mITT only	8 (5%)	3 (5%)
ITT and PP	56 (33%)	17 (28%)
ITT and mITT	3 (2%)	2 (3%)
ITT and as-treated	4 (2%)	4 (7%)
ITT and other definition	6 (4%)	2 (3%)
PP and mITT	17 (10%)	9 (15%)
PP and other definition	4 (2%)	2 (3%)
mITT and as-treated	0 (0%)	1 (2%)
mITT and other definition	1 (1%)	1 (2%)
ITT, PP and mITT	1 (1%)	1 (2%)
ITT, PP and as-treated	0 (0%)	1 (2%)
ITT, PP and other definition	5 (3%)	5 (8%)
mITT, PP and other definition	4 (2%)	0
Unclear	2 (1%)	1 (2%)

Figure 1a: Chosen analysis by primary or secondary analysis



NB: One study performed ITT and PP analyses but it was unclear which of the two was taken as primary and secondary

Table 1b: Definition of analysis

Analysis	Definition	n (%)
ITT		**129**
	All patients randomised	68 (53%)
	all patients randomised who received at least one dose of treatment/intervention	21 (16%)
	All patients randomised excluding missing data	7 (5%)
	All patients randomised excluding errors in randomisation	3 (2%)
	All patients randomised who received at least one dose of treatment/intervention, excluding missing data	1 (1%)
	All patients randomised with exclusions from one centre which was removed due to misconduct	1 (1%)
	Other	17 (13%)
	Unclear	1 (1%)
	Not defined	10 (8%)
PP		**90**
	Patients who received allocated treatment/intervention	8 (9%)
	Excluding patients with major protocol violations	5 (6%)
	Patients who completed allocated treatment/intervention as intended	4 (4%)
	Patients who adhered to treatment	2 (2%)
	Excluding patients with protocol deviations	2 (2%)
	Patients with no exclusion criteria and who received specific amount of treatment/intervention	2 (2%)
	Patients who received allocated treatment/intervention, no major protocol violations with outcome	2 (2%)
	Excluding patients who switched treatment	1 (1%)
	Patients who received at least one dose of treatment/intervention	1 (1%)
	Patients who adhered to the protocol	1 (1%)
	Patients who completed the assigned study regimen or adhered to treatment before an event	1 (1%)
	Patients who received correctly allocated treatment/intervention excluding withdrawals	1 (1%)
	Patients who received specific amount of treatment/intervention and adhered to protocol	1 (1%)
	Patients who received allocated treatment/intervention, excluding non-adherence	1 (1%)
	Patients who adhered to protocol excluding withdrawals	1 (1%)
	Excluded patients with protocol deviations in addition to mITT definition	1 (1%)
	excluded patients that received rescue medication and protocol violations	1 (1%)
	Patients who received at least one dose of drug/intervention and received allocated treatment/intervention excluding missing outcome data	1 (1%)
	All patients who received at least one dose of treatment/intervention and did not have major protocol violations and were followed for event while receiving drug	1 (1%)

	All patients who received at least one dose of treatment/intervention and did not have major protocol violations	1 (1%)
	Excluding patients who were ineligibile, excluding patients who were administered the incorrect dose of medication and excluding patients who were allocated the incorrect treatment	1 (1%)
	All patients randomised who received at least one dose of treatment/intervention with an outcome, completed the study and complied with protocol	1 (1%)
	Non-adherence, patients who declined follow up, errors in randomisation, recurrent atrial fibrillation before randomisation were excluded	1 (1%)
	The per-protocol population (which consisted of the modified intention-to-treat population with the exclusion of patients with major protocol deviations and a compliance rate of <80%) was of primary interest, since a noninferiority analysis that is based on the modified intention-to-treat population is deemed to be not conservative	1 (1%)
	Patients were not eligible for per-protocol analysis for the following reasons: no follow-up visit; systemic treatment with other antimicrobial drugs up to day 28 (visit three); or missing more than one dose of the study drug during the first week of treatment or more than two doses during the whole treatment period	1 (1%)
	Excluded missing inclusion criteria; incorrect dosing; received prohibited medication; missing assessments	1 (1%)
	Per-protocol analyses excluded participants who had missing data at 1 month or who had major protocol violations (e.g., death, pregnancy, withdrawal from the study, loss to follow-up, or noncompliance). NB: Two results were presented for PP where compliance was included and excluded.	1 (1%)
	Per-protocol prespecified analyses included children with complete follow-up or a confirmed treatment failure, and excluded those treated for malaria without confirmatory microscopy, those for whom the alternative Plasmodium species was detected, and those who defaulted from follow-up despite repeated attempts at contact Flow chart includes: “and followed protocol”	1 (1%)
	Patients who, during the intended treatment period, had a venogram adjudicated as assessable, who developed confirmed deep vein thrombosis or pulmonary embolism, or who died from any cause); patients who had important protocol violations were excluded from the per-protocol analysis.	1 (1%)
	The per-protocol population was defined as all patients included in the ITT analysis, excluding those who did not receive the regimen as prescribed. These were patients who received less than 6 weeks of treatment (42 days of daily treatment or 36 days of 6-days-a-week treatment) or more than 9 weeks of treatment (63 days of daily treatment or 54 days of 6-days-a-week treatment) in the intensive phase and those who received less than 42 doses (ie, 4 weeks of missed treatment) or more than 60 doses (ie, 2 weeks of extra treatment) in the continuation phase (the protocol requirement is that patients receive 18 weeks of 3- times-weekly treatment, ie, 54 doses). Also excluded were patients whose treatment was modified for reasons other than bacteriological failure or relapse (including patients changing treatment for adverse drug reactions, following return after default, or attributable to concomitant HIV infection).	1 (1%)
	Per-protocol snapshot analysis, which included all participants who were enrolled, received at least one dose of study drug, and did not meet any of the following prespecifi ed criteria: discontinuation of study drug before week 48 or HIV RNA data missing in week 48 analysis window (accounting for 80% of excluded patients), and adherence in the bottom 2.5th percentile (accounting for 20% of the excluded patients)	1 (1%)
	The perprotocol group consisted of all patients who were enrolled, had no major protocol deviation, received the full treatment, and were assessed at day 15 or 31,	1 (1%)

	day 45, and 6 months (-2 to +6 weeks).	
	Criteria to exclude patients from this set were violation of major in- or exclusion criteria, change of treatment arm, early treatment discontinuation or relevant dose deviations of chemo- or radiotherapy unless caused by death or progression, radiotherapy without PET panel recommendation or omission of radiotherapy against recommendation, PET panel decision to take the patient off protocol treatment, or missing documentation of treatment	1 (1%)
	The per-protocol analysis set additionally excludes patients with change of treatment arm, early treatment discontinuation or relevant dose deviations of chemo- or radiotherapy unless caused by death or progression, or missing documentation of treatment	1 (1%)
	The perprotocol analysis was based on all participants who received 3 doses of vaccine according to 1 of the study's vaccine dosing schedules, were seronegative to the relevant HPV type at baseline, and had a valid serology result after the third dose of the HPV vaccine	1 (1%)
	Not defined. Taken from flow chart: Patients not meeting the definition of having received adequate treatment provided they have not already had an unfavourable response to treatment. Other exclusions done as well, but are not defined in flow chart	1 (1%)
	All patients who underwent randomization, completed a full treatment course or had early treatment failure before treatment was completed, had outcome data for the primary efficacy end point on day 28, and complied with the protocol to the extent that would allow efficacy evaluation	1 (1%)
	We also conducted a perprotocol analysis, which included those who completed the 2-month visit while receiving treatment (108 oral, 113 intratympanic) because intention-to-treat analyses may bias toward noninferiority. Flow chart also shows patients who withdrew before the 2m follow up, those who discontinued treatment but completed follow up and those who completed treatment but missed 2m follow up were excluded.	1 (1%)
	Which consisted of participants who received all three doses of vaccine within 1 year, did not have the HPV type being analyzed (i.e., were seronegative on day 1 and PCR-negative from day 1 through month 7), and had no protocol violations	1 (1%)
	A total of 12 (10%) patients in each group did not undergo PEG for anatomical reasons. Between the PEG procedure and the follow-up visit, five patients died, one patient pulled out the PEG catheter without ensuing complications, three patients were lost to follow-up, and one patient who was randomised to cefuroxime received co-trimoxazole instead.	1 (1%)
	Will include all subjects in the MITT population grouped by randomized treatment assignment regardless of treatment received with the exception of the following additional exclusions 1. Subjects not meeting the definition of having received an adequate amount of their allocated study regimen (see below for definition), provided they have not already been classified as having an unfavourable outcome 2. Subjects lost to follow-up or withdrawn before the Month 6 visit, unless they have already been classified as having an unfavourable outcome. 3. Subjects whose treatment was modified or extended for reasons (e.g. an adverse drug reaction or pregnancy) other than an unfavourable therapeutic response to treatment, unless they have already been classified as having an unfavourable outcome 4. Subjects who are classified as "major protocol violations" (see section 6.5), unless they have already been classified as having an unfavourable outcome on the basis of data obtained prior to the protocol violation	1 (1%)
	The per-protocol analysis excluding the 6 patients who were lost to follow-up and the 3 patients who received postoperative corticosteroids (including the 4 patients who experienced primary bleeding events)	1 (1%)

Excluded patients who received a platelet transfusion for reasons not recommended in the protocol	1 (1%)
We also did a per-protocol analysis of the medical outcomes, excluding outpatients discharged more than 24 h after randomisation and inpatients discharged 24 h or less after randomisation.	1 (1%)
The perprotocol population was defined as intention-to-treat patients with (1) successful procedure outcome, (2) treatment solely with the zotarolimus-eluting stent, (3) dual antiplatelet therapy according to randomization, and (4) complete clinical follow-up information.	1 (1%)
Not defined. Flow chart shows the following exclusions: had another histology or malignancy; withdrew informed consent; had an allergic reaction on first rituximab infusion and consecutively other treatment; only had radiotherapy; received incorrectly allocated treatment; did not meet inclusion or exclusion criteria; no therapy; death before therapy	1 (1%)
Not defined. Flow chart suggests patients were excluded if they did not receive the protocol and withdrawals	1 (1%)
Censoring of events if any component of the initial randomised trial treatment was stopped	1 (1%)
Not defined. Flow chart shows inclusion/exclusion criteria violated, non-adherence, prohibited medication and missing results were excluded	1 (1%)
Participants who did not follow protocol and/or were seropositive or polymerase chain reaction-positive for HPV-16, HPV- 18, HPV-6, or HPV-11 at enrolment were excluded from the per-protocol population analysis but retained for the intention-to-treat population analysis. Participants were eligible to continue with the 18- and 36-month follow-up if they had all of their doses of vaccine and a 7-month blood sample collected. If participants were excluded from the per-protocol population analysis at 7 months, they remained excluded for the remainder of the study but were retained for intention- to-treat analysis.	1 (1%)
The per-protocol population included all patients who completed the study (1 year), and for whom the second reading of a CT-scan confirmed the diagnosis of uncomplicated appendicitis.	1 (1%)
For analyses based on the per-protocol population, patients were analysed according to their randomly assigned treatment group. To be included in the perprotocol population, a patient was required to meet the following criteria: Had a mean baseline hemoglobin ≥ 8.0 and < 11.0 g/dl; Completed the study through at least week 36, and at least 5 hemoglobin values were obtained during the evaluation period; Had no missing administrations of study medication between weeks 21 and 35, inclusive; Had not received any RBC or whole blood transfusions within the 12 weeks prior to randomization; Had not received any RBC or whole blood transfusions for reasons other than lack of effect of study medication (lack of effect of study medication was documented as “Anemia of CRF” on the case report form) between weeks 21 and 35, inclusive; Had not received any ESA other than the assigned study treatment between weeks 21 and 35, inclusive; Had adequate iron status at baseline and during the evaluation period (defined as serum ferritin ≥100 ng/ml and TSAT ≥ 20% during weeks 24, 28, and 32)	1 (1%)
Not defined. Flow chart shows exclusions: caesarean section or forceps; short umbilical cord or nuchal cord; need for resuscitation; team became unavailable; weight scale malfunctioned; parent withdrew consent	1 (1%)
Completers (observed cases; included patients in the full analysis set who did not have important protocol violations, completed at least 684 days of treatment, and had HbA1c measured at week 104)	1 (1%)
For analyses based on the per-protocol population, patients were analyzed according to their randomly assigned treatment group. To be included in the per-protocol population, a patient was required to meet the following criteria: Had a mean baseline hemoglobin ≥ 10.0 and ≤ 12.0 g/dl; Completed the study through at least week 36, and at least six haemoglobin values were obtained during the	1 (1%)

	evaluation period.; Received ≥ 75% of total prescribed (i.e., expected) doses of study medication between weeks 25 and 35, inclusive (detailed algorithms for this determination were specified in the Statistical Analysis Plan).; Had not received any RBC transfusions within the 12 weeks prior to randomization.; Had not received any RBC transfusions for reasons other than lack of effect of study medication (lack of effect of study medication was documented as "Anemia of CRF" on the case report form) between weeks 25 and 36, inclusive.; Had not received any ESA other than the assigned study treatment between weeks 25 and 35, inclusive.; Had adequate iron status at baseline and at week 36 (defined as serum ferritin ≥ 100 ng/ml and TSAT ≥ 20%).	
	This population included all patients who underwent randomisation and who completed the study procedures to month 6.	1 (1%)
	We also performed a per-protocol analysis, which notably excluded patients in the antibiotic group who had been switched from amoxicillin plus clavulanic acid to another antibiotic.	1 (1%)
	We did a per-protocol snapshot analysis, which included all participants who were randomly assigned treatment, received at least one dose of study drug, and did not meet any of the following prespecifi ed criteria: discontinuation of study drug before week 48 or HIV RNA results missing in the week 48 analysis window, and adherence in the bottom 2.5th percentile.	1 (1%)
	Patients were included in the per-protocol population if they met the criteria for inclusion in the modified intention-to-treat population, underwent an adequate assessment of venous thromboembolism not later than 2 days after administration of the last dose of study drug, and had no major protocol violations.	1 (1%)
	The perprotocol population comprised patients in the modified intention-to-treat group who received treatment for at least 3 days (in the case of patients with treatment failure) or at least 8 days (in the case of patients with clinical cure), had documented adherence to the protocol, and underwent an end-of-therapy evaluation.	1 (1%)
	The per-protocol analysis set consisted of participants with exposure to treatment for at least 12 weeks who did not have any major protocol violations that could affect the primary endpoint and had a valid glycated haemoglobin (HbA1c) assessment at baseline and at (or after) 12 weeks.	1 (1%)
	Not defined	11 (12%)
mITT		**34**
	All patients randomised who received at least one dose of treatment/intervention	10 (29%)
	All patients randomised who received at least one dose of treatment/intervention, excluding missing data	6 (18%)
	All patients randomised with at least one dose of treatment/intervention excluding patients/site with violations of GCP	2 (6%)
	All randomised patients who received at least one dose of treatment/intervention excluding patients without disease or excluding patients resistant to one of the drug combinations. Excluding patients whose death was not related to the disease or had reinfection after being cured or patients who were classed as unassessable at the endpoint	1 (3%)
	Patients were excluded if they were resistant to two of the treatment combinations and patients who were unassessable and had not reached endpoint	1 (3%)
	On-treatment which included events that occurred within 30 days after the last dose of study medication was administered	1 (3%)
	Patients were excluded if they had missing/contaminated outcome data or could not produce an assessment or were lost to follow up or had death not related to disease or had confirmed reinfection	1 (3%)

	Excluded if consent withdrawn, non-compliance, moved and other (other not defined)	1 (3%)
	Other	11 (32%)
As-treated		**4**
	All patients randomised who received intervention	1 (25%)
	Not defined	3 (75%)
Other		**20**
	Full analysis set	4 (20%)
	On treatment analysis	3 (15%)
	Complete follow up data	1 (5%)
	ITT efficacy	1 (5%)
	PP and modified PP	1 (5%)
	Should be classed as PP. All patients who completed study with no major protocol deviations	1 (5%)
	Should be classed as mITT	2 (10%)
	Should be classed as mITT (ITT with no exclusion criteria)	1 (5%)
	Should be as treated (treatment received)	1 (5%)
	Other	5 (25%)
Unclear		**2**

Study conclusions

Of the articles that were designed as non-inferiority trials, two articles stated the trial was non-inferiority, but had drawn equivalence graphs with two margins; one article stated the trial was for non-inferiority but states the sample size calculation is to determine equivalence; one article concluded that their study did not show equivalence; one concluded equivalence; one article stated that the margin was an equivalence margin; one stated that they would test for equivalence; one concluded non-inferiority as the confidence interval was within ±margin; one concluded equivalence in the abstract but non-inferiority in the main paper; one stated that "results were consistent with showing non-inferiority (i.e. equivalence)".

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	5
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	NA
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	6
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	6, 7
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	6
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	6,7
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	6,7
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	6, 7 and supplement
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	NA
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	NA
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	NA

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	NA
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	6
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	8
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	8,18,19
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	NA
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	NA
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	NA
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	NA
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	9
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	10
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	13
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	13
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	NA

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: **www.prisma-statement.org**.

BMJ Open

Non-inferiority trials: are they inferior? A systematic review of reporting in major medical journals

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2016-012594.R1
Article Type:	Research
Date Submitted by the Author:	29-Jul-2016
Complete List of Authors:	Rehal, Sunita; MRC Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology; MRC Clinical Trials Unit at UCL, London Hub for Trials Methodology and Research Morris, Tim; MRC Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology; MRC Clinical Trials Unit at UCL, London Hub for Trials Methodology Research Fielding, Katherine; MRC Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology; London School of Hygiene & Tropical Medicine, MRC Tropical Epidemiology Group, Department of Infectious Disease Epidemiology Carpenter, James; MRC Clinical Trials Unit at UCL, London Hub for Trials Methodology Research; London School of Hygiene & Tropical Medicine, Department of Medical Statistics Phillips, Patrick; MRC Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology
Primary Subject Heading:	Medical publishing and peer review
Secondary Subject Heading:	Research methods
Keywords:	non-inferiority, systematic review, randomised controlled clinical trials, clinical trial

Non-inferiority trials: are they inferior?

A systematic review of reporting in major medical journals

*Sunita Rehal, statistician[1,2], Tim P. Morris, statistician[1,2], Katherine Fielding, reader in medical statistics and epidemiology[1,3], James R. Carpenter, professor of medical statistics[1, 2,4], Patrick P.J. Phillips, senior statistician[1]

[1]MRC Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology, London, UK

[2]London Hub for Trials Methodology Research, MRC Clinical Trials Unit at UCL, London, UK

[3] MRC Tropical Epidemiology Group, Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, UK

[4] Department of Medical Statistics, London School of Hygiene & Tropical Medicine, LondonWC1E 7HT, UK.

Correspondence to Sunita Rehal:*

MRC Clinical Trials Unit at UCL

Institute of Clinical Trials & Methodology

Aviation House

125 Kingsway

London WC2B 6NH

e-mail: s.rehal@ucl.ac.uk

Telephone number: 02076704702

Key words: non-inferiority, systematic review, clinical trial, randomised controlled clinical trial

Word count:

4689

ABSTRACT

Objective

To assess the adequacy of reporting of non-inferiority trials alongside the consistency and utility of current recommended analyses and guidelines.

Design

Review of randomised clinical trials that used a non-inferiority design published between January 2010 and May 2015 in medical journals that had an impact factor greater than 10 (*JAMA Internal Medicine, Archives Internal Medicine, PLOS medicine, Annals of Internal Medicine, BMJ, JAMA, Lancet and New England Journal of Medicine*).

Data sources

Ovid (MEDLINE).

Methods

We searched for non-inferiority trials and assessed the following: choice of non-inferiority margin and justification of margin; power and significance level for sample size; patient population used and how this was defined; any missing data methods used and assumptions declared; and any sensitivity analyses used.

Results

A total of 168 trial publications were included. Most trials concluded non-inferiority (132; 79%). The non-inferiority margin was reported for 98% (164) but less than half reported any justification for the margin (77; 46%). While most chose two different analyses (91; 54%) the most common being intention-to-treat or modified intention-to-treat and per-protocol, a large number of articles only chose to conduct and report one analysis (65; 39%), most commonly the intention-to-treat analysis. There was lack of clarity or inconsistency between the type I error rate and corresponding confidence intervals for 73 (43%) articles. Missing data were rarely considered with (99; 59%) not declaring whether imputation techniques were used.

Conclusion

Reporting and conduct of non-inferiority trials is inconsistent and does not follow the recommendations in available statistical guidelines, which are not wholly consistent themselves. Authors should clearly describe the methods used, and provide clear descriptions

of and justifications for their design and primary analysis. Failure to do this risks misleading conclusions being drawn, with consequent effects on clinical practice.

Strengths and limitations of this study

- This research clearly demonstrates the inconsistency in recommendations for non-inferiority trials provided by guidelines for researchers and this is reflected within this review
- Highlights missing data and sensitivity analyses in the context of non-inferiority trials
- Provide recommendations using examples for researchers using the non-inferiority design
- Justification of the choice of the margin was recorded as such if any attempt was made to do so. And so one could argue that inadequate attempts were counted as a 'justification', however there was good agreement between reviewers when independently assessed.
- Only one reviewer extracted information from all articles and therefore assessments may be subjective. However, there was good agreement when a random 5% of papers were independently assessed.

INTRODUCTION

Non-inferiority trials assess whether a new intervention is not much worse when compared to a standard treatment or care. These trials answer whether we are willing to accept a new intervention that may be clinically worse, yet still be beneficial for patients while having another advantage, such as less intensive treatment, lower cost or fewer side effects(1). Non-inferiority and equivalence are sometimes, mistakenly, used interchangeably. Equivalence trials are designed to show that a new intervention performs not much worse and not much better than a standard intervention. Both trial designs are different to superiority trials, which aim to show that a new intervention performs better when compared to a control.

Poor trial quality can bias trial results towards concluding no difference between treatments(2). This creates more challenges in non-inferiority trials than superiority trials as such bias can produce false positive results for non-inferiority(3-5). The increasing use of this design(6-8) means it is even more important for trialists to understand the issues around quality in the design and analysis of non-inferiority trials.

There are several guidelines available to aid researchers using a non-inferiority design, where various considerations of the design are explained and discussed (table 1).

1) The CONSORT extension statements(1, 9) focus on the reporting of non-inferiority trials, with the most recent 2012 statement being an elaboration of the 2006 statement.
2) The draft FDA 2010(2) document focus on all aspects and issues relative to non-inferiority trials and gives general guidance.
3) The EMEA 2000 guideline(10) discusses switching between non-inferiority and superiority designs and the EMEA 2006(11) guideline discusses the choice of the non-inferiority margin, taking into account two- and three-arm trials.
4) The ICH E9 and E10 guidelines(12, 13) are general statistical guidance documents addressing issues for all clinical trials and designs.
5) SPIRIT(14) is a guidance document for protocols for all trial designs and includes discussions of recently developed methodology.

Table 1: Summary of guidelines

	Justification of margin	Who is included in analysis	Confidence interval	Missing data	Sensitivity analyses
CONSORT 2006(1)	"Margin should be specified and preferably justified on clinical grounds"	"Non-ITT analyses might be desirable as a protection from ITTs increase in type I error. There is greater confidence in results when the conclusions are consistent." Intent-to-treat: "Analysing all patients within their randomized groups, regardless of whether they completed allocated treatment is recommended" Per-protocol: "Alternative analyses that exclude patients not taking allocated treatment or otherwise not protocol-adherent could bias the trial in either direction. The terms on-treatment or per-protocol analysis are often used but may be inadequately defined."	"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 noninferiority trials. If a 1-sided 5% significance level is deemed acceptable for the noninferiority hypothesis test (a decision open to question), a 90% 2-sided CI could then be used."		
CONSORT 2012(9)		"Should be indicated if conclusions are related to PP analysis, ITT analysis or both and if the conclusions are stable between them."	"The two-sided CI provides additional information, in particular for the situation in which the new treatment is superior to the reference treatment"		"Sensitivity analysis is discussed through an example: Study endpoints were analysed primarily for the per protocol population and repeated, for sensitivity reasons, for the intention-to-treat (ITT) population."
Draft FDA 2010(2)	"Whether M1 (the effect of the active control arm relative to placebo) is based on a single study or multiple studies, the observed (if there were multiple studies) or anticipated (if there is only one study) statistical variation of the treatment effect size should contribute to the ultimate choice of M1, as should any concerns about constancy. The selection of M2 (the largest clinically acceptable difference of the test treatment compared to the active control) is then based on clinical judgment regarding how much of the M1 active comparator treatment effect can be lost. The exercise of clinical judgment for the determination of M2 should be applied after the determination of M1 has been made based on the historical data and subsequent analysis"	"It is therefore important to conduct both ITT and 'as-treated' analyses in non-inferiority studies." Intent-to-treat: "preserve the principle that all patients are analyzed according to the treatment to which they have been randomized even if they do not receive it"	"Typically, the one-sided Type I error is set at 0.025, by asking that the upper bound of the 95% CI for control-treat be less than the NI margin. If multiple studies provide very homogeneous results for one or more important endpoints it may be possible to use the 90% lower bound rather than the 95% lower bound of the CI to determine the active control effect size"	"Poor quality can reduce the drug's effect size and undermine the assumption of the effect size of the control agent, giving the study a 'bias towards the null'."	

ICH E9(12)	"This margin is the largest difference that can be judged as being clinically acceptable"	"In confirmatory trials it is usually appropriate to plan to conduct both an analysis of the full analysis set and a per protocol analysis… In an equivalence or non-inferiority trial use of the full analysis set is generally not conservative and its role should be considered very carefully." Intent-to-treat: "subjects allocated to a treatment group should be followed up, assessed and analysed as members of that group irrespective of their compliance to the planned course of treatment" Full analysis set: "The set of subjects that is as close as possible to the ideal implied by the intention-to-treat principle. It is derived from the set of all randomised subjects by minimal and justified elimination of subjects." Per-protocol: "The set of data generated by the subset of subjects who complied with the protocol sufficiently to ensure that these data would be likely to exhibit the effects of treatment, according to the underlying scientific model. Compliance covers such considerations as exposure to treatment, availability of measurements and absence of major protocol violations."	"For non-inferiority trials a one-sided interval should be used. The choice of type I error should be a consideration separate from the use of a one-sided or two-sided procedure."	"Imputation techniques, ranging from LOCF to the use of complex mathematical models may be used to compensate for missing data"	"An investigation should be made concerning the sensitivity of the results of analysis to the method of handling missing values, especially if the number of missing values is substantial."
ICH E10(13)	"The determination of the margin in a non-inferiority trial is based on both statistical reasoning and clinical judgment"				
SPIRIT(14)		Use an example where "non-inferiority would be claimed if both ITT and PP analysis show conclusions of NI." Intent-to-treat: "In order to preserve the unique benefit of randomisation as a mechanism to avoid selection bias, an "as randomised" analysis retains participants in the group to which they were originally allocated. To prevent attrition bias, out-come data obtained from all participants are included in the data analysis, regardless of protocol adherence." Per-protocol and modified intention-to-treat: "Some trialists use other types of data analyses (commonly labelled as "modified intention to treat" or "per protocol") that exclude data from certain participants—such as those who are found to be ineligible after randomisation or who deviate from the intervention or follow-up protocols. This exclusion of data from protocol non-adherers can introduce bias, particularly if the frequency of and the reasons for non-adherence vary between the study		"Multiple imputation can be used to handle missing data although relies on untestable assumptions"	"Sensitivity analyses are highly recommended to assess the robustness of trial results under different methods of handling missing data"

		groups."			
EMEA 2006(11)	"The choice of delta must always be justified on both clinical and statistical grounds"		"A two-sided 95% CI (or one-sided 97.5% CI) is constructed. The interval should lie entirely on the positive side of the margin. Statistical significance is generally assessed using the two-sided 0.05 level of significance (or one-sided 0.025)"		
EMEA 2000(10)		"ITT and PP analyses have equal importance and their use should lead to similar conclusions for robust interpretation"	"A two-sided confidence interval should lie entirely to the right of delta. If one-sided confidence is used then 97.5% should be used"		"It will be necessary to pay particular attention to demonstrating the sensitivity of the trial by showing similar results for the full analysis set and PP analysis set"

There is some inconsistency between these guidelines regarding the conduct of non-inferiority trials (table 1) which may adversely affect the overall quality and reporting of non-inferiority trials. Non-inferiority trials require more care around certain issues, and so clear guidance on how to design and analyse these trials are necessary. Some of these issues that can influence inferences made about non-inferiority are outlined below.

First, the non-inferiority margin – the value that allows for a new treatment to be "acceptably worse"(1) – is used as a reference for conclusions about non-inferiority. It is recommended by all guidelines that this margin is chosen on a clinical basis, meaning the maximum clinically acceptable extent to which a new drug can be less effective than the standard of care and still show evidence of an effect(15). However, it is unclear whether statistical considerations should also impact on the choice of an appropriate margin, as recommended by the Draft FDA 2010, ICH E10 and EMEA 2006 guidelines(2, 11, 13) (table 1). Ignoring statistical evidence from meta-analyses or systematic reviews could have the potential for clinicians to choose an unrealistic margin.

Second, it is important to choose who is included in analyses for non-inferiority trials. The intention-to-treat analysis (includes all randomised patients irrespective of post-randomisation occurrences) is preferred for superiority trials as it is likely to lead to a treatment effect closer to having no effect, and so is conservative(16). For non-inferiority trials, the intention-to-treat (ITT) analysis can bias towards the null, which may lead to false claims of non-inferiority(17). The alternative per-protocol (PP) analysis is often considered instead. But as the PP analysis allows for the exclusion of patients, it fails to preserve a balance of patient numbers between treatment arms (i.e. randomisation) that ITT analysis does and can cause bias in either direction, depending on who the analysis excludes(18). Guidelines often recommend performing both the ITT and PP analyses, although definitions are inconsistent (table 1). In particular, the CONSORT 2006 guidelines describe the PP analysis as excluding patients not taking allocated treatment or otherwise not protocol-adherent(1), whereas the ICH E9 guidelines state that the PP analysis is a "subset of patients who complied sufficiently with the protocol, such as exposure to treatment, availability of measures and absence of major protocol violations"(19). These obscure definitions could lead researchers to arbitrarily exclude patients from analyses. The draft FDA guidelines recommend researchers to use an ITT and as-treated analysis, although it is unclear what is meant by 'as-treated' as this is not defined within the guidelines. Other frequently used classifications such as modified intention-to-treat (mITT), which aims to contain 'justifiable' exclusions (e.g. patients who never had the disease of interest) from the ITT analysis, are also defined inconsistently(20).Third, while two-sided 95% confidence intervals are widely used for superiority trials, there is some inconsistent advice as whether to calculate 90% or 95% confidence intervals for non-inferiority trials and whether these should be presented as one-sided or two-sided intervals (Table 1).

Fourth, the handling of missing data is generally discussed for all trials but rarely in the specific context of non-inferiority trials. Methods recommended to handle missing data vary between guidelines. The ICH E9 guidelines recommend using a last observation carried forward imputation method(19), and the more recent SPIRIT guidelines recommend multiple imputation, but caution the reader that it relies on untestable assumptions(14) (table 1). Methods to handle missing data often contain untestable assumptions and so sensitivity

analyses are essential to test the robustness of conclusions under different assumptions(12). However, it is unclear what sensitivity analyses are appropriate for non-inferiority trials.

Given the inconsistency between guidelines, we hypothesised that poor conduct and reporting would be associated with demonstrating non-inferiority. This review investigates the quality of conduct and reporting for non-inferiority trials in a selection of high-impact journals over a five-year period. We also provide recommendations to aid trialists who may consider a non-inferiority design.

METHODS

Medical journals (general & internal medicine) with an impact factor greater than 10 according to the ISI web of knowledge(21) were included in the review (correct at time of search on 31st May 2015), the rationale being that articles published in these journals are likely to have the highest influence on clinical practice and be the most rigorously conducted and reported due to the thorough editorial process. We searched Ovid (Medline) using the search terms "noninferior", "non-inferior", "noninferiority" and "non-inferiority" in titles and abstracts between 1st January 2010 and 31st May 2015 in *New England Journal of Medicine*, *Lancet*, *JAMA*, *British Medical Journal*, *Annals of Internal Medicine*, *PLOS Medicine* and *Archives of Internal Medicine* (descending impact order). From 2013, Archives of Internal Medicine was renamed JAMA Internal Medicine, and therefore both journals have been included in this review. All journals refer authors to the CONSORT statement and checklist when reporting. Eligibility of articles was assessed via abstracts by two reviewers (SR and TM). Articles included were non-inferiority randomised controlled clinical trials. Articles were excluded if the primary analysis was not for non-inferiority. Systematic reviews, meta-analyses and commentaries were also excluded. Few trials were designed and analysed using Bayesian methods, and were therefore excluded for consistent comparability in frequentist methods.

Before performing the review, a data extraction form was developed to extract information from articles. Information extracted was with regards to the primary outcome. The form was standardised to collect information on year of publication, non-inferiority margin (and how the margin was justified), randomisation, type of intervention, disease area, sample size, analyses performed (how this was defined and what was classed as primary/secondary), primary outcome, p-values (and whether this was for a superiority hypothesis), significance level of confidence intervals (and whether both bounds were reported), imputation techniques for missing data, sensitivity analyses, conclusions of non-inferiority and whether a test for superiority was pre-specified. Justifications for the choice of the non-inferiority margin were reviewed by two reviewers (SR and PP). See supplement for further details on methods.

A quality grading system was developed based on whether the margin was justified (yes vs. no/poor), how many analyses were performed on the primary outcome (<2 vs. ≥2) and whether the type I error rate was consistent with the significance level of the confidence interval (yes vs. no/unclear). Articles were classed as "excellent" if all these criteria were fulfilled and were classed as "poor" if none were fulfilled. Articles which satisfied one criterion were classed as "fair" and articles that provided two of the three criteria were classed as "good". The results of this grading were compared to inferences on non-inferiority to assess if the quality of reporting was associated with concluding non-inferiority at the 5% significance level.

Additional published supplementary content was only accessed if it specifically referred to the information we were extracting within articles. As a sub-study, all statistical methods, outcomes and sample sizes from protocols and/or supplementary content were reviewed from

New England Journal of Medicine as the journal is known to specifically request and publish protocols and statistical analysis plans alongside accepted publications.

Assessments were carried out by one reviewer (SR), with a random selection of 5% independently reviewed (PP). Any assessments that required a second opinion were independently reviewed (TM). Any discrepancies were resolved by discussion between reviewers.

All analyses were conducted using Stata version 14.

RESULTS

Our search found 252 articles. After duplicate publications were removed, 217 were screened for eligibility using their titles and abstracts. A total of 46 articles were excluded leaving 171 articles to be reviewed. A further three articles were excluded during the full-text review leaving 168 articles (figure 1).

Figure 1: flow chart of eligibility of articles

General characteristics of the included studies are shown in table 2.

Table 2: General characteristics

	All articles (n=168)	Including NEJM protocols (n=61)
Characteristics	n (%)	n(%)
Journal		
NEJM	61 (36%)	61
Lancet	64 (38%)	
JAMA	19 (11%)	
BMJ	8 (5%)	
Annals of Internal Medicine	5 (2%)	
PLOS Medicine	7 (4%)	
Archives of Internal Medicine	2 (1%)	
JAMA of Internal Medicine	2 (1%)	
Year of publication		
2010	26 (15%)	9 (15%)
2011	27 (16%)	9 (15%)
2012	29 (17%)	8 (13%)
2013	39 (23%)	19 (31%)
2014	27 (16%)	10 (16%)
2015	20 (12%)	6 (10%)
Type of intervention		
Drug	112 (67%)	44 (72%)
Surgery	22 (13%)	7 (11%)
Other	34 (20%)	10 (16%)
Randomisation		
Patient	163 (97%)	59 (97%)
Cluster	5 (3%)	2 (3%)
Power		

80%	6 (36%)	19 (31%)
85%	11 (7%)	5 (8%)
90%	65 (39%)	26 (43%)
71 to 99% (Excluding the above)	21 (12%)	11 (18%)
Not reported/unclear	10 (6%)	0
Composite outcome		
Yes	78 (46%)	37 (61%)
No	90 (54%)	24 (39%)
Disease		
Heart disease	30 (18%)	13 (21%)
Blood disorder	19 (11%)	6 (10%)
Cancer	16 (10%)	8 (13%)
Diabetes	11 (7%)	2 (3%)
Thromboembolism	6 (4%)	6 (10%)
Skin infection (non-contagious)	3 (2%)	2 (3%)
Urinary tract infection	3 (2%)	0
Arthritis	3 (2%)	1 (2%)
Opthomology	3 (2%)	1 (2%)
Pneumonia	3 (2%)	1 (2%)
Complications in pregnancy	3 (2%)	0
Stroke	3 (2%)	2 (3%)
Testing method	3 (2%)	1 (2%)
Appendicitis	2 (1%)	1 (2%)
Depression	2 (1%)	0
Other non-infectious disease	18 (11%)	7 (11%)
HIV	18 (11%)	2 (3%)
Tuberculosis	6 (4%)	4 (7%)
Malaria	4 (2%)	1 (2%)
Skin infection (contagious)	2 (1%)	0
Hepatitis C	2 (1%)	2 (3%)
Other infectious disease	8 (5%)	1 (2%)

Margin

The non-inferiority margin was specified in 164(98%) articles and was justified in less than half of articles 76(45%). The most common justification was on a clinical basis (29 (17%)) which was often worded ambiguously and with little detail. A total of 14(8%) used previous findings from past trials or statistical reviews to justify the choice of the margin (table 3).

Table 3: Justification of choice of margin, total number of patient populations considered for analyses and patient population included in analysis

	All articles (N=168)	**Including NEJM protocols (N=61)**
	n (%)	n (%)
Justification of NI margin		
Made no attempt for justification	90 (54%)	22 (36%)
Clinical basis. No evidence for consultation with external expert group, and no reference to previous trials of the control arm	32 (19%)	11 (18%)
Preservation of treatment effect based on estimates of control arm effect from previous trials	13 (8%)	14 (23%)
Expert group external to the authors. No reference to previous trials of the control arm	6 (4%)	3 (5%)
The same margin as was used in other similar trials	5 (3%)	2 (3%)
10-12% recommended by disease specific FDA guidelines	4 (2%)	1 (2%)
General comment that margin was decided according to FDA/regulatory guidance.	4 (2%)	0
Clinical basis and based on previous similar trial. No evidence for consultation with external expert group, and no reference to previous trials of the control arm	3 (2%)	0
Based on registry/development program	0	2 (3%)
Other*	11 (7%)	6 (10%)
Number of analyses		
One	65 (39%)	15 (25%)
Two	91 (54%)	38 (62%)
Three	10 (6%)	7 (11%)
Not defined	2 (1%)	1 (2%)
Analysis		
ITT	129 (77%)	44 (72%)
PP	90 (54%)	35 (57%)

mITT	34 (20%)	17 (28%)
As-treated	4 (2%)	6 (10%)
Other	20 (12%)	10 (16%)
Unclear	2 (1%)	2 (3%)

*See supplement

Patients included in analysis

Over a third of articles 65(39%) declared only one analysis (table 3 and 1a of supplement). The majority of trials classed ITT analysis as primary and PP analyses as secondary (figure 1a online supplement). PP analyses were performed in 90(54%) trials of which 11(12%) did not define what was meant by "per-protocol" (table 3 and table 1b online supplement). Definitions of the PP population contained various exclusions, mostly regarding errors in randomised treatment or treatment received.

Type I error rate

Consistency between the type I error rate and confidence intervals reported was moderate at 95(57%) (table 4). Most articles, 69(41%), used a one-sided 2.5% or (numerically equivalent) two-sided 5% significance level (table 5) and some used a one-sided 5% significance level 46(27%). The majority of articles presented two-sided confidence intervals (147; 88%) and 19(11%) articles presented one-sided confidence intervals. Most two-sided confidence intervals were at the 95% significance level: 125(74%).

Table 4: Consistency of type I error rate with significance levels of confidence intervals over year of publication

	Year of publication						
	2010	**2011**	**2012**	**2013**	**2014**	**2015**	**Total**
All articles (N=168)							
Yes	11 (42%)	15 (56%)	15 (52%)	24 (62%)	19 (70%)	11 (55%)	95 (57%)
No	5 (19%)	4 (15%)	4 (14%)	5 (13%)	5 (19%)	3 (15%)	26 (15%)
Not reported	10 (38%)	8 (30%)	10 (34%)	10 (26%)	3 (11%)	6 (30%)	47 (28%)
NEJM subgroup (N=61)							
Yes	7 (78%)	6 (67%)	5 (63%)	14 (74%)	8 (80%)	4 (67%)	44 (72%)
No	1 (11%)	2 (22%)	2 (25%)	3 (16%)	2 (20%)	1 (17%)	11 (18%)
Not reported	1 (11%)	1 (11%)	1 (13%)	2 (11%)	0	1 (17%)	6 (10%)

Table 5: Significance level of a) type I error rate and b) confidence intervals for all articles by whether confidence interval was one or two-sided

a) Type I error rate (%)			
	One sided	Two sided	Not reported
0.8	0	1 (1%)	0
1.25	3 (2%)	0	0
2.45	1 (1%)	0	0
2.5	40 (24%)	2 (1%)	2 (1%)
5	46 (27%)	29 (17%)	15 (9%)
10	1 (1%)	2 (1%)	0
Not reported	3 (2%)	0	23 (14%)
b) **Significance level of confidence interval (%)**			
	One sided	Two sided	Not reported
90	1 (1%)	14 (8%)	1 (1%)
95	14 (8%)	125 (74%)	0
97.5	4 (2%)	7 (4%)	0
Other	0	1 (1%)	0
Not reported	0	0	1 (1%)

Missing data and sensitivity analyses

99(59%) trials did not report whether or not any imputation was done and only 12(7%) explicitly declared that no imputation was used. Assuming a worst-case scenario or multiple imputation were the most common methods used (table 6). The number of imputations used for multiple imputation was specified in 8/11 articles and 4/11 stated at least one of the assumptions from Rubin's rules(22). 64(38%) trials reported using sensitivity analyses to test robustness of conclusions of the primary outcome; of these 27/64 (42%) were related to assumptions about the missing data (table 6).

Table 6: Reporting of a) missing data and b) sensitivity analyses

	n (%)
a) Imputation performed	
Yes	**56 (33%)**
Worst case scenario	19 (34%)
Multiple imputation	11 (20%)
Last observation carried forward	8 (14%)
Complete case analysis	6 (11%)
Best case scenario	2 (4%)

Last observation carried forward and worst case scenario	2 (4%)
Best case/worst case scenario	3 (5%)
Mean imputation	1 (2%)
Complete case analysis, multiple imputation using propensity scores and multiple imputation using regression modelling	1 (2%)
Other and worst case scenario	1 (2%)
Other	1 (2%)
No	**12 (7%)**
Not reported	**99 (59%)**
Unclear	**1 (1%)**
Including NEJM protocols (N=61)	
Yes	22 (36%)
No	7 (11%)
Not reported	31 (51%)
Unclear	1 (2%)
b) Sensitivity analyses performed	
Yes	**64 (38%)**
Patient population	13 (20%)
Competing risks	2 (3%)
Statistical modelling	2 (3%)
Adjusted for baseline variables	1 (2%)
Excluded protocol violations	1 (2%)
On-treatment	1 (2%)
Patient population/other	1 (2%)
Unclear	2 (3%)
Other	15 (23%)
Missing data	**27 (42%)**
Best case/worst case scenario	5
Complete case analysis	3
Imputation of missing values	3
Multiple imputation	3
Worst case scenario	3
Baseline observation carried forward	1
Baseline observation carried forward and complete case analysis	1
Complete case analysis, multiple imputation using propensity scores and multiple imputation using regression modelling	1
Complete case analysis and missing not at random	1
Complete case analysis and best case scenario	1
Different methods	1
Last observation carried forward	1
Modelling	1

Observed-failure	1
Worst case scenario and last observation carried forward	1
No	**103 (61%)**
Unclear	**1 (1%)**
Including NEJM protocols	
Yes	38 (62%)
No	23 (38%)

Study conclusions

There were 7(4%) articles that could not make definitive conclusions. For example, if all analyses conducted had to demonstrate non-inferiority to conclude a treatment was non-inferior, and only one of the analyses did, then non-inferiority could not be concluded and could not be rejected. Non-inferiority was declared in 132(79%) articles. 10 of these had made some reference with equivalence studies within the article (See supplement for details),

Superiority analyses were performed in 37(22%) trials after declaring non-inferiority, of which 27/37 (73%) had explicitly pre-planned for superiority analyses. P-values were reported in 98(58%) articles, of which 29/98 (30%) were testing a superiority hypothesis.

Subgroup of trials with published protocols

Additional information from protocols published by NEJM was extracted for 57 of 61 articles. Including this additional information provided by NEJM improved reporting of results across all criteria: 39(64%) articles justified the choice of the non-inferiority margin compared to 19(31%); most planned two or more analyses 45(74%) compared to 37(61%) (there were a couple of cases where two analyses were planned in the protocol but only one was stated in the published article); consistency between type I error rates and confidence intervals was 44(72%) compared with 36(59%); imputation techniques were considered in 29(48%) compared with 17(28%) articles and sensitivity analyses were considered in 38(62%) articles compared with 25(41%). The majority of articles concluded non-inferiority with 8(13%) not determining non-inferiority. Most articles that concluded superiority 14(23%) pre-planned for it 9/14 (64%). Few articles 8/40 (20%) presented superiority p-values.

Association between quality of reporting and conclusions

Trials that were classed as having some 'other' conclusion about non-inferiority were excluded from the analysis. Overall, there was a suggestive difference between the quality of reporting and concluding non-inferiority: χ^2_1=3.76; p=0.05 (Cochran-Armitage test; table 7). Trials that were poorly reported were less likely to conclude non-inferiority than those that satisfied two or all criteria from justifying the choice of the margin, reporting two or more analyses or reporting a confidence interval consistent with the type I error rate.

Table 7: Quality of reporting of trials associated with conclusions of non-inferiority

Grade	Concluded non-inferiority			
	Yes (N=132)	No (N=29)	Other (N=7)	**Total (N=168)**
	n (%)	n (%)	n (%)	**n (%)**
Excellent[1]	11 (73%)	2 (13%)	2 (13%)	**15**
Good [2]	55 (86%)	9 (14%)	0 (0%)	**64**
Fair [3]	48 (80%)	8 (13%)	4 (7%)	**60**
Poor[4]	18 (62%)	10 (34%)	1 (3%)	**29**

*Excluding trials that concluded 'other': χ_1^2=3.76; p=0.05 (Cochran-Armitage test)

[1] Excellent if margin justified, ≥2 analyses on patient population performed, type I error rate consistent with significance level of confidence interval

[2] Good if fulfilled two of the following: margin justified, ≥2 analyses on patient population performed, type I error rate consistent with significance level of confidence interval

[3] Fair if fulfilled one of the following: margin justified, ≥2 analyses on patient population performed, type I error rate consistent with significance level of confidence interval

[4] Poor if margin not justified, <2 analyses on patient population performed, type I error rate not consistent with significance level of confidence interval

DISCUSSION

Reporting of non-inferiority trials is poor and is perhaps partly due to disagreement between guidelines on vital issues. There are some aspects that guidelines agree on, such as a requirement for the non-inferiority margin to be justified, but we find that this recommendation is neglected by the majority of authors. It is remarkable that several authors performed only one analysis for the primary outcome and the lack of consistency between the significance level chosen in sample size calculations and the confidence interval reported further highlights confusion of non-inferiority trials. Not knowing how to deal with missing data nor appropriate sensitivity analyses, also adds to the confusion. The combination of these recent findings assessed from high impact journals and the inconsistency in guidelines indicate: 1)the non-inferiority design is not well understood by those using the design and 2) methods for non-inferiority designs are yet to be optimised.

We anticipated that poor reporting of articles would bias towards concluding non-inferiority, however, the poorly reported trials were less likely to demonstrate non-inferiority. This is somewhat reassuring. Nevertheless, it is essential to ensure that what is reported at the end of a trial was pre-specified before the start of a trial: scientific credibility and regulatory acceptability of a non-inferiority trial rely on the trial being well-designed and conducted according to the design(23). It is possible that the quality of a trial may also depend on the quality of the outcome; unresponsive outcomes that miss important differences between treatments may be intentionally or unintentionally chosen to demonstrate non-inferiority. Therefore it is also important that the outcome chosen is robust.

Almost 80% of studies concluded non-inferiority, although it is unclear whether this is due to the reporting in articles or publication bias. It appears that positive results (i.e. alternative hypotheses) are published more often, regardless of trial design, as this number is consistent with other studies that found that more than 70% of published superiority trials demonstrated superiority(24, 25).

More than half of articles reported p-values, of which approximately a third reported p-values for a two-sided test for superiority. P-values, if reported, should be calculated for one-sided tests corresponding to the non-inferiority hypothesis; that is, with H_0: δ = margin. P-values for superiority should not be presented unless following demonstration of non-inferiority, where a pre-planned superiority hypothesis is tested(26).

Comparison with other studies

The value of the non-inferiority margin was almost always reported but more than half of articles made no attempt to explain how the choice was justified. While justification of the margin is low, this is actually an improvement from Schiller et al who reported 23% articles made a justification(27), although this difference could be because only high impact journals were included in this review. There were equally as many articles that planned and reported an ITT analysis compared with articles that performed ITT and PP analyses. This is surprising given that CONSORT 2006 state that an ITT analysis can bias non-inferiority trials towards showing non-inferiority(1). These results were lower than found by Wangge et al(28) who

reported 55% used either an ITT or PP and 42% used both ITT and PP. Most articles presented two-sided 95% confidence intervals which is consistent with results from Le Henanff et al(29).

There were very few articles that referred to preserving the treatment effect based on estimates of the standard of care arm from previous trials. It is vital that authors acknowledge this to ensure the standard of care is effective. If the control were to have no effect at all in the study then finding a small difference between the standard of care and new intervention would be meaningless(2).

Clinical considerations(1, 2, 9, 11-13) to justify the choice of the margin often had inadequate justifications, such as "deemed appropriate" or "consensus among a group of clinical experts". Non-inferiority is only meaningful if it has strong justification in the clinical context and so should be reported. If the justification includes a measurable reduction in adverse events, these should be measured and the benefit should be demonstrated. Guidelines recommend that the choice of margin should be justified primarily on clinical grounds, however, previous trials and historical data should also be considered if available. As an example, Gallagher et al(30) justify the choice of the margin providing as much information as possible by including references to all published reports and providing data from the institution where the senior author is based.

A statement often used in articles reviewed was "the choice of the margin was clinically acceptable". This statement does not contain enough information to justify the choice of the non-inferiority margin. If the choice of the margin is based on a group of clinical experts, authors should provide information on how many experts were involved and how many considered the choice of the margin being acceptable: a consensus among a group of 3 clinicians from one institution is different from a consensus of 20 clinicians representing several institutions. Radford et al(31) justify the choice of the non-inferiority margin after performing a delegate survey at a symposium. This method may be a way forward for researchers to obtain clinical assessment from a large group of clinicians. Even better would be to obtain formal assessments, using for example the Delphi method(32) which has been used in the COMET initiative(33), after presenting the proposed research at a conference or symposium for clinicians to really engage with the question at hand.

Definitions provided by authors were inconsistent under what they classed as ITT, PP, mITT and as-treated, for example "all patients randomised who received at least one dose of treatment" was defined at least once in each classification. According to the guidelines, the PP definition excludes patients from the analysis but it is unclear what those exclusions are. The ambiguity of how per-protocol is defined was evident in this review as definitions provided by authors could not be succinctly categorised.

Many articles presented only one analysis, despite most guidelines recommending at least two analyses(1, 2, 9, 10, 12). Unfortunately, guidelines differ in their advice on which of the two analyses should be chosen to base conclusions on. This regrettable, state of affairs was clearly reflected in our review.

Both the ITT and PP analyses have their biases and so neither can be taken as a "gold standard" for non-inferiority trials. The analysis of the primary outcome is the most important result for any clinical trial. It should be pre-defined in the protocol what patients should adhere to and

should be considered at the design stage what can be done to maximize adherence. It should be made clear exactly who is included in analyses given the variety of definitions provided by various authors, particularly for PP analyses where definitions are subjective. Most authors included treatment related exclusions such as "received treatment", "completed treatment" or "received the correct treatment". Such differences in definitions may be may be superficially small but could in fact make critical differences to the results of a trial.

Poor reporting of whether the hypothesis test was one-sided or two-sided or absence of the type I error rate in the sample size calculation meant over a quarter of articles were not clearly consistent with regards to the type I error rate and corresponding confidence interval.

Most guidelines advise presenting two-sided 95% confidence intervals and this is what most articles presented. However, this recommendation may cause some confusion between equivalence and non-inferiority trials. A 5% significance level is maintained using 95% confidence intervals in equivalence trials for two-sided hypotheses whereas non-inferiority takes a one-sided hypothesis and so a two-sided 90% confidence interval should be calculated. If a one-sided type I error rate of 2.5% is used in the sample size calculation then this corresponds to the stricter two-sided 95% confidence intervals, not a one-sided 95% confidence interval(34).

The power and type I error rate should be clearly reported within sample size calculations and whether the type I error rate is for a one-sided or two-sided test. For example, the CAP-START trial used a one-sided significance test of 0.05 with two-sided 90% confidence intervals and the authors provide exact details of the sample size calculation in the supplementary appendix(35). If presenting one bound of the confidence interval throughout an article, this must be done clearly and consistently as described by Schulz-Schupke et al, Lucas et al, Gulmezoglu et al(36-38). Recently, JAMA have introduced a policy to present the lower bound of the confidence interval with the upper bound tending towards infinity(39) and this has been put into practice in recent non-inferiority trials(40-43).

It is unclear whether the potential issues surrounding missing data is well recognised for non-inferiority studies given that the majority of articles did not explicitly state whether or not methods to handle missing outcome data would be considered. Most trials that used multiple imputation stated the number of imputations used but few discussed the assumptions made, which are particularly critical in this context. Some missing data are inevitable, but naïve assumptions and/or analysis threaten trial validity for both ITT and PP analyses(14), particularly in the non-inferiority context where more missing data can bias towards demonstrating non-inferiority(44).

It is recommended for trials to clearly report whether imputation methods to handle missing data was or was not performed. If imputation was used it should be clearly stated what method was used along with any assumptions made, following the guidelines of Sterne et al(45).

Only about a third of articles reviewed reported using sensitivity analyses. There was some confusion between sensitivity analyses for missing data, and secondary analyses. Sensitivity

analyses for missing data should keep the primary analysis model, but vary the assumptions about the distribution of the missing data, to establish the robustness of inference for the primary analysis to the inevitably untestable assumptions about the missing data. By contrast, secondary analysis with regards to excluding patients for the primary outcome is attempting to answer a separate, secondary question(46). Thus, while EMEA 2000 and CONSORT 2012 describe this as sensitivity analysis (and many papers we reviewed followed this), in general this will not be the case, and conflating the two inevitably leads to further confusion.

The focus of the analysis for non-inferiority trials should be on patients who behaved as they were supposed to within a trial, i.e. the per-protocol population. But rather than excluding patients from the PP analyses, an alternative approach would be to make an assumption about the missing data for patients who do not adhere to the pre-defined PP definition and then impute missing outcomes for these patients as if they had continued in the trial without deviating. Sensitivity analyses should then be used to check robustness of these results. However, currently, it is unclear what methods are appropriate to achieve this goal.

Subgroup of trials with published protocols

The mandatory publication of protocols taken from NEJM publications improved results for all criteria assessed. This reiterates the findings from Vale et al who evaluated the risk of bias assessments in systematic reviews assessed from published reports, but had also accessed protocols directly from the trial investigators and found that deficiencies in the medical journal reports of trials does not necessarily reflect deficiencies in trial quality(47). Given this, it is clear that a major improvement in the reporting of non-inferiority trials would result if all journals followed the practice. Since publication of e-supplements is very cheap, there appears to be no reason not to do this.

Strengths and limitations

This research demonstrates the inconsistency in the recommendations for non-inferiority trials provided by the available guidelines, which was also reflected within this review. We have provided several recommendations using examples for researchers wishing to use the non-inferiority design and have outlined the most important recommendations that we hope will be taken up in future guidelines (table 8). We have also highlighted the importance of missing data and using sensitivity analyses specific to non-inferiority trials. There are also some limitations in this review. Firstly, a justification of the choice of the margin was recorded as such if any attempt was made to do so. And so one could argue that inadequate attempts were counted as a 'justification', however there was good agreement between reviewers when independently assessed. Secondly, only one reviewer extracted information from all articles and therefore assessments may be subjective. However, there was good agreement when a random 5% of papers were independently assessed, and the categorisation of the justification of the non-inferiority margin was also independently assessed in all papers where a justification was given. Thirdly, an update of the CONSORT statement for non-inferiority trials was published during the

period of the search in 2012(9), which could improve the reporting of non-inferiority trials over the next few years. However, the first CONSORT statement for non-inferiority trials published in 2006(1) was released well before the studies included in our search and we have found that reporting of non-inferiority trials remains poor.

Table 8: Recommendations

Recommendations
Justification of the margin should be a made mandatory in journals
Authors should make reference to preserving the treatment effect based on estimates of the standard-of-care arm from previous trials
Presentation of the confidence interval should be consistent with the type I error rate used in sample size calculations
Analyses should be performed to answer the question of interest (i.e. the primary outcome) using additional analyses to test the robustness of that definition, rather than to heedlessly satisfy ITT and PP definitions
Missing data and sensitivity analyses should be considered to test assumptions of missing data made on the primary analysis
Protocols should always be published as online supplements and authors should make use of online supplementary content to include additional detail on methods (such as details for justifying the choice of the non-inferiority margin and full definition of analyses conducted) so that a word limit for a published article should not be an excuse for poor reporting

CONCLUSION

Our findings suggest clear violations of available guidelines, including the CONSORT 2006 statement (published four years before the first paper in our review) which concentrates on improving how non-inferiority trials are reported and is widely endorsed across medical journals.

There is some indication that the quality of reporting for non-inferiority studies can affect the conclusions made and therefore the results of trials that fail to clearly report the items discussed above should be interpreted cautiously. It is essential that justification for the choice of the non-inferiority margin becomes standard practice, providing the information early on when planning a study including as much detail as possible. If the choice of the non-inferiority margin changes following approval from an ethics committee, justification for the change and changes to the original sample size calculation should be explicit. If journals enforced a policy where authors must justify the choice of the non-inferiority margin prior to accepting publication, this would encourage authors to provide robust justifications for something so critical given that clinical practice may be expected to change if the margin of non-inferiority is met.

Sample size calculations include consideration of the type I error rate, which should be consistent with the confidence intervals as these provide inferences made for non-inferiority when compared against the margin. Inconsistency between the two may distort inferences made, and stricter confidence intervals may lack power to detect true differences for the

original sample size calculation. If any imputation was performed then this should be detailed along with its underlying assumptions, supplemented with sensitivity analyses under different assumptions about the missing data. There is an urgent need for research into appropriate ways of handling missing data in the per-protocol analysis for non-inferiority trials; once resolved, this analysis should be the primary analysis.

Information that is partially pre-specified before the conduct of a trial may inadvertently provide opportunities to modify decisions that were not pre-specified at the time of reporting without providing any justification. It is therefore crucial for editors to be satisfied that criteria are defined *a priori*. A compulsory requirement from journals to publish protocols as e-supplements and even statistical analysis plans along with the main article would avoid this ambiguity.

References

1. Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ, Group C. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. Jama. 2006;295(10):1152-60.
2. Food DA, H.H.S. Draft Guidance for Industry Non-Inferiority Clinical Trials. 2010.
3. Schumi J, Wittes JT. Through the looking glass: understanding non-inferiority. Trials. 2011;12:106.
4. Treadwell JR, Uhl S, Tipton K, Shamliyan T, Viswanathan M, Berkman ND, et al. Assessing equivalence and noninferiority. Journal of clinical epidemiology. 2012;65(11):1144-9.
5. Snapinn SM. Noninferiority trials. Curr Control Trials Cardiovasc Med. 2000;1(1):19-21.
6. Murthy VL, Desai NR, Vora A, Bhatt DL. Increasing proportion of clinical trials using noninferiority end points. Clinical cardiology. 2012;35(9):522-3.
7. Suda KJ, Hurley AM, McKibbin T, Motl Moroney SE. Publication of noninferiority clinical trials: changes over a 20-year interval. Pharmacotherapy. 2011;31(9):833-9.
8. Vermeulen L. Gain in Popularity of Noninferiority Trial Design: Caveat Lector. Pharmacotherapy. 2011;31(9):2.
9. Piaggio G, Elbourne DR, Pocock SJ, Evans SJ, Altman DG, Group C. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. Jama. 2012;308(24):2594-604.
10. Committee for Medicinal Products for Human Use (CHMP) CfPMP. Points to Consider on Switching Between Superiority and Non-inferiority London, England2000 [cited 2015 November 3rd]. Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003658.pdf.
11. Committee for Medicinal Products for Human U, Efficacy Working P, Committee for Release for C. Committee for Medicinal Products for Human Use (CHMP) guideline on the choice of the non-inferiority margin. Statistics in medicine. 2006;25(10):1628-38.
12. International conference on harmonisation; guidance on statistical principles for clinical trials; availability--FDA. Notice. Federal register. 1998;63(179):49583-98.
13. Food, Drug Administration HHS. International Conference on Harmonisation; choice of control group and related issues in clinical trials; availability. Notice. Federal register. 2001;66(93):24390-1.
14. Chan AW, Tetzlaff JM, Gotzsche PC, Altman DG, Mann H, Berlin JA, et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. Bmj. 2013;346:e7586.
15. GAO. New drug approval. FDA's Consideration of Evidence from Certain Clinical Trials 2010 [cited 2016 11th April]. Available from: http://www.gao.gov/assets/310/308301.pdf.
16. Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. Bmj. 2010;340:c869.
17. D'Agostino RB, Sr., Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues - the encounters of academic consultants in statistics. Statistics in medicine. 2003;22(2):169-86.
18. Matsuyama Y. A comparison of the results of intent-to-treat, per-protocol, and g-estimation in the presence of non-random treatment changes in a time-to-event non-inferiority trial. Statistics in medicine. 2010;29(20):2107-16.
19. ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. International Conference on Harmonisation E9 Expert Working Group. Statistics in medicine. 1999;18(15):1905-42.

20. Abraha I, Montedori A. Modified intention to treat reporting in randomised controlled trials: systematic review. Bmj. 2010;340:c2697.
21. ISI Web of Knowledge [cited 2015 May 31st]. Available from: http://admin-apps.webofknowledge.com/JCR/JCR.
22. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons; 1987.
23. Hwang IKM, T. Design Issues in Noninferiority Equivalence Trials. Drug Information Journal. 1999;33:1205-18.
24. Vickers A, Goyal N, Harland R, Rees R. Do certain countries produce only positive results? A systematic review of controlled trials. Control Clin Trials. 1998;19(2):159-66.
25. Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. The Cochrane database of systematic reviews. 2009(1):MR000006.
26. Tunes da Silva G, Logan BR, Klein JP. Methods for equivalence and noninferiority testing. Biol Blood Marrow Transplant. 2009;15(1 Suppl):120-7.
27. Schiller P, Burchardi N, Niestroj M, Kieser M. Quality of reporting of clinical non-inferiority and equivalence randomised trials--update and extension. Trials. 2012;13:214.
28. Wangge G, Klungel OH, Roes KC, de Boer A, Hoes AW, Knol MJ. Room for improvement in conducting and reporting non-inferiority randomized controlled trials on drugs: a systematic review. PloS one. 2010;5(10):e13550.
29. Le Henanff A, Giraudeau B, Baron G, Ravaud P. Quality of reporting of noninferiority and equivalence randomized trials. Jama. 2006;295(10):1147-51.
30. Gallagher TQ, Hill C, Ojha S, Ference E, Keamy DG, Williams M, et al. Perioperative dexamethasone administration and risk of bleeding following tonsillectomy in children: a randomized controlled trial. Jama. 2012;308(12):1221-6.
31. Radford J, Illidge T, Counsell N, Hancock B, Pettengell R, Johnson P, et al. Results of a trial of PET-directed therapy for early-stage Hodgkin's lymphoma. The New England journal of medicine. 2015;372(17):1598-607.
32. Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. Journal of advanced nursing. 2000;32(4):1008-15.
33. .
34. Lesaffre E. Superiority, equivalence, and non-inferiority trials. Bulletin of the NYU hospital for joint diseases. 2008;66(2):150-4.
35. Postma DF, van Werkhoven CH, van Elden LJ, Thijsen SF, Hoepelman AI, Kluytmans JA, et al. Antibiotic treatment strategies for community-acquired pneumonia in adults. The New England journal of medicine. 2015;372(14):1312-23.
36. Schulz-Schupke S, Helde S, Gewalt S, Ibrahim T, Linhardt M, Haas K, et al. Comparison of vascular closure devices vs manual compression after femoral artery puncture: the ISAR-CLOSURE randomized clinical trial. Jama. 2014;312(19):1981-7.
37. Lucas BP, Trick WE, Evans AT, Mba B, Smith J, Das K, et al. Effects of 2- vs 4-week attending physician inpatient rotations on unplanned patient revisits, evaluations by trainees, and attending physician burnout: a randomized trial. Jama. 2012;308(21):2199-207.
38. Gulmezoglu AM, Lumbiganon P, Landoulsi S, Widmer M, Abdel-Aleem H, Festin M, et al. Active management of the third stage of labour with and without controlled cord traction: a randomised, controlled, non-inferiority trial. Lancet. 2012;379(9827):1721-7.
39. Kaji AH, Lewis RJ. Noninferiority Trials: Is a New Treatment Almost as Effective as Another? Jama. 2015;313(23):2371-2.
40. Lee CH, Steiner T, Petrof EO, Smieja M, Roscoe D, Nematallah A, et al. Frozen vs Fresh Fecal Microbiota Transplantation and Clinical Resolution of Diarrhea in Patients With Recurrent Clostridium difficile Infection: A Randomized Clinical Trial. Jama. 2016;315(2):142-9.

41. Rahman NM, Pepperell J, Rehal S, Saba T, Tang A, Ali N, et al. Effect of Opioids vs NSAIDs and Larger vs Smaller Chest Tube Size on Pain Control and Pleurodesis Efficacy Among Patients With Malignant Pleural Effusion: The TIME1 Randomized Clinical Trial. Jama. 2015;314(24):2641-53.
42. Stevenson AR, Solomon MJ, Lumley JW, Hewett P, Clouston AD, Gebski VJ, et al. Effect of Laparoscopic-Assisted Resection vs Open Resection on Pathological Outcomes in Rectal Cancer: The ALaCaRT Randomized Clinical Trial. Jama. 2015;314(13):1356-63.
43. Fleshman J, Branda M, Sargent DJ, Boller AM, George V, Abbas M, et al. Effect of Laparoscopic-Assisted Resection vs Open Resection of Stage II or III Rectal Cancer on Pathologic Outcomes: The ACOSOG Z6051 Randomized Clinical Trial. Jama. 2015;314(13):1346-55.
44. Wiens BL, Rosenkranz, G.K. Missing Data in Noninferiority Trials. Statistics in Biopharmaceutical Research. 2013;5:383-93.
45. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj. 2009;338:b2393.
46. Morris TP, Kahan BC, White IR. Choosing sensitivity analyses for randomised trials: principles. BMC medical research methodology. 2014;14:11.
47. Vale CL, Tierney JF, Burdett S. Can trial quality be reliably assessed from published reports of cancer trials: evaluation of risk of bias assessments in systematic reviews. Bmj. 2013;346:f1798.

Data sharing statement

No additional data available

Competing interests

The authors declare that they have no competing interests

Contributions

SR: Conception, data extraction, analysis, wrote the manuscript.

TM: Data extraction, critical revision of the manuscript

KF: Critical revision of the manuscript

JC: Critical revision of the manuscript

PP: Data extraction, analysis, critical revision of the manuscript

Acknowlegements

Sunita Rehal is supported by a Medical Research Council studentship. Other authors received no specific funding for this work

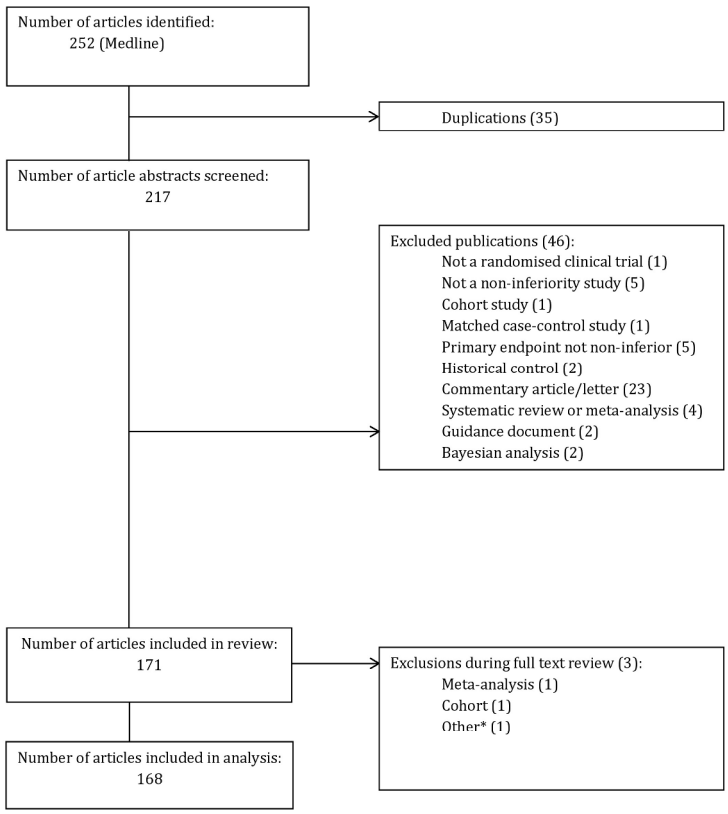


Figure 1: flow chart of eligibility of articles
figure 1
210x297mm (300 x 300 DPI)

Supplement

Methods

Data extraction form

The form was tested by two reviewers (SR & TM) on articles, included in this review, until agreement was achieved between both reviewers. Justifications for the choice of the non-inferiority margin were reviewed by two reviewers due to its complexity. The power from the planned sample size calculation was recorded from the methods section. We recorded what analyses was used for the primary outcome and we noted how this was defined according to authors. This was either extracted from the main text or from the CONSORT flow chart. Definitions that were provided but not classed as ITT, PP, mITT or as-treated were categorised accordingly.

Definition of patient population

If definitions were provided on what patient population was included in analyses but were not classed by authors, then the definitions were categorised as follows:

- All patients randomised into the study were analysed was classed as an *intention-to-treat* analysis
- Patients who were excluded after administration of treatment (e.g. withdrawals, loss to follow up, compliance) was classed as a *per-protocol* analysis
- Patients who were excluded after administration of treatment, but the exclusion was not treatment related (e.g. patients who did not have the disease of interest) was classed as a *modified intention-to-treat* analysis
- Analysis based on what treatment patients actually received as opposed to the treatment that was allocated at the time of randomisation was classed as an *as-treated* analysis

Determining whether the analysis of the patient population was primary or secondary

Information on whether a patient population was considered as a primary analysis or secondary analysis (for the same primary outcome) was collected. The population was assumed primary if only one analysis was reported. If more than one analysis was performed but it was not clearly described which was to be taken as the primary and/or secondary analysis, the primary analysis was assumed to be whatever was presented in the results section of the abstract and secondary if not presented in the abstract but stated elsewhere within the article. If all results were presented for all populations in the abstract, then both were assumed as primary unless non-inferiority was concluded on only one patient population. Analysis was assumed secondary if the patient population was stated but not defined or if the results of the analysis were not presented in the article.

Results

Reasons for "Other" justification of non-inferiority margin

For all articles

There were 12(7%) justifications classed as "other":

- Based on previous trial. No evidence for consultation with external expert group, and no reference to previous trial of the control arm
- Based on unpublished data. No evidence for consultation with external expert group, and no reference to previous trials of the control arm
- Clinical basis and based on previous trials and guidelines. No evidence for consultation with external expert group, and no reference to previous trials of the control arm
- Clinical basis. Attempted to justify based on preservation of treatment effect, but were unable to do so due to paucity of previous trials.
- Expert group external to the authors and previous trial. No reference to previous trial of the control arm
- Justified based on treatment effect of control, but margin actually bigger than control arm treatment effect
- Placebo controlled study. Clinical basis, previous trials and literature review
- Preservation of treatment effect. Reference to separate paper justifying margin
- Regulatory guidelines (WHO), but recommendation is for superiority. No evidence for consultation with external expert group, and no reference to previous trials of the control arm
- Synthesis approach
- Unclear

NEJM protocols

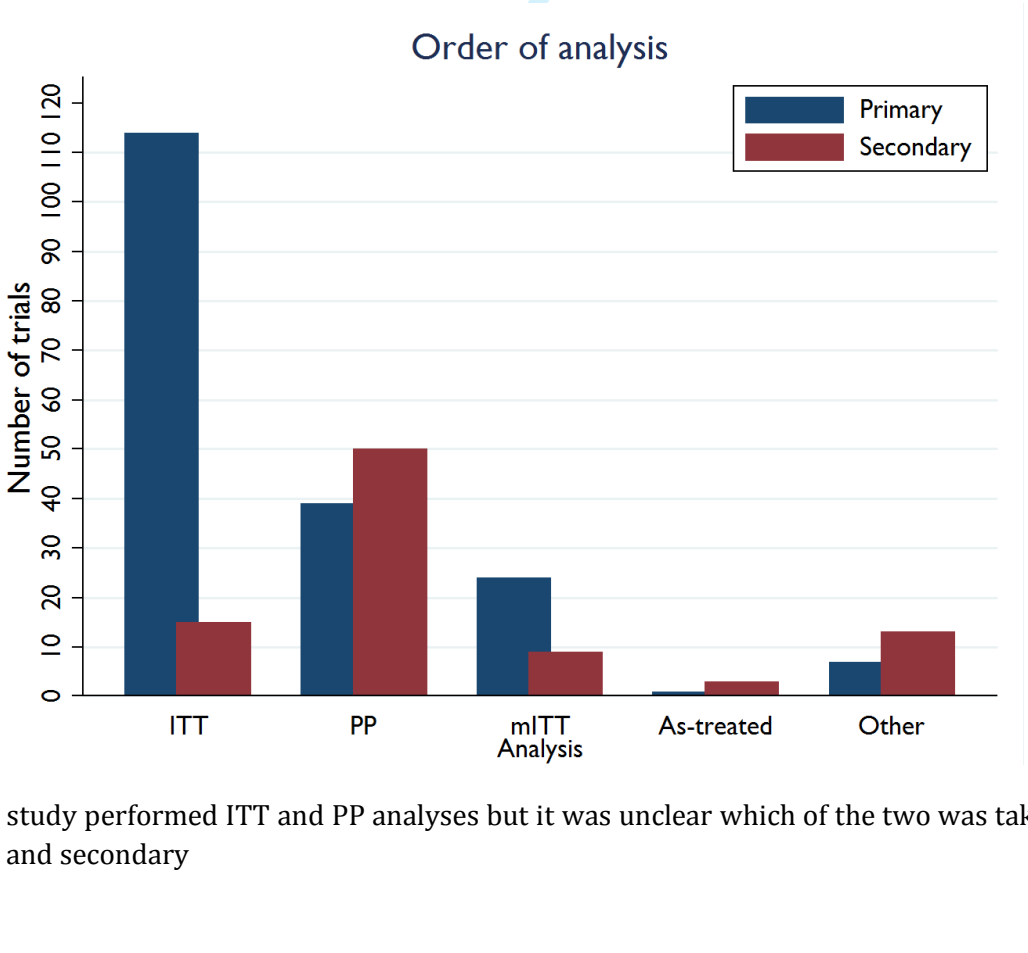
There were 6 (10%) justifications classed as "other":

- Based on previous trial. No evidence for consultation with external expert group, and no reference to previous trial of the control arm
- General comment that margin was decided according to FDA request
- Justified based on treatment effect of control, but margin actually bigger than control arm treatment effect
- Preservation of treatment effect based on estimates of control arm effect from previous trials and clinical basis
- Preservation of treatment effect based on estimates of control arm effect from previous trials, clinical basis and according to FDA guidelines
- Preservation of treatment effect. Reference to separate paper justifying margin

Table 1a: Type of analysis chosen

	All articles	NEJM protocols
Analysis	n (%)	n (%)
ITT only	54 (32%)	12 (20%)
PP only	3 (2%)	0
mITT only	8 (5%)	3 (5%)
ITT and PP	56 (33%)	17 (28%)
ITT and mITT	3 (2%)	2 (3%)
ITT and as-treated	4 (2%)	4 (7%)
ITT and other definition	6 (4%)	2 (3%)
PP and mITT	17 (10%)	9 (15%)
PP and other definition	4 (2%)	2 (3%)
mITT and as-treated	0 (0%)	1 (2%)
mITT and other definition	1 (1%)	1 (2%)
ITT, PP and mITT	1 (1%)	1 (2%)
ITT, PP and as-treated	0 (0%)	1 (2%)
ITT, PP and other definition	5 (3%)	5 (8%)
mITT, PP and other definition	4 (2%)	0
Unclear	2 (1%)	1 (2%)

Figure 1a: Chosen analysis by primary or secondary analysis



NB: One study performed ITT and PP analyses but it was unclear which of the two was taken as primary and secondary

Table 1b: Definition of analysis

Analysis	Definition	n (%)
ITT		**129**
	All patients randomised	68 (53%)
	all patients randomised who received at least one dose of treatment/intervention	21 (16%)
	All patients randomised excluding missing data	7 (5%)
	All patients randomised excluding errors in randomisation	3 (2%)
	All patients randomised who received at least one dose of treatment/intervention, excluding missing data	1 (1%)
	All patients randomised with exclusions from one centre which was removed due to misconduct	1 (1%)
	Other	17 (13%)
	Unclear	1 (1%)
	Not defined	10 (8%)
PP		**90**
	Patients who received allocated treatment/intervention	8 (9%)
	Excluding patients with major protocol violations	5 (6%)
	Patients who completed allocated treatment/intervention as intended	4 (4%)
	Patients who adhered to treatment	2 (2%)
	Excluding patients with protocol deviations	2 (2%)
	Patients with no exclusion criteria and who received specific amount of treatment/intervention	2 (2%)
	Patients who received allocated treatment/intervention, no major protocol violations with outcome	2 (2%)
	Excluding patients who switched treatment	1 (1%)
	Patients who received at least one dose of treatment/intervention	1 (1%)
	Patients who adhered to the protocol	1 (1%)
	Patients who completed the assigned study regimen or adhered to treatment before an event	1 (1%)
	Patients who received correctly allocated treatment/intervention excluding withdrawals	1 (1%)
	Patients who received specific amount of treatment/intervention and adhered to protocol	1 (1%)
	Patients who received allocated treatment/intervention, excluding non-adherence	1 (1%)
	Patients who adhered to protocol excluding withdrawals	1 (1%)
	Excluded patients with protocol deviations in addition to mITT definition	1 (1%)
	excluded patients that received rescue medication and protocol violations	1 (1%)
	Patients who received at least one dose of drug/intervention and received allocated treatment/intervention excluding missing outcome data	1 (1%)
	All patients who received at least one dose of treatment/intervention and did not have major protocol violations and were followed for event while receiving drug	1 (1%)

	All patients who received at least one dose of treatment/intervention and did not have major protocol violations	1 (1%)
	Excluding patients who were ineligibile, excluding patients who were administered the incorrect dose of medication and excluding patients who were allocated the incorrect treatment	1 (1%)
	All patients randomised who received at least one dose of treatment/intervention with an outcome, completed the study and complied with protocol	1 (1%)
	Non-adherence, patients who declined follow up, errors in randomisation, recurrent atrial fibrillation before randomisation were excluded	1 (1%)
	The per-protocol population (which consisted of the modified intention-to-treat population with the exclusion of patients with major protocol deviations and a compliance rate of <80%) was of primary interest, since a noninferiority analysis that is based on the modified intention-to-treat population is deemed to be not conservative	1 (1%)
	Patients were not eligible for per-protocol analysis for the following reasons: no follow-up visit; systemic treatment with other antimicrobial drugs up to day 28 (visit three); or missing more than one dose of the study drug during the first week of treatment or more than two doses during the whole treatment period	1 (1%)
	Excluded missing inclusion criteria; incorrect dosing; received prohibited medication; missing assessments	1 (1%)
	Per-protocol analyses excluded participants who had missing data at 1 month or who had major protocol violations (e.g., death, pregnancy, withdrawal from the study, loss to follow-up, or noncompliance). NB: Two results were presented for PP where compliance was included and excluded.	1 (1%)
	Per-protocol prespecified analyses included children with complete follow-up or a confirmed treatment failure, and excluded those treated for malaria without confirmatory microscopy, those for whom the alternative Plasmodium species was detected, and those who defaulted from follow-up despite repeated attempts at contact Flow chart includes: “and followed protocol”	1 (1%)
	Patients who, during the intended treatment period, had a venogram adjudicated as assessable, who developed confirmed deep vein thrombosis or pulmonary embolism, or who died from any cause); patients who had important protocol violations were excluded from the per-protocol analysis.	1 (1%)
	The per-protocol population was defined as all patients included in the ITT analysis, excluding those who did not receive the regimen as prescribed. These were patients who received less than 6 weeks of treatment (42 days of daily treatment or 36 days of 6-days-a-week treatment) or more than 9 weeks of treatment (63 days of daily treatment or 54 days of 6-days-a-week treatment) in the intensive phase and those who received less than 42 doses (ie, 4 weeks of missed treatment) or more than 60 doses (ie, 2 weeks of extra treatment) in the continuation phase (the protocol requirement is that patients receive 18 weeks of 3- times-weekly treatment, ie, 54 doses). Also excluded were patients whose treatment was modified for reasons other than bacteriological failure or relapse (including patients changing treatment for adverse drug reactions, following return after default, or attributable to concomitant HIV infection).	1 (1%)
	Per-protocol snapshot analysis, which included all participants who were enrolled, received at least one dose of study drug, and did not meet any of the following prespecifi ed criteria: discontinuation of study drug before week 48 or HIV RNA data missing in week 48 analysis window (accounting for 80% of excluded patients), and adherence in the bottom 2.5th percentile (accounting for 20% of the excluded patients)	1 (1%)
	The perprotocol group consisted of all patients who were enrolled, had no major protocol deviation, received the full treatment, and were assessed at day 15 or 31,	1 (1%)

	day 45, and 6 months (-2 to +6 weeks).	
	Criteria to exclude patients from this set were violation of major in- or exclusion criteria, change of treatment arm, early treatment discontinuation or relevant dose deviations of chemo- or radiotherapy unless caused by death or progression, radiotherapy without PET panel recommendation or omission of radiotherapy against recommendation, PET panel decision to take the patient off protocol treatment, or missing documentation of treatment	1 (1%)
	The per-protocol analysis set additionally excludes patients with change of treatment arm, early treatment discontinuation or relevant dose deviations of chemo- or radiotherapy unless caused by death or progression, or missing documentation of treatment	1 (1%)
	The perprotocol analysis was based on all participants who received 3 doses of vaccine according to 1 of the study's vaccine dosing schedules, were seronegative to the relevant HPV type at baseline, and had a valid serology result after the third dose of the HPV vaccine	1 (1%)
	Not defined. Taken from flow chart: Patients not meeting the definition of having received adequate treatment provided they have not already had an unfavourable response to treatment. Other exclusions done as well, but are not defined in flow chart	1 (1%)
	All patients who underwent randomization, completed a full treatment course or had early treatment failure before treatment was completed, had outcome data for the primary efficacy end point on day 28, and complied with the protocol to the extent that would allow efficacy evaluation	1 (1%)
	We also conducted a perprotocol analysis, which included those who completed the 2-month visit while receiving treatment (108 oral, 113 intratympanic) because intention-to-treat analyses may bias toward noninferiority. Flow chart also shows patients who withdrew before the 2m follow up, those who discontinued treatment but completed follow up and those who completed treatment but missed 2m follow up were excluded.	1 (1%)
	Which consisted of participants who received all three doses of vaccine within 1 year, did not have the HPV type being analyzed (i.e., were seronegative on day 1 and PCR-negative from day 1 through month 7), and had no protocol violations	1 (1%)
	A total of 12 (10%) patients in each group did not undergo PEG for anatomical reasons. Between the PEG procedure and the follow-up visit, five patients died, one patient pulled out the PEG catheter without ensuing complications, three patients were lost to follow-up, and one patient who was randomised to cefuroxime received co-trimoxazole instead.	1 (1%)
	Will include all subjects in the MITT population grouped by randomized treatment assignment regardless of treatment received with the exception of the following additional exclusions 1. Subjects not meeting the definition of having received an adequate amount of their allocated study regimen (see below for definition), provided they have not already been classified as having an unfavourable outcome 2. Subjects lost to follow-up or withdrawn before the Month 6 visit, unless they have already been classified as having an unfavourable outcome. 3. Subjects whose treatment was modified or extended for reasons (e.g. an adverse drug reaction or pregnancy) other than an unfavourable therapeutic response to treatment, unless they have already been classified as having an unfavourable outcome 4. Subjects who are classified as "major protocol violations"□ (see section 6.5), unless they have already been classified as having an unfavourable outcome on the basis of data obtained prior to the protocol violation	1 (1%)
	The per-protocol analysis excluding the 6 patients who were lost to follow-up and the 3 patients who received postoperative corticosteroids (including the 4 patients who experienced primary bleeding events)	1 (1%)

	Excluded patients who received a platelet transfusion for reasons not recommended in the protocol	1 (1%)
	We also did a per-protocol analysis of the medical outcomes, excluding outpatients discharged more than 24 h after randomisation and inpatients discharged 24 h or less after randomisation.	1 (1%)
	The perprotocol population was defined as intention-to-treat patients with (1) successful procedure outcome, (2) treatment solely with the zotarolimus-eluting stent, (3) dual antiplatelet therapy according to randomization, and (4) complete clinical follow-up information.	1 (1%)
	Not defined. Flow chart shows the following exclusions: had another histology or malignancy; withdrew informed consent; had an allergic reaction on first rituximab infusion and consecutively other treatment; only had radiotherapy; received incorrectly allocated treatment; did not meet inclusion or exclusion criteria; no therapy; death before therapy	1 (1%)
	Not defined. Flow chart suggests patients were excluded if they did not receive the protocol and withdrawals	1 (1%)
	Censoring of events if any component of the initial randomised trial treatment was stopped	1 (1%)
	Not defined. Flow chart shows inclusion/exclusion criteria violated, non-adherence, prohibited medication and missing results were excluded	1 (1%)
	Participants who did not follow protocol and/or were seropositive or polymerase chain reaction-positive for HPV-16, HPV- 18, HPV-6, or HPV-11 at enrolment were excluded from the per-protocol population analysis but retained for the intention-to-treat population analysis. Participants were eligible to continue with the 18- and 36-month follow-up if they had all of their doses of vaccine and a 7-month blood sample collected. If participants were excluded from the per-protocol population analysis at 7 months, they remained excluded for the remainder of the study but were retained for intention- to-treat analysis.	1 (1%)
	The per-protocol population included all patients who completed the study (1 year), and for whom the second reading of a CT-scan confirmed the diagnosis of uncomplicated appendicitis.	1 (1%)
	For analyses based on the per-protocol population, patients were analysed according to their randomly assigned treatment group. To be included in the perprotocol population, a patient was required to meet the following criteria: Had a mean baseline hemoglobin ≥ 8.0 and < 11.0 g/dl; Completed the study through at least week 36, and at least 5 hemoglobin values were obtained during the evaluation period; Had no missing administrations of study medication between weeks 21 and 35, inclusive; Had not received any RBC or whole blood transfusions within the 12 weeks prior to randomization; Had not received any RBC or whole blood transfusions for reasons other than lack of effect of study medication (lack of effect of study medication was documented as “Anemia of CRF” on the case report form) between weeks 21 and 35, inclusive; Had not received any ESA other than the assigned study treatment between weeks 21 and 35, inclusive; Had adequate iron status at baseline and during the evaluation period (defined as serum ferritin ≥100 ng/ml and TSAT ≥ 20% during weeks 24, 28, and 32)	1 (1%)
	Not defined. Flow chart shows exclusions: caesarean section or forceps; short umbilical cord or nuchal cord; need for resuscitation; team became unavailable; weight scale malfunctioned; parent withdrew consent	1 (1%)
	Completers (observed cases; included patients in the full analysis set who did not have important protocol violations, completed at least 684 days of treatment, and had HbA1c measured at week 104)	1 (1%)
	For analyses based on the per-protocol population, patients were analyzed according to their randomly assigned treatment group. To be included in the per-protocol population, a patient was required to meet the following criteria: Had a mean baseline hemoglobin ≥ 10.0 and ≤ 12.0 g/dl; Completed the study through at least week 36, and at least six haemoglobin values were obtained during the	1 (1%)

	evaluation period.; Received ≥ 75% of total prescribed (i.e., expected) doses of study medication between weeks 25 and 35, inclusive (detailed algorithms for this determination were specified in the Statistical Analysis Plan).; Had not received any RBC transfusions within the 12 weeks prior to randomization.; Had not received any RBC transfusions for reasons other than lack of effect of study medication (lack of effect of study medication was documented as "Anemia of CRF" on the case report form) between weeks 25 and 36, inclusive.; Had not received any ESA other than the assigned study treatment between weeks 25 and 35, inclusive.; Had adequate iron status at baseline and at week 36 (defined as serum ferritin ≥ 100 ng/ml and TSAT ≥ 20%).	
	This population included all patients who underwent randomisation and who completed the study procedures to month 6.	1 (1%)
	We also performed a per-protocol analysis, which notably excluded patients in the antibiotic group who had been switched from amoxicillin plus clavulanic acid to another antibiotic.	1 (1%)
	We did a per-protocol snapshot analysis, which included all participants who were randomly assigned treatment, received at least one dose of study drug, and did not meet any of the following prespecifi ed criteria: discontinuation of study drug before week 48 or HIV RNA results missing in the week 48 analysis window, and adherence in the bottom 2.5th percentile.	1 (1%)
	Patients were included in the per-protocol population if they met the criteria for inclusion in the modified intention-to-treat population, underwent an adequate assessment of venous thromboembolism not later than 2 days after administration of the last dose of study drug, and had no major protocol violations.	1 (1%)
	The perprotocol population comprised patients in the modified intention-to-treat group who received treatment for at least 3 days (in the case of patients with treatment failure) or at least 8 days (in the case of patients with clinical cure), had documented adherence to the protocol, and underwent an end-of-therapy evaluation.	1 (1%)
	The per-protocol analysis set consisted of participants with exposure to treatment for at least 12 weeks who did not have any major protocol violations that could affect the primary endpoint and had a valid glycated haemoglobin (HbA1c) assessment at baseline and at (or after) 12 weeks.	1 (1%)
	Not defined	11 (12%)
mITT		**34**
	All patients randomised who received at least one dose of treatment/intervention	10 (29%)
	All patients randomised who received at least one dose of treatment/intervention, excluding missing data	6 (18%)
	All patients randomised with at least one dose of treatment/intervention excluding patients/site with violations of GCP	2 (6%)
	All randomised patients who received at least one dose of treatment/intervention excluding patients without disease or excluding patients resistant to one of the drug combinations. Excluding patients whose death was not related to the disease or had reinfection after being cured or patients who were classed as unassessable at the endpoint	1 (3%)
	Patients were excluded if they were resistant to two of the treatment combinations and patients who were unassessable and had not reached endpoint	1 (3%)
	On-treatment which included events that occurred within 30 days after the last dose of study medication was administered	1 (3%)
	Patients were excluded if they had missing/contaminated outcome data or could not produce an assessment or were lost to follow up or had death not related to disease or had confirmed reinfection	1 (3%)

	Excluded if consent withdrawn, non-compliance, moved and other (other not defined)	1 (3%)
	Other	11 (32%)
As-treated		**4**
	All patients randomised who received intervention	1 (25%)
	Not defined	3 (75%)
Other		**20**
	Full analysis set	4 (20%)
	On treatment analysis	3 (15%)
	Complete follow up data	1 (5%)
	ITT efficacy	1 (5%)
	PP and modified PP	1 (5%)
	Should be classed as PP. All patients who completed study with no major protocol deviations	1 (5%)
	Should be classed as mITT	2 (10%)
	Should be classed as mITT (ITT with no exclusion criteria)	1 (5%)
	Should be as treated (treatment received)	1 (5%)
	Other	5 (25%)
Unclear		**2**

Study conclusions

Of the articles that were designed as non-inferiority trials, two articles stated the trial was non-inferiority, but had drawn equivalence graphs with two margins; one article stated the trial was for non-inferiority but states the sample size calculation is to determine equivalence; one article concluded that their study did not show equivalence; one concluded equivalence; one article stated that the margin was an equivalence margin; one stated that they would test for equivalence; one concluded non-inferiority as the confidence interval was within ±margin; one concluded equivalence in the abstract but non-inferiority in the main paper; one stated that "results were consistent with showing non-inferiority (i.e. equivalence)".

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	5
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	NA
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	6
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	6, 7
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	6
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	6,7
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	6,7
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	6, 7 and supplement
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	NA
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	NA
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	NA

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	NA
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	6
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	8
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	8,18,19
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	NA
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	NA
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	NA
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	NA
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	9
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	10
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	13
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	13
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	NA

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: **www.prisma-statement.org**.