Assessing Equivalence and Non-Inferiority

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
http://www.ahrq.gov

This information is distributed solely for the purposes of pre-dissemination peer review. It has not been formally disseminated by the Agency for Healthcare Research and Quality. It does not represent and should not be construed to represent an Agency for Healthcare Research and Quality or Department of Health and Human Services (AHRQ) determination or policy.

Contract No. XXX-XX-XXXX-XX

Prepared by:

AHRQ Publication No. xx-EHCxxx
September 2011

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies and strategies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To improve the scientific rigor of these evidence reports, AHRQ supports empiric research by the EPCs to help understand or improve complex methodologic issues in systematic reviews. These methods research projects are intended to contribute to the research base in and be used to improve the science of systematic reviews. They are not intended to be guidance to the EPC program, although may be considered by EPCs along with other scientific research when determining EPC program methods guidance.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers as well as the health care system as a whole by providing important information to help improve health care quality. The reports undergo peer review prior to their release as a final report.

We welcome comments on this Methods Research Project. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by e-mail to epc@ahrq.hhs.gov.

Carolyn M. Clancy, M.D.
Director
Agency for Healthcare Research and Quality

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Stephanie Chang M.D., M.P.H.
Director, EPC Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Elisabeth Kato, M.D., M.R.P.
Task Order Officer
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Acknowledgements

Contents

Introduction 1
Methods 1
Workgroup Composition 1
Methods Projects 2
Guidance Development 3
Guidance 3
Section 1: Unique Risk of Bias Issues for Trials Designed as EQ-NI 3
Performance 6
Detection 6
Attrition 6
Similarity of EQ-NI Trials to Trials that Established Efficacy of the Active Comparator 7
Completeness of Reporting 8
Section 2: Setting the Reviewer's Minimum Important Difference (MID) 8
Ways to Determine MID 9
MID for Specified Outcomes 10
Section 3: Analytic Foundations for Concluding EQ or NI 12
Whether one treatment was believed in advance to be better on some outcomes 12
Whether the reviewer's MID was pre-specified by the reviewer 13
Whether the reviewer's MID was justified by the reviewer 13
Whether the meta-analytic model was random-effects 13
Whether the confidence interval was narrow enough to rule out an MID 13
Other Aspects of Strength-of-Evidence 14
Whether trials described themselves as EQ-NI trials 15
Bayesian Analysis 15
Non-Meta-Analytic Situations 16
Section 4: Language Considerations When Concluding EQ or NI 18
Summary of Recommendations 20
References 22

Tables

Table 1. Issues Related to Risk of Bias in EQ-NI Trials 5
Table 2. Methods for Identifying Appropriate MID 11
Table 3. Recommended Non-Inferiority Margins in Different Areas of Biomedical Research 11
Table 4. EPC Guidance on Equivalence and Non-Inferiority 20

Figures

Figure 1. Example Scenarios Comparing the Confidence Interval to the MID 14
Figure 2. Bayesian Posterior Distribution Compared to the MID 16

Appendixes

Appendix A. Example MIDs for Specific Clinical Topics

Introduction

This chapter provides guidance to EPCs about several issues on equivalence and non-inferiority. This guidance is urgent for many reasons. First, comparative effectiveness research (CER) involves *comparing active treatments*, and these comparisons often suggest equivalence. What kinds of data permit a valid conclusion of equivalence? As CER receives greater prominence in critical medical decisions, evidence reviewers need clear and consistent guidelines for concluding equivalence. The same is true for individual trialists. However, our perspective is that of a systematic reviewer confronted with multiple trials making the same active comparison.

A second reason for urgency is that the medical literature has seen a recent increase in the number of trials actually defining themselves as "equivalence trials" or "non-inferiority trials." How should evidence reviewers incorporate such trials? Third, the wider field of systematic review has no guidance on equivalence and non-inferiority.[1-3] Some guidance exists from regulatory agencies and academia,[4-15] but these are targeted to individual trialists, not reviewers. Fourth, systematic reviewers vary greatly in their choice of language for concluding equivalence or non-inferiority (e.g., "similar effects," "no evidence of a difference," "evidence of no difference," "evidence does not suggest a difference," "treatment A is not worse," "treatment A is not superior," etc.); this variation is confusing and possibly misleading to users.

Before presenting our methods and guidance, we briefly discuss the difference between "equivalence" trials (EQ) and "non-inferiority" (NI) trials. These trials share the concept of ruling out the possibility of an important effect. They differ, however, in the typical clinical context and the permissible conclusions. Equivalence trials aim to determine whether a new treatment is therapeutically similar to a standard treatment within a predefined margin of equivalence (e.g., an odds ratio from 0.80 to 1.25 is used by the FDA to establish bioequivalence).[16]

In contrast, NI trials are conducted in a clinical context of prior knowledge that a newer treatment (which we call the "test intervention") is superior to an older treatment (which we call the "active comparator") on certain outcomes (e.g., fewer side effects, lower cost, and/or greater convenience). This prior knowledge sets the stage for being willing to accept a small inferiority regarding the effectiveness of the test intervention. Thus, NI trials aim to determine whether the test intervention is *not less effective* than the active comparator by a pre-specified amount.

Despite the contextual differences between EQ and NI trials, many details about their design and analysis are similar (e.g., the pre-specification of a decision threshold). Thus, for ease of exposition, this guidance document refers to them collectively as "EQ-NI" trials, or refers to reviewers' conclusions as "EQ-NI" conclusions. Any areas where systematic reviewers should treat them differently are delineated in the pertinent sections of the guidance.

Methods

Workgroup Composition

The workgroup for this chapter included 13 individuals from seven EPCs and AHRQ. All members of the workgroup had specifically expressed interest in working on the guidance, and many had prior expertise in the analysis and interpretation of equivalence and non-inferiority trials. The project was led by the ECRI Institute EPC. Project leadership involved setting the scope and timeline, scheduling conference calls, devising and assigning subgroups, participating in all subgroups, contributing to the writing of sections of the draft guidance chapter, assembling

documents for group-wide review, and writing the first drafts of the Introduction and Methods sections.

Methods Projects

Prior to developing any guidance, the Lead EPC also performed two methods projects intended to assist the workgroup in clarifying the context, prioritizing the issues, targeting the scope, and summarizing the state-of-the-art. The first project involved a review of 12 existing guidance documents pertaining to equivalence and non-inferiority (see supplemental file, Methods Project 1). These guidance documents (10 from regulatory agencies and two from academia) were all intended for primary researchers designing and interpreting equivalence trials. Major insights from this project were:

- EQ-NI trials are conducted in many contexts, such as (1) placebo-controlled trials are unethical because a proven treatment exists, (2) the advantage of the test intervention (e.g., safety, cost, and/or convenience) may counterbalance the reduced efficacy, or (3) there is a general interest in comparative efficacy or effectiveness, or (4) a general interest in comparative efficacy or effectiveness, or (5) the infeasibility of a superiority trial.
- The guidance unanimously emphasized *a priori* specification of the decision threshold.
- The guidance also emphasized that researchers should *justify* the chosen threshold.
- The guidance was unclear on *how* researchers should determine the threshold. The documents suggested a focus on "clinical" impact, but the specific meaning of this was unclear. The guidance suggested researchers should also consider statistical considerations, how well the active comparator works, historical data, safety, cost, acceptability, adherence, independent expert consensus, and regulatory requirements.
- Regarding risk-of-bias, the documents mentioned nine areas that can contribute to underestimates of the difference between active treatments. Most of these areas can also contribute to overestimates, depending on the specifics of the situation.
- The concerns of regulatory agencies are different from those of systematic reviewers. For example, systematic reviewers are concerned about any direction of bias, whereas regulatory agencies are primarily concerned about bias *in favor of the sponsor's product*. Also, regulatory agencies generally assume that a trial should stand on its own to demonstrate a finding, whereas systematic reviewers view a trial as one in a larger set of trials that, taken together, may or may not demonstrate a consistent finding.

The second methods project involved assessing methodology used within published systematic reviews that contain conclusions that could be interpreted as conclusions of EQ-NI between two or more treatments (see supplemental file, Methods Project 2). This project focused on methodology related to the following areas within systematic reviews: assessing risk of bias, defining the minimum important difference (MID), analytical basis for drawing conclusions of EQ-NI, and wording of conclusions of EQ-NI. Major insights from this project were:

- Authors of reviews rarely address how risk of bias factors known to impact studies of EQ or NI differently than studies of superiority will be assessed and taken into account when drawing conclusions of EQ-NI.
- Authors of reviews rarely define or use a decision threshold such as a MID.

- Authors of reviews rarely pre-specify how they will handle findings of no difference or similarity. In many of the reviews assessed, meta-analytic findings of no statistically significant difference were naively interpreted as demonstrating equivalence.
- Authors of reviews typically use indirect language (e.g., "There is no evidence that [video assisted thoracoscopic surgery] VATs is more effective than fibrinolytic treatment.") to express conclusions of EQ-NI instead of using more direct terms, such as "equivalent to," "similar to," "comparable to," and "not inferior to" to express conclusions of EQ-NI.

Guidance Development

We split the workgroup into four subgroups, each assigned to a specific section of the guidance. Each workgroup member participated in one or more subgroups. The subgroups were:

- *Unique risk of bias issues for trials calling themselves EQ-NI trials.*
- *Setting the reviewer's Minimum Important Difference (MID).*
- *Analytic foundations for concluding EQ or NI.*
- *Language considerations when concluding EQ or NI.*

Each subgroup devised guidance on their topic based on telephone and email communications. A first draft of a combined guidance document (containing all four sections) was then reviewed by the full workgroup. Based on comments and suggestions received, the Lead EPC made revisions to the combined document.

Guidance

The four sections below contain the guidance. After Section 4, we present a tabular summary of all the recommendations in Table 4.

Section 1: Unique Risk of Bias Issues for Trials Designed as EQ-NI

In this section, we consider the unique aspects of risk-of-bias that require particular attention when assessing trials that define themselves as EQ-NI trials. According to Sanchez and Chen (2006), EQ-NI trials "are not conservative in nature."[17] When a trial defines itself as EQ or NI, the trial authors have stated upfront what they hope their data will show. This tacit admission can steer a systematic reviewer's attention to specific directions of potential bias. Specifically, an "equivalence trial" may have been conducted in ways (intentionally or unintentionally) that underestimated the difference between treatments. Or, a "non-inferiority trial" comparing a new treatment to an established treatment may have been conducted in ways (intentionally or unintentionally) that tended to bias results towards the new treatment.

Like superiority trials, the sources of bias in EQ-NI trials commonly include selection, performance, detection, and attrition bias.[18] In the first methods project to inform this guidance, Treadwell (2011) highlighted several risk of bias issues that are of particular concern to EQ/NI trials.[19] These issues include poorly implemented entry criteria, poor compliance, use of concomitant treatments, protocol violations, and inadequate measurement techniques. In Table 1, we list the risk of bias issues reported in Treadwell's review and briefly summarize the implications of these issues on the findings of EQ-NI trials (e.g., whether the bias would tend to lead to underestimates or overestimates (or both) of the difference between groups). We also present questions that reviewers might want to consider when assessing EQ-NI trials to detect bias.

When assessing the validity of the results of EQ-NI trials, it is important to keep in mind the assumption of the null hypothesis for these trials. Unlike superiority trials in which the null hypothesis assumes no difference between treatment arms, the null hypothesis in EQ-NI trials assumes a difference between arms.[15] The difference in the nature of the null hypothesis between superiority trials and EQ-NI trials has an impact on what sources of bias are more or less relevant for reviewers to consider. For instance, sources of bias that result in an inappropriate finding of difference (e.g., test intervention received concomitant treatment that the active comparator did not) may be more important to consider than sources that result in no difference (e.g., Hawthorne effect, when both arms respond to being observed rather than to the treatment itself) when assessing superiority trials.[20] The null hypothesis of EQ-NI trials requires more attention to sources of bias that mute the difference between treatment arms.

Table 1. Issues Related to Risk of Bias in EQ-NI Trials

Source of Bias	Risk of Bias Issue	Implication on Findings of Equivalence/Non-Inferiority Trials	Question to Identify Bias
Selection	Inconsistent application of inclusion/exclusion criteria	When inclusion criteria are inconsistently applied across intervention and comparator arms, the study may be biased toward an under-or overestimate of the difference between groups because of confounding.	Was the inclusion/exclusion criteria clearly stated and implemented consistently across all study participants?
	Patients selected for anticipated nonresponse or good response	This would result in an underestimate of the difference between groups, regardless of the trial intent: If patients are selected for anticipated nonresponse, then both treatments will appear ineffective, or if patients are selected for anticipated good response, then both treatments will appear effective.	Were participants selected for non- response or good response?
Performance	Poor compliance	If compliance is poor within both treatment groups, the difference between groups would be underestimated. If compliance varies by treatment group, the difference between groups would be overestimated.	Was compliance with treatment ≥ 85% in both of the study's groups and across all subgroups?
	Use of concomitant treatments	Use of concomitant treatments by both treatment groups can mask the difference between groups and lead to an underestimate of the difference. Use of concomitant treatment more often in one treatment group than the other can lead to an overestimate of the difference.	Did researchers rule out any impact from a concurrent intervention or an unintended exposure that might bias or confound results?
	Protocol violations	Any deviation from the study protocol (e.g., intended treatment regimen, schedule, and manner measuring outcomes, etc.) can reduce the sensitivity of a trial and lead to an underestimate of the difference between groups and a higher likelihood of a conclusion of EQ-NI.	Did the study vary from the protocol in treatment assignments? Did researchers deliver the assigned treatment appropriately in terms of dose, schedule, and duration?
Detection	Inadequate outcome measurement techniques	Use of non-valid instruments to measure outcomes could lead to an under-or overestimate of the difference between groups. Use of a data collection method/mode that can influence the likelihood of outcome could mute or lead to an underestimate of the difference.	Was the outcome measured using a validated instrument? Was the outcome measure of interest objective and was it objectively measured?
	Lack of blinding outcomes assessor	Outcome assessors may be biased toward finding no difference if they know that all groups received active treatment.	Were those who assessed the patient's outcomes blinded to the group to which patients were assigned?
Attrition	Drop out, loss to follow-up	Intention-to-treat (ITT) analysis may underestimate the difference. Study should present both ITT and results with dropouts excluded (or per protocol analysis).	Was there a high rate of differential or overall attrition? Were intention-to-treat and per-protocol analyses used?

When assessing EQ-NI trials, reviewers should keep in mind the following issues related to performance, detection, and attrition that make EQ-NI trials especially vulnerable to bias.

Performance

Protocol violations: In EQ-NI trials, deviations from the inclusion criteria, from the intended treatment regimen, from the schedule, and from the manner and precision of measuring outcomes can reduce the sensitivity of a trial and make a conclusion of EQ-NI more likely, even in situations where the deviations are of an unsystematic or random nature.[6] Thus, reviewers need to assess these trials carefully for any violations in the intended protocol.

Treatment compliance: EQ-NI trials require a high degree of patient compliance in both the new and active comparator groups. For instance, use of concomitant medications in both groups can produce a ceiling effect that can mask differences between the two treatments.[14] Use of concomitant medication more often in one group than the other can bias the trial toward finding a favorable difference for one treatment over the other.

Detection

***Blinding*:** Wangge et al. (2010) suggest that in a superiority trial, a blinded outcome assessor who has a preliminary belief in the superiority of the one of the treatments cannot manipulate the results to support his/her belief. Not knowing the treatment status of the patients in a superiority trial prevents the outcome assessor from assigning more positive ratings to one group of patients. In EQ-NI trials, however, the value of blinding outcome assessors is debatable, especially if the end points are subjective.[21] In an EQ-NI trial, the blinded outcome assessor with a preliminary belief in equivalence or non-inferiority of the test intervention can still bias the results by assigning similar ratings to the treatment response of all patients.

***Duration of treatment and evaluations*:** In EQ-NI trials, the randomized treatments need to be given for long enough and the patient response evaluated over a long enough period so that any potential treatment differences have a realistic opportunity to reveal themselves.[22] Thus reviewers should pay particular attention to these factors in EQ-NI trials, and use the duration of treatment and length of follow-up in previous trials demonstrating efficacy of the active comparator as a reference (assuming the same outcome measures were used in those previous trials).

Attrition

Intention-to-treat analysis (ITT): An ITT analysis includes all participants according to the treatment which they have been randomized to even if they do not receive the treatment. Protocol violators, patients that miss one or more visits, patients that dropout, and patients that were randomized into the wrong group are analyzed according to the planned treatment. In superiority trials, the ITT approach is considered the most appropriate approach because it tends to avoid overly optimistic estimates of treatment efficacy (non-completers included in the analysis will generally diminish the estimated treatment effect).[17] In EQ-NI trials, however, the ITT approach does not have the same conservative effect. Use of an ITT analysis in these trials could lead to a false conclusion of EQ-NI by diluting any real treatment differences.

In EQ-NI trials, a per-protocol analysis, which considers outcomes only among patients who complete the trial, is thought to be a more appropriate analytical approach. However, according to Sanchez and Chen, this analytical approach is not without problems as completers are a select group of patients who may bias the trial in favor of one treatment over the other.[17] These and other authors suggest that EQ-NI trials include both ITT and per-protocol approaches as there is no single ideal approach in situations where there is substantial non-compliance or missing data.

Similarity of EQ-NI Trials to Trials that Established Efficacy of the Active Comparator

EQ-NI trials rely on the assumption that assay sensitivity has been established. Assay sensitivity refers to the findings of previous trials that demonstrated the superior efficacy of the active comparator over placebo.[12] It is also assumed that this efficacy will be preserved under the conditions of the EQ-NI trial.[14]

Thus, an EQ-NI trial should not only be assessed for factors that might obscure differences between treatments, but also for factors that might make the trial different from the trials that demonstrated efficacy of the active comparator.[5] When assessing the validity of EQ-NI trials, reviewers should therefore determine if the participants and outcomes are similar to those in trials that established efficacy of the active comparator. If patients in an EQ-NI trial deviate from the patient population in whom superiority over placebo had been established, then any claim about the efficacy of the new treatment could not be distinguished between true EQ-NI and inappropriate selection of patients. Similarly, outcome measures used in EQ-NI trials should be the same as those used in trials to demonstrate superiority of the active comparator in order to appropriately conclude EQ-NI. Finally, the active comparator in EQ-NI trials should be given in the same form, dose, and quality as was previously used to demonstrate the efficacy of that treatment over placebo. Presence of a distinct difference in these characteristics may decrease the strength of inference.

A recent example in which the claims of a NI trial were questioned is the ROCKET trial comparing the efficacy of the new anticoagulant medication, rivaroxaban, to the commonly used standard medication, warfarin.[23,24] In their review of the trial for regulatory approval, the Food and Drug Administration (FDA) questioned the authors' conclusion that rivaroxaban was non-inferior to warfarin for the prevention of stroke or systemic embolism in patients with nonvalvular atrial fibrillation. The reviewers of the study expressed concern about the dosage of warfarin used in the ROCKET trial compared to doses used in other recently published warfarin-controlled trials. The FDA reviewers also questioned why the study demonstrated superiority of rivaroxaban over warfarin in the as-treated safety population, but not in the intention-to-treat analysis. Regulatory approval of rivaroxaban depends on further review of the FDA.

To determine whether the conditions in an EQ-NI trial are similar to those in previous trials that demonstrated superiority of the active comparator, we recommend that reviewers consider the following questions suggested by Piaggio and colleagues:[15]

- Is the active comparator a well-established, effective standard therapy that has predictable and consistent treatment effects?
- Is the active comparator in the EQ-NI trial the same or very similar to that in trials that established its efficacy in terms of form, dose, and quality?
- Are participants in the EQ-NI trial the same or very similar to those in trials that established the efficacy of the active comparator?
- Are the outcomes in the EQ-NI trial the same or very similar to those in trials that established the efficacy of the active comparator?

The questions listed above are likely to be relevant at multiple stages in the review process and should be considered early during the development of the PICOTS (patients, interventions, controls, outcomes, treatments, and settings) and inclusion/exclusion criteria. Narrow and clear specification of inclusion/exclusion criteria may help to restrict the review to studies with PICOTS that are comparable to those in studies that established the efficacy of the active comparator. For more inclusive reviews, these questions will continue to be relevant and will

need to be further considered when assessing the 1) risk of bias of included studies (e.g., do deviations from the study protocol increase the risk of bias?); 2) the applicability of included studies (e.g., do populations, interventions, comparators, or outcomes that are not broadly generalizable reduce applicability?); and 3) the overall strength of evidence (e.g., does indirect or inconsistent evidence reduce the strength of evidence?). Protocols for reviews that include EQ-NI studies should clarify when these questions will be addressed in the review process. Reviewers should also specify how included studies will be compared to placebo-controlled trials that established the efficacy of the active comparator (e.g., technical expert panel input, comparison with previous reviews of comparisons of the active comparator with placebo-control).

Completeness of Reporting

Assessing risk of bias of any study requires adequate reporting of the study.[18] Standards for reporting are now available for EQ-NI trials. In 2006, Piaggio and colleagues extended the CONSORT (Consolidated Standards of Reporting Trials) statement and reporting checklist for randomized superiority trials to accommodate EQ-NI trials.[15] The extension encompasses the following issues: 1) the rationale for adopting a non-inferiority or equivalence design; 2) how study hypotheses were incorporated into the design; 3) choice of participants, interventions, and outcomes; 4) statistical methods; and 5) how the design affects interpretation and conclusions.

However, in a recent systematic review to identify how NI trials were conducted and reported, Wangge et al. (2010) found poor reporting.[21] Only 3.0% of the 232 trials reviewed reported the similarity of the inclusion/exclusion criteria with previous trials on the effect of the active comparator, 5.6% of the trials reported the similarity of the type of intervention with previous trials, and 3.4% reported the similarity of outcomes. Further, the authors found no improvement of reporting after the release of the extension of the CONSORT statement for EQ-NI trials.

Thus, when assessing the risk of bias in study designs, we recommend that EPCs focus primarily on the design and conduct of studies and not on the quality of reporting.[18] EPCs should set up clearly stated and consistent standards within their own reviews for how they will deal with the issue of poor reporting.

Section 2: Setting the Reviewer's Minimum Important Difference (MID)

The aim of equivalence trials is to show that a new treatment is therapeutically similar to a standard treatment within a predefined margin of equivalence. Similarly, the aim of non-inferiority trials is to show that a new treatment has at least as much efficacy as the standard treatment or is worse by an amount less than a pre-specified margin. Thus, conclusions of EQ-NI should be based on where the confidence interval for the treatment effect falls relative to the pre-specified margin. Determining an appropriate margin is therefore extremely important within the context of a systematic review in which it is possible to draw a conclusion of EQ or NI.

In this guidance document, we use the term minimum important difference (MID) to refer to the selected margin of EQ-NI (sometimes referred to by others as Δ). Other terms, including minimal clinically important difference (MCID) and minimal clinically significant difference (MCSD) have also been used. Our concern with the word "clinically" is the possibility of shifting one's focus away from the patient's perspective; some users may interpret a "clinical" difference strictly from a clinician's perspective or a policymaker's perspective. Thus, we

recommend the generic term MID, and we use the definition provided by Schunemann et al. (2005):[25] the smallest difference in score in the outcome of interest that informed what patients or proxies perceive as important, and which would lead the patient or clinician to consider a change in the management.

Ways to Determine MID

Having a pre-specified MID not only helps to guide the interpretation of the findings of a systematic review, but also facilitates the evaluation of statistical significance in the context of clinical relevance.[26] For instance, "a reviewer may identify a statistically significant difference between treatments, but this statistical difference may not inform clinicians or policy makers as to whether patients will perceive the [treatment] effects as a benefit or whether the effect is of any clinical relevance."[27]

However, determining an appropriate MID to use as the EQ-NI margin can be challenging. It might seem fitting to select a value for the MID after the data have been extracted for a review. However, this process of selection (using knowledge about trial findings to inform the decision) can lead to or be viewed as a source of bias, and potentially decrease the validity of the review's results and conclusions. The overall consensus from guidance documents and reviewers is that no overarching rule can be provided for determining MID for all clinical areas. Nevertheless, defining the MID *a priori* and providing clinical and statistical justification for the selected value and/or margin is essential.[9] The process of this determination will vary and will need to be tailored to the specific topic of interest.

Ideally, specification of the margin should be blind to existing evidence on the effect of the test intervention versus the active comparator. The following list provides suggested ways of determining and defining the MID—one or a combination of these suggestions may be used by reviewers:

- Use an already-conducted empirical study. For example, one study calculated the minimal clinically important difference for the Oswestry Disability Index (ODI) and Visual Analog Scale (VAS) of back pain using linear regression analysis of score change compared to pre-treatment scores.[28] The authors determined that the minimal clinically important difference for the ODI was 10, and for the VAS of back pain it was 18-19.[28]
- Use a number suggested by a prominent authority. For example, the FDA has used 5% body weight loss as the definition of clinically significant weight loss. Also, the FDA has defined therapeutic equivalence of odds ratios or relative risks as the range from 0.8 to 1.25.[16]
- Use a number suggested by a clinical reviewer who is specialized in that clinical area. Such a reviewer may tailor their numerical recommendation to the specific clinical area.
- Use a number suggested by a general reviewer who is *not* specialized in that clinical area but is familiar with the outcome measure.
- Use a number suggested by one or more of the studies being examined in the report, if a study asserts that the number is the minimum important difference or uses other such language. Some studies assert this in the methods section when describing their power analysis, but see further discussion of this point below.
- Use a number that was determined specifically for the review; one that ideally was suggested by the review's Technical Expert Panel or Key Informants.
- Use a number that was used in previous reviews examining this outcome measure.

- Use an MID based on Cohen's book Statistical Power Analysis,[29] which suggests definitions of small/moderate/large effects for different effect size metrics. For example, for the standardized mean difference, Cohen defined a "small" standardized mean difference as 0.2, moderate as 0.5, and large as 0.8.[26]

To strengthen the justification for the selected MID, more than one definition may be necessary. Reviewers may need to examine the margin reported in the included trials and evaluate the justifications for the margin. The difference in outcome for the active comparator compared to the test intervention is a preferable measure of efficacy and comparative effectiveness. Determining margin in NI trials, reviewers need to define the MID as the difference between treatment groups and not the changes from baseline. The pre-defined margin quantifies the important difference between treatment groups, rather than the uncontrolled changes from baseline.[30] Soliciting patient opinions about important differences in the outcome of interest may also be necessary. Finally, review authors may include non-inferiority trials in pooled analyses that tested superiority hypothesis. In such cases, reviewers should conduct subgroup/sensitivity analysis by primary trial design in meta-analyses for all outcomes.

Some trial authors state an MID when describing their power analyses. Other authors mention an effect size with this power discussion, but instead of an MID, they use an effect size they are specifically looking for or an effect size they anticipate to occur. For example, if a study simply says that "our study had a 90% power to detect a standardized mean difference of 0.38," that is not necessarily a statement about an MID, but rather an anticipated finding. The European Medicines Agency (EMEA) suggests that the choice of margin should be independent of considerations of power as the size of the clinically important difference is not altered by the size of the study.[31]

MID for Specified Outcomes

Systematic reviewers are charged with the task of identifying an appropriate MID for the outcomes of interest. For some outcomes, any difference identified between treatment groups would seem to be clinically relevant. For example, if the treatment under consideration is used for the prevention of death or irreversible morbidity and there is no second chance for treatment, it can be very difficult to justify a NI margin of any size. However, if an infinitesimally small difference is still considered "important," then an EQ-NI conclusion can never be reached, because one can never rule out the possibility of an "important" difference. Reviewers might want to consider the information in Table 2 when defining the MID for outcomes of interest. The table lists information for objectively measured vs. subjectively measured (e.g., patients self-reported) outcomes. Additionally, organizations such as the FDA and Committee for Proprietary Medicinal Products have previously published recommended NI margins for various areas of biomedical research (Table 3). These suggested MIDs may provide reviewers with a starting point as they develop the framework of a systematic review. We have also included tables in Appendix A of MID examples that have been suggested or used for various clinical topics. As previously mentioned, the MID will typically vary depending on the topic assessed and reviewers should carefully determine and justify the chosen MID.

Table 2. Methods for Identifying Appropriate MID

Outcome	Method of Identifying Potential MID	Clinical Judgment	Criteria
Objectively measured	Continuous outcomes - literature review, guidelines, achieving consensus with several rounds of expert surveys using Delhi methods[32]	Important for clinical management	Strong consistent association with mortality, morbidity, and/or quality of life[33]
		Important for prognosis	Criteria of surrogate end points[34,35]
	Events data - literature review, guidelines, achieving consensus with several rounds of expert surveys using Delhi methods[32]	Magnitude of the effect (e.g., 25% relative difference or 25% absolute risk difference) Clinical importance of the effect (e.g., 100 attributable to active comparator prevented disability events per 1000 treated or 1.0% reduction in mortality[36]	Identified based on low risk of bias efficacy placebo-controlled trials with target population similar for NI trial. Event rate with placebo established in efficacy trials should be taken into account Can be supported by dose response or active comparator studies Smallest number of the events should be prespecified
Subjectively measured (e.g., patient self-report)	Anchor method[1] (preferable) - literature review, guidelines[37]	Clinical anchors	Strong association (correlation) with important for management or prognosis clinical outcomes
		Patient based anchors: perceived improvement in the disease, perceived improvement in quality of life, perceived treatment satisfaction	A single MID may be insufficient Consider several MIDs
	Achieving consensus with several rounds of expert surveys using Delhi methods)[32]	Expert opinion about clinical importance of the difference in patient reported outcomes	Not applicable

[1] Anchor method compares patient opinion with scale score

Table 3. Recommended Non-Inferiority Margins in Different Areas of Biomedical Research

Indication	Year	Authority Outcome	Estimation	MID, Reference
Anti-microbial	1992	FDA	Response rate	Absolute Risk Difference 10–20%1 (FDA, 1992)[38]
	1997	Committee for Proprietary Medicinal Products	Response rate	Absolute Risk Difference 10% (CPMP, 1997)[38]
Urinary tract infections	1998	FDA	Response rate	Absolute Risk Difference 15% (FDA, 1998)[38]
Anti-epileptic	1998	International League Against Epilepsy	Response rate	Absolute Risk Difference 20% (Anonymous, 1998)[38]
Vaccines	1999	Committee for Proprietary Medicinal Products	Protection rate	Absolute Risk Difference 10% (CPMP, 1999)[38]
Anti-retroviral	1999	FDA	Success rate ~	Absolute Risk Difference 10% (FDA, 1999)[38]
Thrombolytics	2000	FDA	Short-term mortality	Relative risk 1.14 (FDA, 2000)[38]
Anti-inflammatory and anti-rheumatic drugs	1988	FDA	Amelioration (quantitative)**	Ratio 0.6 (FDA, 1988)[38]

Indication	Year	Authority Outcome	Estimation	MID, Reference
Dentistry	1995	American Dental Association	Deterioration (quantitative)**	Ratio 1.1 (Proskin et al., 1995)[38]
Anti-hypertensives	1998	Committee for Proprietary Medicinal Products	Reduction of diastolic BP	Difference 2 mm Hg (CPMP, 1998)[38]

~ proportion of patients with plasma HIV RNA level below the detection limit
** e.g., score

In summary, there is no recognized gold standard method for selecting an MID, as its magnitude may be influenced by several factors (e.g., efficacy, safety, cost, acceptability, and adherence).[15] To adequately choose a value, an informed decision must be taken, supported by evidence of what is considered an important difference in the particular disease area.[31] Ultimately, the pre-determined MID will help authors interpret the results and determine whether an EQ-NI conclusion is warranted.

A priori specification of a MID may help to limit any bias of a reviewer's conclusions. EPCs should remain flexible in their approach to selecting the MID as multiple measures may be necessary. Piaggio et al. listed one required reporting item (item 7) that authors should specify the margin of equivalence as well as the rationale for its choice.[15] EPCs should "clearly specify and demonstrate that a systematic approach has been taken in search for relevant and appropriate references to support the nominated threshold."[10]

Section 3: Analytic Foundations for Concluding EQ or NI

Conclusions of EQ-NI should be considered within the wider context of rating the strength-of-evidence (SOE) using the EPC guidance of Owens et al. (2009).[39] The SOE rating can be High, Moderate, Low or Insufficient. An Insufficient rating means that the evidence does not permit a conclusion. To arrive at a rating, four core domains (risk-of-bias, consistency, directness, and precision) as well as four optional domains (large magnitude of effect, publication bias, all-plausible-confounders-would-reduce-the-effect, and dose-response association) are considered. The rating can vary by outcome or by timepoint, because any of the underlying domains can vary accordingly (e.g., consistent data on one outcome but not another).

When drawing conclusions of EQ-NI, reviewers also need to consider the following unique factors: (1) Whether one treatment was believed in advance to be better on some outcomes; (2) Whether both interventions had demonstrated superiority over placebo (or other inactive treatment); (3) Whether the reviewer's MID was pre-specified by the reviewer; (4) Whether the reviewer's MID was justified by the reviewer; (5) Whether the meta-analytic model was random-effects; and (6) Whether the confidence interval was narrow enough to rule out a MID. Below, we briefly discuss these factors along with others that the reviewer should consider when drawing a conclusion of EQ-NI.

Whether one treatment was believed in advance to be better on some outcomes

As noted in the Introduction, this situation is a prerequisite for drawing a conclusion of non-inferiority regarding an important effectiveness outcome. The typical situation is when test intervention is less costly, more convenient, or has fewer adverse effects, and potentially the clinical community would be willing to accept a small decrement in effectiveness in exchange for the known advantages. Thus, a conclusion that the test intervention is non-inferior to the

active comparator could apply. However, the reverse conclusion (that the active comparator is non-inferior to the test intervention) would not make sense.

If such prior knowledge on certain outcomes exists, the systematic reviewer should lay the groundwork for a possible conclusion of non-inferiority by wording the Key Question accordingly. For example, the reviewer could phrase the Key Questions as: Is the efficacy of treatment A not worse than treatment B? Such wording reminds the user that the topic was approached *a priori* with a notion of non-inferiority in a specific direction.

Whether the reviewer's MID was pre-specified by the reviewer

The importance of pre-specifying the MID was discussed above. The problem with post-hoc specification involves the possibility of reviewer bias, whether intentional or unintentional. Specifically, a reviewer could set an MID that allows (or precludes) a conclusion of EQ or NI based on viewing the results of the studies. The same concern underlies the regulatory guidance statements that individual trial authors should pre-specify their MIDs.

Whether the reviewer's MID was justified by the reviewer

In an NI study, the test intervention is believed to be non-inferior to an active comparator if the effect of the intervention versus comparator is better than the value specified as the MID (i.e., the non-inferiority margin). The choice of the non-inferiority margin is thus crucial to establishing non-inferiority, and has an important implication to the interpretation of non-inferiority. For instance, an intervention that is claimed to be non-inferior to an active comparator may have been clinically worse than the active comparator if the MID value is too lenient. A stringent non-inferiority margin should exclude the possibility of an important inferiority.

Whether the meta-analytic model was random-effects

The EPC guidance chapter on quantitative methods[40] recommends that meta-analyses employ a random-effects model rather than a fixed-effects model. One reason for this is that the random-effects confidence interval incorporates heterogeneity, whereas the fixed-effects model ignores it. This recommendation may be even more appropriate in the context of concluding equivalence or non-inferiority. Under heterogeneity, a fixed effects model may lead to an inappropriate EQ-NI conclusion due to its inappropriately narrow interval.

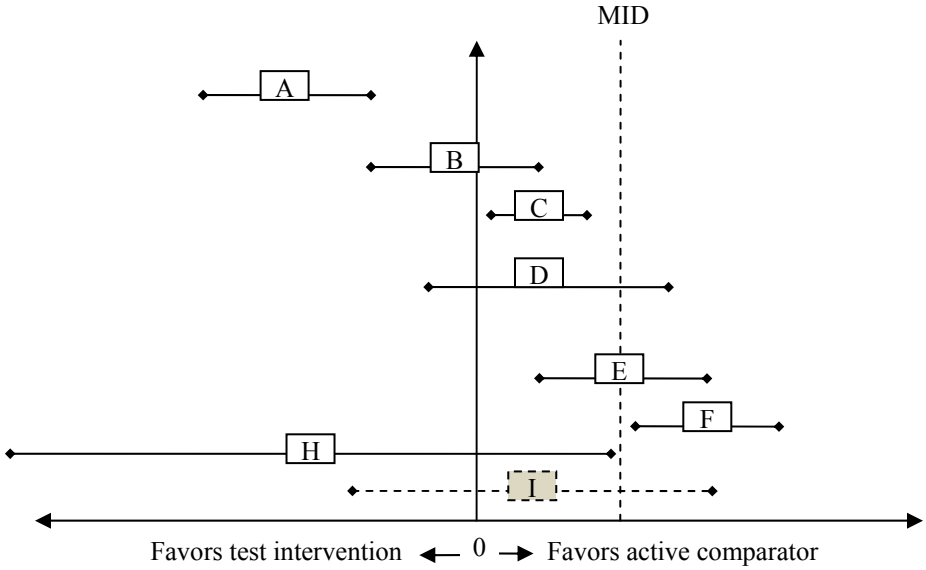
Whether the confidence interval was narrow enough to rule out an MID

For demonstrating equivalence or non-inferiority, the confidence interval should be narrow enough to exclude the MID. Figure 1 (see below) demonstrates different scenarios of effect estimates in a non-inferiority framework. Scenarios B and C in the figure represent situations in which a new intervention is non-inferior to an active comparator. In scenario A, the test intervention is significantly better than the active comparator. If the magnitude of effect is larger than a clinically important difference for superiority, one may further claim of superiority of test intervention over the active comparator. In scenarios D and E, the confidence interval is too wide to include the non-inferiority margin line, and no conclusive decision can be made. Scenario F represents a situation in which the test intervention is inferior to the active comparator. Scenario H is a special case in which the test intervention appears non-inferior to the active comparator. However, substantial uncertainty exists due to a small number of studies and/or events. In this

situation, the confidence interval will likely cross the non-inferiority line when new studies (e.g. study I in the figure) are added to the existing evidence body.

When further evidence appears, reviewers should always consider the possibility of a shift in the confidence interval resulting in a change of conclusion. Reviewers should discuss in their reports the extent to which their conclusion of non-inferiority is susceptible to new evidence.

Figure 1. Example Scenarios Comparing the Confidence Interval to the MID



Notes: A: comparison showing superiority; B and C: comparisons showing non-inferiority; D and E: comparison showing inconclusive evidence, and non-inferiority cannot be claimed; F: comparison showing inferiority; H: comparison apparently showing non-inferiority; due to the small events or sample size, however, the conclusion is unstable, and could easily be overturned by new evidence (e.g., I).

Other Aspects of Strength-of-Evidence

We examined the eight domains of SOE described by Owens et al. (2009),[39] and considered which domains should be treated differently if the evidence indicates EQ-NI. Section 1 of this guidance document already discussed how risk-of-bias assessment would be altered. For three other domains (consistency, directness, and publication bias), an EQ-NI situation does not appear to introduce unique considerations. For the domain of precision, a conclusion of EQ-NI depends on sufficient precision because the idea is to rule out the possibility of a minimum important difference on that outcome. In many cases, the evidence is too imprecise to rule out such a difference, resulting in a rating of Insufficient, and no conclusion to be drawn about that outcome.

Three other domains (*magnitude of effect, all-plausible-confounders-would-reduce-the-effect,* and *dose-response association*) appear to work in the opposite direction for EQ-NI conclusions as compared to superiority conclusions. The SOE rating system described by Owens et al. (2009)[39] was designed with the assumption of a superiority conclusion. These three domains underscore that assumption. For a superiority conclusion, the SOE was stated to increase if the following were observed: 1) a large magnitude of effect, or 2) a difference between treatments despite studies being biased against an effect, or 3) a positive dose-response association. If the evidence actually suggests equivalence, than the following observations would increase the SOE:

- The evidence for equivalence might be considered stronger if the effect is *very small*. Thus the magnitude-of-effect domain works in the opposite direction than it does for a superiority conclusion.
- The evidence for equivalence might be considered stronger if the studies were biased *in favor* of one of the two treatments (and yet the evidence found equivalence). Again, the domain works in the opposite direction than it does for a superiority conclusion.
- The evidence for equivalence might be considered stronger if the studies showed a clear *lack of a dose-response association*. Again, the domain works in the opposite direction than it does for a superiority conclusion.

Whether trials described themselves as EQ-NI trials

In the most straightforward circumstances, an EQ-NI conclusion can be drawn solely from trials describing themselves as EQ-NI trials. This may provide more reviewer confidence in the EQ-NI conclusion. Often, however, studies do not define themselves in terms of EQ-NI, and may even define themselves as superiority trials, and yet the accumulated evidence on one or more outcomes suggests equivalence. One example is the outcome of mortality when comparing bare metal stents and drug-eluting stents in the treatment of coronary artery disease. A systematic review published in 2007[41] analyzed 17 trials, all of which were designed as superiority trials. Generally the trials did show superiority on the rate of target lesion revascularization (evidence favoring drug-eluting stents), however the evidence on mortality led the reviewers to conclude equivalence (based on a summary hazard ratio of 1.03, 95% CI 0.84 to 1.22). Given the narrow confidence interval and the clear interpretation of the outcome, equivalence was a reasonable conclusion, even though the studies were designed as superiority trials. Overall, confidence in EQ-NI findings from superiority trials are enhanced if reviewers pre-specify all critical components in their study protocol (e.g., the MID). If reviewers do so, the confidence in the subsequent decisions on the conclusions from the data need not depend on trial authors' intentions to show EQ-NI. Reviewers should exercise their own independent judgments on EQ-NI.

In addition to author intent, what about reviewer intent? In some reviews, a systematic reviewer may state a priori that for a given treatment comparison, no EQ-NI could ever be drawn for any outcome, and then obviously such a conclusion cannot be drawn. Each outcome would either 1) indicate superiority, or 2) constitute inconclusive evidence. Alternatively, a systematic reviewer could a priori prohibit a superiority conclusion for all outcomes, which would limit the permissible output to either EQ-NI conclusions or inconclusive evidence. These situations would be rare, however. Both cases would be related to the availability of critical background information and prior analyses informing the review (resulting in stricter hypothesis testing) or alternatively, the lack of critical information (resulting for example, from concern about the precision of scale values or self-reported measures). More commonly, a CER reviewer will interpret the data without an a priori set of prohibited conclusions. This allows a straightforward unbiased reflection of the evidence.

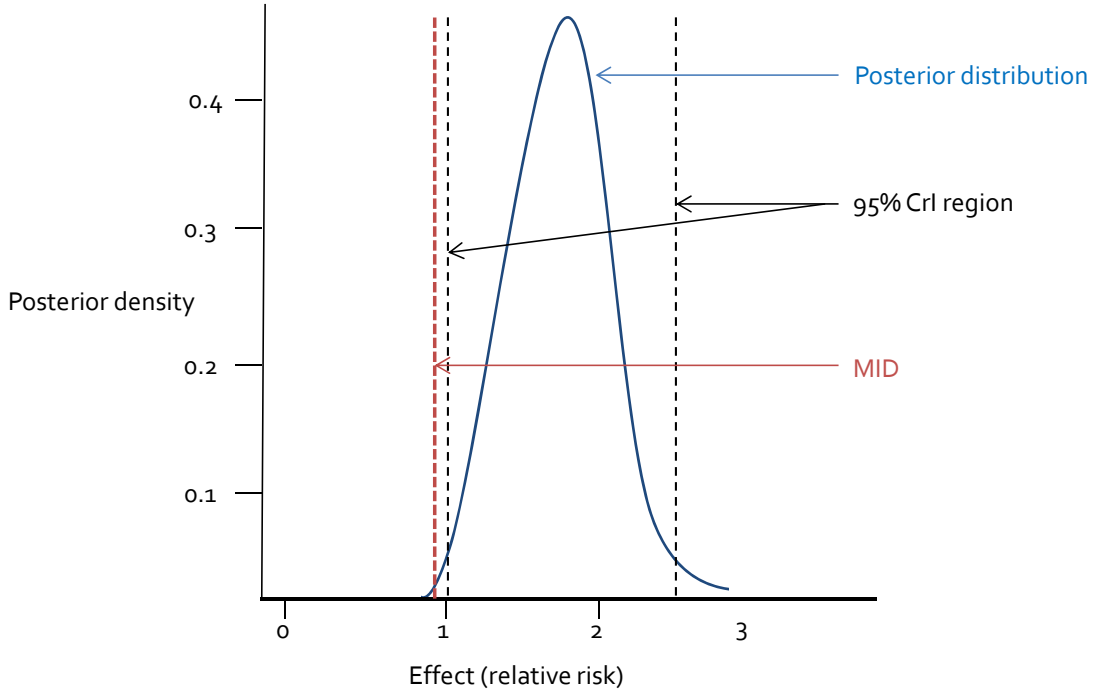
Bayesian Analysis

Bayesian analysis methods in medicine involve the statistical combination of a 'prior (probability) distribution' with a likelihood distribution based on observed data from one or more clinical trials or a meta-analysis.[42] When used to compare two or more medical interventions, the result is an updated 'posterior (probability) distribution' that reflects a true absolute or relative

difference between them along with 95% credible intervals (CrIs) (e.g., the 2.5 and 97.5 percentiles of the posterior distribution). In addition to reporting an effect size and 95% CrIs, the posterior distribution can also be interpreted in terms of an estimate of the probability that a given treatment is better than its comparator(s) (e.g., "There is an x% probability that an intervention results in a greater response than another intervention"). For this reason, many believe that Bayesian methods are more directly relevant and easily interpretable to medical decision-makers and healthcare professionals.

Bayesian methods can also be used in the determination of non-inferiority or equivalency of two or more medical interventions. After the posterior probabilities reflecting the difference between the two (or more) interventions are constructed, the observation that the 95% credible interval (CrI) falls entirely to one side of the MID would define NI (see Figure 2). A simple extension of this methodology could be used to determine EQ. As with frequentist approaches to clinical trial interpretation, *a priori-specified* margin should be determined using both statistical and clinical reasoning.[9]

Figure 2. Bayesian Posterior Distribution Compared to the MID



Note: This figure has been adapted from Quilici et al. (2008).[43] The hypothetical bell-shaped curve is the posterior probability distribution (i.e., after the data have been incorporated). The leftmost vertical dashed line indicates a hypothetical MID of about 0.90 relative risk. The Bayesian 95% credible interval (the other two vertical dashed lines) is fully above this line, which means that the evidence is sufficiently precise to permit a conclusion of non-inferiority.

Non-Meta-Analytic Situations

A conclusion of EQ or NI requires the ability to rule out the possibility of a MID (see above), and this is typically done by examining the confidence interval around the summary effect of a meta-analysis (or if there is only a single study, then one examines the single-study confidence interval). Some multiple-study situations, however, may not be appropriate candidates for meta-analysis.

The decision to combine studies in a meta-analysis depends on the clinical and methodological similarity of the studies. As described in Chapter 9 of the Methods Guide for Effectiveness and Comparative Effectiveness Reviews, there is no commonly accepted standard of defining which studies are "similar enough."[18] The decision to combine studies depends on the focus of the research question and on the extent of clinical, methodological, and statistical heterogeneity present among the included studies. The reviewer must decide how much heterogeneity is excessive. In general, meta-analysis is considered inappropriate when there is considerable clinical heterogeneity among factors such as patient populations, treatment implementation, and/or outcome measures. Meta-analysis may also be inappropriate in cases where studies do not report the necessary information to allow meta-analysis, or such information cannot be calculated from reported information (for example, when the Ns at follow-up are not reported).

The research question and the overall strength of the evidence, not just the appropriateness of meta-analysis, are what should form the basis of any conclusion within a systematic review. The presence of clinical heterogeneity among studies indicates that the studies are really addressing different questions. Thus, each study should be considered individually (i.e., the single-study CI) when determining whether its clinical question warrants a conclusion. For example, if three studies are included for a review's "Key Question 2" and the three studies are clinically heterogeneous, then they address clinically distinct issues, and should be treated as three sub-questions (i.e., Key Question 2a, 2b, and 2c). Any one of these three evidence bases could permit a conclusion of EQ-NI, as long as the strength of the evidence was not judged to be Insufficient. As always, whether the evidence is sufficient to permit a conclusion depends on many factors such as the risk of bias, directness of the evidence, consistency, and precision (e.g., does the study's CI fall within the margin of EQ-NI). See Owens et al. (2009)[39] for how to assess these factors.

Thus, the lack of meta-analysis does not necessarily preclude a conclusion of EQ-NI, just as it does not preclude an evaluation of the strength of evidence in relation to a particular outcome. The same approach should also be used for determining if studies appropriately form a body of evidence for evaluating one outcome in relation to one key question. Reviewers should still evaluate the analytic foundations discussed above, as they would prior to conducting a meta-analysis. In particular, the MID should be determined *a priori*. However, the reviewer will need to qualitatively determine what the MID should be for each study and what would be considered comparable across studies. In the case of different measures of the effectiveness of an intervention or outcome (such as the use of different measurement scales or instruments), the reviewer will need to be knowledgeable of what might be considered a MID for each of the included measures. The reviewer will also need to compare studies in relation to their choice of potential confounding variables controlled in analyses because these could result in differences in effect size.

Studies rarely report effect sizes and confidence intervals, but these can usually be calculated based on reported information. For example, for dichotomous outcomes, studies rarely report a relative risk and its standard error, but these are easily calculated from the reported rates and Ns. For continuous outcomes, studies often fail to report measures of dispersion (e.g., standard deviations or SDs). It may be appropriate to impute SDs for those studies based on other studies that did report SDs.[44,45] If such imputation is too uncertain, and no confidence interval can be determined, then the reviewer cannot rule out the possibility of an MID; thus no conclusion of EQ-NI can be drawn.

Section 4: Language Considerations When Concluding EQ or NI

Clear communication can be particularly difficult when drawing conclusions of EQ-NI, and a key emphasis should be on preventing misinterpretations. For example, concluding only that "there was no statistically significant difference" can be factually correct, but highly misleading. The problem is that the statement fails to distinguish between two very different scenarios: 1) the evidence shows equivalence, and 2) the evidence is inconclusive due to low statistical power. The first scenario should be expressed more directly (e.g., "the treatments had similar mortality rates"), whereas the second scenario indicates insufficient evidence (i.e., no conclusion).

Similar criticism can be leveled at the wording "no evidence of a difference." This can be misinterpreted by users that the reviewer has concluded "evidence of no difference." Other indirect wordings to be avoided include "studies failed to show a difference," "studies did not find/suggest/show/indicate a difference," "trials have not found a difference." None of these wordings distinguish between a *conclusion of equivalence* and a *non-conclusion due to low precision*. The processes detailed in the other three sections of this guidance, involving comparing the confidence interval to the MID as well as assessing all other aspects of the strength of evidence, together determine whether the evidence is sufficient to permit a conclusion of equivalence.

The main question in an equivalence study is whether a new treatment is therapeutically similar to the current standard of care. As indicated above, the most common problem observed among studies that claim equivalence is the misinterpretation that lack of a statistically significant difference between treatments is evidence of equivalence. This guidance document highlights that conclusions of equivalence should be based on whether the confidence intervals fall within the pre-specified margin of EQ. If the confidence intervals indicate EQ, reviewers (as well as trial authors) could use the following terminology to phrase their conclusion: "Treatment A is *equivalent to* Treatment B," "Treatment A is *similar to* Treatment B," or "Treatment A improves recovery [or whatever outcome] *as much as* Treatment B."

In contrast, the main question in a non-inferiority study is whether a new treatment is not worse than the current standard of care. Thus, when drawing a conclusion of NI whether in an individual trial or a systematic review, the conclusion should be worded to indicate that the test intervention is non-inferior to the active comparator. And not, as previously mentioned in this guidance document, phrased in the reverse—the active comparator is non-inferior to the test intervention. Phrases that are commonly used to express non-inferiority include: "Treatment A is *not inferior to* Treatment B," or "Treatment A is *at least as effective as* Treatment B."

To be complete, a conclusion of EQ-NI should include a description of the study objectives. For example, Bingham et al conducted two identical non-inferiority trials in which the primary purpose was to determine whether etoricoxib 30 mg daily was as effective as the recommended dose of celecoxib 200 mg daily in patients with osteorarthritis.[46] The authors of the study stated their conclusion as follows: "Etoricoxib 30 mg qd was at least as effective as celocoxib 200 mg and had similar safety in the treatment of knee and hip OA [osteorarthritis]." Similarly, ECRI Institute conducted two systematic reviews comparing the efficacy and safety of inhaled insulin to short-acting injected insulin—one addressing type 1 diabetes[47] and the other type 2 diabetes.[48] The primary purpose of the reviews was to determine if inhaled insulin provided similar glucose control to short-acting injected insulin. In both reviews, the authors concluded that "inhaled insulin and injected insulin provided similar levels of glycated hemoglobin" in the short-term (12 and 24 weeks).

Some reviews use qualifiers (e.g., "may," "suggest") when drawing conclusions. These words are intended to convey degrees of reviewer confidence in the evidence. Strength-of-

evidence ratings, similarly, are intended to communicate one's confidence. Thus, they can replace the need for language qualifiers. Instead of using a language qualifier ("Survival rates after treatments A and B *may be* similar"), one could make a simple declarative statement and pair it with a strength rating (e.g., "Survival rates after treatments A and B are similar [*Strength of Evidence: Low*]") Some EPCs may choose to use both the rating and a language qualifier. If so, reviewers should be aware of potential inconsistency (e.g., "The evidence *suggests that* quality-of-life after Treatment A is similar after Treatment B [*Strength of Evidence: High*]").

As described in the previous section, Bayesian analysis can permit other options for communicating conclusions of EQ-NI. For example, based on an appropriate posterior distribution, a reviewer might state "there is a 96% chance that the true difference between the treatments is less than the MID," or "the chance is 98% that treatment A is not inferior to treatment B on this outcome."

Sometimes, the evidence is so precise that it is simultaneously compatible with multiple conclusions. For example, Weng et al. (2010)[49] meta-analyzed 11 trials comparing the percentage change in LDL after atorvastatin 10 mg vs. simvastatin 20 mg. The summary confidence interval was +1.2% to +3.1%, in favor of atorvastatin. This was statistically significant, but less than the reviewers' MID of 7%. What should the reviewer do?

Technically, the data are consistent with four conclusions:

1. Atorvastatin is superior to simvastatin (because the CI was above 0);
2. Atorvastatin and simvastatin are approximately equivalent (because the CI was within the range -7% to +7%);
3. Atorvasatin is non-inferior to simvastatin (because the CI was above -7%); and
4. Simvastatin is non-inferior to atorvastatin (because the CI was below +7%).

The latter two conclusions, which are NI conclusions, are not relevant to these drugs because neither drug was known in advance to have advantages on other outcomes. The real question is whether to conclude superiority (conclusion 1) or equivalence (conclusion 2).

One option would have been to present only the superiority conclusion, based on the fact that the MID itself is a subjective judgment, and some patients or clinicians may have lower MIDs for LDL. One might argue that a reviewer's responsibility is to report all observed differences, no matter how small. Another option is to present only the equivalence conclusion, based on the fact that users of the review need not be bothered by such small differences among treatments, differences that were clearly considered unimportant by the reviewers. A third option is to combine the conclusions into one, as done by Weng et al. (2010): "Meta-analysis indicated a statistically significant but clinically minor difference (<7%) between statins in cholesterol lowering effect."[49] This option fully describes the data, and absolves the reviewer from having to prioritize one conclusion over the other.

Choosing among these options is a reviewer judgment based on the context of the review, and one's certainty in the MID for that outcome. In the statin example, the abstract's Results section contained the joint statement quoted above, but the abstract's Conclusion section stated more simply "At comparable doses, statins are therapeutically equivalent in reducing LDL-C."[49] Thus, the authors prioritized the equivalence conclusion over the superiority conclusion. This suggests they were fairly confident about their 7% MID for percentage change in LDL. Other clinical situations may warrant other approaches.

Summary of Recommendations

Table 4. EPC Guidance on Equivalence and Non-Inferiority

Area of Consideration	Recommendations ("Do's")	Issues to Avoid ("Don'ts")
Unique risk of bias issues for trials designed as EQ-NI	• Assess studies for the following areas of particular concern: poorly implemented entry criteria, poor compliance, use of concomitant treatments, protocol violations, and inadequate measurement techniques. • Assess studies for similarity to trials that established efficacy of the active comparator. • Specify how studies included in a review will be compared to placebo-controlled trials that established the efficacy of the active comparator (e.g., technical expert panel input, comparison with previous reviews of comparisons of the active comparator with placebo-control). • Focus on the design and conduct of the studies. • Set up clearly stated and consistent standards for how to deal with the issue of poor reporting.	• Do not focus exclusively on the quality of reporting
Setting the reviewer's Minimum Important Difference (MID)	• Define the margin of EQ-NI based on the MID, which incorporates the patient and clinical perspective. • Pre-specify MID and justify choice of the value in a systematic review, ideally using prior research on what matters to patients. • Examine the MID for reported in trials of EQ-NI for primary outcomes and examine the trial authors' justification for their choice. • If an author's MID is used as justification for the reviewer 's MID, ensure that the author stated it to be the "minimum" difference considered important • Consider more than one definition of MID if necessary. • For subjective outcomes, consider using anchor-based margins. Anchor method compares patient opinion with scale score and may include perception of improvement in disease status, function, disability or quality of life. • Define MID as the difference between treatment groups, not as change from baseline. • Include MIDs in review for the same primary outcomes reported in individual trials of EQ-NI.	• Do not ignore MID in individual trials of EQ-NI
Analytic foundations for concluding EQ or NI	• Drawing conclusions of EQ-NI should be considered within the wider context of rating the strength-of-evidence (SOE). • Non-inferiority conclusions require prior knowledge that one of the treatments was better on some outcomes (e.g., safety, cost, convenience) in order to justify a small sacrifice in effectiveness in a predefined direction. • Consider including a supplemental evidence review on the comparison of the active comparator vs. placebo (or inactive control) to establish that the active comparator is itself effective. • Consult confidence interval to determine if it is narrow enough to exclude the MID. • Consider that other aspects of SOE need to be treated in the opposite manner, such as magnitude of effect, all-plausible-confounders-would-reduce-the-effect, and dose-response association. • Consider other analytic approaches (e.g., Bayesian methods)	• Conclusions of EQ-NI should not be based solely on the lack of statistical significance between two treatments. • Do not require meta-analysis

Area of Consideration	Recommendations ("Do's")	Issues to Avoid ("Don'ts")
Language considerations when concluding EQ or NI	• Avoid confusion or misinterpretation when wording conclusions of EQ-NI. • Use direct phrasing to express an EQ-NI conclusion, or if the evidence is inconclusive, state that. • For examples of direct phrasing, consider phrasing such as "not inferior to," "similar to," "comparable to," or "at least as effective as" when expressing EQ or NI. • Include a description of the study objectives when concluding EQ or NI. • Consider using SOE rating in place of language qualifiers to express uncertainty. If both are used, be aware of potential inconsistencies.	• Avoid indirect wording that fail to distinguish between an *EQ-NI conclusion* and *inconclusive evidence*. Examples to avoid are "no evidence of a difference," "there was no statistically significant difference," "studies failed to show a difference," "studies did not find/suggest/show/indicate a difference," "trials have not found a difference."

References

1. Witte S, Victor N. Some problems with the investigation of noninferiority in meta-analysis. Methods Inf Med 2004;43(5):470-4. PMID: 15702203

2. Lange S, Biester K. The equivalence of non-inferiority problem in systematic reviews [abstrcat P172]. In: XIII Cochrane colloquium; 2005 Oct 22-26; Melbourne, Australia. 2005.

3. Prins H, de Haan R. Formal criteria for establishing equivalence in meta-analysis of randomized clinical trials [abstract PB34]. In: 8th international Cochrane colloquium; 2000; Cape Town, South Africa. 2000.

4. ICH Expert Working Group. ICH harmonised tripartite guideline: statistical principles for clinical trials. Geneva: International Conference on Harmonisation (ICH); 1998 Feb 5. 39 p. Also available: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf.

5. Committee for Proprietary Medicinal Products (CPMP). ICH topic E 10: choice of control group in clinical trials. London (UK): European Medicines Agency (EMEA); 2001 Jan. 30 p.

6. Committee for Proprietary Medicinal Products (CPMP). Points to consider on switching between superiority and non-inferiority. London (UK): European Medicines Agency (EMEA); 2000 Jul 27. 11 p. Also available: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003658.pdf.

7. Committee for Proprietary Medicinal Products (CPMP). Note for guidance on the investigation of bioavailability and bioequivalence. London (UK): European Medicines Agency (EMEA); 2000 Dec 14. 19 p.

8. Committee for Proprietary Medicinal Products (CPMP). Common technical document for the registration of pharmaceuticals for human use. Clinical overview and clinical summary of module 2. Module 5: study reports. London (UK): European Medicines Agency (EMEA); 2003 Jul. 44 p. Also available: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002723.pdf.

9. Committee for Proprietary Medicinal Products (CPMP). Guideline on the choice of the non-inferiority margin. London (UK): European Medicines Agency (EMEA); 2005 Jul 27. 11 p. Also available: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003636.pdf.

10. Pharmaceutical Benefits Advisory Committee. Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee (Version 4.3). Barton ACT: Commonwealth of Australia; 2008 Dec. 300 p. Also available: http://www.pbs.gov.au/industry/listing/elements/pbac-guidelines/PBAC4.3.2.pdf.

11. Points to be considered by the review staff involved in the evaluation process of new drug. Final. Tokyo, Japan: Pharmaceuticals and Medical Devices Agency (PMDA); 2008 Apr 17. 6 p. Also available: http://www.pmda.go.jp/english/service/pdf/points.pdf.

12. Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER. Guidance for industry: non-inferiority clinical trials. Rockville (MD): U.S. Food and Drug Administration (FDA); 2010 Mar. 66 p. Also available: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM202140.pdf.

13. Guidance for industry on preparation of common technical document for import / manufacture and marketing approval of new drugs for human use (New Drug Application – NDA). New Delhi: Central Drugs Standard Control Organization (CDSCO); 2010 Nov. 110 p. Also available: http://cdsco.nic.in/CTD_Guidance%20-Final.pdf.

14. Gomberg-Maitland M, Frison L, Halperin JL. Active-control clinical trials to establish equivalence or noninferiority: methodological and statistical concepts linked to quality. Am Heart J 2003 Sep;146(3):398-403. PMID: 12947355

15. Piaggio G, Elbourne DR, Altman DG, et al. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. JAMA 2006 Mar 8;295(10):1152-60. PMID: 16522836

16. Guidance for industry: statistical approaches to establishing bioequivalence. Rockville (MD): U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research; 2001 Jan. 48 p. Also available: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm070244.pdf.

17. Matilde Sanchez M, Chen X. Choosing the analysis population in non-inferiority studies: per protocol or intent-to-treat. Stat Med 2006 Apr 15;25(7):1169-81. Also available: http://rds.epi-ucsf.org/ticr/syllabus/courses/26/2007/01/23/Other/readings/ITT%20vs%20PP%20non-inf%20trials.pdf. PMID: 16397861

18. Assessing the risk of bias of individual studies when comparing medical interventions. Rockville (MD): Agency for Healthcare Research and Quality (AHRQ); 2011 Jun 9. 30 p. Also available: http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?productid=714&pageaction=displayproduct.

19. Treadwell JR. Methods project 1: existing guidance for individual trials. Plymouth Meeting (PA): ECRI Institute; 2011 Jun 29. 39 p.

20. Shuttleworth M. Hawthorne effect. [internet]. Experiment Resources; 2009 [accessed 2011 Sep 8]. [3 p]. Available: http://www.experiment-resources.com/hawthorne-effect.html.

21. Wangge G, Klungel OH, Roes KC, et al. Room for improvement in conducting and reporting non-inferiority randomized controlled trials on drugs: a systematic review. PLoS ONE 2010;5(10):e13550. Also available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2965079. PMID: 21048948

22. Scott IA. Non-inferiority trials: determining whether alternative treatments are good enough. Med J Aust 2009 Mar 16;190(6):326-30. Also available: http://www.mja.com.au/public/issues/190_06_160309/sco10995_fm.html. PMID: 19296815

23. Sutter S. FDA panel to consider rivaroxaban's comparative efficacy. In: Family practice news [database online]. Rockville (MD): International Medical News Group, LLC.; 2011 Sep 7 [accessed 2011 Sep 8]. [4 p]. Available: http://www.familypracticenews.com/index.php?id=2934&type=98&tx_ttnews[tt_news]=62576&cHash=da03e20e36.

24. Patel MR, Mahaffey KW, Garg J, et al. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. N Engl J Med 2011 Aug 10; Also available: http://www.nejm.org/doi/pdf/10.1056/NEJMoa1009638. PMID: 21830957

25. Schunemann HJ, Guyatt GH. Commentary--goodbye M(C)ID! Hello MID, where do you come from?. Health Serv Res 2005 Apr;40(2):593-7. Also available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1361157/pdf/hesr_00374.pdf. PMID: 15762909

26. Lemieux J, Beaton DE, Hogg-Johnson S, et al. Three methods for minimally important difference: no relationship was found with the net proportion of patients improving. J Clin Epidemiol 2007 May;60(5):448-55. Also available: http://www.sciencedirect.com/science?_ob=MImg&_imagekey=B6T84-4MJC1Y5-7-5&_cdi=5076&_user=851995&_pii=S0895435606003234&_origin=&_coverDate=05/31/2007&_sk=999399994&view=c&wchp=dGLzVzz-zSkWB&_valck=1&md5=944e41c59e1e716b5d451a64b1bed015&ie=/sdarticle. PMID: 17419955

27. Troosters T. How important is a minimal difference. Eur Respir J 2011 Apr;37(4):755-6. Also available: http://erj.ersjournals.com/content/37/4/755.full.pdf. PMID: 21454895

28. Hagg O, Fritzell P, Nordwall A, et al. The clinical importance of changes in outcome scores after treatment for chronic low back pain. Eur Spine J 2003 Feb;12(1):12-20. PMID: 12592542

29. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale (NJ): Lawrence Erlbaum Associates; 1988. 567 p.

30. ICH Expert Working Group. ICH Harmonized Tripartite Guideline: Choice of control group and related issues in clinical trials E10. 2000 Jul 20. 35 p. Also available: http://www.sourcesolution.com/speakup/resources/download/E10-_Choice_of_Control_Group_and_Related_Issues_in_Clinical_Trials.pdf.

31. Committee for medicinal products for human use (CHMP) guideline on the choice of the non-inferiority margin. Stat Med 2006 May 30;25(10):1628-38. PMID: 16639773

32. Wells G, Beaton D, Shea B, et al. Minimal clinically important differences: review of methods. J Rheumatol 2001 Feb;28(2):406-12. PMID: 11246688

33. Beninato M, Gill-Body KM, Salles S, et al. Determination of the minimal clinically important difference in the FIM instrument in patients with stroke. Arch Phys Med Rehabil 2006 Jan;87(1):32-9. PMID: 16401435

34. Garattini S, Bertele' V. Ethics in clinical research. J Hepatol 2009 Oct;51(4):792-7. PMID: 19664839

35. Gentile I, Borgia G. Surrogate endpoints and non-inferiority trials in chronic viral hepatitis. J Hepatol 2010 May;52(5):778. PMID: 20347500

36. Siegel JP. Equivalence and noninferiority trials. Am Heart J 2000 Apr;139(4):S166-70. PMID: 10740125

37. Revicki D, Hays RD, Cella D, et al. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol 2008 Feb;61(2):102-9. PMID: 18177782

38. Lange S, Freitag G. Choice of delta: requirements and reality--results of a systematic review. Biom J 2005 Feb;47(1):12-27; discussion 99-107. PMID: 16395993

39. Owens DK, Lohr KN, Atkins D, et al. Grading the strength of a body of evidence when comparing medical interventions-Agency for Healthcare Research and Quality and the Effective Health Care Program. J Clin Epidemiol 2010 May;63(5):513-23. PMID: 19595577

40. Fu R, Gartlehner G, Grant M, et al. Conducting quantitative synthesis when comparing medical interventions: AHRQ and the effective health care program. J Clin Epidemiol 2011 Apr 6. PMID: 21477993

41. Stettler C, Wandel S, Allemann S, et al. Outcomes associated with drug-eluting and bare-metal stents: a collaborative network meta-analysis. Lancet 2007 Sep 15;370(9591):937-48. PMID: 17869634

42. O'Hagan L, Luce BR. A primer on Bayesian statistics in health economics and outcomes research. Bethesda (MD): MEDTAP International, Inc.; 2003. 72 p. Also available: http://www.shef.ac.uk/content/1/c6/02/55/92/primer.pdf.

43. Quilici S, Abrams KR, Nicolas A, et al. Meta-analysis of the efficacy and tolerability of pramipexole versus ropinirole in the treatment of restless legs syndrome. Sleep Med 2008 Oct;9(7):715-26. PMID: 18226947

44. Furukawa TA, Barbui C, Cipriani A, et al. Imputing missing standard deviations in meta-analyses can provide accurate results. J Clin Epidemiol 2006 Jan;59(1):7-10. PMID: 16360555

45. Thiessen Philbrook H, Barrowman N, Garg AX. Imputing variance estimates do not alter the conclusions of a meta-analysis with continuous outcomes: a case study of changes in renal function after living kidney donation. J Clin Epidemiol 2007 Mar;60(3):228-40. PMID: 17292016

46. Bingham CO 3rd, Sebba AI, Rubin BR, et al. Efficacy and safety of etoricoxib 30 mg and celecoxib 200 mg in the treatment of osteoarthritis in two identically designed, randomized, placebo-controlled, non-inferiority studies. Rheumatology (Oxford) 2007 Mar;46(3):496-507. Also available: http://rheumatology.oxfordjournals.org/content/46/3/496.full.pdf. PMID: 16936327

47. ECRI Institute. Inhaled insulin for type 1 diabetes. Plymouth Meeting (PA): ECRI Institute Health Technology Assessment Information Service; 2007 Sep. 147 p. (ECRI Institute evidence report; no. 149).

48. ECRI Institute. Inhaled insulin for the treatment of type 2 diabetes. Plymouth Meeting (PA): ECRI Institute Health Technology Assessment Information Service; 2007 Jun 1. 141 p. (Windows on medical technology; no. 146). Also available: http://www.ecri.org.

49. Weng TC, Yang YH, Lin SJ, et al. A systematic review and meta-analysis on the therapeutic equivalence of statins. J Clin Pharm Ther 2010 Apr;35(2):139-51. PMID: 20456733