York Health Economics Consortium

Providing consultancy and research in health economics for the NHS, pharmaceutical and health care industries since 1986

NICE DECISION SUPPORT UNIT

Clinical Equivalence and Non-Inferiority

Draft Report

JAMES MAHON, Associate
MIKE CHAMBERS, Associate
MATTHEW TAYLOR, Director

04 February 2020

UNIVERSITY of York

Contents

Section 1: Introduction

1.1 BACKGROUND AND AIMS

In the assessment of the cost-effectiveness of technologies by NICE, when a technology can be robustly evidenced to generate at least the same quality-adjusted life years (QALYs) as an existing technology (or technologies) the opportunity can arise to undertake a cost minimisation analysis (CMA) as opposed to a cost-utility analysis (CUA).

CMA is primarily used in two circumstances by NICE:

- In the Fast Track Appraisal (FTA) process as part of the Technology Appraisal (TA) process
- In the Medical Technologies Evaluation Programme (MTEP), for devices and diagnostic technologies

This paper provides recommendations how NICE can assess the plausibility of claims of non-inferiority (NI) or equivalence for drugs in the TA programme, medical devices or diagnostic technologies for MTEP and potential diagnostics in the Diagnostics Assessment Programme (DAP), covering:

- Evidence that can be used to assess non-inferiority or equivalence
- How the quality of non-inferiority and equivalence studies and the choice of non-inferiority margins should be assessed
- Interpretation of evidence from non-inferiority and equivalence studies

The paper is *not* designed to provide a detailed statistical methodology on how non-inferiority and equivalence trials should be designed and analysed.

Occasions can arise where a submission is made to NICE claiming superiority for a technology with a cost-utility analysis, but the statistical evidence from clinical trials is inconclusive as to whether the technology is superior. This paper explicitly does not make recommendations in such a situation, but the recommendations made in this paper may be used as a basis to reappraise the evidence presented by the company within a non-inferiority or equivalence framework. In addition, the paper does not specifically deal with considerations for the assessment of biosimilar medicines.

Activities undertaken to meet the aims of the paper included:

- A review of approaches suggested by the FDA, EMA and MHRA on how non-inferiority and equivalence trials should be undertaken and analysed
- A targeted review of published literature on how non-inferiority and equivalence trials are being undertaken and reported in practice and on specific methodology issues related to diagnostics and indirect treatment comparisons
- Interviews with key stakeholders within NICE and on NICE TA committees, along with health economists, medical statisticians and clinical trialists and an expert on MTEP submissions to verify findings and help formulate potential recommendations

1.2 BASIC CONCEPTS IN NON-INFERIORITY AND EQUIVALENCE TRIALS

Equivalence trial

A trial where the aim is to show that the difference in effectiveness between two or more technologies differs by a clinically unimportant amount.

Non-inferiority trial

A trial where the aim is to show that the effectiveness of one technology is not inferior to a comparator technology by a clinically important amount.

Non-inferiority margin (δ)

The non-inferiority margin is the 'clinically important/unimportant amount' referred to in the descriptions of equivalence and non-inferiority trials. It is the percentage of the effect size of the current or comparator technology against placebo (M1) that is clinically acceptable to be maintained for non-inferiority to be assumed (M2, the 'preserved fraction' or 'degree of inferiority').

Establishing M1 and M2 allows a sample size can be drawn to ensure there is sufficient power to conclude that the intervention technology is non-inferior or equivalent to the comparator trial.

Important considerations of the choice of M1 and M2 are:

- If M1 is overestimated, then there is a danger that the trial will incorrectly conclude that a new drug is effective
- If M1 is underestimated, then the power calculation for the trial may be too small to conclude the intervention technology is non-inferior or equivalence

- The smaller M2 is (i.e. the smaller the acceptable loss in efficacy) the larger the sample size will need to be
- The larger M2 is, the greater the risk of concluding that an inferior intervention technology is non-inferior or equivalent to a comparator technology that is more effective

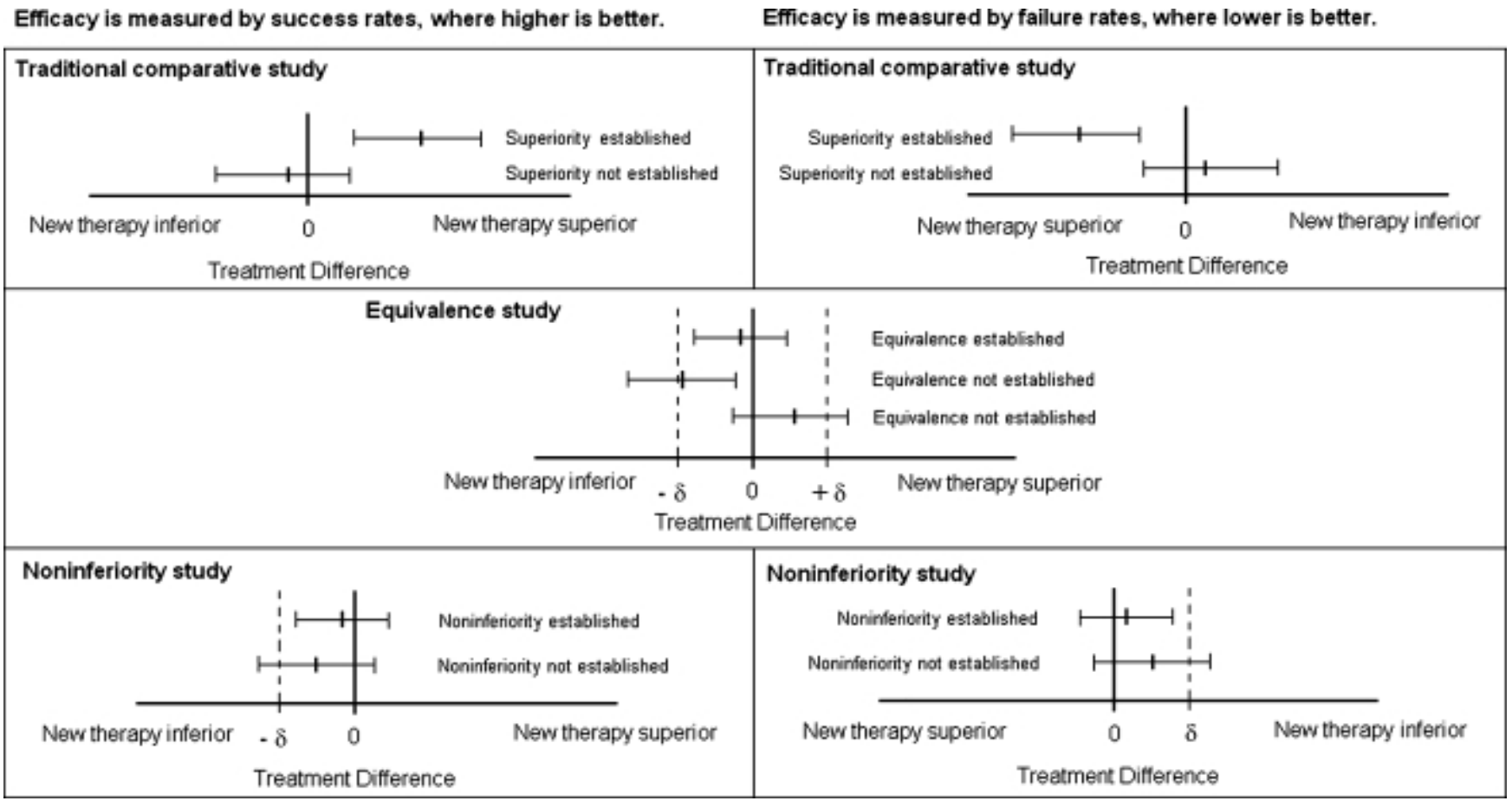
Placebo or bio-creep

Because non-inferiority trials, by design, allow for a new technology to be less effective than an existing technology but still be assessed as non-inferior, there is a possibility that increasingly inferior treatments become standard of care until a treatment that is no more effective than placebo is standard of care, even though effective treatments are available. This is called placebo or bio-creep and is a particular issue if M1 is overestimated or M2 is too large.

Interpretation of non-inferiority and equivalence results

Statistical inference in non-inferiority and equivalence trials is based upon the confidence intervals generated from the trial results. This is summarised in Figure 1 where δ is the non-inferiority margin and a comparison between inference in superiority, equivalence and non-inferiority studies is shown.

Figure 1: Confidence intervals and the equivalence margin in equivalence/non-inferiority studies



Source: Walker and Nowacki 2011

There are occasions where the confidence interval could lie between the non-inferiority margin and no treatment difference. In such a case, the conclusion should be that a new technology is statistically inferior (or potentially superior in an equivalence study) to an existing technology but that the inferiority (or superiority) is not clinically significant.

Section 2: Guidance on non-inferiority and equivalence from regulators

Published guidelines and methods of regulatory bodies (including the MHRA, the FDA and the EMA) on how they make decisions on non-inferiority/equivalence were reviewed. Table 2.1 summarises key information from these guidelines and the sections in each document where this information can be found. In addition, a CONSORT checklist specifically for non-inferiority and equivalence studies was published in 2012 was also reviewed. The full checklist can be found in Appendix A.

Limited guidance on studies demonstrating non-inferiority or clinical equivalence was identified from HTA organisations, with the exception of IQWiG in Germany. However, the German HTA system focuses the primary benefit assessment on clinical effectiveness and, as such, the IQWiG guidelines were very similar to those published by regulatory agencies.

No guidelines could be found providing methods on how to determine ‘similarity’ of technologies except as a synonym of equivalence. No guidelines could be found specifically on non-inferiority and equivalence for diagnostic tests. Where information taken from observational studies was discussed in the guidelines, it was only in the context of estimating M1, the effect size over placebo of the comparator technology.

The published guidance identified focuses on the specific requirements for assessing the design quality and interpreting the results of non-inferiority and equivalence trials, in particular the differences between these and superiority trials. Much of the guidance concerns use of equivalence and non-inferiority studies to demonstrate ‘absolute’ efficacy (assay sensitivity, determination of NI margin). The guidelines all make it clear that the quality of trials is especially important for non-inferiority studies as poor quality trials are more likely to lead to a failure to reject the null hypothesis that a new technology is not inferior to or is equivalent to an existing technology even if the opposite is true.

Whilst the guidelines are different in the detail that they provide on how non-inferiority and equivalence studies should be conducted (with the FDA in 2016 being the most comprehensive), the detail contained within can be summarised into several key areas on how margins should be set and how to conduct analysis.

For medical devices, formal regulations in US (510K) and EU (MDR) for medical devices focus on determinations of 'substantial equivalence': evidence is required to

demonstrate ‘equivalence' to a reference or predicate technology, for which clinical data are available and which has been accepted for clinical use. Such evidence covers technical aspects, biological impact, and clinical aspects including (same) intended use and measurement of intended effect. On the basis of these determinations it may not be necessary to generate new clinical evidence, however for certain technologies such as implantable devices new clinical data will usually be required.
In general a wider range of study designs (than for new pharmaceuticals) may be considered acceptable. Where new clinical trials are required the same considerations of study design and hypothesis testing for non-inferiority or statistical equivalence (or superiority) as for studies of new pharmaceuticals (and reviewed in this report) will be relevant.

How the non-inferiority margin should be set

Whilst all guidelines discussed in general terms how margins should be set and that there should be clinical input, especially to determine M2, the FDA guideline provided the most detail on actual methods that can be employed to select the margin. The FDA suggested two potential methods:

- *The fixed-margin method*: Effectiveness of the active comparator (M1) is determined from the smallest effect size of the active comparator from historical evidence, determined by the confidence limit in published effectiveness trials of the active comparator closest to the null effect. M1 can be the non-inferiority margin, or clinical judgement can be used to determine the preserved fraction (M2).
- *The synthesis method*: Data from the non-inferiority trials are combined to estimate the effect of the new technology against placebo. M1 is not specified but, rather, M2 is specified as a percentage of the effect against placebo of the existing technology that the new technology must maintain.

Whilst the synthesis method has advantages over the fixed margin method in terms of a potential for a smaller sample size in the non-inferiority trial, the FDA recommend the use of a fixed-margin method as it separates the choice of M1 and M2 from the analyses of the study, allows clarity to clinicians on how big the effect size M1 is and how much they would want to preserve and provides certain statistical advantages in terms of accounting for some differences in trials included in the analysis that generates M1.

Justification of margin

The CONSORT checklist is clear that the margin should always be specified and justified and there was consensus amongst guidelines that the non-inferiority margin should be 'acceptably worse' and that this should be determined through clinical input. Whilst not required in the CONSORT checklist, the majority of guidelines specify that statistical considerations should also be taken into account when determining the margin (such as variation in the comparator effect size M1). There was some variation in how 'acceptably worse' could be defined, with notably the FDA stating that a larger M2 could be acceptable if other factors, such as safety and ease of administration, were a benefit of the new technology. However, the EMA states that, in such cases, trial evidence should be presented that show non-inferiority for efficacy and superiority for other outcomes.

Who should be included in the analysis

All guidelines highlighted that, for non-inferiority and equivalence trials, intention to treat (ITT) analysis does not lead to a conservative estimate as is the case for superiority trials. Instead, an ITT analysis can mean that the effect of the existing technology is 'watered down', resulting in a failure to correctly identify that the new technology is inferior. However, a per protocol (PP) analysis can introduce bias in the same way as a superiority trial. Guidelines therefore suggest actions be taken to minimise protocol violations and drop outs within the trial and that both ITT and PP analyses should be reported.

How confidence intervals should be used in inference

All guidelines provided recommendations on whether one or two-sided confidence intervals should be used for non-inferiority trials, although there was consensus that the type 1 error should be set to 0.025 which is the equivalent of a 95% 2 sided interval or 97.5% one-sided interval.

Dealing with uncertainty

Methods for dealing with uncertainty in results from non-inferiority and equivalence trials, outside of ensuring ITT and PP analyses were performed, was largely lacking from guidelines. Only ICH E9 mentioned any additional sensitivity analyses should be performed, to analyse the impact of different ways to handle missing data.

Non-inferiority and clinical equivalence in formal (regulatory and HTA) guidance documents

Agency: US FDA (Nov 2016)

Document: Non-Inferiority (NI) clinical trials to establish effectiveness. Guidance for Industry

Section	Guidance
IIIA (p3-7)	• Null hypothesis for NI studies • Importance of the NI margin
IIIB (p7)	• Reasons for NI design: ethical concerns for placebo use, increasing interest in comparative effectiveness
IIIC (pp7-11)	• NI margin may be based on 'clinically acceptable difference' (M_2) rather than 'active control effect' (M_1)
IIID (pp11-14)	• Need for assay sensitivity - active control performed as expected (evidence of effect of test drug vs placebo) • Need for (a) historical evidence of sensitivity to drug effects (HESDE); (b) similarity of trial to historic trials; and (c) good study quality (poor quality may bias results towards non-inferiority)
IIIE (pp14-15)	• Analytical approaches include the fixed margin method and the synthesis method • Caution with sequential testing of different hypotheses for multiple study endpoints, to control Type 1 error rate. Nevertheless a planned NI trial may generally be used to test for superiority (p31)

Agency: EMEA (Jan 2006)

Document: Guideline on the choice of the non-inferiority margin

Section	Guidance
Intro (p3)	• Possible rationale for NI studies: (a) essential similarity but bioequivalence studies not possible (modified release, topical preparations); (b) potential safety advantage but

	requires efficacy comparison for risk-benefit assessment; (c) direct comparison against active comparator needed for assessing risk-benefit; (d) loss of efficacy compared to active comparator is unacceptable; (e) use of placebo arm not possible in disease area
2 (pp4-5)	• Non-inferiority margins for trials based on combination of statistical reasoning and clinical judgement, independent of statistical power considerations) • Recommend 3 arm trials to establish efficacy in relation to placebo and active comparators
3.2 (pp6-7)	• 2-arm trials: need systematic review of studies comparing reference with placebo. Possible issues: a) selection bias; (b) constant trial designs and clinical practice over time; c) constant effects over time; (d) publication bias
4 (pp8-9)	• NI Margin (delta) for studies showing no important loss of efficacy for test product vs reference cannot be based only on past trials of reference vs placebo • Wider non-inferiority margins for efficacy may be acceptable if there are other advantages (e.g. safety, administration, posology, superiority on secondary efficacy endpoint, etc.)

Agency: EMEA ICD E3 (Jul 1996)

Document: Structure and content of clinical study reports

Section	Guidance
9.2 (p11)	• For studies demonstrating efficacy by showing equivalence (absence of a specified degree of inferiority), problems associated with such study designs should be addressed, so there is a basis for considering the study capable of distinguishing active from inactive therapy
11.4.2.7 (p23)	• For active control studies intended to show equivalence the analysis should show the confidence interval for the comparison between the two agents for critical endpoints and the relation of that interval to the prespecified degree of inferiority that would be considered unacceptable

Agency: EMEA ICD E9 (Sep 1998)

Document: Statistical principles for clinical trials

Section	Guidance
3.3.2 (pp17-18)	• Lack of measure of internal validity in equivalence or NI trials, so external validation is required • NI trials are non-conservative: trial flaws tend to bias conclusion towards equivalence or NI • Need clinical justification of equivalence margins • One-sided statistical tests suitable for NI
3.5 (p20)	• Base sample size for equivalence or NI trials on showing that treatments differ at most by a ‘clinically acceptable’ difference (generally smaller than a 'clinically relevant' difference used to show superiority)
5.2.3 (p26)	• In equivalence or NI trials use of ‘full analysis set‘ (ITT) may not be conservative, needs careful consideration

Agency: ICH E9 (R1) (Nov 2019)

Document: Addendum on estimands and sensitivity analysis in clinical trials

Section	Guidance
A3.4 (pp12-13)	• The considerations informing the construction of estimand to support regulatory decision making based on a non-inferiority or equivalence objective may differ to those for the choice of estimand for a superiority objective. • NI or equivalence trials are not conservative in nature: it is important to minimise the protocol violations and deviations, non-adherence and study withdrawals, and the result of the Full Analysis Set should be considered very seriously

Agency: EMEA ICH E10 (Jan 2001)

Document: Choice of control group in clinical trials

Section	Guidance
1.4.1 (p9)	• 'Equivalence' trials designed to show similar efficacy to a standard agent are often actually NI trials, aiming to show new drug is not less effective than a control by more than a defined amount (the margin)
1.4.3 (p10)	• In equivalence or NI trials a fair effectiveness comparison with control is needed: design aspects that could unfairly favour one treatment are choice of dose or patient population and selection and timing of endpoints
1.5.1 (p11)	• Assay sensitivity in NI or equivalence trials is determined from historical evidence of sensitivity to drug effects, and evidence of appropriate trial conduct (and study conduct was similar to previous trials)
1.5.1.1 (pp12-13)	• Determining the margin in an NI trial based on statistical reasoning and clinical judgment. It should reflect uncertainties in the evidence and be suitably conservative • In practice, the NI margin chosen will usually be smaller than the smallest expected effect size of the active control, to ensure that some clinically acceptable effect size (fraction of control drug effect) is maintained
1.5.1.2 (pp14-15)	• In NI trials there may be a weaker stimulus to ensure study quality, which will help ensure that differences will be detected (demonstrate assay sensitivity) • Trial conduct should be reviewed for factors that might (a) obscure differences between treatments and (b) make the trial different from those on which NI margin was based
2.4.7.2 (p25)	• Because the choice of the margin in NI trials generally needs to be conservative, sample sizes may be large
2.5 (pp25-28)	• Inability to control bias is the major limitation of externally controlled trials, often sufficient to make the design unsuitable • Potential persuasiveness of findings from such trials depends on obtaining higher much levels of statistical significance and larger estimated differences

	• Matching on selection criteria or adjustments to account for population differences should be specified prior to selection of control and performance of the study

Agency: EMEA (Jul 2000)

Document: Points to consider on switching between superiority and non-inferiority

Section	Guidance
II.2-3 (pp3-4)	• In an equivalence trial for the two treatments are to be declared equivalent, the two-sided 95% confidence interval (defining the range of plausible differences between the two treatments) should lie entirely within the interval -Δ to +Δ (margin). In an NI trial the two-sided 95% confidence interval should lie entirely to the right of the value -Δ.
IV.1.4 (p6)	• In an NI trial, the full analysis set and the PP analysis set have equal importance and their use should lead to similar conclusions for a robust interpretation • Switching the objective of a trial from NI to superiority is feasible provided (a) the trial has been properly designed and carried out in accordance with the strict requirements of a non-inferiority trial; (b) Actual p-values for superiority are presented to allow independent assessment of the strength of the evidence; (c) Analysis according to the intention-to-treat principle is given greatest emphasis
IV.2.5 (p8-9)	• In any superiority trial where NI may be an acceptable outcome (for licensing purposes), a non-inferiority margin should be specified in the protocol to avoid difficulties arising from later selection • Switching the objective of a trial from superiority to NI may be feasible provided (a) the NI margin was pre-defined or can be justified post-hoc (which is difficult); (b) analysis according to ITT and PP, showing confidence intervals and p-values for the null hypothesis of inferiority, gives similar findings; (c) trial was properly designed and carried out according to the strict requirements of a NI trial; (d) sensitivity of trial is high enough to

	ensure that it can detect relevant differences if they exist; (e) there is direct or indirect evidence that the control treatment is showing its usual level of efficacy
VII (p10)	• Switching to wider acceptable margins when examining study results is generally problematical, however data that satisfy narrower equivalence margins may be safely interpreted as such

Agency: MHRA (Sep 2017)

Document: Clinical invest-igations of medical devices - statistical considerations

Section	Guidance
2.3 (p8)	• Tests for non-inferiority, which are one-sided, would typically be performed at the 2.5% level

Agency: EU MDR (Apr 2017)

Document: Medical device regulation 2017/745

Section	Guidance
Article 61 (pp55-56) and Annex XIV (pp168-9)	• Clinical evaluations may be based on clinical data relating to a device to which equivalence can be demonstrated. Equivalence covers 1. Technical aspects: similar design, conditions of use, similar specifications and properties, deployment methods and principles of operation; 2 Biological aspects (where appropriate): same materials in contact with same human tissues (similar kind/duration of contact), similar release characteristics of substances. 3. Clinical aspects: same clinical condition or purpose (similar severity and stage of disease), same body site, similar population (same kind of user), similar relevant critical performance for the expected clinical effect.

Agency: FDA 510(k) (Jul 2014)

Document: Guidance (1766) on evaluating equivalence of medical devices

Section	Guidance

Appendix A (p27)	• Assessment of 'substantial equivalence' with predicate devices is based on determinations of 1. legal marketing status of predicate, 2. same intended use, 3. same technological characteristics. Where technological characteristics are different an assessment is required of whether safety or effectiveness issues are raised, and if so the acceptability of methods to investigate these and the results ('clinical data') of these methods • Valid scientific evidence is defined in 21 CFR 860.7(c)(2): '*from well-controlled investigations, partially controlled studies, studies and objective trials without matched controls, well-documented case histories conducted by qualified experts, and reports of significant human experience with a marketed device, from which it can fairly and responsibly be concluded by qualified experts that there is reasonable assurance of the safety and effectiveness of a device under its conditions of use'*

Agency: EUNetHTA (2015)

Document: Internal validity of randomised controlled trials

Section	Guidance
2.2.1 (p11)	• In a superiority trial a replacement strategy (for ITT analysis) which diminishes group differences (e.g. by assigning all lost patients to 'failures' or 'successes') may lead to a conservative estimate (in favour of the null hypothesis), while the same strategy may lead to an anti-conservative estimate (in favour of the alternative hypothesis) in NI or equivalence trials

Agency: IQWiG (Jul 2015)

Document: General Methods Version 5

Section	Guidance
9.3.6 (p178-9)	• In practice, the requirement is generally not a demonstration of exact equivalence, but that the difference between 2 groups is "at most irrelevant". This depends on first defining what an irrelevant difference is (i.e. specifying an equivalence range). The range can be two-

	sided, resulting in an equivalence interval, or one-sided in terms of an “at most irrelevant difference” or “at most irrelevant inferiority” (referred to as a “non-inferiority hypothesis”) • A frequently used technique to analyse equivalence studies is the confidence interval approach, where equivalence is demonstrated if the confidence interval lies entirely within the equivalence range, defined a priori. To maintain the level of $\alpha = 0.05$, calculating a 90% confidence interval is sufficient, but generally 95% confidence intervals are used, to follow the international approach • Caution is necessary in the evaluation of equivalence studies, which show specific methodological problems: (a) it is often difficult to provide meaningful definitions of equivalence ranges; (b) usual study design criteria, such as randomization and blinding, no longer sufficiently protecting from bias; (c) the ITT principle should be applied carefully, as inappropriate use may falsely indicate equivalence

Section 3: Literature review

The regulatory information identified in the previous section provided substantial background to developing recommendations on how non-inferiority and equivalence studies should be assessed for their quality. Therefore, the focus of the pragmatic literature review was to identify published literature on how non-inferiority and equivalence trials are being undertaken and reported in practice and on specific methodology issues related to diagnostics and indirect treatment comparisons.

A preliminary search of the literature in PubMed highlighted that the volume of literature on methods employed in non-inferiority and equivalence trials and the quality of such studies was substantial, and that a pragmatic approach to literature identification was required. The preliminary search identified a number of systematic reviews of methods employed, reporting and analysis in non-inferiority and equivalence trials as well as studies that summarised approaches and challenges to undertaking non-inferiority and equivalence trials. Many of the studies provided the same information and so, to make the review tractable, a recent systematic review on non-inferiority and equivalence trials (Althunian 2017) was used as the basis for this pragmatic review, with studies published from 2010 until the end of 2019 only included if they provided additional information on specific methods or challenges with methods not covered in the review.

For issues related to diagnostics, only one study published after 2010 was identified that summarised the approach to non-inferiority testing whilst a second provided an alternative approach to testing sensitivity and specificity jointly.

One study published after 2010 reported how indirect treatment comparisons could be used in non-inferiority testing where no head to head trial existed.

3.1 QUALITY OF NON-INFERIORITY/EQUIVALENCE STUDIES

Althunian (2017) conducted a systematic review to assess whether non-inferiority studies were adequately reporting the non-inferiority margin and the methods that were being employed to determine the margin. The study in part was driven by an extension to the 2010 CONSORT checklist for reporting of non-inferiority and equivalence trials which stated that the non-inferiority margin (see Appendix A) and methods to determine the margin should be reported was published in 2012 (Piaggio 2012), and also driven by the FDA guidelines for industry published in 2016 on how non-inferiority studies should be conducted.

The literature searches identified 273 articles on randomised non-inferiority trials from 1966 to 2015 to understand how non-inferiority margins were defined and reported. Only trials that could influence pharmaceutical regulatory decisions were included, so only blinded, randomised, controlled drug trials were eligible for the review.

Within the review, Althunian described the main ways that non-inferior margins can be determined as well as providing information on less common ways that they identified from studies included in the review. The methods to determine this margin discussed by Althunian and identified in his literature review were:

- Based upon historical evidence of M1:
  - *The point effect method*: Effectiveness of the active comparator is determined from the point estimate of effectiveness of the active comparator from historical evidence (either meta-analyses or a pivotal trial). Clinical judgement determines the fraction of the effectiveness of the active comparator that must be maintained by the trial drug (the 'preserved fraction')
  - *The fixed-margin method*
  - *The synthesis method*
- Expert opinion: Clinical opinion determines the efficacy of the active comparator and the preserved fraction based upon evidence from literature
- Historical margins: The margin was chosen based upon other non-inferiority trials with the same indication
- Regulatory consultation/guideline: The margin is determined by guidelines or recommendations from regulatory bodies
- Efficacy of the trial drug: The margin is based upon the efficacy of the drug itself based upon previous trials.

In both the point effect and fixed-margin methods, Althunian states that clinical judgement will weigh what loss in effectiveness from the trial drug is acceptable against gains in other advantages from the trial drug such as cost or safety. This statement is based directly on the FDA guidelines published in 2016 on how to conduct non-inferiority trials.

Althunian found that 17.2% of trials had used historical evidence to determine non-inferiority margins. A wide range of preserved fractions were reported (0 to 85%) with a median of 50%. The rational for the margin was not reported in over 50% of studies. The authors concluded that the methods for defining margins is not well reported and where methods are reported only a small proportion used historical evidence as preferred by the FDA.

Aupiais (2018) undertook a review of non-inferiority and equivalence trials in the context of paediatrics. Whilst finding similar issues of poor quality of studies and reporting to Althunian, Aupias highlighted that, in non-inferiority trials, sample sizes have to be larger than superiority trials which could be challenging in paediatrics where interventions are often in small and vulnerable populations. They highlight that this may have led to the rareness of the condition influencing the non-inferiority margin as where sample size is small, non-inferiority margins tended to be larger.

Rehal (2016) conducted a systematic review of non-inferiority trials again looking at quality indicators in 168 trials between 2010 and 2015. As with Apuiais and Althunian they found quality overall to be poor and in addition to the quality issues identified in these reviews Rehal reported that the majority of studies did not undertake both a per protocol and ITT analysis and the methods of dealing with loss to follow up were poorly described. In only 57% of studies did they find consistency between the type 1 error rate although only 27% used a one sided 5% significance level rather than 2.5%.

Galdstone (2014) undertook a review of non-inferiority margins in clinical trials and estimated the likelihood of degradation of effect based upon the likelihood a treatment is actually harmful if it was declared non-inferior. Based on a detailed statistical assessment of 112 non-inferiority trials, the authors found that the median likelihood of degradation was 56% and only two fifths had a likelihood less than 50%. On the back of these findings, the authors propose a third margin on top of M1 and M2, a margin that means there is less than a 50% chance of a degradation of effect.

3.2 METHODOLOGICAL ISSUES NOT COVERED IN PUBLISHED GUIDELINES

3.2.1 Non-inferiority or equivalence in diagnostic testing

Ahn (2013) looked at non-inferiority and equivalence testing in radiology and had a focus on diagnostic non-inferiority testing. The paper sets out a method to undertake non-inferiority and equivalence testing that is the same for other technologies but looking at sensitivity or specificity of a test as binary effectiveness outcomes. Ahn discusses non-inferiority testing with receiver operating characteristic (ROC) curves and points out that methods have been developed to statistically test for differences between curves, although it is not made clear how a non-inferiority margin would be established in such a test. Shan (2016) describes statistical approaches in testing for non-inferiority or equivalence in diagnostic tests where sensitivity and specificity are jointly tested for non-inferiority as opposed to what, as they describe, has traditionally been the case of testing sensitivity and specificity separately. The paper

is methodological in nature and summarises the statistical approaches that have been tried and proposed to jointly test for noninferiority between two diagnostic technologies or procedures. Whilst the approaches discussed by Shen are statistically interesting, the arguments for testing for sensitivity and specificity jointly as opposed to separately are not made on statistical or interpretation grounds.

3.2.2 Indirect treatment comparison

Julious (2011) undertook a statistical analysis to highlight how indirect treatment comparison (ITC) could be used to undertake non-inferiority testing between two interventions where no direct head to head trial exists. Julius first points out that this requires assay sensitivity to exist for the placebo controlled trials that actually established effectiveness of the comparator(s), that bias is minimised by ensuring that the efficacy endpoints and patient populations are similar across trials and that the effect seen using an intervention in one trial is consistent with the effect of the same intervention seen in another trial. Julius highlights that placebo creep can be a particular problem in ITCs, and so a more conservative M2 may be required than in a direct head to head trial. Julius also identifies an issue with ITC where due to the difference in time between trials included in the ITC there may be a shift in population characteristics or other treatments offered, available and used that mean that populations within the trials in the ITC may not be comparable.

Section 4: Interviews with stakeholders

Consultations were held with NICE staff, committee members and academic staff to understand whether the findings from the international guidelines and from the literature search resonated, and to understand the potential implications of findings on the TA and MTEP processes. Given the lack of evidence found on how to establish non-inferiority or equivalence with a weak evidence base in terms of evidence from clinical trials, as often is the case for MTEP, views were also collected on how non-inferiority or equivalence could be established in the absence of robust clinical trial evidence.

Overall the views of those spoken to aligned with what had been identified in published guidelines and studies. It was felt that, with the large literature on non-inferiority and clinical equivalence studies (powering, determining margins, study quality etc.) that exists, NICE should not attempt to deviate from what has been established in this area. There was also a call for clarity in terminology: 'non-inferiority' and 'clinical equivalence' have technical meanings associated with trial design/hypotheses. A view was expressed that terms such as ‘equivalence' or ‘similarity' are being used more loosely in NICE appraisal contexts, particularly in MTEP, without clear guidelines on what these terms mean and how they should be assessed.

For MTEP, an Assessment Group consultee felt that appraisals were increasingly leading to research recommendations for technology submission due to uncertainty over non-inferiority or equivalence to gather more evidence rather than recommendations to adopt or not adopt a technology.

One committee consultee stated that there was a need for transparency of process to develop margins, study quality and an analysis plan, so that these are open to review/critique by analysts (ERG/AG) and clinicians.

In terms of diagnostics, whilst no consultee raised that diagnostics should be considered differently in terms of non-inferiority or equivalence, concerns were raised by some consultees that joint testing of sensitivity and specificity was being seen in submissions and there was uncertainty around how this should be interpreted. In cases where a diagnostic test is based upon categorical variables, a medical statistician suggested that Cohen’s kappa is generally used but has limitations. For example, unless agreement is perfect, if one or two categories are small compared to the other, kappa will usually be small, no matter how good the agreement is. This means that it may be difficult to show non-inferiority or

equivalence with Cohen's kappa. For non-categorical diagnostic tests, either Passing and Bablok regression or Bland-Altman plots were considered acceptable approaches for testing of non-inferiority or equivalence provided that the hypothesis tests were correctly specified. Importantly, the statistician made it clear that, whatever approach was used, as is the case in all non-inferiority and equivalence testing, the non-inferiority margin has to be set a priori and clearly justified.

A consensus opinion across consultees was that, before an appraisal of non-inferiority or equivalence trial evidence, there should be a clear, evidence-based rationale provided as to why non-inferiority or equivalence should be assumed, and that non-inferiority and 'equivalence' should be for of a range of study endpoints: safety, HRQoL etc. as well as efficacy. This was recognised as challenging in practice, because most non-inferiority or equivalence studies only considered one efficacy end point. Building on this, it was felt by some consultees that the pathway in which the technology sits and the potential impact on all direct aspects of patient outcome from the technology needs to be fully understood before even considering trial evidence.

Several consultees expressed the view that there may be no need to have robust clinical evidence of non-inferiority, or there may be flexibility in what non-inferiority margin may be considered acceptable, depending on:

- How complicated a pathway is
- The impact a new technology will have on the pathway
- The impact on patients if a new technology is actually inferior
- The evidence on why a new technology should be at worst non-inferior

Section 5: Discussion

The evidence gathered through published guidelines on non-inferiority shows there is a clear body of evidence on how non-inferiority or equivalence trials should be conducted, including on how non-inferiority margins should be set and interpreted. The guidelines and evidence from literature reviews is clear that non-inferiority and equivalence trials are not an easier route to undertaking a trial compared to a superiority trial. The opposite is true: non-inferiority and equivalence trials are more complex in design, in delivery, in analysis and in interpretation of results.

Published evidence found during this process focussed on the assessment of non-inferiority and equivalence being made from a clinical trial. Nothing in the research for this paper suggested that different rules for non-inferiority or equivalence should apply to trials of medical devices going through MTEP as opposed to other technologies being assessed through other routes. However, consultees did point out that the evidence accepted for decision making for an MTEP submission is often of lower quality than the TA process, frequently being from non-randomised single arm observational studies reflecting the difficulty that device manufacturers can have in undertaking RCTs and also reflecting the fact that recommendations from MTEP are not binding on NHS Trusts and CCGs as opposed to a TA recommendation. However, from a population health perspective it will not be beneficial for MTEP to be recommending devices that are inferior to existing technology on the belief that they are non-inferior or equivalent, even if the recommendation is not binding.

This presents a dichotomy. On one hand, there is no evidential reason to assess the non-inferiority or equivalence of a device differently to a drug and a public health interest to expect the same level of evidence but, in practice, devices are already assessed differently with a lower threshold of evidence expected than for a drug. However, evidence from consultees makes it clear that more evidence than just a clinical trial should feed into any health technology assessment where non-inferiority or equivalence is being asserted. This applies whether the technology is going through MTEP or the TA process, and the final judgement of non-inferiority or equivalence has to be made by a committee taking on board the totality of evidence.

Specifically, academic consultees felt that an assessment of the non-inferiority or equivalence should not start with an assessment of trial evidence. Instead, there needs to be a rationale presented as to why a new technology is likely to produce patient outcomes that are no worse than existing technology. This rationale would be presented in conjunction with an assessment of where the technology sits in the patient pathway and on what outcomes and aspects of the pathway the technology is

likely to impact and the extent of the impact if the technology is actually inferior. It may be at this stage a decision is made by a committee that no trial evidence for inferiority is required, although evidence of the technology in use in practice and the views of clinicians and/or patients on the use of the technology may be sought.

Where a committee decides that evidence from a clinical non-inferiority or equivalence trial is required, then the committee can then appraise such evidence ensuring non-inferiority or equivalence trial evidence for all patient outcomes that the committee feels could impact on the cost-effectiveness of the new technology.

The choice of inferiority margin requires particular consideration by the committee as not only is it the foundation of the statistical underpinnings of the trial (non-inferiority or equivalence) but it also can explicitly define what is considered to be an acceptable loss of effectiveness where a preserved fraction approach is used (M2). An argument can be made that the preserved fraction approach itself is not suitable for a HTA process, as any loss in effectiveness will have an impact on patient outcomes that is important in an assessment of cost-effectiveness. In any case, it is doubtful that any margin that allows a lower preserved fraction because of other potential benefits of a new technology, as suggested is acceptable by the FDA, would be acceptable for HTA decision making as this implies that the margin actually represents a difference in patient outcome. There is, therefore, a rationale that for HTA the correct approach to setting a margin should always follow the fixed-margin method where M1 is determined by the lower bound of the current technologies effect against placebo, preferably identified from published trials, and M2 should always be set to zero. In practice, the committee will always have to appraise how the non-inferiority bounds have been set and be satisfied that the bound genuinely represents what the committee considers to be a clinically insignificant difference.

In cases where no non-inferiority trial exists but evidence from a superiority trial is available, there is no methodological reason why the data from the superiority trial cannot be used to test for non-inferiority or equivalence – it is a case of how inference from the trial data is drawn.
A superiority test requires the null hypothesis that a new technology is the same as an existing technology to be rejected which becomes increasingly less likely as sample size reduces. Therefore, if the failure to reject the null hypothesis of a superiority test is to be used as evidence that a new technology is the same as an existing technology, non-inferiority and equivalence trials would be replaced with very small superiority trials. As such, in no circumstances should a failure to provide statistical evidence of superiority for a new technology over an existing technology be used as statistical evidence that the new technology is the same as, similar to or equivalent to an existing technology.

Whilst failure to show superiority in a superiority trial should not be interpreted as non-inferiority or equivalence, the data from the superiority trial can be used to statistically test for non-inferiority or equivalence as if the data came from a non-inferiority trial. However, to avoid bias non-inferiority margins should be set before a trial commences which us unlikely to have been the case where the data is drawn from a superiority trial. In situations where margins have been set after the trial, particular scrutiny by a NICE committee on whether the margins truly represent a clinically insignificant difference will be required. In addition, sample sizes in a superiority trial may be too small to be properly powered to test for non-inferiority or equivalence and less care may have been taken to minimise drop out and protocol violations. Given that for these reasons statistical evidence from a superiority trial on non-inferiority or equivalence is unlikely to be as robust as evidence from a non-inferiority trial, submissions claiming non-inferiority or equivalence for a new technology based upon superiority trial data are likely to be more dependent on other non-trial evidence to support the claim.

In appraising the quality of non-inferiority or equivalence trials, or superiority trials repurposed to test for non-inferiority or equivalence, the starting point should be the CONSORT checklist for non-inferiority and equivalence trials. Quality of non-inferiority or equivalence trials can be seen to be even more important than for superiority trials because a poor quality trial can lead to the conclusion that an inferior technology is actually non-inferior or equivalent. It is, therefore, concerning that the majority of non-inferiority or equivalence trials identified in systematic reviews have significant quality issues. It is likely that superiority trials will perform poorly against aspects of the checklist, especially around areas of sample size calculation, definition of the non-inferiority margin and patient flow. Any quality issues raised from the CONSORT checklist should be an indication to a committee that the trial evidence presented is likely to be unsuitable for decision making.

Whilst the CONSORT checklist favours randomised controlled trials for establishing non-inferiority, no compelling reason arose from the published literature and guidelines or from consultees as to why, for observational studies, a comparator could not be drawn from case-control evidence, a retrospective cohort or the use of patient level data through a matched adjusted indirect comparison (MAIC). Indeed, the FDA guideline on non-inferiority studies makes it clear that observational studies can establish the active control treatment effect, M1, when other data on effect is not available so observational data per se are acceptable as part of the non-inferiority process. However, observational studies by their nature are more prone to bias than RCTs and depending on the assessment of the pathway described above a committee may decide that RCT evidence is required.

The methodological literature on diagnostics compared to pharmaceuticals of statistical testing was limited but in large part non-inferiority and equivalence for diagnostics can be seen as no different as is the case for other technologies. The only difference of note would be that a non-inferiority margin for sensitivity and specificity will be required or on correlation coefficients or the slopes and intercepts of regression analysis. Whilst methods for joint testing of sensitivity and specificity of have been described in the literature, or for differences in ROC curves the reasons for doing this may be limited to some advantages in sample size for the trial but potentially this could be at the expense in clarity of what the non-inferiority bound represents. In addition, non-inferiority or equivalence assessed through ROC curves may not be acceptable in HTA where a thorough understanding of the consequences of false negatives and false positives are required to compare outcomes from two diagnostic tests.

Section 6: Recommendations

1. The terminology in NICE guidelines requires more precision around non-inferiority and equivalence. The word ‘similar’ should not be used and the terms ‘non-inferiority’ and ‘equivalence’ should not be used interchangeably.

2. Before looking at non-inferiority evidence, the patient pathway in which the technology will be placed and all aspects of the pathway where the technology will impact on patient outcomes and costs needs to be determined. Evidence on non-inferiority needs to be provided against all aspects of the patient pathway where the technology could change patient outcomes negatively that would have a material impact on costs or QALYs. Where evidence against any of these aspects suggests that an assessed technology is inferior to existing technology, or evidence is absent or inconclusive, a conclusion of non-inferiority or equivalence should not be drawn.

3. Assessment of non-inferiority and equivalence should not be based on trial evidence alone, but should first be based upon an assessment of the technological, biological and/or pharmacokinetic reasons to support the assumption that a new technology would not be inferior to the existing comparator. Without this rationale, the reasons for undertaking a non-inferiority trial should be challenged. For devices or diagnostics, where the impact of the technology may be on only one or a small number of patient outcomes, then the innovations in the technology itself over existing technologies coupled with evidence from how the technology has been successfully introduced in clinical practice may be sufficient to conclude non-inferiority or equivalence without a controlled trial.

4. The assessment of evidence presented to NICE on non-inferiority or equivalence should be undertaken with the same, if not greater, rigour as an assessment of superiority in the TA or MTEP assessment processes. An assessment of non-inferiority or equivalence should not be seen as an easier evidential hurdle than superiority.

5. In circumstances where a non-inferiority or equivalence study is required, a poor quality and/or under powered study is likely to lead to inconclusive results or incorrectly conclude the new technology is, or is not, inferior to an existing technology. Observational studies with a comparator drawn, for example, from the population through a MAIC may be used where an RCT

was not possible but the methods used to undertake the analysis must be rigorous and transparent.

6. Analysis of the findings from a non-inferiority or equivalence trial has to ensure that steps are taken to minimise bias arising from issues such as from loss to follow up (such as using ITT analysis), do not increase the likelihood of concluding that treatments are equivalent or the study treatment is non-inferior to the comparator treatment by reducing the effect size of the comparator treatment. Acceptance of poor quality evidence of non-inferiority or equivalence for treatments that are not actually non-inferior or equivalent will result in a degradation of effect in NHS funded treatments with increasingly ineffective technologies being approved until a technology is recommended that is no more effective than placebo.

7. Although they are similar in some respects, equivalence and non-inferiority studies are not the same. Both the similarities and the differences in the interpretation of the statistics generated by an equivalence and non-inferiority trial needs to be understood and recognised to correctly interpret findings and ensure the conclusions drawn on trial evidence correctly reflect trial design. Statistical inference from non-inferiority and equivalence trials should be based upon an assessment of confidence intervals and whether they do or do not cross the non-inferiority/equivalence margins. In cases where either a clinical equivalence or non-inferiority study shows that trial technology is statistically inferior (compared to the non-inferior bound) then clinical equivalence or non-inferiority can be ruled out.

8. Failure to reject the null hypothesis of no difference between a new and existing technology from a superiority trial should in no circumstances be interpreted as evidence that the new technology is non-inferior or equivalent to an existing technology. Data from superiority trials can be repurposed to test for non-inferiority but particular scrutiny will be required on how post trial non-inferiority margins have been set given the bias that could arise from not setting the margins pre-trial. Non-trial evidence on non-inferiority or equivalence will likely be more important to the assessment of non-inferiority or equivalence than may be the case if a non-inferiority or equivalence trial had been undertaken. This should be clearly reflected in the characterisation and assessment of uncertainty in the economic evaluation.

9. Reporting of non-inferiority or equivalence trials should be assessed against the CONSORT checklist for such studies published in 2012. Given the importance of quality of non-inferiority and equivalence trials, studies that do not provide all the information required on the CONSORT checklist

should be considered low quality. Where data to test for non-inferiority or equivalence is taken from a superiority trial, the CONSORT checklist for non-inferiority or equivalence trials should also be used, noting that the trial may perform poorly against elements on the checklist relating to setting of non-inferiority margins, sample size and participant flow.

10. Where evidence synthesis across trials is performed in the form of a meta-analysis or some form of an indirect treatment comparison, superiority, non-inferiority and equivalence trials may be included in the synthesis, but a test for non-inferiority or equivalence between the new technology and comparator technologies will still require a non-inferiority margin to be defined.

11. The method of determining a non-inferiority margin should be based upon a historical appraisal of evidence on the effectiveness of the active comparator, preferably through a statistical appraisal of variance in effectiveness. The preserved fraction of effectiveness, if a factor in the margin, should be determined by clinicians. For example, the EVEREST II trial for the MitraClip device versus mitral valve surgery for mitral regurgitation used clinical judgement to determine an acceptable inferiority margin for MitraClip for the primary outcome (death, future surgery or continued mitral regurgitation) over the superiority of mitral valve surgery over placebo.

12. Regardless of the method to derive the margin, the margin needs to have been set so that it reflects no clinically meaningful change in patient outcomes. There should be no trade-off in setting the margin or preserved fraction against perceived other benefits of the new technology. It may be that the preserved fraction is considered too small or too large or it may be considered by a committee that any preserved fraction is unacceptable. In such cases, the margin will need to be redefined based upon what is considered a true clinically insignificant change in outcomes which may in turn mean that the sample size is insufficiently powered to detect whether the new technology is actually non-inferior.

13. In assessing non-inferiority and equivalence evidence on diagnostic tests, the same considerations on equivalence and non-inferiority apply as for other technologies. However, evidence on the equivalence and non-inferiority for both sensitivity and specificity or tests will be required. The only circumstances where this would not be the case would be where either specificity or sensitivity had no bearing on outcomes for patients beyond incurring potential further test costs. Statistical methods are available to

report the non-inferiority of sensitivity and specificity jointly. Unless a compelling reason can be presented as to why joint testing should be undertaken and, in the case of an assessment of a ROC curve, the non-inferiority margin can be transparently described and justified in terms of false negatives and false positives, separate tests of sensitivity and specificity are by design more transparent and easier to interpret and should be preferred to joint tests.

14. For non-inferiority or equivalence testing of diagnostic technologies with categorical or continuous results, it would be acceptable to use for categorical tests Cohen's kappa or for continuous tests Passing-Bablok regression or Bland and Altman plots. However, as in all other cases of non-inferiority and equivalence testing the non-inferiority margins applied must represent a clinically acceptable difference and be justified.

15. Due consideration should be given to the uncertainty in non-inferiority or equivalence evidence presented. Both an ITT and PP analysis of non-inferiority and equivalence trials should be performed and the impact of different methods of dealing with missing data should be reported.

References

Ahn, S; Park, SH; Lee, KH (2013). How to demonstrate similiarity by using non-inferiority and equivalence statistical testing in radiology research. Radiology 267, 2

Althunian, T. A.; de Boer, A.; Klungel, O. H.; Insani, W. N.; Groenwold, R. H. (2017), Methods of defining the non-inferiority margin in randomized, double-blind controlled trials: a systematic review, Trials 18, 1:107

Aupiais, C.; Zohar, S.; Taverny, G.; Le Roux, E.; Boulkedid, R.; Alberti, C. (2018), Exploring how non-inferiority and equivalence are assessed in paediatrics: a systematic review, Arch Dis Child 103, 11:1067-1075

Dunn, D. T.; Copas, A. J.; Brocklehurst, P. (2018), Superiority and non-inferiority: two sides of the same coin?, Trials 19, 1:499

Gladstone, B. P.; Vach, W. (2014), Choice of non-inferiority (NI) margins does not protect against degradation of treatment effects on an average--an observational study of registered and published NI trials, PLoS One 9, 7:e103616

Julious, S. A. (2011), The ABC of non-inferiority margin setting from indirect comparisons, Pharm Stat 10, 5:448-53

Mauri, L., Garg, P., Massaro, J. M., Foster, E., Glower, D., Mehoudar, P., . . . Feldman, T. (2010). The EVEREST II trial: Design and rationale for a randomized study of the evalve mitraclip system compared with mitral valve surgery for mitral regurgitation. The American Heart Journal, 160(1), 23-29.

Rehal, S.; Morris, T. P.; Fielding, K.; Carpenter, J. R.; Phillips, P. P. (2016), Non-inferiority trials: are they inferior? A systematic review of reporting in major medical journals, BMJ Open 6, 10:e012594

Shan, G., Amei, A., & Young, D. (2015). Efficient Noninferiority Testing Procedures for Simultaneously Assessing Sensitivity and Specificity of Two Diagnostic Tests. Computational and mathematical methods in medicine, 2015, 128930. doi:10.1155/2015/128930

Committee for Medicinal Products for Human Use (CHMP) (2006). “Guideline on the Choice of the Non-Inferiority Margin.” Statistics in Medicine 25, 1628–1638.

Committee for Proprietary Medicinal Products (CPMP) (2000). “Points to consider on switching between superiority and non-inferiority. CPMP/EQP/482/99

EMEA (1996). ICH Topic E3: Structure and content of clinical study reports

International Conference on Harmonisation: Choice of Control Group and Related Issues in Clinical Trials (ICH E10), Food and Drug Administration, DHHS, July 2000.

International Conference on Harmonisation: Statistical Principles for Clinical Trials (ICH E-9), Food and Drug Administration, DHHS, 1998.

IQWIG (2015). General Methods 5.0

MHRA (2017) Statistical investigations of medical devices – statistical considerations

Official Journal of the European Union (2017). Regulation (EU) 2017/745 of the European Parliament and of the Council Annex 14

Piaggio G, Elbourne DR, Pocock SJ, Evans SJW, Altman DG, CONSORT Group FT. Reporting of Noninferiority and Equivalence Randomized Trials: Extension of the CONSORT 2010 Statement. JAMA. 2012;308(24):2594–2604. doi:10.1001/jama.2012.87802

U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER) (2016). Non-Inferiority Clinical Trials to Establish Effectiveness: Guidance for Industry

U.S. Department of Health and Human Services, Food and Drug Administration (2014). The 510(k) Program: Evaluating Substantial Equivalence in Premarket Notifications [510(k)] Guidance for Industry and Food and Drug Administration Staff

Appendix A

CONSORT Checklist for Non-inferiority and Equivalence Trials

Items to include when reporting a non-inferiority or equivalence randomized trial

PAPER SECTION **And topic**	**Item**	**Descriptor**	**Reported on Page #**
TITLE AND ABSTRACT	1	How participants were allocated to interventions (*e.g.*, "random allocation", "randomized", or "randomly assigned"), *specifying that the trial is a non-inferiority or equivalence trial.*	
INTRODUCTION Background	2	Scientific background and explanation of rationale, *including the rationale for using a non-inferiority or equivalence design.*	
METHODS Participants	3	Eligibility criteria for participants *(detailing whether participants in the non-inferiority or equivalence trial are similar to those in any trial(s) that established efficacy of the reference treatment)* and the settings and locations where the data were collected.	
Interventions	4	Precise details of the interventions intended for each group *detailing whether the reference treatment in the non-inferiority or equivalence trial is identical (or very similar) to that in any trial(s) that established efficacy,* and how and when they were actually administered.	
Objectives	5	Specific objectives and hypotheses, *including the hypothesis concerning non-inferiority or equivalence*.	
Outcomes	6	Clearly defined primary and secondary outcome measures *detailing whether the outcomes in the non-inferiority or equivalence trial are identical (or very similar) to those in any trial(s) that established efficacy of the reference treatment* and, when applicable, any methods used to enhance the quality of measurements (*e.g.*, multiple observations, training of assessors).	
Sample size	7	How sample size was determined *detailing whether it was calculated using a non-inferiority or equivalence criterion and specifying the margin of equivalence with the rationale for its choice*. When applicable, explanation of any interim analyses and stopping rules (*and whether related to a non-inferiority or equivalence hypothesis*).	
Randomization -- Sequence generation	8	Method used to generate the random allocation sequence, including details of any restrictions (*e.g.*, blocking, stratification)	
Randomization -- Allocation concealment	9	Method used to implement the random allocation sequence (*e.g.*, numbered containers or central telephone), clarifying whether the sequence was concealed until interventions were assigned.	

Randomization -- Implementation	10	Who generated the allocation sequence, who enrolled participants, and who assigned participants to their groups.	
Blinding (masking)	11	Whether or not participants, those administering the interventions, and those assessing the outcomes were blinded to group assignment. If done, how the success of blinding was evaluated.	
Statistical methods	12	Statistical methods used to compare groups for primary outcome(s), *specifying whether a one or two-sided confidence interval approach was used*. Methods for additional analyses, such as subgroup analyses and adjusted analyses.	
RESULTS Participant flow	13	Flow of participants through each stage (a diagram is strongly recommended). Specifically, for each group report the numbers of participants randomly assigned, receiving intended treatment, completing the study protocol, and analyzed for the primary outcome. Describe protocol deviations from study as planned, together with reasons.	
Recruitment	14	Dates defining the periods of recruitment and follow-up.	
Baseline data	15	Baseline demographic and clinical characteristics of each group.	
Numbers analyzed	16	Number of participants (denominator) in each group included in each analysis and whether the analysis was *"intention-to-treat" and/or alternative analyses were conducted*. State the results in absolute numbers when feasible (*e.g.*, 10/20, not 50%).	
Outcomes and estimation	17	For each primary and secondary outcome, a summary of results for each group, and the estimated effect size and its precision (*e.g.*, 95% confidence interval). *For the outcome(s) for which non-inferiority or equivalence is hypothesized, a figure showing confidence intervals and margins of equivalence may be useful*.	
Ancillary analyses	18	Address multiplicity by reporting any other analyses performed, including subgroup analyses and adjusted analyses, indicating those pre-specified and those exploratory.	
Adverse events	19	All important adverse events or side effects in each intervention group.	
DISCUSSION Interpretation	20	Interpretation of the results, taking into account the *non-inferiority or equivalence hypothesis and any other* study hypotheses, sources of potential bias or imprecision and the dangers associated with multiplicity of analyses and outcomes.	
Generalizability	21	Generalizability (external validity) of the trial findings.	
Overall evidence	22	General interpretation of the results in the context of current evidence.	