Methods Research Report

Assessing Equivalence and Noninferiority

Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • **www.ahrq.gov**

Methods Research Report

Assessing Equivalence and Noninferiority

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

Contract No. 290-2007-10063

Prepared by:
EPC Workgroup:

ECRI Institute Evidence-based Practice Center, Plymouth Meeting, PA
Johns Hopkins University Evidence-based Practice Center, Baltimore, MD
McMaster University Evidence-based Practice Center, Hamilton, Ontario, Canada
Oregon Evidence-based Practice Center, Portland, OR
Research Triangle Institute Evidence-based Practice Center, Research Triangle Park, NC
University of Connecticut Evidence-based Practice Center, Hartford, CT
University of Minnesota Evidence-based Practice Center, Minneapolis, MN
The University of Adelaide, Australia
Agency for Healthcare Research and Quality, Rockville, MD

Investigators:
Jonathan Treadwell, Ph.D.
Stacey Uhl, M.S.S.
Kelley Tipton, M.P.H.
Sonal Singh, M.D., M.P.H.
Lina Santaguida, Ph.D.
Xin Sun, Ph.D.
Nancy Berkman, Ph.D.
Meera Viswanathan, Ph.D
Craig Coleman, Ph.D
Tatyana Shamliyan, M.D.
Shi-Yi Wang, M.D.
Rema Ramakrishnan
Adam Elshaug, M.P.H., Ph.D.

AHRQ Publication No. 12-EHC045-EF
June 2012

This report is based on research conducted by the EPC Workgroup under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-2007-10063). The findings and conclusions in this document are those of the author(s), who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help health care decisionmakers—patients and clinicians, health system leaders, and policymakers, among others—make well-informed decisions and thereby improve the quality of health care services. This report is not intended to be a substitute for the application of clinical judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical reference and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances presented by individual patients.

This report may be used, in whole or in part, as the basis for development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

The investigators have no relevant financial interests in the report. The investigators have no employment, consultancies, honoraria, or stock ownership or options, or royalties from any organization or entity with a financial interest or financial conflict with the subject matter discussed in the report.

Suggested citation: Treadwell J, Uhl S, Tipton K, Singh S, Santaguida L, Sun X, Berkman N, Viswanathan M, Coleman C, Shamliyan T, Wang S, Ramakrishnan R, Elshaug A. Assessing Equivalence and Noninferiority. Methods Research Report. (Prepared by the EPC Workgroup under Contract No. 290-2007-10063.) AHRQ Publication No. 12-EHC045-EF. Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov.

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 email 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

Investigator Affiliations

Jonathan Treadwell, Ph.D.[a]
Stacey Uhl, M.S.S.[a]
Kelley Tipton, M.P.H [a]
Sonal Singh, M.D., M.P.H.[b]
Lina Santaguida, Ph.D.[c]
Xin Sun, Ph.D.[d]
Nancy Berkman, Ph.D.[e]
Meera Viswanathan, Ph.D[e]
Craig Coleman, Ph.D.[f]
Tatyana Shamliyan, M.D.[g]
Shi-Yi Wang, M.D.[g]
Rema Ramakrishnan[g]
Adam Elshaug, M.P.H., Ph.D.[h]

[a] ECRI Institute Evidence-based Practice Center (Lead EPC), Plymouth Meeting, PA
[b] Johns Hopkins University Evidence-based Practice Center, Baltimore, MD
[c]McMaster University Evidence-based Practice Center, Hamilton, Ontario, Canada
[d]Oregon Evidence-based Practice Center, Portland, OR
[e]Research Triangle Institute Evidence-based Practice Center, Research Triangle Park, NC
[f]University of Connecticut Evidence-based Practice Center, Hartford, CT
[g] University of Minnesota Evidence-based Practice Center, Minneapolis, MN
[h] Agency for Healthcare Research and Quality, Rockville, MD, and The University of Adelaide, Australia

Acknowledgements

We appreciate the comments and suggestions on an earlier draft of this guidance from several peer reviewers and public commenters, as well as Karen Schoelles, M.D., S.M., FACP, Wendy Bruening, Ph.D., and Irena Kirman, M.D., Ph.D.

Peer Reviewers

Jeffrey Andrews, M.D., FRCSC
Vanderbilt Evidence-based Practice Center
Nashville, TN

Ethan Balk, M.D., M.P.H.
Tufts Evidence-based Practice Center
Boston, MA

Jesse Berlin, Sc.D.
Johnson & Johnson Pharmaceutical
New Brunswick, NJ

Tianjing Li, M.D., Ph.D., M.H.S.
Johns Hopkins Bloomberg School of Public Health
Baltimore, MD

James Reston, Ph.D., M.P.H.
ECRI Institute
Plymouth Meeting, PA

Norma Terrin, Ph.D.
Tufts Medical Center
Boston, MA

Lucy Turner, M.Sc.
Ottawa Health Research Institute
Ottawa, Ontario, Canada

Assessing Equivalence and Noninferiority

Structured Abstract

Objective: To provide guidance on how to manage the concepts of equivalence and noninferiority in the context of systematic reviews.

Methods: This guidance was developed by a workgroup consisting of 13 individuals from seven Evidence-based Practice Centers (EPCs) and the Agency for Healthcare Research and Quality, under the leadership of the ECRI EPC. Prior to developing any guidance, the lead EPC also performed two methods projects intended to assist the workgroup. The first was a review of 12 existing guidance documents pertaining to equivalence and noninferiority, all of which were intended for primary researchers. The second project assessed the methodology used for a random sample of 50 recent systematic reviews that concluded equivalence or noninferiority between two or more treatments. Based on the previous experience and knowledge of the workgroup members, as well as insights from the two methods projects, guidance was developed and underwent posting for public comment and peer review.

Results: No guidance documents specifically addressing equivalence and noninferiority in the context of systematic review were identified. The workgroup developed a list of recommendations for four areas. First, how to assess the unique risk of bias for trials self-identifying as equivalence or noninferiority trials. Second, how to set Minimum Important Difference for a systematic review. Third, the analytic foundations for concluding equivalence or noninferiority in a systematic review. Fourth, language considerations when concluding equivalence or noninferiority in a systematic review.

Conclusions: Systematic reviewers need to adopt a consistent and conceptually sound approach to interpreting, concluding, and expressing equivalence or noninferiority in the context of systematic reviews. This paper provides preliminary guidance in that endeavor.

Contents

Appendixes

Introduction

This chapter provides guidance to EPCs about several issues on equivalence and noninferiority. 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 "noninferiority trials." How should evidence reviewers incorporate such trials? Third, the wider field of systematic review has no guidance on equivalence and noninferiority.[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 noninferiority (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 "noninferiority" (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. EQ trials aim to determine whether a new treatment is therapeutically similar to a standard treatment within a predefined margin of equivalence (e.g., a ratio of plasma drug levels from 80 percent to 125 percent is used by FDA to establish bioequivalence).[16]

In contrast, NI trials are conducted in a clinical context of the assumption 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 unrelated to effectiveness (e.g., fewer side effects, lower cost, and/or greater convenience). This assumption sets the stage for being willing to accept a small decrement 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 prespecified amount. For example, Aujensky et al.[17] conducted a noninferiority trial comparing outpatient to inpatient care for patients with acute pulmonary embolism. Outpatient treatment entails fewer costs, and thus some minimal reduced effectiveness may be acceptable (which the authors defined as no more than 4 percent increased rate of symptomatic recurrent venous thromboembolism within 90 days).

Despite the contextual differences between EQ and NI trials, many details about their design and analysis are similar (e.g., the prespecification 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 noninferiority 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 groupwide 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 (identified by a medical librarian using a targeted literature search) pertaining to equivalence and noninferiority (see Appendix B, Methods Project 1). These guidance documents (10 from regulatory agencies and 2 from academia) were all intended for primary researchers designing and interpreting equivalence trials. Major insights from this project (according to the author of the 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.
- 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 (this point about overestimation was not made by the regulatory documents, but rather by the author of Methods Project 1).
- 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 50 recent systematic reviews that contain conclusions that could be interpreted as conclusions of EQ-NI between two or more treatments (see Appendix C, Methods Project 2). The reviews were randomly selected from 235 reviews identified in a targeted literature search by a medical librarian. 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 (according to the authors of the 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 prespecify 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 simply interpreted as demonstrating equivalence, without any mention of whether an MID was employed in the process.
- 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 self-identifying as 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. Given the lack of empirical guidance, the recommendations were opinion-based, and were informed by the two Methods projects described above as well as the previous experience and knowledge of the workgroup members.

A first draft of a combined guidance document (containing all four sections) was reviewed by the full workgroup. Based on comments and suggestions received, the lead EPC made revisions to the combined document. The document was posted for public comment for one month (November 2011), and five public commenters provided feedback. In addition, we relied on seven solicited peer reviewers with expertise in systematic review (e.g., EPC directors and associate directors) and/or equivalence/noninferiority trials (e.g., an EPC statistician, and an FDA statistician). Further, the AHRQ Task Order Office and associate editor provided additional comments and feedback. Two workgroup calls near the end of the project were used to clarify responses to the major comments received. Overall, the reviews resulted in numerous edits, additions, and clarifications.

Guidance

Section 1: Unique Risk of Bias Issues for Trials Self-Identifying 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, EQ-NI trials "are not conservative in nature."[18] In other words, EQ-NI trials are vulnerable to certain biases that require systematic reviewers to pay particular attention to the direction of potential biases. Specifically, an "equivalence trial" may have been conducted in ways (intentionally or unintentionally) that underestimated the difference between treatments. Or, a "noninferiority trial" comparing a new treatment to an established treatment may have been conducted in ways (intentionally or unintentionally) that tended to bias results in favor of the new treatment.

Like superiority trials, the sources of bias in EQ-NI trials commonly include selection, performance, detection, and attrition bias.[19] In the first methods project to inform this guidance, Treadwell highlighted several risk of bias issues that are of concern to EQ/NI trials.[20] Many of these issues are also of concern in superiority trials. However, their impact on the results of a trial is likely to differ between superiority and EQ-NI trails. The issues include poorly implemented entry criteria, poor adherence, 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 of the difference between groups). We also present questions that reviewers might want to consider when assessing EQ-NI trials to detect bias.

Table 1. Issues related to risk of bias in EQ-NI trials

Source of Bias	Risk of Bias Issue	Implication on Findings of Equivalence/Noninferiority 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 patients in the treatment groups may vary in ways that impact response to treatment.	Were the inclusion/exclusion criteria clearly stated and implemented consistently across all study participants?
	Patients selected for anticipated nonresponse or good response in one arm	If patients are selectively recruited for anticipated response in one arm, the study may make a treatment seem significantly worse or better than the other.	Were participants selected for nonresponse or positive response?
Performance	Poor adherence	If adherence is poor within both treatment groups, the difference between groups would be underestimated. If adherence varies by treatment group, the difference between groups may be overestimated.	Was adherence with treatment sufficient 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?

Table 1. Issues related to risk of bias in EQ-NI trials (continued)

Source of Bias	Risk of Bias Issue	Implication on Findings of Equivalence/Noninferiority Trials	Question To Identify Bias
Detection	Inadequate outcome measurement techniques	Use of nonvalid instruments or improper use of valid instruments to measure outcomes could lead to an underestimate 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? Was the outcome measure properly administered (e.g., interviewers properly trained on interview protocol)?
	Lack of blinding outcomes assessor	Outcome assessors may be biased (consciously or unconsciously) 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 followup	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 and reported?

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. A conservative approach to evaluating the null hypothesis in a superiority trial requires paying attention to sources of bias that may artificially inflate the differences between study arms. A similarly conservative approach to evaluating the null hypothesis for EQ-NI trials requires paying close attention to sources of bias that may artificially reduce the differences between study arms.

The guidance in this section refers primarily to risk of bias when assessing RCTs. Observational studies can also test assertions of EQ/NI; evaluation of risk of bias from such studies would need to account for confounding and other selection biases in addition to the biases listed below. When assessing EQ-NI trials, reviewers should keep in mind the following issues related to selection, performance, detection, and attrition that make EQ-NI trials especially vulnerable to bias. These sources of bias are applicable to superiority trials as well. Table 1 describes their application to EQ/NI trials.

Selection

Application of Inclusion/Exclusion Criteria

As indicated in Table 1, inclusion/exclusion criteria that are inconsistently applied across intervention and comparator arms may bias the study toward an under- or overestimate of the difference between groups. This is because inconsistent application of inclusion/exclusion criteria may result in differences between patients that can impact their response to treatment.

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 to detect real differences. This has the potential to underestimate the different between treatments, 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 Adherence

EQ-NI trials require a high degree of patient adherence to treatment 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. suggest that in a superiority trial, a blinded outcome assessor who has a preliminary belief in the superiority of the one of the treatments is unable (due to blinding) to 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 noninferiority 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] Conversely, researchers may not have measured short-term outcomes that would have shown a difference (e.g., short-term pain after laparoscopic surgery vs. open surgery). Thus reviewers should pay particular attention to these factors in EQ-NI trials, and use the duration of treatment and length of followup in previous trials demonstrating efficacy of the active comparator as a reference (assuming the same outcome measures were used in those previous trials). In addition, reviewers can consult the general literature on the appropriate measuring timing for those outcomes.

Attrition

Intention-to-Treat Analysis (ITT)

An ITT analysis includes all participants according to the treatment to which they have been randomized, even if they do not receive the treatment. Protocol violators, patients who miss one or more visits, patients who drop out, and patients who 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 (noncompleters included in the analysis will generally diminish the estimated treatment effect).[18] 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, may be more appropriate than it is in superiority trials.[12] 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.[18] 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-adherence or missing data. We recommend that systematic reviewers consider both per-protocol and intention-to-treat analyses, and also evaluate reasons for drop-out, as a means for understanding the potential for bias, especially if one of the treatments is more toxic than the other.

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

EQ-NI trials rely on two assumptions: assay sensitivity and constancy. Assay sensitivity assumes that the superiority of the active control over placebo has been established and will be preserved under the conditions of the EQ-NI trial.[12] Constancy assumes that a EQ-NI trial is sufficiently similar to previous trials that showed the superior efficacy of the active comparator to placebo.[23] These assumptions, together with a supportive finding in an EQ-NI trial, can permit the indirect inference that the new treatment works, in the absence of direct placebo-controlled trials of that new treatment.

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 versus placebo. If patients in an EQ-NI trial deviate in ways that may affect their response to treatment from the patients in trials that demonstrated superiority of the active treatment control, 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 (and the timing of measurement) used in EQ-NI trials should be similar to 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. One subtle possibility concerns the development of antibiotic resistance; an active comparator may be less efficacious than it used to be due to a fundamental change in the disease being treated.

Typically, EQ-NI trials do not include a comparison of the test treatment with placebo. In a systematic review, however, both treatments may have been compared to the placebo, and the choice of test treatment versus active comparator may sometimes depend on the framing of the research question. Demonstrating the superiority of both treatments over placebo may strengthen the inference of EQ/NI in the context of systematic review, and if two treatments are concluded

to yield similar outcomes, the reviewer should determine that both treatments actually benefitted patients (see Section 3 of this guidance for other factors affecting this 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.[24,25] In their review of the trial for regulatory approval, the Food and Drug Administration (FDA) questioned the authors' conclusion that rivaroxaban was noninferior 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. Specifically, FDA was concerned because the "mean time in therapeutic range (or TTR) in the warfarin arm was 55 percent, which is well below the TTR in other recent warfarin-controlled studies (63 percent to 73 percent)."[26] Regulatory approval of rivaroxaban depends on further review by 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, timing, and settings) and inclusion/exclusion criteria. Interpreting answers to these questions requires clinical judgment, particularly when the response to one or more of the questions listed above is "no" or when comparators, participants, and outcomes are "very similar" but not the same as in the efficacy trials. Systematic reviews that make conclusions about EQ/NI (discussed in more detail in Section 3 of this guidance) have the responsibility to ensure that these conclusions build on evidence of the efficacy of the active comparator. This evidence may come from pre-existing reviews or the evidence may be generated at the same time as conclusions about EQ/NI as part of the same review.

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.[19] 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 noninferiority 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. found poor reporting.[21] Only 3.0 percent 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 percent of the trials reported the similarity of the type of intervention with previous trials, and 3.4 percent 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.[19] EPCs should set up clearly stated and consistent standards within their own reviews for how they will deal with the issue of poor reporting. Given the variety of possibilities, more specific guidance is difficult; making reviewer judgments transparent allows readers to evaluate the reasonableness of those judgments as well as the implications for the review.

Section 2: Setting the Reviewer's Minimum Important Difference

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 noninferiority trials is to show that a new treatment has at least as much efficacy as the standard treatment or is no worse by an amount less than a prespecified margin. Thus, conclusions of EQ-NI should be based on where the confidence interval for the treatment effect falls relative to the prespecified margin or value. Determining an appropriate margin or value 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 or value 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.: 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.[27]

Ways To Determine MID

A Priori Versus Post-Protocol Specified MID

Having a prespecified 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.[28] 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."[29]

Determining an appropriate MID to use can be challenging. It might seem fitting to select a margin or value for the MID after the data have been extracted (post-protocol) for a systematic 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.

Members of ECRI Institute's EPC are currently working on a comparative effectiveness review (CER) for the topic of inguinal hernia. After reviewing the included literature, it was apparent that the prespecified MID for the outcome measure of return to daily activities after surgery (a 7-day MID) to daily activities (because the typical time before returning to daily activities after surgery was only 10 days). Working with the project's task order officer, the authors decided to change the MID (to 1 day instead of 7) for this outcome measure, provided justification for the new MID selection, and made a protocol amendment. Another example of a post-protocol MID selection or change includes the identification of an included trial's validated MID that differed from the prespecified MID. In this example, a sensitivity analysis could be performed with the reviewer's prespecified MID and the individual trial's MID to determine if the results and potential conclusions differed on the outcome measure of interest.

The overall consensus from guidance documents and reviewers is that no overarching rule can be provided for determining MID for all clinical areas. Ideally, defining the MID a priori (during protocol development) 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. There are several ways for a systematic reviewer to determine an MID. Some processes may be more familiar to those involved with clinical trials vs. the overall assessment of a literature base. "Anchor-based methods for MID selection relate changes in scores on a measure to a standard that is different from the specific measure itself, whereas distribution-based methods use statistical parameters associated with the measure (e.g., effect size, standard error of measurement) to interpret the magnitude of changes in the measure's scores over time."[30] The Delphi approach for defining MID may be used when systematic reviewers are assessing new treatments or treatments that are purported to improve patient-centered outcomes such as quality of life.

Another method that may be unfamiliar to the systematic review process has been recommended by FDA. For NI trials, FDA recommends "first determining the margin for the treatment effect of the active comparator (M1) and then calculate the margin for the test intervention (M2) by taking a percentage (e.g., 50 percent) of M1."[12] However, the European Medicines Agency (EMEA) recommended against this overall approach, stating: "It is not appropriate to define the [NI] margin as a proportion of the difference between active comparator and placebo. Such ideas were formulated with the aim of ensuring that the test product was superior to . . . placebo."[9] Systematic reviewers may acknowledge the MID that original trials aimed to examine, but justify the selection of a different MID and reevaluate the individual trial conclusions.

Reviewers should be careful when using the trials' descriptions of statistical power as a basis for an MID. Some individual trial authors clearly state an MID when describing their power analyses (e.g., "For LDL, the minimum effect deemed important is seven percentage points"). However, other authors simply mention an effect size with their discussion of power, specifically an effect size they are looking for or an effect size they anticipate to occur. These are not necessarily MIDs. For example, if a study simply says that "our study had a 90 percent 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 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]

The following list provides suggested ways of determining and defining the MID. To strengthen the justification for MID selection—one or a combination of these suggestions may be used by systematic reviewers. Please note, this list is not exhaustive and reviewers may have other ways that are appropriate for the topic of interest:

- 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 pretreatment scores.[32] 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.[32]
- Use a number suggested by a prominent authority and incorporated into trial's design. For example, FDA has used 5 percent body weight loss as the definition of clinically significant weight loss. Also, FDA has defined therapeutic equivalence of plasma drug levels as the range from 80 percent to 125 percent.[16]
- FDA recommends first determining the margin for the treatment effect of the active comparator (M1) and then calculating the margin for the test intervention (M2) by taking a percentage (e.g., 50 percent of M1).[12]
- 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 general reviewers, policy makers, public health professionals who are not specialized in that clinical area but are familiar with the outcome measure and the importance for individuals and public health.
- 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 derived from evidence based guidelines, literature reviews, or suggested by the review's Technical Expert Panel or Key Informants.
- Use a number that was used in previous high-quality reviews examining this outcome measure.
- Use an MID based on Cohen's book Statistical Power Analysis,[33] 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.[28]
- Use anchor-based or distribution-based methods for determining or defining the MID for patient-reported outcome measures.

MID for Specified Outcomes

Systematic reviewers are charged with the task of identifying an appropriate MID for one or more outcome measures of interest. For continuous outcomes, the idea is that very small differences (such as 1 point on a 0 to 100 scale) are negligible and can be ignored for the purposes of drawing a clear conclusion.

For dichotomous outcomes, however, the notion of an MID is more challenging. Any difference between treatment groups might seem to be important. 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, even one additional event seems important, because it happened to one additional patient. 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.

Systematic reviewers may be criticized for the MID value selected for dichotomous outcomes, especially if this value differs from previous systematic reviews. The selection may also be challenged when there is some uncertainty amongst reviewers, clinicians, experts; if no one has previously set the MID for the outcome; or if studies are using different MIDs for this outcome.

When uncertainty plays a role in MID selection, performing a sensitivity analysis can help systematic reviewers compare MID values and identify any changes in conclusions based on these values. Although it may be harder to defend the MID selection for dichotomous outcomes, not taking a stand is as problematic as the selection of some arbitrary MID. Without an MID, systematic reviewers may also draw unfounded conclusions.

Systematic reviewers will want to consider the importance of patient opinions about important difference in outcome measures (e.g., pain). Table 2 lists information for objectively measured vs. subjectively measured (e.g., patients self-reported) outcomes. "It is important to note that an important change in individual patients cannot be directly applied to the evaluation of clinically important group differences."[30] Also, systematic reviewers need to define the MID as the identified difference between treatment groups and not the changes in each group from baseline.[5,30] The predefined margin quantifies the important difference between treatment groups, rather than the uncontrolled changes from baseline.[34] In Appendix A, we included tables of MID examples that have been suggested or used for various clinical topics. These suggested MIDs may provide reviewers with a starting point as they develop the framework of the systematic review. 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 of Importance of MID	Suggested Criteria To Use When Confirming MID Selection
Objectively measured	Continuous outcomes (mean difference)- literature review, guidelines, achieving consensus with several rounds of expert surveys using Delphi methods[35]	Important for clinical management	Strong consistent association with mortality, morbidity, and/or quality of life[36]
		Important for prognosis	Criteria of surrogate end points[37,38]
	Events data (relative risk, hazard ratio, odds ratio, absolute risk difference) - literature review, guidelines, achieving consensus with several rounds of expert surveys using Delphi methods[35]	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[39]	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* (preferable) - literature review, guidelines[40]	Clinical anchors	Strong association (correlation) with importance 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 Delphi methods)[35]	Expert opinion about clinical importance of the difference in patient reported outcomes	Not applicable

*Anchor method compares patient opinion with scale score

EPCs should remain flexible in their approach to selecting the MID as multiple methods processes may be necessary. To adequately choose a MID for an outcome measure, an informed decision must be made, supported by evidence of what is considered an important difference in the particular disease area.[31] Ultimately, the predetermined MID will help authors interpret the results and determine whether an EQ-NI conclusion is warranted. 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.[41] 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).

The sections below list numerous issues to be considered before concluding equivalence or noninferiority.

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 noninferiority 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 noninferior 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 noninferiority by wording the Key Question accordingly. For example, the reviewer could phrase the Key Questions as: Is the efficacy of treatment A not lower than treatment B? Such wording reminds the user that the topic was approached a priori with a notion of noninferiority in a specific direction.

Whether Trials Self-Identified 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 self-identify as EQ-NI trials, and may even self-identify 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[42] analyzed 17 trials, all of which were designed as superiority trials. Generally the trials did show superiority on the primary outcome 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 for mortality, even though the studies were designed as superiority trials. Overall, confidence in EQ-NI findings from superiority trials are enhanced if reviewers prespecify 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.

Another question is whether to combine data from trials with a mix of author intentions: some trials self-identify as EQ, some as NI, some as superiority, and some do not specify. From the reviewer's standpoint, the key concern is whether the various trials addressed the same Key Question. If studies enrolled similar patients, made the same treatment comparison, measured the same outcome in the same way at similar time points, then there is little reason not to consider the data together (either qualitatively or quantitatively), regardless of the trial authors' intentions.

Risk of bias, however, may be assessed differently based on author intentions (see Section 1 above).

In addition to author intent, what about reviewer intent? Reviewers should generally approach a Key Question without an a priori intent of what could or should be concluded. Instead, they should undertake a straightforward summary of the evidence. Thus, reviewers would generally not categorize their review as an "equivalence CER," a "superiority CER," or a "noninferiority CER." Different outcomes could warrant different types of conclusions, depending on the data.

Whether the Reviewer's MID Was Prespecified and Justified by the Reviewer

The importance of prespecifying 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 prespecify their MIDs. A justification of the chosen MID is also important (also discussed above). If the MID is too large, an intervention could be claimed to be noninferior to an active comparator yet may actually have been clinically worse.

Some reviews may not have prespecified an MID, possibly because the review was initiated before this guidance was completed, or reviewers felt that any MID would be too uncertain. Such reviews may still be able to conclude equivalence or noninferiority for an outcome, but these conclusions would be based on a post-hoc MID or some unspecified MID in the mind of the reviewer. Therefore, these conclusions need to be made sparingly, with adequate documentation of the reasons for the conclusion (such as consistency, directness, low risk of bias, and large sample size). Reviewers need to be transparent and clear about these judgments; conclusions of EQ/NI may require additional studies following the failure of superiority trials to show a difference.

Whether the Meta-Analytic Model Was Random Effects

The EPC guidance chapter on quantitative methods[43] 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 noninferiority. Under heterogeneity, a fixed effects model may lead to an inappropriate EQ-NI conclusion due to its inappropriately narrow interval. As discussed in the EPC guidance chapter on pooling,[44] however, a random-effects model does not actually explain the heterogeneity, and meta-regressions or subgroup analyses can be used for this purpose.

One technical point involving meta-analysis: Noninferiority trials may report a one-tailed confidence interval because of the one-sided nature of the authors' hypothesis. If the systematic reviewer also chose a priori to approach the question from a one-sided perspective, this interval can be used for the purpose of inclusion in a one-tailed meta-analysis (which may require customization of meta-analysis software). However, if instead the reviewer intended to compare the treatments from a two-sided perspective (i.e., either direction of effect is possible), then the one-sided interval should be used to determine the standard error of the study's effect size, which

should then be used as a measure of within-study error in a standard two-sided meta-analysis. Any conclusion of equivalence should be supported by a two-tailed confidence interval.

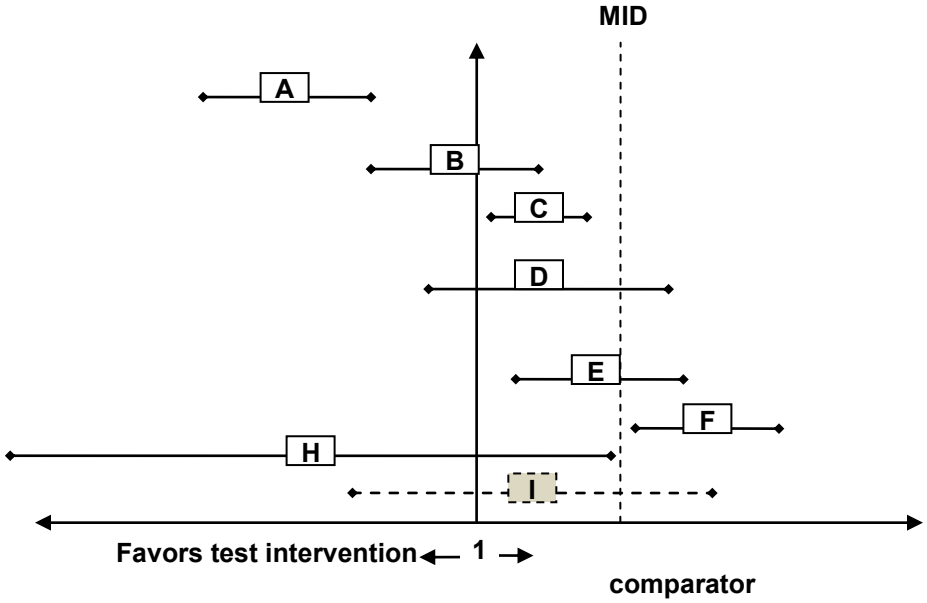
Whether the Confidence Interval Was Narrow Enough To Rule Out an MID

For demonstrating equivalence or noninferiority, the confidence interval should be narrow enough to exclude the MID. Figure 1 demonstrates different scenarios of effect estimates in a noninferiority framework. 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 the superiority of the test intervention over the active comparator. Scenarios B and C in the figure represent situations in which a new intervention is non-inferior to an active comparator. Scenario C represents a situation of extreme precision that could permit multiple conclusions, and we discuss these in detail in Section 4: Language Considerations When Concluding EQ or NI. In scenarios D and E, the confidence interval is too wide to exclude the MID, 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, when a new study is published (e.g., study I in the figure), the resulting summary confidence interval could then cross the noninferiority line, thereby eliminating the conclusion. When further evidence appears, reviewers should 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 noninferiority is susceptible to new evidence.

As an example, Meier et al.[45] performed a meta-analysis of defibrillation-first versus compression-first strategy for patients with cardiac arrest. One of the measured outcomes was survival until hospital discharge, and the minimum important difference was an odds ratio of 1.25. The summary odds ratio of defibrillation versus was 1.10, with a 95 percent confidence interval from 0.70 to 1.70. This falls into scenario D, thus no conclusion can be reached, because the evidence was consistent with both (1) a clinical advantage of the active comparator and (2) an advantage of the test intervention. If the confidence interval had ranged from 0.8 to 1.15, the comparison would falls into scenario B, permitting a noninferiority conclusion.

Figure 1. Example scenarios comparing the confidence interval with the MID



Notes: The figure shows a noninferiority comparison of test intervention to active comparator. Only the right-side MID line is shown because it is noninferiority; if it had been equivalence, both the right-side MID and the left-side MID would appear. Also the example involves a dichotomous outcome and a relative effect size metric (e.g., relative risk). If the outcome were continuous, the vertical line of no difference w be 0 and the MID would be set at on that scale.
Scenario A: comparison showing superiority; B and C: comparisons showing noninferiority; D and E: comparison showing inconclusive evidence, and noninferiority cannot be claimed; F: comparison showing inferiority; H: comparison apparently showing noninferiority; due to the small events or sample size, however, the conclusion is unstable, and could easily be overturned by new evidence (e.g., I).

Strength-of-Evidence Considerations

We examined the eight domains of SOE described by Owens et al.[41] 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. They should be considered regardless of the type of conclusion. 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.[41] 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, then 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 (when the comparison is drug vs. drug).

For example, if two drugs appear to have similar effects regardless of the dose of either, that may increase the strength of evidence for the conclusion of equivalence. Again, the domain works in the opposite direction than it does for a superiority conclusion.

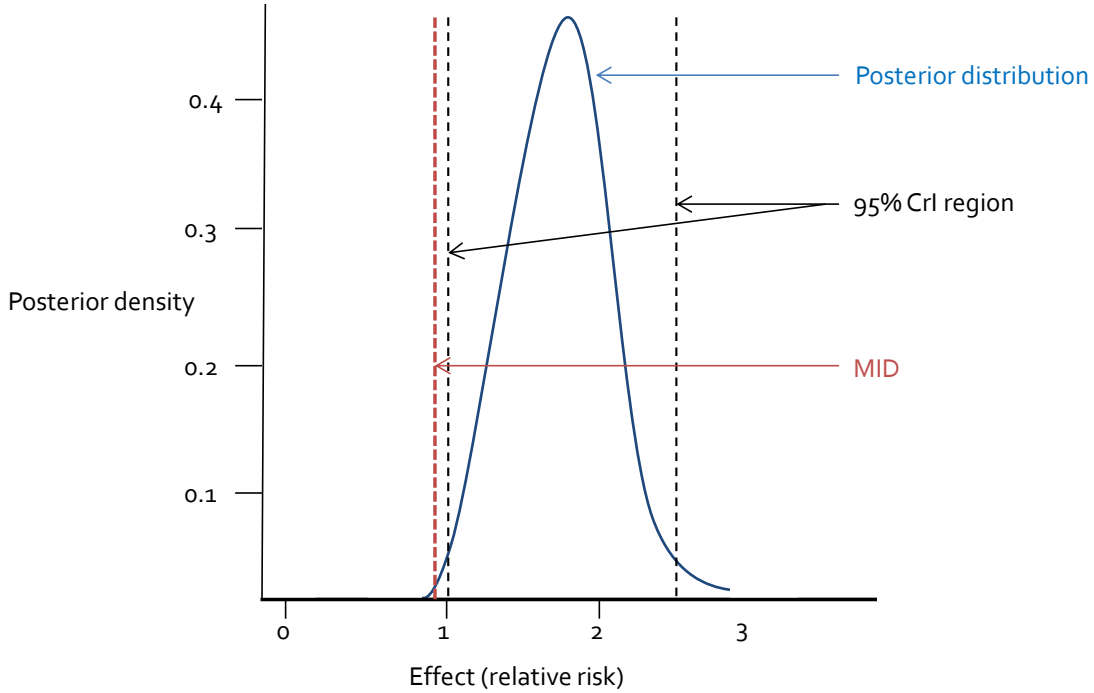
Bayesian Analysis

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.[46] 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 percent 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 percent 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 percent 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 decisionmakers and health care professionals.

Bayesian methods can also be used in the determination of noninferiority 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 percent 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, an a priori-specified MID should be determined using both statistical and clinical reasoning.[9]

Figure 2. Bayesian posterior distribution compared with the MID



Note: This figure has been adapted from Quilici, et al.[47] 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 noninferiority.

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."[19] 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 measured the same outcome in disparate ways (e.g., different quality-of-life measures), or 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 followup 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, suppose three studies are included for a review's "Key Question 2" on treatments for diabetes. Further suppose the three studies were clinically heterogeneous in that they enrolled different populations (one study enrolling children with type 1 diabetes, another study enrolling adults with type 1 diabetes, and a third study enrolling adults with type 2 diabetes). The three studies address clinically distinct issues, and should be treated as three subquestions (i.e., Key Questions 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.[41] 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 sometimes do not report effect sizes and confidence intervals, but these can usually be calculated based on reported information. For example, for dichotomous outcomes, studies sometimes do not 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.[48,49] 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 prespecified 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 noninferiority study is whether a new treatment is not worse than the current standard of care by a prespecified amount. 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 noninferior 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 noninferiority 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 noninferiority 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 osteoarthritis.[50] The authors of the study stated their conclusion as follows: "Etoricoxib 30 mg q.d. was at least as effective as celecoxib 200 mg and had similar safety in the treatment of knee and hip OA [osteoarthritis]." 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[51] and the other type 2 diabetes.[52] 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 percent chance that the true difference between the treatments is less than the MID," or "the chance is 98 percent 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.[53] 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 percent to +3.1 percent, in favor of atorvastatin. This was statistically significant, but less than the reviewers' MID of 7 percent. 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 percent to +7 percent);
3. Atorvastatin is noninferior to simvastatin (because the CI was above -7 percent); and
4. Simvastatin is noninferior to atorvastatin (because the CI was below +7 percent).

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.: "Meta-analysis indicated a statistically significant but clinically minor difference (<7 percent) between statins in cholesterol lowering effect."[53] 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."[53] Thus, the authors prioritized the equivalence conclusion over the superiority conclusion. This suggests they were fairly confident about their 7 percent MID for percentage change in LDL. Other clinical situations may warrant other approaches.

This section has addressed language for when the evidence permits a conclusion of EQ-NI. Language for phrasing a superiority conclusion is straightforward. If no conclusion is possible from the evidence on that outcome (i.e., one cannot conclude equivalence or noninferiority or superiority), then care should be taken to use balanced and non-suggestive wording, such as "the evidence was insufficient to permit a conclusion for this outcome."

Examples

Beauchamp, et al.,[54] performed a CER comparing two types of exercise training for chronic obstructive pulmonary disease:

- For the Chronic Respiratory Questionnaire domains of dyspnea and total score (on which scores range from 1 to 7), authors chose 0.5 as the minimal important difference. In support of this value, they cited an earlier systematic review from other authors that had delved into existing research on the questionnaire itself, and found converging evidence to support the choice of 0.5. For the outcome of functional exercise capacity, authors chose 54 meters difference in 6-minute walking distance. To support this choice of MID, they cited a study that had empirically determined that 54 meters was the distance at which "the average patient to stop rating themselves as 'about the same' and start rating themselves as either 'a little bit better' or 'a little bit worse'". Systematic reviews only rarely have such sound foundations for the chosen MIDs, but it can happen.
- Of the four trials reporting these outcomes, one self-identified as a noninferiority trial, and the other three did not self-identify as either superiority/noninferiority/equivalence trials. Instead, the three studies' stated goal was generally to compare the two treatments. Risk-of-bias assessment methods were not altered for the one trial self-identifying as a noninferiority trial. Thus, authors did not feel that the risk of bias of the NI trial should be measured differently because it was an NI trial.
- The reviewers performed a meta-analysis including all four trials (one NI trial and three non-self-identifying trials), because of their similarities in patients, treatments, outcomes, and time points. The 95 percent confidence interval was fully within one MID of 0 for both outcomes, providing statistical support for a conclusion of equivalence.
- Authors concluded: "The results of this meta-analysis suggest that there are no differences between the effect of interval and continuous training on measures of exercise capacity or on health-related quality of life in individuals with moderate to severe COPD."[54] This wording clearly indicates equivalence, and successfully avoids the impression that reviewers believed the evidence was insufficient to permit a conclusion.
- Eyawo et al.[55] compared three different types of proglandin therapy in the treatment of open-angle glaucoma and ocular hypertension:
- For the efficacy outcome of intraocular pressure, authors set the MID at 1.5 mmHG. They stated that this was a "commonly used margin of equivalence," and they cited a trial that addressed the same medical condition but was not included in the review.
- Seven out of 16 studies had set their own MIDs. The review did not report whether the seven MIDs matched the authors' selected MID.
- One of the 16 trials mentioned in its abstract that authors tested noninferiority; none of the other 15 studies' titles or abstracts had self-identified as superiority trials, equivalence trials, or noninferiority trials. The risk of bias was assessed in the same way for all trials.
- Authors performed pairwise comparisons of three drugs, and reported the three meta-analytic confidence intervals. Two were fully within 1.5 of 0 (i.e., permitting a conclusion of equivalence), and the third was slightly over (0.13 to 1.63 mmHg). They did find drug differences in rates of conjunctival hyperemia.
- Authors concluded: "Randomized head-to-head evaluations of prostaglandin therapy demonstrate similar efficacy effects but different hyperemia effects." This is a straightforward conclusion of equivalence on efficacy.

Summary of Recommendations

The two tables below summarize the recommendations. Table 3 involves risk of bias considerations for individual trials that self-identify as equivalence trials or noninferiority trials. Table 4 involves systematic review conclusions of either equivalence or noninferiority, specifically the a priori specification setting of the minimum important difference (MID), analytic considerations to support a reviewer conclusion of EQ or NI, and language considerations when a reviewer does conclude EQ or NI.

Table 3. EPC guidance on risk-of-bias considerations for trials self-identifying as equivalence or noninferiority trials

Recommendations ("Do's")	Issues to Avoid ("Don'ts")
• Assess studies for the following areas of particular concern: poorly implemented entry criteria, poor adherence, 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

Table 4. EPC Guidance on drawing conclusions of equivalence and noninferiority

Area of Consideration	Recommendations ("Do's")	Issues to Avoid ("Don'ts")
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. • Prespecify MID and justify choice of the value in a systematic review, ideally using prior research on what matters to patients. • Examine the reported MID 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 sensitivity analysis using multiple definitions of MID, if uncertainty about the chosen MID is substantial. • 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 the authors' stated 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). • Noninferiority 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. • Determine whether the confidence interval around the effect size 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.

Table 4. EPC Guidance on drawing conclusions of equivalence and noninferiority (continued)

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. • If no conclusion is warranted (i.e., the evidence is insufficient to conclude neither superiority, nor equivalence, nor noninferiority), use balanced and non-suggestive wording, such as "the evidence was insufficient to permit a conclusion for this outcome."	• Avoid indirect wording that fails to distinguish between an EQ-NI conclusion and insufficient 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 [abstract 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; 1998 Feb 5:39. 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; January 2001:30.
6. Committee for Proprietary Medicinal Products (CPMP). Points to consider on switching between superiority and non-inferiority. London, UK: European Medicines Agency; July 27 2000:11. 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; 2000 December 14:19.
8. Committee for Proprietary Medicinal Products. 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; July 2003:44. 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; July 27, 2005:11. 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; December 2008:300. 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); April 17, 2008:6 . www.pmda.go.jp/english/service/pdf/points.pdf.
12. Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research. Guidance for industry: non-inferiority clinical trials [draft guidance]. Rockville, MD: U.S. Food and Drug Administration; March 2010:66. 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; November 2010:110 .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; January 2001:48 . www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm070244.pdf.

17. Aujensky D, Roy PM, Verschuren F, et al. Outpatient versus inpatient treatment for patients with acute pulmonary embolism: an international, open-label, randomised, non-inferiority trial. Lancet. 2011 Jul 2;378(9785):41-8. PMID: 21703676.

18. 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. http://rds.epi-ucsf.org/ticr/syllabus/courses/26/2007/01/23/Other/readings/ITT%20vs%20PP%20non-inf%20trials.pdf. PMID: 16397861

19. Assessing the risk of bias of individual studies when comparing medical interventions. Rockville, MD: Agency for Healthcare Research and Quality; June 9, 2011:30. www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?productid=714&pageaction=displayproduct.

20. Treadwell JR. Methods project 1: existing guidance for individual trials. Plymouth Meeting, PA: ECRI Institute; June 29, 2011:39.

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. 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. www.mja.com.au/public/issues/190_06_160309/sco10995_fm.html. PMID: 19296815.

23. D'Agostino RB Sr, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues—the encounters of academic consultants in statistics. Stat Med. 2003 Jan;22(2):169-86. PMID: 12520555.

24. 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 7:4. www.familypracticenews.com/index.php?id=2934&type=98&tx_ttnews[tt_news]=62576&cHash=da03e20e36. Accessed September 8 2011.

25. Patel MR, Mahaffey KW, Garg J, et al. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. N Engl J Med. www.nejm.org/doi/pdf/10.1056/NEJMoa1009638. 2011; 365: 1557-9. PMID: 21830957.

26. Morrissey JP, Dalton KM, Steadman HJ, et al. Assessing gaps between policy and practice in Medicaid disenrollment of jail detainees with severe mental illness. Psychiatr Serv. 2006 Jun;57(6):803-8. PMID: 16754756.

27. 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. www.ncbi.nlm.nih.gov/pmc/articles/PMC1361157/pdf/hesr_00374.pdf. PMID: 15762909.

28. 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. 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.

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

30. Dworkin RH, Turk DC, Wyrwich KW, et al. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. J Pain. 2008 Feb;9(2):105-21. PMID: 18055266.

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. 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.

33. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988:567.

34. ICH Expert Working Group. ICH Harmonized Tripartite Guideline: choice of control group and related issues in clinical trials E10. July 20, 2000: 35. www.sourcesolution.com/speakup/resources/download/E10-_Choice_of_Control_Group_and_Related_Issues_in_Clinical_Trials.pdf. Accessed March 16, 2012.

35. 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.

36. 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.

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

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

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

40. 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.

41. 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.

42. 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.

43. 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 Nov;64(11):1187-97. Epub 2011 Apr 7. PMID: 21477993.

44. Fu R, Gartlehner G, Grant M, et al. Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care Program. In: Methods Guide for Comparative Effectiveness Reviews. Rockville, MD: Agency for Healthcare Research and Quality; October 2010:25. http://effectivehealthcare.ahrq.gov.

45. Meier P, Baker P, Jost D, et al. Chest compressions before defibrillation for out-of-hospital cardiac arrest: a meta-analysis of randomized controlled clinical trials. BMC Med. 2010;8:52. PMID: 20828395.

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

47. 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.

48. 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.

49. 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.

50. 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. http://rheumatology.oxfordjournals.org/content/46/3/496.full.pdf. PMID: 16936327.

51. ECRI Institute. Inhaled insulin for type 1 diabetes. ECRI Institute evidence report No. 149. Plymouth Meeting, PA: ECRI Institute Health Technology Assessment Information Service; September 2007:147.

52. ECRI Institute. Inhaled insulin for the treatment of type 2 diabetes. Windows on Medical Technology No. 146. Plymouth Meeting, PA: ECRI Institute Health Technology Assessment Information Service; June 1, 2007. 141. http://www.ecri.org.

53. 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.

54. Beauchamp MK, Nonoyama M, Goldstein RS, et al. Interval versus continuous training in individuals with chronic obstructive pulmonary disease—a systematic review. Thorax. 2010 Feb;65(2):157-64. PMID: 19996334.

55. Eyawo O, Nachega J, Lefebvre P, et al. Efficacy and safety of prostaglandin analogues in patients with predominantly primary open-angle glaucoma or ocular hypertension: a meta-analysis. Clin Ophthalmol. 2009;3:447-56. PMID: 19684868.

Appendix A. Example MIDs for Specific Clinical Topics

Table 1. Minimal important differences in outcomes in patients with COPD reported in Make B. (2007)[1]

Name	Domain	MID	Worst to Best
Forced expiratory volume in 1 second	Single domain/measure	100 ml	NA
Maximal exercise test	Single domain/measure	10 watts 3–5 watts for severe COPD	NA
Submaximal exercise endurance test	Single domain/measure	1.75 minutes	NA
Six-minute walk test	Single domain/measure	37–71 meters 24–28 for severe COPD	NA
Transition Dyspnea index	Functional impairment (FI) Magnitude of task (MT) Magnitude of effort (ME)	1 unit	-9 to +9
UCSD Shortness of Breath Questionnaire	Severity of shortness of breath during 21 activities of daily living, and three additional questions about fear of harm from overexertion, limitations, and fear caused by shortness of breath	5–7 units	120 to 0
Borg scale of perceived dyspnea	Single domain/measure	2 units	20 to 6
Visual analog scale of dyspnea	Single domain/measure	10–20	100 to 0
Breathlessness diary	Single domain/measure	0.2	4 to 0
St. George's Respiratory Questionnaire	Symptoms Activity Impacts	4 units	100 to 0
Chronic Respiratory Disease Questionnaire	Dyspnoea Fatigue Emotional functioning Mastery	0.5 units	1 to 7
Quality of Well-Being Scale	Mobility Physical activity Social activity 25 symptoms/problem complexes	0.03 units	0 to 1
Exacerbations	Single domain/measure	22% change; 1 exacerbation/year	NA

Table 2. Minimal important differences reported in the Western Ontario McMaster Universities Osteoarthritis Index

Author, Year	Method	Worst to Best Scale	Reference	Definition of Minimally Important Differences
Dougados, 2000[2]	Anchor	(Varies) WOMAC, Lequesne Functional Severity Index, Global VAS	OARSI Responder Criteria—Proposition A: This emphasizes the domain 'pain'. A 'high' improvement in pain was sufficient to define a responder. However, using this set of criteria, a patient can be also considered as a responder if an improvement of 'moderate' magnitude is observed in two of the three domains, i.e., pain, function and patient's global assessment. OARSI Responder Criteria—Proposition B: This scenario applies equal importance to 'pain' and 'function', requiring a 'high' response of one OR the other. Alternatively, a 'moderate' magnitude of response could be present in two of the three domains..	If there was a 'high' improvement in pain: improvement of at least 40% was required (ranging from 40% to 60%) together with an absolute improvement of at least 20 normalized units ranging from 20 to 30. If there was moderate improvement in pain, function, and patient's global assessment: a relative improvement ranging from 15 to 35% and an absolute improvement from 10 to 20 normalized units.
Angst, 2001[3]	Anchor	10 to 0 (for each of the 24 items) WOMAC pain scale	The transition questionnaire was used to gather data from the patients about their current subjective health status in relation to the OA joint in terms of their general health. At the 3-month follow up, patients had to compare their general health status with that of 3 months earlier, i.e., with that at baseline examination, using the assessment categories "much worse," "slightly worse," "equal," "slightly better," and "much better."	The mean score difference between the "equal" group and the "slightly better" group = 0.67 was the MID for improvement

Author, Year	Method	Worst to Best Scale	Reference	Definition of Minimally Important Differences
Tubach, 2005[4]	Anchor	100-0 WOMAC: function scale	At the final visit, patients assessed their response to NSAID treatment on a five point Likert scale (none = no good at all, ineffective drug; poor = some effect but unsatisfactory; fair = reasonable effect but could be better; good = satisfactory effect with occasional episodes of pain or stiffness; excellent = ideal response, virtually pain free).The MCII was determined in patients whose assessment of response to treatment was measured on a five point Likert scale and who had completed the final visit. The MCII was estimated for both the absolute (final value-baseline value) and the relative ((final value-baseline value)/baseline value) changes in each patient reported outcome. It was estimated by constructing a curve of cumulative percentages of patients as a function of the change in score (for example, difference in pain score) among patients whose final evaluation of response to treatment was ‘‘good, satisfactory effect with occasional episodes of pain or stiffness’’.	Patients with knee OA considered themselves clinically improved if the decrease in function score exceeded 9.1 on the WOMAC function scale
Tubach, 2005[5]	Anchor	100-0 WOMAC: function subscale	(1) “What is the level of pain above which you experience difficulties?” (This could be considered close to the external anchor for the PASS.) (2) “What is the level of pain above which you would consider taking a pain killer drug?” (This could be considered close to the external anchor for the LDAS.)	The MID in the high tertile of score is -20 (absolute change)

Author, Year	Method	Worst to Best Scale	Reference	Definition of Minimally Important Differences
Weigl, 2006[6]	Anchor	Varies WOMAC; Transition scale (that investigates the current state of health of the OA joint at the 6 months follow-up compared to its state 6 months earlier(baseline examination)	The transition scale investigates the current state of health of the OA joint at the 6-month follow-up compared to its state 6 months earlier (at baseline examination).	Three different definitions of responder: (1) For the WOMAC global score, a percentage change (100*(change of score/baseline score)) greater or equal to 18% represents an MID in improvement; (2) patients who reported a slightly or a much better health status on the transition scale were classified as responders; (3) responders had to show an MID in improvement on the WOMAC global score and report a health improvement on the transition scale
Stratford, 2007[7]	Distribution	4-0 for each of the 5 items WOMAC LK 3.1	The five pain items of WOMAC that were analyzed were: (1) walking on flat ground; (2) going up or down stairs; (3) at night while in bed; (4) sitting or lying; and (5) standing upright.	90% of stable patients will display random fluctuations equal to or less than 3.94 when assessed on multiple occasions
Bieleman, 2009[8]	Anchor	68-0 WOMAC (Dutch versions)function scale	Functional Capacity Evaluation (FCE)	The cut-off point for the WOMAC scale the cut-off point was ≥21 where subjects had work limitations that corresponded to the physical work limitations on the FCE scale
White, 2010[9]	Anchor	WOMAC: physical function	The definitions of [MID] were that they were anchored to patient-based indicators of improvement and defined meaningful improvement relative to baseline WOMAC physical function scores. The definitions of [MID] 26% and [MID] Tertile were estimated in a group of people with knee pain reporting a “good, satisfactory effect with occasional episodes of pain or stiffness” following a 4-week course of nonsteroidal antiinflammatory drug (NSAID). The [MID] 17% definition was from a group of people with knee OA who underwent 3 to 4 weeks of inpatient rehabilitation.	3 definitions of [MID] for WOMAC physical function: [MID] 26% and [MID] 17% defines meaningful improvement as a 26% and 17% decrease in WOMAC physical function (final value minus baseline value/baseline value), respectively, with a minimum absolute decrease of 2 out of 68.

Author, Year	Method	Worst to Best Scale	Reference	Definition of Minimally Important Differences
Escobar, 2007[10]	Anchor	100 to 0 WOMAC: pain subscale	Patients had to answer a question about improvement in their knee at 6 months and 2 years after intervention. The possible responses were “a great deal better,” “somewhat better,” “equal,” “somewhat worse,” and “a great deal worse.”	At 6 months: Mean change in WOMAC pain score of 37.58(19.71) was equivalent to patient reporting “A great deal better.”
Escobar, 2007[10]	Anchor	100 to 0 WOMAC: function subscale	Patients had to answer a question about improvement in their knee at 6 months and 2 years after intervention. The possible responses were “a great deal better,” “somewhat better,” “equal,” “somewhat worse,” and “a great deal worse.”	At 6 months: Mean change in WOMAC function score of 34.58 (19.33) was equivalent to patient reporting “A great deal better.”
Escobar, 2007[10]	Anchor	100 to 0 WOMAC: stiffness subscale	Patients had to answer a question about improvement in their knee at 6 months and 2 years after intervention. The possible responses were “a great deal better,” somewhat better,” “equal,” “somewhat worse,” and “a great deal worse.”	At 6 months: Mean change in WOMAC stiffness score of 34.74(28.38) was equivalent to patient reporting “A great deal better.”
Escobar, 2007[10]	Anchor	100 to 0 WOMAC: pain subscale	Patients had to answer a question about improvement in their knee at 6 months and 2 years after intervention. The possible responses were “a great deal better,” “somewhat better,” “equal,” “somewhat worse,” and “a great deal worse.”	At 6 months: Mean change in WOMAC pain score of 33.87(18.13) was equivalent to patient reporting “somewhat better.” This was considered the MID.
Escobar, 2007[10]	Anchor	100 to 0 WOMAC: function subscale	Patients had to answer a question about improvement in their knee at 6 months and 2 years after intervention. The possible responses were “a great deal better,” “somewhat better,” “equal,” “somewhat worse,” and “a great deal worse.”	At 6 months: Mean change in WOMAC function score of 19.01(17.48) was equivalent to patient reporting “somewhat better.” This was considered the MID.

Author, Year	Method	Worst to Best Scale	Reference	Definition of Minimally Important Differences
Escobar, 2007[10]	Anchor	100 to 0 WOMAC: stiffness subscale	Patients had to answer a question about improvement in their knee at 6 months and 2 years after intervention. The possible responses were “a great deal better,” “somewhat better,” “equal,” “somewhat worse,” and “a great deal worse.”	At 6 months: Mean change in WOMAC stiffness score of 14.53(26.50) was equivalent to patient reporting “somewhat better.” This was considered the MID.
Quintana, 2006[11]	Distribution	100 to 0 WOMAC: pain subscale	Six months after the intervention, patients were sent another letter with the questionnaires and additional questions on the clinical aspects of their disease and satisfaction with the intervention. The satisfaction question was dichotomized as being satisfied or not. At this time, patients answered a transitional question about their joint improvement after the intervention. The possible responses included “a great deal better,” “somewhat better,” “equal,” “somewhat worse,” or “a great deal worse.”	The [MID] for pain subscale of WOMAC was at 22.85.
Quintana, 2006[11]	Distribution	100 to 0 WOMAC: functional limitation subscale	Six months after the intervention, patients were sent another letter with the questionnaires and additional questions on the clinical aspects of their disease and satisfaction with the intervention. The satisfaction question was dichotomized as being satisfied or not. At this time, patients answered a transitional question about their joint improvement after the intervention. The possible responses included “a great deal better,” “somewhat better,” “equal,” “somewhat worse,” or “a great deal worse.”	The [MID] for functional limitation subscale of WOMAC was at 12.72.

Author, Year	Method	Worst to Best Scale	Reference	Definition of Minimally Important Differences
Quintana, 2006[11]	**Distribution**	100 to 0 WOMAC: stiffness subscale	Six months after the intervention, patients were sent another letter with the questionnaires and additional questions on the clinical aspects of their disease and satisfaction with the intervention. The satisfaction question was dichotomized as being satisfied or not. At this time, patients answered a transitional question about their joint improvement after the intervention. The possible responses included “a great deal better,” “somewhat better,” “equal,” “somewhat worse,” or “a great deal worse.”	The [MID] for stiffness subscale of WOMAC was at 29.14.
Quintana, 2006[11]	**Anchor**	100 to 0 WOMAC: pain subscale	Six months after the intervention, patients were sent another letter with the questionnaires and additional questions on the clinical aspects of their disease and satisfaction with the intervention. The satisfaction question was dichotomized as being satisfied or not. At this time, patients answered a transitional question about their joint improvement after the intervention. The possible responses included “a great deal better,” “somewhat better,” “equal,” “somewhat worse,” or “a great deal worse.”	The MID for pain subscale of WOMAC was at 22.60.
Quintana, 2006[11]	Anchor	100 to 0 WOMAC: functional limitation subscale	Six months after the intervention, patients were sent another letter with the questionnaires and additional questions on the clinical aspects of their disease and satisfaction with the intervention. The satisfaction question was dichotomized as being satisfied or not. At this time, patients answered a transitional question about their joint improvement after the intervention. The possible responses included “a great deal better,” “somewhat better,” “equal,” “somewhat worse,” or “a great deal worse.”	The MID for functional limitation subscale of WOMAC was at 17.67.

Author, Year	Method	Worst to Best Scale	Reference	Definition of Minimally Important Differences
Quintana, 2006[11]	Anchor	100 to 0 WOMAC: stiffness subscale	Six months after the intervention, patients were sent another letter with the questionnaires and additional questions on the clinical aspects of their disease and satisfaction with the intervention. The satisfaction question was dichotomized as being satisfied or not. At this time, patients answered a transitional question about their joint improvement after the intervention. The possible responses included "a great deal better," "somewhat better," "equal," "somewhat worse," or "a great deal worse."	The MID for stiffness subscale of WOMAC was at 12.94.
Ornetti, 2011[12]	PASS based on Anchor	100 to 0 WOMAC: function subscale	All patients had to assess their current global state (global PASS) by answering 'Yes' or 'No' in answer to the question "Taking into account all the activities you have during your daily life, your level of pain, and also your functional impairment, do you consider that your current state is satisfactory?"	Patients considered their global state as satisfactory if the WOMAC function was <28.06 (95% CI, 25.74 to 30.38.Global PASS is defined as the value of measurement beyond which patients consider their global state as satisfactory).
Ornetti, 2011[12]	PASS based on Anchor	100 to 0 WOMAC: function subscale	PASS for functional state: The PASS of each function scale was defined as the 75th centile of the absolute score among patients who considered their final state as satisfactory.	Patients considered their functional state as satisfactory if the WOMAC function was <28.40 (95% CI, 26.03 to 30.78). Function PASS is defined as the value of measurement beyond which patients consider their functional state as satisfactory.
Ornetti, 2011[12]	Anchor	100 to 0 WOMAC: function subscale	All patients had to assess their degree of improvement of global state, on a three-point Likert scale (worsened function, no change, improved function). Among patients who improved, the degree of improvement was scored on a four-point Likert scale (poor, fair, good, excellent).	Patients considered their global state as improved for a change of WOMAC function scale >-17.13 (95% CI, -20.07 to -14.19).Global [MID] is defined as the smallest change in global state that signifies an important improvement in a patient's symptoms.

Author, Year	Method	Worst to Best Scale	Reference	Definition of Minimally Important Differences
Ornetti, 2011[12]	Anchor	100 to 0 WOMAC: function subscale	[MID] for functional state: The [MID] of each function scale was defined as the 75th centile of the absolute change in score among patients whose final evaluation of response to NSAID was improved (improvement good or excellent).	Patients considered their functional state as improved for a change of WOMAC function scale >-17.02 (95% CI, -20.15 to -13.90). Functional [MID] is defined as the smallest change in functional state that signifies an important improvement in a patient's symptoms.

Table 3. Clinically important differences in group outcomes in bone density studies reported in Cranney, et al. (2001)[13]

Classification	Method	Spine	Femoral Neck
Minimum actually detectable beyond error in differences between groups and changes within group[14]	Calculated SD of short-term and long-term intra- and inter-subject variance for rates of change in g/cm^2/year as minimal detectable beyond error in group	Sample size decreased as length of follow-up and frequency of observations increased	NA
Differences between groups and changes within group observed in those estimated to differ/to have improved[15,16]	Effect size = mean in treatment less mean in placebo divided by SD of placebo	Better responsiveness for spine BMD than other sites	NA
Differences between groups and changes within group observed in those estimated to differ/to have improved[17]	Different cut-offs of % change in spine BMD used to define a responder in 3 year RCT	Choice of cut-off affects results of RCT	NA
Differences between groups and changes within group observed in those estimated to differ/to have improved[18]	Treatment response index = difference in % change between groups of RCT divided by standardized precision in 4 year RCT of hormone replacement therapy (HRT) for early, non-osteoporotic post-menopausal women	Treatment response index: DXA spine 10.4(0.5) DXA total hip 3.9(0.4) BUA 3.1(1.2) SOS 0.3(13.7) Stiffness 4.2(0.4)	NA
Differences between groups and changes within group observed in those estimated to differ/to have improved[19]	Compared dose-related % change in BMD in 2 RCTs and differences between treated and placebo groups Correlated changes in new region of interest in forearm to changes in BMD at the spine and hip	New region of interest in forearm is as responsive to change over time during therapy in an RCT as spine or hip BMD	New region of interest in forearm is as capable of detecting differences between groups
Differences between groups and changes within group observed in those estimated to have an important difference/improvement[20]	Clinician judgment that group difference or change should be greater than SD to be important	5%	8%
Changes within group observed in population[21]	Used 3 arbitrary cut-offs (0%, 0–3%, >3% change in spine, hip, and femoral neck BMD)	>3% change predicted fewer vertebral fractures: Spine 6%, 4%, 3.7% Femoral neck 5.5%, 4.2%. 3.1% Total Hip 6.2%, 4.6%, 2.7%	
Differences between groups observed in those estimated to have an important difference/improvement[22]	Working party of 14 European experts in osteoporosis	15% reduction in fracture frequency suggested as MID, depending on unwanted effects	

Table 4. Reduction in bone fractures corresponding to importance of the effect reported in Cranney et al. (2001)[13]

Classification of Discrimination and Changes in Studies of Osteoporosis	Method	Outcomes
Differences between groups and changes within group observed in those estimated to differ /to have improved[15,16]	Effect size = mean in treatment less mean in placebo divided by SD of placebo	Better responsiveness for clinical vertebral in high risk subgroup with femoral neck BMD >2.5 SD
Differences between groups observed in population[23]	Tested if different criteria for diagnosis of vertebral fracture would change results of RCT of fluoride	Found no difference between difference between groups for any of the fracture definition criteria
Differences between groups observed in those estimated to have an important difference/ improvement[24]	Sample size calculation to detect difference in fractures associated with 1 SD decrease in bone mass	30% reduction in incidence of vertebral and non-vertebral fractures associated with low bone mass
Differences between groups observed in those estimated to have an important difference/ improvement[25]	Evaluated 7 methods of classifying/diagnosing vertebral fracture and calculated sample size for RCT	40% risk reduction in vertebral fracture
Differences between groups observed in those estimated to have an important difference/ improvement[26]	Sample size determination for FIT trial	40% risk reduction in vertebral fracture 90% power to detect 32% reduction
Differences between groups observed in those estimated to have an important difference/ improvement[27]	Sample size calculations for different definitions of vertebral fracture, from 5–30% reduction in height	50% risk reduction in vertebral fracture
Differences between groups observed in those estimated to have an important difference/ improvement[27]	Sample size calculation for 90% power, 2 tailed test, $p < 0.05$ MORE RCT of 7705 post-menopausal women with ≥1 vertebral fracture	40% risk reduction in vertebral fractures
Differences between groups observed in those estimated to have an important difference/ improvement[28]	Sample size calculation for risedronate RCT of 2,458 post-menopausal women with ≥1 vertebral fracture	40% risk reduction in vertebral fractures
Differences between groups observed in those estimated to have an important difference/ improvement[26]	Sample size calculation for RCT of HRT, vitamin D/calcium in post-menopausal women	21% hip fracture reduction 20% reduction in combined fractures (vertebra, proximal femur, distal forearm, proximal humerus, pelvis)
Differences between groups observed in those estimated to have an important difference/ improvement[22]	Working party of 14 European experts in osteoporosis	15% reduction in fracture frequency suggested as MCID depending on unwanted effects
Differences between groups observed in those estimated to have an important difference/ improvement[29]	Cost per averted hip fracture used to determine important difference needed for policy to change	10% difference in hip fracture incidence with thiazides is cost-neutral

References

1. Make B. How can we assess outcomes of clinical trials: the MCID approach? COPD 2007 Sep;4(3):191-4. PMID: 17729062.

2. Dougados M, Leclaire P, van der Heijde D, et al. Response criteria for clinical trials on osteoarthritis of the knee and hip: a report of the Osteoarthritis Research Society International Standing Committee for Clinical Trials response criteria initiative. Osteoarthritis Cartilage 2000 Nov;8(6):395-403. PMID: 11069723.

3. Angst F, Aeschlimann A, Stucki G. Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the low. Arthritis Rheum 2001 Aug;45(4):384-91. PMID: 11501727.

4. Tubach F, Ravaud P, Baron G, et al. Evaluation of clinically relevant changes in patient reported outcomes in knee and hip osteoarthritis: the minimal clinically important improvement. Ann Rheum Dis 2005 Jan;64(1):29-33. PMID: 15208174.

5. Tubach F, Wells GA, Ravaud P, et al. Minimal clinically important difference, low disease activity state, and patient acceptable symptom state: methodological issues. J Rheumatol 2005 Oct;32(10):2025-9. PMID: 16206363.

6. Weigl M, Angst F, Aeschlimann A, et al. Predictors for response to rehabilitation in patients with hip or knee osteoarthritis: a comparison of logistic regression models with three different definitions of responder. Osteoarthritis Cartilage 2006 Jul;14(7):641-51. PMID: 16513373.

7. Stratford PW, Kennedy DM, Woodhouse LJ, et al. Measurement properties of the WOMAC LK 3.1 pain scale. Osteoarthritis Cartilage 2007 Mar;15(3):266-72. PMID: 17046290.

8. Bieleman HJ, Reneman MF, van Ittersum MW, et al. Self-reported functional status as predictor of observed functional capacity in subjects with early osteoarthritis of the hip and knee: a diagnostic study in the CHECK cohort. J Occup Rehabil 2009 Dec;19(4):345-53. PMID: 19557505.

9. White DK, Keysor JJ, Lavalley MP, et al. Clinically important improvement in function is common in people with or at high risk of knee OA: the MOST study. J Rheumatol 2010 Jun;37(6):1244-51. PMID: 20395640.

10. Escobar A, Quintana JM, Bilbao A, et al. Responsiveness and clinically important differences for the WOMAC and SF-36 after total knee replacement. Osteoarthritis Cartilage 2007 Mar;15(3):273-80. PMID: 17052924.

11. Quintana JM, Escobar A, Arostegui I, et al. Health-related quality of life and appropriateness of knee or hip joint replacement. Arch Intern Med 2006 Jan 23;166(2):220-6. PMID: 16432092.

12. Ornetti P, Dougados M, Paternotte S, et al. Validation of a numerical rating scale to assess functional impairment in hip and knee osteoarthritis: comparison with the WOMAC function scale. Ann Rheum Dis 2011 May;70(5):740-6. PMID: 21149497.

13. Cranney A, Welch V, Wells G, et al. Discrimination of changes in osteoporosis outcomes. J Rheumatol 2001 Feb;28(2):413-21. PMID: 11246689.

14. Nguyen TV, Sambrook PN, Eisman JA. Sources of variability in bone mineral density measurements: implications for study design and analysis of bone loss. J Bone Miner Res 1997 Jan;12(1):124-35. PMID: 9240735.

15. Wells G, Cranney A, Shea B, et al. Responsiveness of endpoints in osteoporosis clinical trials. J Rheumatol 1997 Jun;24(6):1230-3. PMID: 9195542.

16. Cranney A, Welch V, Tugwell P, et al. Responsiveness of endpoints in osteoporosis clinical trials−an update. J Rheumatol 1999 Jan;26(1):222-8. PMID: 9918268.

17. Oppenheimer L, Kher U. The impact of measurement error on the comparison of two treatments using a responder analysis. Stat Med 1999 Aug 30;18(16):2177-88. PMID: 10441772.

18. Sahota O, San P, Cawte SA, et al. A comparison of the longitudinal changes in quantitative ultrasound with dual-energy X-ray absorptiometry: the four-year effects of hormone replacement therapy. Osteoporos Int 2000;11(1):52-8. PMID: 10663359.

19. Ravn P, Hosking D, Thompson D, et al. Monitoring of alendronate treatment and prediction of effect on bone mass by biochemical markers in the early postmenopausal intervention cohort study. J Clin Endocrinol Metab 1999 Jul;84(7):2363-8. PMID: 10404804.

20. Eastell R. Treatment of postmenopausal osteoporosis. N Engl J Med 1998 Mar 12;338(11):736-46. PMID: 9494151.

21. Hochberg MC, Ross PD, Black D, et al. Larger increases in bone mineral density during alendronate therapy are associated with a lower risk of new vertebral fractures in women with postmenopausal osteoporosis. Fracture Intervention Trial Research Group. Arthritis Rheum 1999 Jun;42(6):1246-54. PMID: 10366118.

22. Kanis JA, Geusens P, Christiansen C. Guidelines for clinical trials in osteoporosis. A position paper of the European Foundation for Osteoporosis and Bone Disease. Osteoporos Int 1991 Jun;1(3):182-8. PMID: 1790407.

23. Melton LJ 3d, Egan KS, O'Fallon WM, et al. Influence of fracture criteria on the outcome of a randomized trial of therapy. Osteoporos Int 1998;8(2):184-91. PMID: 9666944.

24. Seeley DG, Browner WS, Nevitt MC, et al. Which fractures are associated with low appendicular bone mass in elderly women? The Study of Osteoporotic Fractures Research Group. Ann Intern Med 1991 Dec 1;115(11):837-42. PMID: 1952469.

25. Black DM, Palermo L, Nevitt MC, et al. Defining incident vertebral deformity: a prospective comparison of several approaches. The Study of Osteoporotic Fractures Research Group. J Bone Miner Res 1999 Jan;14(1):90-101. PMID: 9893070.

26. Black DM, Reiss TF, Nevitt MC, et al. Design of the fracture intervention trial. Osteoporos Int 1993;3 Suppl 3:S29-39. PMID: 8298200.

27. Kleerekoper M, Nelson DA, Peterson EL, et al. Outcome variables in osteoporosis trials. Bone 1992;13 Suppl 1:S29-34. PMID: 1581116.

28. Harris ST, Watts NB, Genant HK, et al. Effects of risedronate treatment on vertebral and nonvertebral fractures in women with postmenopausal osteoporosis: a randomized controlled trial. Vertebral Efficacy With Risedronate Therapy (VERT) Study Group . JAMA 1999 Oct 13;282(14):1344-52. PMID: 10527181.

29. Torgerson DJ, Ryan M, Ratcliffe J. Economics in sample size determination for clinical trials. QJM 1995 Jul;88(7):517-21. PMID: 7633878.

Appendix B. Methods Project 1: Existing Guidance for Individual Trials

Jonathan R. Treadwell, Ph.D., ECRI Institute, June 29, 2011

Table of Contents

Purpose

This document is intended to inform the development of EPC guidance on equivalence and non-inferiority. It summarizes publically available guidance (mostly from regulatory agencies) for *individual trials* that describe themselves as equivalence trials or non-inferiority trials. Thus, it will be most relevant to how EPCs might 1) adjust their risk-of-bias assessment methods when confronted with these types of trials, and 2) determine an appropriate reviewer threshold (such as the minimum important difference) for deciding whether the included evidence permits a conclusion of equivalence or non-inferiority within the context of systematic reviews.

Methods

A searcher from the Scientific Resource Center (SRC) searched numerous sources for any formal guidance on the planning, conduct, analysis, and interpretation of clinical trials described as equivalence or non-inferiority (EQ-NI) trials. One reviewer (JRT) read through the search results for relevant guidance documents. The same reviewer read each relevant document in detail, capturing statements in five categories:

- *Definitions/justifications/assumptions,* including definitions of EQ-NI trials, appropriate justifications for their use, and assumptions made
- *Planning* EQ-NI trials, including pre-specification of trial intent, choice of control groups, threshold for decision making, and sample size calculations
- *Conducting* EQ-NI trials, particularly avoiding bias
- *Analyzing* EQ-NI trials, including the use of confidence intervals, the use of intention-to-treat or per-protocol analysis, one-tailed or two-tailed tests, and multiplicity
- *Interpreting/reporting* EQ-NI trials, including the types of permissible conclusions, the conditions under which the conclusion can differ from the original trial intent, and reporting requirements

Results

General description of guidance documents

Searches identified 14 potentially relevant documents, but two (one from Health Canada[1] and one from the European Medicines Agency)[2] were simply endorsements of one of the other documents.[3] Thus, we extracted information from 12 documents. These included 10 documents from 6 regulatory groups, and two documents from collaborative academic groups:

- Two documents from the International Conference on Harmonisation (ICH) of Technical Requirements for Registration of Pharmaceuticals for Human Use: one in 1998,[3] and one in 2000.[4]
- Four documents from European Medicines Agency (EMEA): two in 2000,[5,6] one in 2003,[7] and one in 2005.[8]
- A 2008 document from the Australian Pharmaceutical Benefits Advisory Committee (Australian PBAC) from 2008.[9]
- A 2008 document from the Japan Pharmaceuticals and Medical Devices Agency (Japan PMDA) from 2008.[10]
- A 2010 document from the U.S. Food and Drug Administration (FDA), in 2010.[11]

- A 2010 document from the India Central Drugs Standard Control Organisation (hereafter abbreviated India CDSCO) from 2010.[12]
- One document from Gomberg-Maitlin et al. (2003).[13]
- One document that was an extension of CONSORT guidelines, Piaggio et al. (2006).[14]

Appendix A (starting on page 14) provides the verbatim statements from the 12 documents, using the categories listed in the Methods section above.

Most of the documents were written from the perspective of regulatory agencies, therefore the primary focus was not on comparative effectiveness, but whether a given treatment can be judged as efficacious. To this extent, their usefulness in the context of comparativeness effectiveness may be limited. However, these types of trials involve active-treatment control groups, and such trials provide direct evidence comparing treatments, which is a major interest of comparative effectiveness research. This common thread means that many of regulatory statements can have implications for systematic reviews of comparative effectiveness, particularly in the areas of 1) determining the minimum important difference, 2) assessing unique sources of bias in these types of trials, and 3) drawing conclusions based on these trials.

Definitions/Justification/Assumptions

The stated definitions of an "equivalence trial" were quite consistent (e.g., trials "designed to confirm the absence of a meaningful difference between treatments").[5] Regarding a "non-inferiority" trial, four documents[3,5,9,14] contained statements that the intent was to show that "the proposed drug is no worse than its main comparator"[9] (or used similar wording). Two documents,[8,11] however, pointed out that technically, "only a superiority trial can demonstrate this", because the exclusion of *any* possible inferiority logically requires a demonstration of superiority, and so the definition should include the idea that the intent is to show that the difference is not larger than some "pre-specified, small amount."[8]

Two documents[4,14] criticized the fact that some authors use the term equivalence trial to refer to what was actually a non-inferiority trial, and one document called the phrase "equivalence trial" a "misnomer, as true equivalence (i.e., assurance that the test drug is not *any* less effective than the control), could only be shown by demonstrating superiority."[11] This latter document appeared to argue against future use of the phrase "equivalence trial" for this reason.

Regarding valid justifications for performing EQ-NI trials, the documents cited four general reasons:

- Ethical problems with placebo-controlled trials and/or the existence of a proven treatment,[11,13,14] (i.e., if an effective treatment exists, the key issue from a patient's perspective is how the newer treatment compares to it, rather than whether a newer treatment works at all)
- Safety/cost/convenience advantages of the newer treatment that could counterbalance the reduced efficacy,[8,13,14]
- General interest in comparative efficacy or effectiveness,[10,11] and
- Infeasibility of a superiority trial.[4,8]

Several documents[4,11-13] stressed the importance of an underlying assumption of EQ-NI trials: that the active control treatment actually worked in the trial (e.g., by improving outcomes to the usual degree). This is a necessary assumption because if the active control treatment had not worked, then showing equivalence or noninferiority to it would not constitute evidence of efficacy of the new treatment being submitted for regulatory clearance. The relevant phrase "assay sensitivity" was used by some documents[4,11,12] to denote the ability of a trial to

distinguish efficacious from inefficacious treatments. A key component of assay sensitivity was a demonstration that the active control actually had its intended effect in the non-inferiority trial (possibly via comparison to data from prior studies).

Planning

The importance of pre-specifying the intent of the trial was a common refrain.[3,5,8,14] Some documents,[5,11,13,14] however, did state that it can be quite appropriate to draw conclusions from a trial even when the trial was not designed for that conclusion (see details on this point in the section below on Interpreting/reporting). One document[5] stated that the trial *intent* can be changed midstream, but optimally the trial would be designed with all reasonable possibilities in mind. This latter document provided conditions under which changes are acceptable. For example, switching from non-inferiority to superiority was generally more acceptable than the reverse; switching from superiority to non-inferiority would require prior specification of the non-inferiority margin.[5]

Regarding the types of control groups, six documents[3,8,10,11,13] stated that ideally the trial would contain at least three arms: the reference treatment, the control treatment, *and* a placebo treatment. One stated that a control group may actually not be necessary.[10] Several documents asserted the importance of using an appropriate drug for the active control treatment,[3,10,11] and at the appropriate dose.[4,9,10,13]

Trials measure multiple outcomes, so an outstanding question is whether a trial's initial intent applies to just the primary outcome, or to all outcomes. The two documents[8,14] mentioning this issue agreed that trials can have different intentions for different outcomes, and that the research protocol would state for "which outcomes noninferiority or equivalence hypotheses apply and for which superiority hypotheses apply".[14]

A priori power calculations utilizing the pre-specified threshold was mentioned by many documents,[3,5,9,11,13,14] but the key focus was on producing a confidence interval sufficiently narrow to permit a conclusion, regardless of whether an a priori power calculation had been performed. One document stated that *after a trial is complete*, "power calculations are of relatively little interest", because of the paramount importance of the post-trial confidence interval.[5]

The rest of this section is devoted to the threshold for decision making, which was referred to by the documents as the "margin" or "Δ" (delta). We first discuss the various definitions; then the importance of a priori specification; and finally the documents' advice for determining/justifying the chosen threshold.

Various definitions of the threshold revealed a subtle difference in wording. The two journal publications[13,14] described the threshold in terms of the *smallest value that would be clinically important*. Three regulatory documents[3,5,11] described it as the *largest difference that is clinically acceptable*. The importance of this difference in perspective is unclear, and may be due to the differing informational foci of academic authors and regulatory agencies. The ICH E10 document did not take either of these perspectives, instead defining the threshold generically as "the degree of inferiority of the test treatments to the control that the trial will attempt to exclude statistically."[4] None of the documents clarified what was meant by the word "clinically"; whether this meant a difference that would be *noticed by patients*, or a difference that would be *noticed by physicians*, or a difference that would *actually result in a change in clinical management*, or a difference in *other important outcomes*, or other possible interpretations.

The documents agreed on the importance of *a priori specification* of the decision threshold, whatever its definition.[3-5,7,8,10,11,13,14] The rationale is clear: post-hoc specification leaves "ample room for bias."[5] Specifically, if a researcher can choose the decision threshold after analyzing the data, then the researcher might choose it in such a way to bolster his own interpretation of the data. Of course, this post-hoc activity would not at all influence the actual *data*, but would influence the author's personal *interpretation* of the data.

Definitions and timing aside, the documents consistently acknowledged the difficulty faced by researchers in determining and justifying the chosen threshold. Mention of a focus on "clinical" impact was common,[3,4,8,9,11,13,14] but again it was unclear what exactly was meant by "clinical." Some mentioned that "statistical" considerations play a role in determining the threshold.[4,8,11] Other influencing factors mentioned included how well the active-control treatment works,[4,11,13] historical data,[4] safety,[14] cost,[14] acceptability,[14] adherence,[14] independent expert consensus,[13] and regulatory requirements.[13] One document[8] specifically stated that the choice of threshold should *not* be influenced by the anticipated sample size (because of the obvious bias). The documents were consistent in recommending that research justify the chosen threshold.[3,4,8,9,11-14]

Several specific statements were made about whether the threshold should be on a relative scale (e.g., relative risk or odds ratio) or an absolute scale (e.g., rate difference). One recommended the use of a relative scale,[13] another recommended that researchers can use either,[14] another recommended that they use both.[9] For continuous measures, one document[8] specifically recommended against the use of the standardized mean difference as the metric for the threshold. The reason was that "this statistic provides information on how difficult a difference would be to detect, but does not help justify the clinical relevance of the difference, and does not ensure that the test product is superior to placebo".[8]

Two documents disagreed on the appropriateness of the following strategy for setting the threshold. The FDA document[11] recommended that the researcher first set the value for M1, defined as "treatment effect of the active comparator" (i.e., how well the active control treatment is expected to work), and second define the value for M2, defined as the critical decision threshold for the non-inferiority trial, "by taking a percentage or fraction of M1."[11] For example, "A typical value for M2 is often 50% of M1, at least partly because the sample sizes needed to rule out a smaller loss become impractically large."[11] One of the EMEA documents,[8] however, recommended against this overall approach, stating:

"It is not appropriate to define the non-inferiority margin as a proportion of the difference between active comparator and placebo. Such ideas were formulated with the aim of ensuring that the test product was superior to (a putative) placebo; however they may not achieve this purpose. If the reference product has a large advantage over placebo this does not mean that large differences are unimportant, it just means that the reference product is very efficacious"[8]

Conducting

In this section, we list the unique aspects of risk-of-bias that, according to the reviewed documents, require particular attention in trials that define themselves as EQ-NI trials. In superiority trials, poor conduct tends to preclude any conclusion of superiority. But in EQ-NI trials, the documents agreed that a sloppily-run trial can actually bolster an incorrect conclusion of EQ or NI; if the trial had been better conducted, an important difference might have been observed. Below are the nine specific potential sources of bias mentioned:

- PATIENTS

    - Vague or inconsistently applied patient enrollment criteria or diagnostic criteria[3-5,11]
    - Patients selected for anticipated nonresponse or relatively high response rate[4,14]
- TREATMENTS
    - Poor compliance with treatments[3,4,11,14]
    - Concomitant treatments that mask true differences by producing a ceiling effect[4,11,13]
    - Inappropriate dosing[4,9,13]
    - Patients did not receive the treatment as assigned, or patients crossed over to the other treatment[5,11,14]
- MEASUREMENT
    - Dropouts[3,4,11,14]
    - Inadequate outcome measurement techniques[5,11]
    - Outcome assessor bias because they know that all groups received active treatment[4,11]

Many of the above sources of risk-of-bias also apply to trials that call themselves superiority trials. Below, Table 1 places each of these nine issues according to this concern as well as whether the potential would tend to lead to underestimates or overestimates (or both) of the difference between groups. We suspect that six of the nine issues could lead to either overestimates or underestimates. Further, it seems that all of these risk-of-bias issues would also be potential concerns in trials that call themselves superiority trials.

Table 1. Risk of bias issues mentioned in the documents

Issue	Would this tend to lead to underestimates or overestimates of the difference between groups?	For *SUPERIORITY* trials, is this also a potential source of bias?	Comments
Vague or inconsistently applied patient enrollment criteria or diagnostic criteria	Over or under	Yes	This adds noise to the data.
Patients selected for anticipated nonresponse or high response	Under	Yes	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 high response, then both treatments will appear effective.
Poor compliance	Over or under	Yes	If compliance is consistently poor, the difference would be underestimated. If compliance varies by treatment, the difference would be overestimated.
Concomitant treatments mask true differences	Under	Yes	In a superiority trial, this could result in a nonsignificant difference, even if the treatments truly differ
Inappropriate dosing	Over or under	Yes	The impact depends on whether dosing was too high or too low, and whether it was inappropriate in the same way for both groups
Patients did not receive the treatment as assigned, or patients crossed over to the other treatment	Under	Yes	In a superiority trial, this could result in a nonsignificant difference, even if the treatments truly differ
Dropouts	Over or under	Yes	The impact depends on whether dropouts were more common in one group than the other, and also the reasons for dropouts.
Inadequate outcome measurement techniques	Over or under	Yes	This would adds noise to the data.
Outcome assessor bias	Over or under	Yes	If assessors believe a priori that one treatment is better, then unblinded outcome measurement will be biased accordingly, regardless of the trial intent.

Analyzing

Several documents agreed that researchers should not solely rely on intention-to-treat analysis in EQ-NI trials.[3,5,9,11,13,14] The reason is that ITT analyses have the potential to obscure real differences between treatments. Instead they recommend that EQ-NI trials perform *both* intention-to-treat analysis and per-protocol analysis.[3,5,9,11,13,14] Ideally, the results of these two analyses would indicate the same conclusion (or lack thereof). If they disagree, one document instructed researchers to use the one that provided the "least positive" results.[13]

There was wide agreement that researchers should compute 95% confidence intervals (CI) around the difference between groups, instead of simply computing p values.[3,5-11,13,14] Some of the reasons provided were that the CI:

- "provides more information",[13]
- is the "most straightforward",[5] and
- "because the CI not only evaluates the null hypothesis but indicates the effect size and the lower and upper bounds of the this estimate".[13]

On whether to compute a one-tailed or two-tailed confidence interval in a non-inferiority trial, there was some disagreement among the documents. Two stated that one-sided comparisons should be used,[3,13] but these did not specify whether the interval should be 95% one-sided or 97.5% one-sided. Two other documents[5,11] argued for "two-sided 95% confidence intervals are to be used for all clinical trials whatever their objective."[5] The stated reason was "preserved consistency between significance testing and subsequent estimation …" If one-sided intervals are used, then they should be used with a coverage probability of 97.5%".[5] A fifth document[14] took a more middle-ground approach, suggesting that "2-sided CIs are appropriate in most noninferiority trials. If a 1-sided 5% significance level is deemed acceptable for the noninferiority hypothesis test (a decision open to question), a 90% 2-sided CI could then be used."[14]

Regarding multiplicity (i.e., statistical tests of multiple outcomes inflating the overall Type I error rate), one document[11] advocated statistical adjustment in order to preserve the overall Type I error rate at 5%.

Interpreting/Reporting

Three documents specifically stated that it would be inappropriate to draw a conclusion of equivalence or noninferiority solely on the basis of a non-significant difference.[3,9,13] The obvious reason is that a non-significant difference alone does not discriminate between two very different situations: 1) when the confidence interval is narrow enough to exclude the possibility of a meaningful difference, and 2) when the confidence interval is too wide to permit any conclusion about whether the treatments differed.

Several documents discussed the possibility of drawing a type of conclusion that was not originally intended in the design of the trial:

- Agreement among four documents[5,11,13,14] that it can be perfectly appropriate to draw a conclusion of superiority from a non-inferiority study, and two further stated that this situation requires no adjustment for multiplicity.[5,11] One document[13] made the additional point that the study must first establish non-inferiority before attempting to show superiority.
- Disagreement about the reasonableness of concluding equivalence from a superiority trial.[11,13] One document essentially said that this would be inappropriate,[13] but another[11] said that this would occur "only rarely," and the document outlined a specific example.

- Disagreement about the reasonableness of concluding non-inferiority from a superiority trial. Two documents said this can be acceptable if the decision threshold was pre-specified.[5,14] A third, however, stated that "Seeking an NI conclusion in the event of a failed superiority test would almost never be acceptable …. If it is clear that an NI conclusion is a possibility, the study should be designed as an NI study."[11]

The two academic papers listed several items that should be reported by studies calling themselves EQ or NI. Six items were listed by Gomberg-Maitlin et al. (2003)[13] listed six items, and 22 by Piaggio et al. (2006)[14] listed 22 items. Similarities were found between items 1-4 on the listed by Gomberg-Maitlin et al. (2003)[13] and items 3, 7, and 13 by Piaggio et al. (2006)[14]. The other two items listed by, Gomberg-Maitlin et al.(2003)[13] did not explicitly appear in the list by Piaggio et al. (2006).[14] These involved specifying the "minimum requisite number of primary events" and "a comparison of event rates during treatment with the active control in the trial and in the historical trials that established its efficacy compared with placebo." Piaggio et al. (2006)[14] listed numerous additional detailed items that were not listed by Gomberg-Maitlin et al. (2003)[13] (the full list of items appears in Appendix A).

Summary

For the development of EPC guidance on equivalence and non-inferiority in the context of systematic reviews, the three most important aspects of this methods project are: 1) the various justifications for performing these kinds of trials; 2) strategies for determining the threshold for decision-making; and 3) the unique areas of potential bias in trials that call themselves EQ-NI. We next discuss these three areas.

The four justifications included ethical problems with placebo-controlled trials and/or the existence of a proven treatment, safety/cost/convenience advantages of the newer treatment that could counterbalance the reduced efficacy, a general interest in comparative efficacy or effectiveness, and the infeasibility of a superiority trial. These ideas could assist systematic reviewers in devising the Key Questions (including items in PICOTS), and/or the study inclusion criteria.

The documents consistently acknowledged the difficulty faced by researchers in determining the chosen threshold. The guidance was limited to generalities about what to consider when determining the threshold, and systematic reviewers might use these to determine what threshold(s) the review will employ. The documents consistently mentioned the need to focus on "clinical" impact, but the specific meaning of this was unclear. Additional influencing factors included statistical considerations, how well the active-control treatment works, historical data, safety, cost, acceptability, adherence, independent expert consensus, and regulatory requirements. One benefit of the reviewer deciding the threshold(s) to be used for analysis is that interpretation of the conclusiveness of the data (based on confidence intervals) would be more straightforward. The documents were consistent in recommending that the researcher justify the chosen threshold (as well as that it should be chosen a priori), and these recommendations also seem to apply to thresholds chosen by systematic reviewers.

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. Further, almost all of the nine areas would also be potential sources of bias in trials that call themselves superiority trials.

Finally, we note that most of these documents originated from regulatory agencies, and such agencies have different informational needs than systematic reviewers. Systematic reviewers are

concerned about any direction of bias, whereas regulatory agencies are primarily concerned about bias *in favor of the sponsor's product*. In the context of an EQ or NI trial, this bias can act to produce an incorrect conclusion of EQ or NI, when in fact there is a meaningful difference between the treatments being compared. This is evidenced in the regulatory focus on minimizing the chance of a type I error, since from a regulatory perspective that would involve providing marketing clearance for an inefficacious drug or device. Systematic reviewers, on the other hand, aim for the best possible estimate of the extent to which two treatments yield different outcomes: too-low estimates of this difference are just as bad as too-high estimates.

A second area of difference involves the fact that regulatory agencies tend to 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. This is evident from the regulatory focus on whether a given trial has "assay sensitivity". The regulatory decision is often based on one or two trials conducted by a single sponsor, whereas systematic reviewers typically review evidence from many different groups. An outgrowth of this difference in perspective involves meta-analysis, which is common among systematic reviews but relatively rare in analyses used by regulatory agencies. Meta-analysis often involves the calculation of effect sizes from reported information, which means that authors' choices in interpreting and summarizing data are relatively unimportant, as long as the necessary underlying information is reported. For example, whether the author chose to perform both intention-to-treat analysis and per-protocol analysis is moot if they reported the underlying data necessary for a reviewer to perform the correct analyses. Also moot is whether the author deemed the evidence conclusive, as long as they reported their data sufficiently.

A third area involves differential interest in the generalizability or applicability of findings. This is important to any investigation of comparative "effectiveness," but a regulatory agency is more focused on a proof of biological impact rather than whether this impact has been realized in multiple populations or multiple settings.

References

1. Drugs and health products: ICH. [internet]. Ottawa (ON): Health Canada; 2011 Feb 15 [accessed 2011 May 19]. [1 p]. Available: http://www.hc-sc.gc.ca/dhp-mps/prodpharma/applic-demande/guide-ld/ich/index-eng.php.

2. Committee for Proprietary Medicinal Products (CPMP). Note for guidance on statistical principles for clinical trials. London (UK): European Medicines Agency (EMEA); 1998 Sep. 37 p. Also available: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002928.pdf.

3. 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.

4. 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.

5. 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.

6. 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.

7. 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.

8. 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.

9. 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.

10. 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.

11. 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.

12. 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.

13. 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

14. 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

Appendix A

This appendix contains verbatim quotations from the source documents that were reviewed. These quotations were selected for the degree of relevance to the EPC workgroup on equivalence and non-inferiority, and they are not intended to be an exhaustive representation of the content of the source documents.

Definitions/Justifications/Assumptions

Definitions of Equivalence Trial

"Although there is some variability in the definitions, equivalence trials are based on an active control, aiming to match the action of an established therapy and prove this to statistical significance within a predetermined range (Δ) in both positive and negative directions (a 2-sided test)."[13]

"An equivalence trial is designed to confirm the absence of a meaningful difference between treatments." [5]

"…equivalence trials aim to determine whether one (typically new) intervention is therapeutically similar to another, usually an existing treatment." [14]

Equivalence trial: "A trial with the primary objective of showing that the response to two or more treatments differs by an amount which is clinically unimportant. This is usually demonstrated by showing that the true treatment difference is likely to lie between a lower and an upper equivalence margin of clinically acceptable differences." [3]

Therapeutic equivalence: "A medicinal product is therapeutically equivalence with another product if it contains the same active substance or therapeutic moiety and, clinically, shows the same efficacy and safety as that product, whose efficacy and safety have been established." [6]

Definitions of Non-inferiority Trial

"In these [non-inferiority trials] we wish to show that a new treatment is no less effective than an existing treatment – it may be more effective or it may have a similar effect."[5]

"A noninferiority trial seeks to determine whether a new treatment is no worse than a reference treatment."[14]

"Noninferiority means that, in terms of effectiveness, the proposed drug is no worse than its main comparator."[9]

Noninferiority trial: "A trial with the primary objective of showing that the response to the investigational product is not clinically inferior to a comparative agent (active or placebo control)."[3]

"Many active control trials are designed to show that the efficacy of an investigational product is no worse than that of the active comparator, and hence fall into the latter category [of non-inferiority trials]."[3]

"The term 'non-inferiority' is now well established, but if taken literally could be misleading. The objective of a non-inferiority trial is sometimes stated as being to demonstrate that the test product is not inferior to the comparator. However, only a superiority trial can demonstrate this. In fact a noninferiority trial aims to demonstrate that the test product is not worse than the comparator by more than a pre-specified, small amount."[8]

"What non-inferiority trials seek to show is that any difference between the two treatments is small enough to allow a conclusion that the new drug has at least some effect or, in many cases, an effect that is not too much smaller than the active control."[11]

Confusion Among Terms

"Not all noninferiority or equivalence trials use these words, and the term equivalence is often inappropriately used when reporting negative results of superiority trials; such trials often lack statistical power to rule out important differences. Identifying noninferiority trials is difficult because they are often labeled as equivalence trials."[14]

"Clinical trials designed to demonstrate efficacy of a new drug by showing that it is similar in efficacy to a standard agent have been called "equivalence" trials. Most of these are actually non-inferiority trials, attempting to show that the new drug is not less effective than the control by more than a defined amount, generally called the margin".[4]

"These active control trials, which are not intended to show superiority of the test drug, but to show that the new treatment is not inferior to an unacceptable extent, were once called equivalence trials, but this is a misnomer, as true equivalence (i.e., assurance that the test drug is not any less effective than the control), could only be shown by demonstrating superiority. Because the intent of the trial is one-sided (i.e., to show that the new drug is not materially worse than the control), they are now called non-inferiority (NI) trials. But that too, is a misnomer, as guaranteeing that the test drug is not any (even a little) less effective than the control can only be demonstrated by showing that the test drug is superior. What non-inferiority trials seek to show is that any difference between the two treatments is small enough to allow a conclusion that the new drug has at least some effect or, in many cases, an effect that is not too much smaller than the active control."[11]

Justification

"Efficacy comparable to the standard strategy may be sufficient to justify the alternative when an advantage in safety, cost, or convenience is demonstrated (Figure 1)."[13]

"…equivalence and noninferiority trials should only be undertaken when a well-proven standard therapy exists"[13]

"Noninferiority trials are intended to show whether a new treatment has at least as much efficacy as the standard or is worse by an amount less than Δ, often on the premise that it has some other advantage, eg, greater availability, reduced cost, less invasiveness, fewer side effects (harms), or greater ease of administration, for instance, one daily dose rather than 2 doses or more than 2 doses."[14]

"A noninferiority or equivalence trial requires that the reference treatment's efficacy is established or is in widespread use so that a placebo or untreated control group would be deemed unethical."[14]

"Evidence for other advantages of the new treatment over the reference treatment, if present, should be given, to justify use of the new treatment, if not inferior"[14]

"When the superiority of a drug is confirmed over placebo, a non-inferiority study with existing drugs may not always be necessary. However, if the clinical significance of a drug is unclear, even in the case when a standard drug has already been established for the indicated disease and the superiority of drug against placebo has been confirmed, it may be appropriate to conduct a non-inferiority study in order to clarify the clinical positioning of a drug with respect to the standard drug (e.g., anti-infective drugs, etc.). Additionally, to clarify the positioning of a drug with respect to the existing drugs, a controlled study with 3 groups including placebo, the

investigational drug, and an existing drug as the control drug, may be useful, even if statistical power is not secured."[10]

"In most cases, evidence of efficacy is most convincingly demonstrated by showing superiority to a concurrent control treatment. If a superiority trial is not feasible or is inappropriate for ethical or practical reasons, and if a defined treatment effect of the active control is regularly seen (e.g., as it is for antibiotics in most situations), a non-inferiority or equivalence trial can be used and can be persuasive."[4]

"There are many situations where a non-inferiority trial might be performed as opposed to, or in addition to, a superiority trial over placebo. These include: • Applications based upon essential similarity in areas where bioequivalence studies are not possible, e.g. modified release products or topical preparations; • Products with a potential safety advantage over the standard might require an efficacy comparison to the standard to allow a risk-benefit assessment to be made; • Cases where a direct comparison against the active comparator is needed to help assess riskbenefit; • Cases where no important loss of efficacy compared to the active comparator would be acceptable; • Disease areas where the use of a placebo arm is not possible and an active control trial is used to demonstrate the efficacy of the test product. In the final 4 situations above a non-inferiority trial would not be necessary if superiority could be shown over the reference product."[8]

"The usual reason for using a non-inferiority active control study design instead of a study design having more readily interpretable results (i.e., a superiority trial) is an ethical one. Specifically, this design is chosen when it would not be ethical to use a placebo, or a treatment control, or a very low dose of an active drug, because there is an effective treatment that provides an important benefit (e.g., life-saving or preventing irreversible injury) available to patients for the condition to be studied in the trial…. There are, however, other reasons for using an active control: (1) interest in comparative effectiveness and (2) assessing the adequacy (assay sensitivity) of a placebo-controlled study."[11]

"If the comparator was suitable for a demonstration of non-inferiority, then there should be well-controlled data to show that it is an effective treatment."[5]

Assumptions

"Equivalence and noninferiority trials rely on the premises that the superior efficacy of the active control over placebo has been previously proven for a given indication, and that this efficacy will be preserved under the conditions of the trial"[13]

"Assay sensitivity is a property of a clinical trial defined as the ability to distinguish an effective treatment from a less effective or ineffective treatment. Assay sensitivity is important in any trial but has different implications for trials intended to show differences between treatments (superiority trials) and trials intended to show non-inferiority. If a trial intended to demonstrate efficacy by showing superiority of a test treatment to control lacks assay sensitivity, it will fail to show that the test treatment is superior and will fail to lead to a conclusion of efficacy. In contrast, if a trial is intended to demonstrate efficacy by showing a test treatment to be non-inferior to an active control, but lacks assay sensitivity, the trial may find an ineffective treatment to be non-inferior and could lead to an erroneous conclusion of efficacy."[4]

"…the NI [non-inferiority] study is dependent on knowing something that is not measured in the study, namely, that the active control had its expected effect in the NI study. This is critical to knowing that the trial had *assay sensitivity* (i.e., could have distinguished an effective from an ineffective drug). A successful superiority trial has, by definition, assay sensitivity. A

"successful" NI trial, one that shows what appears to be an acceptably small difference between treatments, may or may not have had assay sensitivity and may or may not have supported a conclusion that the test drug was effective. Thus, if the active control had no effect at all in the NI trial (i.e., did not have any of its expected effect), then finding even a very small difference between control and test drug is meaningless, providing no evidence that the test drug is effective. Knowing whether the trial had assay sensitivity relies heavily on external (not within-study) information, giving NI studies some of the characteristics of a historical control trial."[11]

"For non-inferiority trials used to demonstrate efficacy, the evidence supporting a determination that the trial had assay sensitivity and justifying the choice of non-inferiority margin."[12]

"For non-inferiority trials used to demonstrate efficacy, the evidence supporting a determination that the trial had assay sensitivity and justifying the choice of non-inferiority margin."[12]

The Japan PMDA[10] stated that regulators should determine an answer to the question "Even though non-inferiority has been confirmed, has superiority to placebo been denied in other clinical studies?."[10]

"If someone proposing to use an active-control or non-inferiority design cannot provide sufficient support for historical evidence of the sensitivity to drug effects of the study with the chosen non-inferiority margin, a finding of non-inferiority cannot be considered informative with respect to efficacy"[4]

"Lack of appropriate use of the control treatment would also make the study unusable as a non-inferiority study, if superiority of the test drug is not shown, because assay sensitivity of the study would not be ensured (see section 1.5.2)."[4]

"There are circumstances in which a finding of non-inferiority cannot be interpreted as evidence of efficacy. Specifically, for a finding of non-inferiority to be interpreted as showing efficacy, the trial needs to have had the ability to distinguish effective from less effective or ineffective treatments."[4]

Planning

Prespecifying the Trial Intent

"It is important to classify the therapeutic profile of the proposed drug in relation to its main comparator (ie whether it is therapeutically superior, inferior or equivalent to the comparator)."[9]

"Pre-definition of a trial as a superiority trial, an equivalence trial or a non-inferiority trial is necessary for numerous reasons including the following: to ensure that comparator treatments, doses, patient populations and endpoints are appropriate; to allow sample size estimates to be based on the correct power calculations; to ensure that equivalence and non-inferiority criteria are pre-defined; to permit appropriate analysis plans to be described in the protocol; to ensure that the trial has sufficient sensitivity to achieve its objectives."[5]

Piaggio et al. (2006)[14] listed one required reporting item "5. Specific objectives and hypotheses, including the hypothesis concerning noninferiority or equivalence."[14]

"It is vital that the protocol of a trial design to demonstrate equivalence or non-inferiority contain a clear statement that this is its explicit intention."[3]

Switching the Trial Intent

"Switching the objective of a trial from non-inferiority to superiority is feasible provided: the trial has been properly designed and carried out in accordance with the strict requirements of a non-inferiority trial; actual p-values for superiority are presented to allow independent assessment of the strength of the evidence; analysis according to the intention-to-treat principle is given greatest emphasis."[5]

"Switching the objective of a trial from superiority to non-inferiority may be feasible provided: the non-inferiority margin with respect to the control treatment was pre-defined or can be justified (the latter is likely to prove difficult and to be limited to rare cases where there is a widely accepted value for Δ); analysis according to the intention-to-treat principle and PP [per protocol] analysis, showing confidence intervals and p-values for the null hypothesis of inferiority, give similar findings; the trial was properly designed and carried out in accordance with the strict requirements of a non-inferiority trial; the sensitivity of the trial is high enough to ensure that it is capable of detecting relevant differences if they exist; there is direct or indirect evidence that the control treatment is showing its usual level of efficacy."[5]

"The problem of switching objectives post hoc can be avoided by designing a trial prospectively in the knowledge that both non-inferiority and superiority are outcomes of potential value."[5]

"For a well-designed and conducted trial, there are few difficulties connected with the change from non-inferiority to superiority that cannot be addressed by appropriate analysis. However, there are more severe difficulties associated with the switch from superiority to non-inferiority because of the possible need to find a basis for, and agree on, a margin of equivalence after seeing the outcome, and because of the inherent difficulties of non-inferiority trials."[5]

Inclusion of a Placebo Group

"A trial with 3 arms (one of which involves administration of a placebo) increases both the complexity and the sample size requirement, and may not be ethical when a superior treatment strategy has been established. However, if it is feasible, it allows the investigators, by use of regression analyses, to compare the investigative agent and active control individually to placebo and compare the investigative agent to active control. Hence equivalence or noninferiority trials can only be undertaken where a predictable control is accepted as the standard of care"[13]

"Where comparative effectiveness is the principal interest, it is usually important—where it is ethical, as would be the case in most symptomatic conditions—to include a placebo control as well as the active control. Trials of most symptomatic treatments have a significant failure rate (i.e., they often cannot show the drug is superior to placebo). Where that is the case in a comparative trial, seeing no difference between treatments is uninformative. Inclusion of a placebo group can provide clear evidence that the study did have assay sensitivity (the ability to distinguish effective from ineffective treatments), critical if a finding of no difference between treatments is to be interpretable. For example, we have seen that approximately 50% of all placebo-controlled antidepressant trials of effective agents cannot distinguish drug from placebo. A trial in which two antidepressants are compared and found to have a similar effect is informative only if we know that the two drugs can be distinguished from the concurrent placebo group."[11]

"When the superiority of a drug is confirmed over placebo, a non-inferiority study with existing drugs may not always be necessary. However, if the clinical significance of a drug is unclear, even in the case when a standard drug has already been established for the indicated

disease and the superiority of drug against placebo has been confirmed, it may be appropriate to conduct a non-inferiority study in order to clarify the clinical positioning of a drug with respect to the standard drug (e.g., anti-infective drugs, etc.). Additionally, to clarify the positioning of a drug with respect to the existing drugs, a controlled study with 3 groups including placebo, the investigational drug, and an existing drug as the control drug, may be useful, even if statistical power is not secured."[10]

"Active control equivalence or non-inferiority trials may also incorporate a placebo, thus pursuing multiple goals in one trial; for example, they may establish superiority to placebo and hence validate the trial design and simultaneously evaluate the degree of similarity of efficacy and safety to the active comparator. There are well-known difficulties associated with the use of the active control equivalence (or non-inferiority) trials that do not incorporate a placebo or do not use multiple doses of the new drug. These relate to the implicit lack of any measure of internal validity (in contrast to superiority trials), thus making external validation necessary."[3]

"A three-armed trial with test, reference and placebo allows some within-trial validation of the choice of non-inferiority margin and is therefore the recommended design; it should be used wherever possible"[8]

"Even where it would be ethical to include a placebo group in addition to the active treatments (e.g., in studies of a symptomatic treatment), one is not necessarily included in these comparative trials. Such omission of a placebo group may render such studies uninformative, however, when they show no difference between treatments, unless assay sensitivity can be supported in some other way."[11]

Choice of Active Control Treatment

"In a disease area where the placebo responder rate is presumed to be constant, results showing the investigational drug's non-inferiority against an existing drug or results from an objective and appropriate clinical study even without a control group may be sufficient for the evaluation."[10]

The Japan PMDA[10] stated that for showing non-inferiority/superiority against an active control, regulators should determine answers to the questions "Is the control drug appropriate?" and "Is the dosage of the control drug appropriate?"[10]

"Active comparators should be chosen with care. An example of a suitable active comparator would be a widely used therapy whose efficacy in the relevant indication has been clearly established and quantified in well designed and well documented superiority trial(s) and which can be reliably expected to exhibit similar efficacy in the contemplated active control trial."[3]

"In practice, an active control equivalence or non-inferiority trial offered as evidence of efficacy also almost always needs to provide a fair effectiveness comparison with the control, because any doubt as to whether the control in the study had its usual effect would undermine assurance that the trial had assay sensitivity (see section 1.5). Among aspects of trial design that could unfairly favor one treatment are choice of dose or patient population and selection and timing of endpoints".[4]

"The active control must be a drug whose effect is well-defined. The most obvious choice is the drug used in the historical placebo-controlled trials. Where studies of several pharmacologically similar drugs have been pooled, which is often done to obtain a better estimate of effect and a narrower confidence interval, and thus a larger M1, the choice may become complicated. In general, if the drugs in a meta-analysis of placebo-controlled trials seem to have similar effects, any of them could be used as an active control. If their observed

treatment effects differ, however, even if not significantly, the one with the highest point estimate of effect should ordinarily be used."[11]

"Another possibility is a trial in which multiple doses of the investigational drug are compared with the recommended dose or multiple doses of the standard drug. The purpose of this design is simultaneously to show a dose-response relationship for the investigational product and to compare the investigational product with the active control."[3]

Multiple Outcomes

"The authors should specify for which outcomes noninferiority or equivalence hypotheses apply and for which superiority hypotheses apply."[14]

"A clinical trial or clinical programme may plan to show noninferiority for certain variables while superiority may be the objective for others"[8]

Power and Sample Size

"A satisfactory approach to this subject requires an understanding of confidence intervals and the manner in which they capture the results of the trial and indicate the conclusions that can be drawn from them. Such an understanding also leads to an appreciation of why power calculations are of relatively little interest *when a trial is complete*."[5][italics added]

"A specifically designed noninferiority direct randomised trial would have specified a noninferiority threshold in its power calculation and so might have provided one or more grounds to justify this threshold as a prespecified minimal clinically important difference (MCID)."[9]

"The acceptable breadth of the CI around the point estimate of the difference between event rates with alternative treatments is one of the principal determinants of sample size"[13]

"The required sample size is calculated using the confidence interval (CI) approach, considering where the CI for the treatment effect lies with respect to both the margin of noninferiority Δ and a null effect."[14]

"The sample size of an equivalence trial or a non-inferiority trial (see Section 3.3.2) should normally be based on the objective of obtaining a confidence interval for the treatment difference that shows that the treatments differ at most by a clinically acceptable difference."[3]

"It is important to plan the sample size for an NI clinical trial so that the trial will have the statistical power to conclude that the NI margin is ruled out if the test drug is truly non-inferior."[11]

Decision Threshold: Definitions

"…the margin of inferiority must be clinically small and not exceed the slightest possible effect of the active control"[13]

"A margin of clinical equivalence (Δ) is chosen by defining the largest difference that is clinically acceptable, so that a difference bigger than this would matter in practice."[5]

"A prestated margin of noninferiority is often chosen as the smallest value that would be a clinically important effect."[14]

"… this margin is the largest difference that can be judged as being clinically acceptable and should be smaller than differences observed in superiority trials of the active comparator".[3]

"This margin is the degree of inferiority of the test treatments to the control that the trial will attempt to exclude statistically."[4]

"It is therefore usual in NI [non-inferiority] studies to choose a smaller margin (M2) that reflects the largest loss of effect that would be clinically acceptable. This can be described as an absolute difference in effect (typical of antibiotic trials) or as a fraction of the risk reduction provided by the control (typical in cardiovascular outcome trials). Note that the clinically acceptable margin could be relaxed if the test drug were shown to have some important advantage (e.g., on safety or on a secondary endpoint)."[11]

Decision Threshold: A Priori Specification

"It is always possible to choose a value of Δ which leads to a conclusion of equivalence or non-inferiority if it is chosen after the data have been inspected. Since the choice of Δ is generally a difficult one, there is ample room for bias here, however well-intentioned the researcher may be. Plausible arguments may often be advanced for a retrospective choice. In the design of equivalence and non-inferiority trials, this reason (amongst others) makes it necessary for the choice of Δ, and the reasoning behind the choice, to be set down in advance by the researcher in the study protocol."[5]

"Although there is some variability in the definitions, equivalence trials are based on an active control, aiming to match the action of an established therapy and prove this to statistical significance within a predetermined range (Δ) in both positive and negative directions (a 2-sided test)."[13]

"The results should be evaluated by using the predefined criteria for defining equivalence or non-inferiority and the rationale for the criteria and support for the determination that the study (studies) had assay sensitivity should be provided (see ICH E10)."[7]

"Because proof of exact equality is impossible, a prestated margin of noninferiority (Δ) for the treatment effect in a primary patient outcome is defined."[14]

The Japan PMDA[10] stated that for showing non-inferiority/superiority against an active control, regulators should determine answers to the question "Is the endpoint appropriate and is the *pre-determined* non-inferiority limit (Δ) appropriate (in the case of a non-inferiority study)?"[10]

"An equivalence margin should be specified in the protocol".[3]

"Prior to the trial, an equivalence or non-inferiority margin, sometimes called Δ, is selected."[4]

"In order to demonstrate non-inferiority, the recommended approach is to pre-specify a margin of noninferiority in the protocol"[8]

"As described above, the NI study seeks to show that the difference in response between the active control (C) and the test drug (T), (C-T), the amount by which the control is superior to test drug, is less than some pre-specified non-inferiority margin (M)."[11]

Decision Threshold: Determining and Justifying

One EMEA document[8] focused entirely on this issue; see http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003636.pdf

It stated "The choice of Δ must always be justified on both clinical and statistical grounds. It always needs to be tailored specifically to the particular clinical context and no rule can be provided that covers all clinical situations. However, certain principles can be used to provide general guidance."[8]

Also see the FDA document[11] section III.A.4 (starting on page 7) and section IV (starting on page 17) that discuss this issue in detail (http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM202140.pdf). This document states that the critical threshold for the non-inferiority (NI) study is M2, which is based on first determining M1, which is the "treatment effect of the active comparator". "The determination of M2 is based on clinical judgment and is usually calculated by taking a percentage or fraction of M1." "Deciding on the NI clinical margin M2 is also a relatively straightforward concept. It is plainly a matter of judgment about how much of the treatment effect must be shown to be preserved, a consideration that may reflect the seriousness of the outcome, the benefit of the active comparator, and the relative safety profiles of the test and comparator."[11]

"A typical value for M2 is often 50% of M1, at least partly because the sample sizes needed to rule out a smaller loss become impractically large."[11]

"…explain and justify on clinical or other grounds the value of the noninferiority threshold difference in treatment effect"[9]

"Demonstrate that a systematic approach has been taken in the search for relevant and appropriate references to support the nominated threshold"[9]

"Discuss the clinical importance of the primary outcome and secondary outcomes listed in Tables B.5.1 and B.5.2. For primary outcomes, this might be informed by the basis given in the trial protocol for the minimal clinically important difference used in the power calculation. Discuss clinical importance in terms of both relative and absolute changes"[9]

"The size of the acceptable margin depends on the smallest clinically significant difference (preferably established by independent expert consensus), expected event rates, the established efficacy advantage of the control over placebo, and regulatory requirements."[13]

"A prestated margin of noninferiority Δ can be specified as a difference in means or proportions or the logarithm of an odds ratio, risk ratio, or hazard ratio."[14]

"When designing an equivalence or noninferiority trial, it may be preferable to express the marginal difference between event rates in terms of constant ratios. When patients at lower risk dilute the target population, event rates should decrease by the same proportion in the groups assigned to each treatment. Even when event rates are variable, it may also be helpful to assess trial results with odds ratios and CI techniques."[13]

"For non-inferiority trials used to demonstrate efficacy, the evidence supporting a determination that the trial had assay sensitivity and justifying the choice of non-inferiority margin."[12]

"There are several techniques to determine Δ, its magnitude being influenced by several factors, e.g., efficacy, safety, cost, acceptability, and adherence."[14]

Piaggio et al. (2006)[14] listed one required reporting item (item 7) that authors should specify the margin of equivalence as well as "the rational for its choice".[14]

"The margin of noninferiority or equivalence should be specified, and preferably justified on clinical grounds".[14]

"The choice of a 'clinically acceptable' difference needs justification with respect to its meaning for future patients, and may be smaller than the 'clinically relevant' difference referred to above in the context of superiority trials designed to establish that a difference exists.[3]

"The choice of equivalence margins should be justified clinically".[3]

"An acceptable non-inferiority margin should be defined, taking into account the historical data and relevant clinical and statistical considerations".[4]

"The determination of the margin in a non-inferiority trial is based on both statistical reasoning and clinical judgment, should reflect uncertainties in the evidence on which the choice is based, and should be suitably conservative."[4]

"The margin chosen for a non-inferiority trial cannot be greater than the smallest effect size that the active drug would be reliably expected to have compared with placebo in the setting of the planned trial. If a difference between active control and the new drug favors the control by as much as or more than this margin, the new drug might have no effect at all."[4]

"The margin generally is identified based on past experience in placebo-controlled trials of adequate design under conditions similar to those planned for the new trial, but could also be supported by dose response or active control superiority studies. Regardless of the control groups used in those earlier studies, the value of interest in determining the margin is the measure of superiority of the active treatment to its control, not uncontrolled measures such as change from baseline."[4]

In the context of bioequivalence only and the use of AUC-ratio [area under the curve) or Cmax ratio as the outcome: "The 90 percent confidence interval for this measure of relative bioavailability should lie within an acceptance interval of 0.80-1.25."[6]

"It is not appropriate to define the non-inferiority margin as a proportion of the difference between active comparator and placebo. Such ideas were formulated with the aim of ensuring that the test product was superior to (a putative) placebo; however they may not achieve this purpose. If the reference product has a large advantage over placebo this does not mean that large differences are unimportant, it just means that the reference product is very efficacious"[8]

"It is not appropriate to use effect size (treatment difference divided by standard deviation) as justification for the choice of non-inferiority margin. This statistic provides information on how difficult a difference would be to detect, but does not help justify the clinical relevance of the difference, and does not ensure that the test product is superior to placebo"[8]

"Where 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 non-inferiority margin of any size. Discussion of the number of extra deaths that are acceptable is ethically very difficult. However it is not in the best interest of public health to reject all choices of margin. Unless a statistically significant difference has been found between treatments, the confidence interval for the difference will not only indicate that the test product has a possible inferiority to the reference, but it will also show that it has possible superiority. Hence even if we think we are not prepared to accept any possible level of inferiority we are accepting some, by continuing to use the currently authorized product. It is important therefore that non-inferiority trials should still be possible in these areas"[5]

"Determining the NI margin is the single greatest challenge in the design, conduct, and interpretation of NI trials"[11]

"As described above, the NI study seeks to show that the difference in response between the active control (C) and the test drug (T), (C-T), the amount by which the control is superior to test drug, is less than some pre-specified non-inferiority margin (M). M can be no larger than the presumed entire effect of the active control in the NI study, and the margin based on that whole active control effect is generally referred to as M1. It is critical to reiterate that M1 is not measured in the NI trial, but must be assumed based on past performance of the active control, the comparison of the current NI study with prior studies, and assessment of the quality of the NI study (see below). The validity of any conclusion from the NI study depends on the choice of M1. If, for example, the NI margin is chosen as 10 (because we are sure the control had an effect

of at least that size), and the study does indeed rule out a difference of 10 (seeming to demonstrate "effectiveness" of T), but the true effect of C in this study was actually less than 10, say 5, T would not in fact have been shown to have any effect at all; it will only appear to have had such an effect. The choice of M1, and assurance that this effect was present in the trial (i.e., the presence of assay sensitivity) is thus critical to obtaining a meaningful, correct answer in an NI study."[11]

"Although the NI margin used in a trial can be no larger than the entire assumed effect of the active control in the NI study (M1), it is usual and generally desirable to choose a smaller value, called M2, for the NI margin. Showing non-inferiority to M1 would provide assurance that the test drug had an effect greater than zero. However, in many cases that would not be sufficient assurance that the test drug had a clinically meaningful effect. After all, the reason for using the NI design is the perceived value of the active control drug. It would not usually be acceptable to lose most of that active control's effect in a new drug. It is therefore usual in NI studies to choose a smaller margin (M2) that reflects the largest loss of effect that would be clinically acceptable. This can be described as an absolute difference in effect (typical of antibiotic trials) or as a fraction of the risk reduction provided by the control (typical in cardiovascular outcome trials). Note that the clinically acceptable margin could be relaxed if the test drug were shown to have some important advantage (e.g., on safety or on a secondary endpoint)."[11]

"As described in section III, the selection of a margin for an NI study is a two-step process. The first step involves making a reasonable assumption about the effect of the active comparator in the NI study. M1 is chosen to equal that treatment effect. If the advantage of the control over the test drug in the NI study is larger than M1, then the test drug has not been shown to have any effect. Effectiveness is therefore demonstrated by showing that the advantage of the control over the test drug (C-T) is smaller than M1. This can be demonstrated by showing that the upper bound of the 95 percent CI of C-T is below M1. This is very similar to testing a superiority finding at $P \leq 0.05$. If we rule out loss of the entire assumed effect of the control, we can conclude that the test drug is superior to placebo. In most situations where active control studies are used, however, assuring some effect greater than zero is not clinically sufficient, and the second step in selecting the NI margin is choosing a specified portion of the control effect (M1) whose loss by the test product must be ruled out. This new non-inferiority margin is called M2, and is based upon clinical judgment. The multiple steps and assumptions that are made in determining an NI margin are all potential sources of uncertainty that may be introduced into the results and conclusions of an NI study. This guidance attempts to identify these sources and suggest approaches to accounting for these uncertainties so that we can reduce the possibility of drawing false conclusions from an NI study."[11]

"The choice of margin should be independent of considerations of power. It should be based upon the clinical and statistical principles noted in later sections of this document and not upon issues of sample size, as the size of the clinically important difference is not altered by the size of the study. A small study is not a justification for a wider non-inferiority margin"[8]

"Let us suppose that a bioequivalence trial finds a 90 percent confidence interval for the relative bioavailability of a new formulation lies that ranges from 0.90 to 1.15. Can we only conclude that the relative bioavailability lies between the conventional limits of 0.80 and 1.25 because these were the pre-defined equivalence margins? Or can we conclude that it lies between 0.90 and 1.15? The narrower interval based on the actual data is the appropriate one to accept…. However, if the trial had resulted in a confidence interval ranging from 0.75 to 1.20, then a post

hoc change of equivalence margins to ±25 percent would not be acceptable because of the obvious conclusion that the equivalence margin was chosen to fit the data."[5]

Conducting

"A variety of study quality deficiencies can introduce what is known as a "bias toward the null," where the observed treatment difference in an NI study is decreased from the true difference between treatments. These deficiencies include imprecise or poorly implemented entry criteria, poor compliance, and use of concomitant treatments whose effects may overlap with the drugs under study, inadequate measurement techniques, or errors in delivering assigned treatments. Many such defects have small (or no) effects on the variability of outcomes (variance) but reduce the observed difference C-T, potentially leading to a false conclusion of non-inferiority. It should also be appreciated that intent-to-treat approaches, which preserve the principle that all patients are analyzed according to the treatment to which they have been randomized even if they do not receive it, although conservative in superiority trials, are not conservative in an NI study, and can contribute to this bias toward the null. It is more important than usual to plan in advance steps to ensure quality during the conduct of an NI study. Finally, it should be recognized that although most investigators seek to carry out high quality trials, the incentives in an NI study are perverse, and quite different from those in superiority trials. In a superiority trial, sloppiness can lead to study failure, and major efforts in trial conduct and monitoring are therefore devoted to avoiding it. In general, sloppiness of any sort obscures true treatment differences. In an NI trial, in contrast, where the goal is to show no difference (or no difference greater than M), poor quality can sometimes lead to an apparent finding of non-inferiority that is incorrect. There is therefore a critical need for particular attention to study quality and conduct when planning and executing an NI study."[11]

"To avoid masking differences between treatments, concomitant medications must be carefully addressed. For example, administration of a standard dose of a beneficial concomitant medication to all patients may produce a ceiling effect that masks differences between the treatments under investigation. Use of this medication more often in one treatment group than the other may also bias the result."[13]

"A trial to show equivalence or non-inferiority must show a high degree of consistency with protocolled plans if it is to be reliable. Deviations from the inclusion criteria, from the intended treatment regimen, from the schedule, manner and precision of taking measurements, and so on, all tend to reduce the sensitivity of a trial and to make a conclusion of 'no difference' more likely, even when the deviations are of an unsystematic or random nature. The size of the bias associated with these and other departures from the protocol is generally unknown and may render such a trial uninterpretable. Failure to show a difference between two treatments can also arise when both treatments are inefficacious, perhaps as a result of being inappropriately administered."[5]

"Trial conduct should closely match any trial that demonstrated efficacy of the reference treatment, provided they were of high quality. One should avoid features that might dilute true differences between treatments, thereby enhancing the risk of erroneously concluding noninferiority, e.g., poor adherence, dropouts, recruitment of patients unlikely to respond, and treatment crossovers."[14]

"The equivalence (or non-inferiority) trial is not conservative in nature, so that many flaws in the design or conduct of the trial will tend to bias the results towards a conclusion of equivalence. For these reasons, the design features of such trials should receive special attention

and their conduct needs special care. For example, it is especially important to minimize the incidence of violations of the entry criteria, non-compliance, withdrawals, losses to follow-up, missing data, and other deviations from the protocol, and also to minimize their impact on the subsequent analyses."[3]

"There are many factors in the conduct of a trial that can reduce the observed difference between an effective treatment and a less effective or ineffective treatment and therefore may reduce a trial's assay sensitivity, such as: 1. Poor compliance with therapy 2. Poor responsiveness of the enrolled study population to drug effects 3. Use of concomitant non-protocol medication or other treatment that interferes with the test drug or that reduces the extent of the potential response 4. An enrolled population that tends to improve spontaneously, leaving no room for further drug-induced improvement 5. Poorly applied diagnostic criteria (patients lacking the disease to be studied) 6. Biased assessment of endpoint because of knowledge that all patients are receiving a potentially active drug, e.g., a tendency to read blood pressure responses as normalized, potentially reducing the difference between test drug and control"[4]

"…in trials intended to show that there is not a difference of a particular size (noninferiority) between two treatments, there may be a much weaker stimulus to engage in many of these efforts to ensure study quality that will help ensure that differences will be detected, i.e., that ensure assay sensitivity. The kinds of trial error that diminish observed differences between treatments (e.g., poor compliance, high placebo response, certain concomitant treatment, misclassification of outcomes) are of particular concern with respect to preservation of assay sensitivity."[4]

"As noted, to determine that a non-inferiority trial had appropriate trial conduct, its conduct should be reviewed not only for the presence of factors that might obscure differences between treatments but also for factors that might make the trial different from the trials that provided the basis for determining the non-inferiority margin. In particular, it should be determined whether any observed differences in the populations enrolled, the use of concomitant therapies, compliance with therapy, and the extent of, and reasons for, dropping out could adversely affect assay sensitivity. Even when the design and conduct of a trial appear to have been quite similar to those of the trials providing the basis for determining the non-inferiority margin, outcomes with the active control treatment that are visibly atypical (e.g., cure rate in an antibiotic trial that is unusually high or low) can indicate that important differences existed."[4]

"Other sources of bias that could occur in any study are also of concern in the NI study and are of particular concern in an open label study. For such open label NI studies, how best to ensure unbiased assessment of endpoints, unbiased decisions about inclusion of patients in the analysis, and a wide variety of other potential biases, need particular attention."[11]

"Selection of subjects for an active control trial can affect outcome; the population studied should be carefully considered in evaluating what the trial has shown. For example, if many subjects in a trial have previously failed to respond to the control treatment, there would be a bias in favor of the new treatment."[4]

One of the reporting items listed by Piaggio et al. (2006)[14] was item 3: "Eligibility criteria for participants (*detailing whether participants in the noninferiority 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."[14]

"In examining the results of a comparison of two treatments, it is important to consider whether an apparently less effective treatment has been used at too low a dose or whether the apparently less well tolerated treatment has been used at too high a dose."[4]

"Any conclusion of noninferiority should be accompanied by a determination of equi-effective doses"[9]

"Calculate equi-effective doses at 'steady state'. In other words, the dose of each drug should be the average dose used by the remaining participants after dose titrations are complete and after excluding participants who discontinue the drug (note that this is similar to the method used to calculate doses from Level 5 evidence). Assess the impact of extrapolating dose titration if there is evidence that the trial was of inadequate duration for the doses to have reached steady state"[9]

"Dosing should be well founded, because too high or low a dose for either or both treatments could fall below or exceed a critical threshold and lead to an erroneous conclusion of equivalency."[13]

Analyzing

Intention-to-treat (ITT)

…"indicate whether the analysis of each trial was conducted on a per protocol basis (which is appropriate for an analysis in support of a conclusion of noninferiority, because it helps examine any impact on the conclusions of losses to follow-up or poor compliance), as well as the standard intention-to-treat (ITT) basis (which is the generally preferred basis for an analysis; see Part II, Subsection B.2)."[9]

"A combination of efficacy and intention-to-treat analyses is necessary to promote full understanding of data from all randomized patients in this type of trial. The investigators must prespecify which method will be primary. If the 2 approaches give marginally different results, the investigators should trust the least positive result as valid."[13]

"In a non-inferiority trial, the full analysis set and the PP [per-protocol] set have equal importance and their use should lead to similar conclusions for a robust interpretation."[5]

"In a superiority trial the full analysis set, based on the ITT (intention-to-treat) principle, is the analysis set of choice, with appropriate support provided by the PP (per protocol) analysis set. In a non-inferiority 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. A switch of objective would require this difference of emphasis to be recognized."[5]

"In noninferiority trials, ITT analysis will often increase the risk of falsely claiming noninferiority (type I error), although not always."[14]

"Alternative analyses that exclude patients not taking allocated treatment or otherwise not protocol-adherent could bias the trial in either direction. The terms on-treatment or per-protocol analysis are often used but may be inadequately defined. Potentially biased non-ITT analysis is less desirable than ITT in superiority trials but may still provide some insight. In noninferiority and equivalence trials, non-ITT analyses might be desirable as a protection from ITT's increase of type I error risk (falsely concluding noninferiority). There is greater confidence in results when the conclusions are consistent."[14]

"Subjects who withdraw or dropout of the treatment group or the comparator group will tend to have a lack of response, and hence the results of using the full analysis set (see Glossary) may be biased toward demonstrating equivalence."[3]

"The full analysis set and the per protocol set play different roles in superiority trials (which seek to show the investigational product to be superior), and in equivalence or non-inferiority trials (which seek to show the investigational product to be comparable, see section 3.3.2). In

superiority trials the full analysis set is used in the primary analysis (apart from exceptional circumstances) because it tends to avoid over-optimistic estimates of efficacy resulting from a per protocol analysis, since the non-compliers included in the full analysis set will generally diminish the estimated treatment effect. However, in an equivalence or non-inferiority trial use of the full analysis set is generally not conservative and its role should be considered very carefully."[3]

"It should also be appreciated that intent-to-treat approaches, which preserve the principle that all patients are analyzed according to the treatment to which they have been randomized even if they do not receive it, although conservative in superiority trials, are not conservative in an NI study, and can contribute to this bias toward the null."[11]

"Intent-to-treat (ITT) analyses in superiority trials are nonetheless preferred because they protect against the kinds of bias that might be associated with early departure from the study. In non-inferiority trials, many kinds of problems fatal to a superiority trial, such as non-adherence, misclassification of the primary endpoint, or measurement problems more generally (i.e., "noise"), or many dropouts who must be assessed as part of the treated group, can bias toward no treatment difference (success) and undermine the validity of the trial, creating apparent non-inferiority where it did not really exist. Although an "as-treated" analysis is therefore often suggested as the primary analysis for NI studies, there are also significant concerns with the possibility of informative censoring in an as-treated analysis. It is therefore important to conduct both ITT and as-treated analyses in NI studies. Differences in results using the two analyses will need close examination."[11]

Use Confidence intervals

"The appropriate comparison to present is the point estimate of the difference with its 95 percent confidence interval. This allows PBAC to assess whether the confidence interval contains the minimal clinically important difference"[9]

"The CI provides more information than the P value in both superiority and equivalence, and noninferiority trials, because the CI not only evaluates the null hypothesis but indicates the effect size and the lower and upper bounds of the this estimate, as well"[13]

About non-inferiority: "Again a confidence interval approach is the most straightforward way of performing the analysis but now we are only interested in a possible difference in one direction."[5]

"The interpretation of superiority trials as non-inferiority trials and vice versa is best approached by expressing the results as a confidence interval for the difference between the test treatment and control. There is no fundamental problem associated with the use of this confidence interval as a basis for either mode of interpretation."[5]

"If the objective of an active control trial was to show equivalence or non-inferiority, the difference or the ratio of outcomes between treatments should be given with the confidence interval."[7]

"Although a modified hypothesis testing framework exists, a more informative CI approach is preferred in the design, analysis, and reporting of noninferiority and equivalence trials."[14]

"Equivalence is inferred when the entire confidence interval falls within the equivalence margins".[3]

In the context of bioequivalence only: "The statistical method for testing bioequivalence is based upon the 90 percent confidence interval for the ratio of the population means (Test/Reference), for the parameters under consideration."[6]

"After study completion, a two-sided 95 percent confidence interval (or one-sided 97.5 percent interval) for the true difference between the two agents will be constructed. This interval should lie entirely on the positive side of the non-inferiority margin."[8]

"Again, non-inferiority is established by showing that the upper bound of the two-sided confidence interval for C-T is < M1."[11]

One-tailed/two-tailed

"…equivalence trials are based on an active control, aiming to match the action of an established therapy and prove this to statistical significance within a predetermined range (d), in both positive and negative directions (a 2-sided test)"[13]

"Noninferiority trials are statistically based on a 1-sided comparison to an active control in the positive direction (a 1-sided test)."[13]

"For non-inferiority trials a one-sided interval should be used".[3]

"…two-sided 95 percent confidence intervals are to be used for all clinical trials whatever their objective. Among other benefits, this preserved consistency between significance testing and subsequent estimation….If one-sided intervals are used, then they should be used with a coverage probability of 97.5 percent. In the special case of bioequivalence studies, two-sided 90 percent confidence intervals have been established as the norm…"[5]

"Again, non-inferiority is established by showing that the upper bound of the two-sided [95%] confidence interval for C-T is < M1."[11]

"Many noninferiority trials based their interpretation on the upper limit of a 1-sided 97.5 percent CI, which is the same as the upper limit of a 2-sided 95 percent CI. Although both 1-sided and 2-sided CIs allow for inferences about noninferiority, we suggest that 2-sided CIs are appropriate in most noninferiority trials. If a 1-sided 5 percent significance level is deemed acceptable for the noninferiority hypothesis test (a decision open to question), a 90 percent 2-sided CI could then be used."[14]

Adjustment for multiple outcomes

"A possibility that has thus far had relatively little attention is to have different endpoints with different goals (e.g., superiority on the composite endpoint of death, AMI, and stroke, but NI on death alone). The multiple endpoints would require some alpha adjustment in such a case, but the procedures here are not well defined. Similarly, if a study had several doses, with interest in NI on each of them and, at the same time, interest in a potential superiority finding for one or more doses, the analytical approach is not yet fully established, although it is clear that some correction for multiplicity would be needed."[11]

"When there are multiple endpoints or multiple doses of the test treatment evaluated in an NI study, the valid statistical decision tree can be very complex. Using the same 95 percent confidence interval to test non-inferiority and superiority at each endpoint level or at each dose may inflate the overall Type I error rate associated with drawing one or more false conclusions from such multiple comparisons, regardless of whether they are non-inferiority or superiority testing. Thus, for any statistical decision tree composed of tests of superiority and non-inferiority in multiple comparison settings, it is imperative to evaluate the overall Type I error rate for all the comparisons involved in the testing and make appropriate statistical adjustments."[11]

Interpreting/reporting

Drawing conclusions

"…it is not adequate to demonstrate the absence of a statistically significant difference between the treatments to claim equivalence; such a lack of a significant difference might occur when the trials are too small to demonstrate a real difference in the effects of the interventions."[9]

"The difference between lack of proof of superiority and "equivalence" is important. In other words, a lack of "evidence of a difference" is not the same as evidence of a lack of difference."[13]

"Concluding equivalence or non-inferiority based on observing a non-significant test result of the null hypothesis that there is no difference between the investigational product and the active comparator is inappropriate".[3]

"With 2-sided equivalence the interpretation is analogous, but both margins Δ and $-\Delta$ need considering, and claiming equivalence requires the CI to lie wholly between $-\Delta$ and Δ."[14]

"The presence of assay sensitivity in a non-inferiority or equivalence trial may be deduced from two determinations: 1) Historical evidence of sensitivity to drug effects, i.e., that similarly designed trials in the past regularly distinguished effective treatments from less effective or ineffective treatments and 2) Appropriate trial conduct, i.e., that the conduct of the trial did not undermine its ability to distinguish effective treatments from less effective or ineffective treatments."[4]

Drawing a superiority conclusion based on data from a non-inferiority trial

"If the 95 percent confidence interval for the treatment effect not only lies entirely above $-\Delta$ but also above zero then there is of superiority in terms of statistical significance at the 5 percent level ($p<0.05$)….In this case it is acceptable to calculate the p-value associated with a test of superiority and to evaluate whether this is sufficiently small to reject convincingly the hypothesis of no difference. There is no multiplicity argument that affects this interpretation because, in statistical terms, it corresponds to a simple closed test proceedure."[5]

"Once noninferiority is evident, it is acceptable to then assess whether the new treatment appears superior to the reference treatment, using an appropriate test or CI (ie, not just the point estimate), preferably defined a priori and with an ITT analysis."[14]

"In some cases, a study planned as an NI study may show superiority to the active control. ICH E-9 and FDA policy has been that such a superiority finding arising in an NI study can be interpreted without adjustment for multiplicity. Showing superiority to an active control is very persuasive with respect to the effectiveness of the test drug, because demonstrating superiority to an active drug is much more difficult than showing superiority to placebo."[11]

"The designer of an NI trial might hope that the test drug is actually superior to the control. It is possible to design the NI study to first test the hypothesis of NI with the pre-specified margin, and then if this test is successful, proceed to analyze the study for a superiority conclusion. This sequential strategy is entirely acceptable. No statistical adjustment is required."[11]

"Noninferiority trials [but not equivalence trials] allow for the possibility of superiority over an active control. Before superiority can be assessed, noninferiority must be established; it would not make sense to start by proving equivalence"[13]

Drawing a superiority conclusion based on data from an equivalence trial

"Noninferiority trials [but not equivalence trials] allow for the possibility of superiority over an active control. Before superiority can be assessed, noninferiority must be established; it would not make sense to start by proving equivalence"[13]

"In most cases a successful NI study supports effectiveness of the test drug, but it only rarely will support a conclusion that the drug is "equivalent" or "similar" to the active control, a concept that has not been well-defined for these situations. Such similarity might be concluded, however, if the point estimate of the test drug favored it over the control and the upper bound of the 95 percent CI for C-T was close to showing superiority. Where the chosen M2 is very small compared to the control drug effect (e.g., a 10% margin in an antibiotic trial in urinary tract infections where response rate is 80%), it might be concluded that the effectiveness of the test drug and control are very similar."[11]

Drawing a non-inferiority conclusion based on data from a superiority trial

"If a superiority trial fails to detect a significant difference between treatments, there may be interest in the lesser objective of establishing non-inferiority. If the results of the superiority trial are summarized by means of a 95 percent confidence interval for the treatment difference, the lower end of that confidence interval provided a quantitative estimate of the minimum estimated effect of the new treatment relative to the comparator. When the study protocol contains an acceptable, prospectively defined margin $-\Delta$ for non-inferiority, downgrading the objective presents less methodological problems…. Although there does not appear to be a statistical multiplicity issue per se related to this switch of objective, that does not diminish the difficulties associated with the post hoc definition of Δ."[5]

"It is inappropriate to claim noninferiority post hoc from a superiority trial unless clearly related to a predefined margin of equivalence. That is, both superiority and noninferiority hypotheses need explicit specification in the trial protocol."[14]

"Seeking an NI conclusion in the event of a failed superiority test would almost never be acceptable. It would be very difficult to make a persuasive case for an NI margin based on data analyzed with study results in hand. If it is clear that an NI conclusion is a possibility, the study should be designed as an NI study."[11]

"A comparator chosen for a demonstration of superiority may not be acceptable for a conclusion of non-inferiority."[5] See further comments on this issue in reference[5] as it related to switching the objective from superiority to non-inferiority.

Reporting

Gomberg-Maitlin et al. (2003)[13] listed 6 items to be reported:

"1. A table comparing the inclusion and exclusion criteria with those of previous trials on which the standard therapy was based."
"2. A flow diagram delineating the number of eligible patients screened, the number randomized, the number of patients assigned to each group, the number of withdrawals and crossovers, and the number of patients in each group who successfully completed the trial on assigned treatment."

"3. A statement of the projected and actual total treatment exposure (patient-years), the minimum per-patient exposure, and the respective impact of withdrawals and crossovers on exposure to initially assigned treatment."
"4. The rationale for setting the margin of acceptable difference with specific reference to the minimum clinically important treatment effect and with the established efficacy advantage for the control over placebo. Where event rate ratios or floating margins are utilized, the rationale for their use, their prespecified criteria for adjustment, and the margin or ratios used to determine sample size should be provided."
"5. The minimum requisite number of primary events should be established at the outset."
"6. A comparison of event rates during treatment with the active control in the trial and in the historical trials that established its efficacy compared with placebo….Though no external method currently exists to evaluate the pragmatic utility of this approach, the 6 elements listed above (taken together) provide a readily accessible index of trial quality"[13]

Piaggio et al. (2006)14 listed 22 items to be reported:

"1. How participants were allocated to interventions (eg, "random allocation," "randomized,"or "randomly assigned"), specifying that the trial is a noninferiority or equivalence trial."
"2. Scientific background and explanation of rationale, including the rationale for using a noninferiority or equivalence design."
"3. Eligibility criteria for participants (detailing whether participants in the noninferiority 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."
"4. Precise details of the interventions intended for each group, detailing whether the reference treatment in the noninferiority 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."
"5. Specific objectives and hypotheses, including the hypothesis concerning noninferiority or equivalence."
"6. Clearly defined primary and secondary outcome measures, detailing whether the outcomes in the noninferiority 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 (eg, multiple observations, training of assessors)."
"7. How sample size was determined, detailing whether it was calculated using a noninferiority 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 noninferiority or equivalence hypothesis)."
"8. Method used to generate the random allocation sequence, including details of any restriction (eg, blocking, stratification)."
"9. Method used to implement the random allocation sequence (eg, numbered containers or central telephone), clarifying whether the sequence was concealed until interventions were assigned."
"10. Who generated the allocation sequence, who enrolled participants, and who assigned participants to their groups."

"11. Whether or not participants, those administering the interventions, and those assessing the outcomes were blinded to group assignment. When relevant, how the success of blinding was evaluated."
"12. Statistical methods used to compare groups for primary outcome(s), specifying whethera 1- or 2-sided confidence interval approach was used. Methods for additional analyses, such as subgroup analyses and adjusted analyses."
"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 trial protocol, and analyzed for the primary outcome. Describe protocol deviations from trial as planned, together with reasons."
"14. Dates defining the periods of recruitment and follow-up."
"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 "intention-to-treat" and/or alternative analyses were conducted. State the results in absolute numbers when feasible (eg, 10/20, not 50%)."
"17. For each primary and secondary outcome, a summary of results for each group and the estimated effect size and its precision (eg, 95% confidence interval). For the outcome(s) for which noninferiority or equivalence is hypothesized, a figure showing confidence intervals and margins of equivalence may be useful."
"18. Address multiplicity by reporting any other analyses performed, including subgroup analyses and adjusted analyses, indicating those prespecified and those exploratory."
"19. All important adverse events or side effects in each intervention group."
"20. Interpretation of the results, taking into account the noninferiority or equivalence hypothesis and any other trial hypotheses, sources of potential bias or imprecision and the dangers associated with multiplicity of analyses and outcomes."
"21. Generalizability (external validity) of the trial findings."
"22. General interpretation of the results in the context of current evidence."[14]

Appendix C. Methods Project 2: Review of Recent Systematic Reviews That Concluded Equivalence or Noninferiority

Stacey Uhl, M.S.S., Kelley Tipton, M.P.H., ECRI Institute, July 15, 2011

Table of Contents

Tables

Purpose

The purpose of this document is to assess current practices used within published systematic reviews that contain conclusions that could be interpreted as conclusions of equivalence or non-inferiority (EQ/NI) between two or more treatments. The information presented in this document is intended to aid in the development of guidance for Evidence-based Practice Centers (EPCs) to follow when drawing conclusions of EQ/NI in the context of comparative effectiveness reviews (CERs). This document focuses on current practices related to the following areas within systematic reviews: assessing risk of bias, defining the minimum important difference (MID, or margin of EQ/NI), analytical basis for drawing conclusions of EQ/NI, and wording of conclusions of EQ/NI.

Methods

To evaluate current practices, we first identified recently published systematic reviews that contained an author's conclusion that could be interpreted as a conclusion of EQ/NI. A searcher from the Scientific Resource Center (SRC) conducted a search of systematic reviews or meta-analyses. The following inclusion criteria were applied to identify potentially relevant documents:

- A systematic review or meta-analysis indexed in MEDLINE in 2010 to present.
- Published in a peer-reviewed journal within one year of the search date. The focus on recent reviews is to give maximal relevance to today's review processes, and also to permit us to finish the project on schedule.
- Compared outcomes after two or more medical treatments.
    - The title or abstract or executive summary contained an author's conclusion that could be interpreted as equivalence or non-inferiority. This was approached leniently. The authors did not have to actually use the word "equivalent" or "non-inferior" or "similar." Wording such as "the treatments appear to not differ substantially" or "there was no statistically significant difference" would qualify because these statements could be interpreted as conclusions of equivalence. But this sort of statement must have appeared in either the title or abstract or executive summary, for two reasons: (1) to ensure that it was truly a conclusion of the authors (and not simply a vague "by the way" buried in the discussion section), and (2) it would be too resource intensive to go through the full text of all of the reviews that may have drawn a conclusion of EQ/NI somewhere within the full text.

The search identified 961 potentially relevant documents. The abstracts were reviewed by three ECRI Institute analysts (KT, SU, and JT). A total of 235 documents met the inclusion criteria described above. The remaining 726 documents did not meet the inclusion criteria for one of the following reasons: (1) the document was not a systematic review/meta-analysis, (2) the document did not assess the efficacy of two or more active treatments, or (3) the document did not contain an author's conclusion that could be interpreted as a conclusion of EQ/NI. From the pool of 235 remaining systematic reviews, a random sample of 50 was selected (using the random numbers function in Excel) for further review and data abstraction. From these reviews, two ECRI Institute analysts (KT and SU) abstracted the information listed in Table 1.

Table 2. Information abstracted from reviews

•	Background information on the reviews, including purpose, medical condition, interventions, and number and sample size of included trials
•	Methodological design of included studies, including if the included trials were designed to be EQ/NI trials or not
•	Primary/secondary outcomes
•	Internal validity of included studies
•	Definition of minimum important difference (MID), if used
•	Whether the MID was pre-specified
•	The use of meta-analysis to support the conclusion
•	Statistical methods used to perform meta-analysis, if conducted
•	Results of meta-analysis, including summary estimates and 95% confidence intervals when available
•	Authors' conclusions
•	Categorization of conclusions (e.g., conclusion contains direct terms of EQ/NI such as the two treatments are "equivalent," "similar," or "not inferior" to each other)
•	Strength-of-evidence ratings, if used

The information described above is presented in detail in the data tables located in the Appendices of this document. The remainder of this document summarizes the information presented in the data tables.

Characteristics of Included Systematic Reviews

Overall, the stated purpose of 46 of the 50 randomly selected systematic reviews was to compare the efficacy and safety of two or more medical treatments. Thus, the authors of the reviews did not generally define their review as a specific type of systematic review such as an "equivalence review." In one review, the authors specifically indicated that the purpose of the review was to determine if one treatment was not inferior to another treatment.[1] In three reviews, the authors specified that the purpose of the review was to determine whether one treatment was superior to or better than another treatment.[2-4] See Table for further details on the characteristics of each review assessed in this document.

The medical conditions and interventions covered in the systematic reviews were wide ranging. The medical conditions included various types of cancers, cancer related pain, heart disease, autoimmune disorders, diabetes, chronic wounds, osteoarthritis, infertility, dental disease, Alzheimer's disease, seizure disorders, and depression. In most of the systematic reviews, a newer or less frequently used treatment was compared to what would be considered the standard or conventional treatment for the medical condition under consideration. The interventions compared included partial breast irradiation versus whole breast irradiation, transdermal fentanyl versus sustained-released oral morphine, drug-eluting stents versus bare metal stents, intensive glucose control strategies versus conventional glucose control strategies, radiofrequency ablation versus surgical resection, and cognitive behavioral therapy plus antidepressant therapy versus antidepressant therapy alone.

In all of the reviews the included studies were comparative studies of two or more treatments, with most reviews including only randomized controlled trials of head-to-head comparisons of treatments (34/50). However, some reviews included non-randomized controlled trials, cohort trials, and/or observational trials. The evidence base of the systematic reviews ranged from 3 to 69 included trials, with the number of included patients ranging from 96 to 28,065. In none of the systematic reviews did the authors report if any of the included studies were designed to be EQ/NI trials.

Assessing Risk of Bias

In the document developed for Methods Project 1, which summarizes existing guidance for individual trials of EQ/NI, the author describes a number of factors related to conducting and analyzing EQ/NI trials that may be important when assessing the risk of bias of studies included in a systematic review of these types of trials. These factors include the following:

- Vague or inconsistently applied patient enrollment criteria
- Biased selection of patients (e.g., selecting patients for anticipated nonresponse or relatively high response)
- Poor compliance with treatment
- Use of concomitant treatments that mask the true treatment effect
- Inappropriate dosage of treatment or differences in amount of treatment
- Patients not receiving treatment as assigned or crossing over to other treatment group
- Dropout/attrition
- Inadequate measurement of outcomes
- Not blinding outcome assessors
- Not performing both intention-to-treat and per-protocol analysis

While none of the authors of the systematic reviews assessed in this document reported if any of the trials included in their review were designed specifically to be EQ/NI trials, most authors (42/50) did report using some method to assess the risk of bias or quality of the included studies (See Table 3).[1] In most of the reviews, the authors reported using a recognized method, instrument, or checklist to assess risk of bias. These include: Jadad scale (12 reviews), the method described in the Cochrane Handbook (4 reviews), Newcastle-Ottawa Quality Scale (3 reviews), PEDro scale (3 reviews), and other (9 reviews).[2] The authors of 11 reviews did not report using any specific method or instrument to measure risk of bias, but did describe aspects of bias within the studies included in their review that could limit the conclusions of the review. These factors include: allocation of concealment; adequacy of randomization; baseline comparability; blinding of patients, clinicians, and outcome assessors; completeness of data reporting; documentation of dropouts/loss to follow-up; use of intention-to-treat analysis; and source of funding.

Interestingly, the use of intention-to-treat (ITT) analysis was considered good practice in the included trials, and yet the reviewers ended up concluding EQ/NI. As detailed in the summary document for Methods Project 1, the use of ITT tends to increase the chance of an incorrect conclusion of EQ-NI where in fact there is a true difference.

Overall, the quality of the controlled trials included in the reviews that reported on overall quality was moderate. The authors of these reviews indicated that the quality of the studies was limited due to lack of concealment of randomization, not blinding or not reporting if the outcome assessors were blinded, and not reporting if an intention-to-treat analysis was performed.

[1] The authors of eight reviews did not report using any method to assess the risk of bias and did not discuss limitations related to bias within the studies included in their review.

[2] In some of the reviews, the authors used more than one instrument, especially in reviews that included studies with different research designs (e.g., controlled trials and observational studies). The "other" category includes methods of assessing risk of bias in which the authors combined questions from various assessment instruments.

Method of Defining Minimum Important Difference

In only three systematic reviews did the authors define an MID for some of the outcomes considered in their review (See Table 4).[5-7] In one review that compared the effects of interval versus continuous exercise training on peak oxygen uptake, peak power, 6 minute walk test (6MWT) distance, and health-related quality of life in individuals with chronic obstructive pulmonary disease, the authors defined the MID for quality of life as measured by the Chronic Respiratory Questionnaire (CRQ) as a difference 0.5 for domains of dyspnea and total score and as a difference of 54 meters for the 6MWT.[5] The definitions of MID used in this review were pre-specified and based on previously published studies assessing the test properties of the CRQ and 6MWT.

In another review that evaluated the beneficial effect of chest compression-first versus defibrillation-first on survival of patients with out-of-hospital cardiac arrest, the authors defined a clinically relevant change for the primary outcome of survival to hospital discharge as a change of 20 percent to 25 percent.[6] The values used in this review were pre-specified and based on previously published studies evaluating interventions for out-of-hospital cardiac arrest, such as pre-defibrillation chest compressions and therapeutic hypothermia. In the final review, which evaluated the comparative data of commonly prescribed monotherapy drug regimens for both manic and depressive phases of bipolar 1 disorder, the authors calculated effect sizes for the number needed to treat and the number needed to harm.[7] For these outcomes, the authors considered an effect size of fewer than 10 patients as clinically meaningful. This value was pre-specified and based on a previous publication by Kraemer and Kupfer (2006) that discusses the size of treatment effects and their importance to clinical research and practice in psychiatry.

The authors of the reviews did not indicate that the MID defined in their review was to be used as a margin for determining if the treatments being evaluated were equivalent or not inferior to each other. However, based on the results of their meta-analysis, the authors of one review concluded that the treatments being evaluated appeared to be equivalent.[6] In this review, the meta-analysis indicated no statistically or clinically significant difference between the treatments being compared. Based on similar meta-analytic results, the authors of the two other reviews concluded that there was no difference between the treatments being compared in their review, but did not specifically indicate that the treatments were equivalent or similar to each other.[5,7]

Analytic Basis of Drawing Conclusion of Equivalence or Non-inferiority

In the majority of the reviews (43/50) the authors based their conclusions of EQ/NI on meta-analytic findings of no statistically significant difference between treatments. Few reviews considered whether the evidence was sufficient to rule out a clinically important difference. In most reviews (39/50), the authors performed meta-analysis using either a fixed (Mantel-Haenszel or Peto method) or random (DerSimonian and Laird method) effects model. The specific model used in most cases depended on the presence of significant heterogeneity, which was typically measured using the I^2 statistic. Individual and pooled summary effect size estimates were calculated with 95 percent confidence intervals (95 percent CIs) using the following metrics depending on the type of data reported: odds ratio, relative risk ratio, hazard ratio, or weighted mean difference. In one review, when appropriate, the authors conducted indirect analyses of the evidence using the method described by Butcher, et al.[8] In four reviews, the authors either didn't report the meta-analytic model used to perform meta-analysis or reported using a method other than that described by Mantel-Haenszel or DerSimonian and Laird. The meta-analytic results

(summary effect size estimate with 95% confidence intervals) of each review are presented in Table 5.

In seven of the systematic reviews no meta-analysis was performed. Instead, the authors of these reviews based their conclusions on a qualitative assessment of the evidence.[9-14] In two reviews, the authors indicated that no meta-analysis was performed due to the presence of clinical heterogeneity (e.g., differences in comparators used and outcomes measured across studies).[10,15] The authors of these reviews reported using vote counting (i.e., counting the number of studies supporting and not supporting the intervention) of the overall trials to draw conclusions. Based on this method, the authors of one review concluded that "the antiseptic effect of iodine is not inferior to that of other (antiseptic) agents and does not impair wound healing."[10] The authors of the other review concluded that "current evidence indicates that there appears to be no [disease free survival] advantage [of the addition of neoadjuvant transarterial chemoembolization for resectable hepatocellular carcinoma] despite its safety and feasibility."[15]

In the remaining five reviews, the authors did not report their reasons for not performing a quantitative analysis of the evidence nor did they report using an alternative method such as vote counting. In these reviews, the findings of the included studies were narratively summarized and the authors based their conclusions on the narrative synthesis of the evidence.

Wording of Conclusions within Systematic Review

In Appendix E

Table 6 (Appendix E), we present the exact wording of the authors' conclusions as reported in the abstract or discussion section of each review. As indicated in previous sections of this document, none of the authors of the reviews used a margin of EQ/NI (as defined as the minimal important difference) to determine if the treatments being considered in their review were equivalent or non-inferior to each other. Thus, the conclusion statements described in this section are based primarily on meta-analytic results that indicated no statistically significant difference (or in a few cases no clinically significant difference) between the treatments being assessed. According to many of the regulatory documents reviewed in Methods Project 1, it is inappropriate to draw a conclusion of EQ/NI solely on the basis of a non-significant difference. The reason being that a non-significant difference alone does not discriminate between two very different situations: (1) "when the confidence interval is narrow enough to exclude the possibility of a meaningful difference, and (2) when the confidence interval is too wide to permit any conclusion about whether the treatments differed."

We assessed the specific wording used in the authors' conclusions of each review to determine how many used direct terms of EQ/NI such as "equivalent to," "similar to," "comparable to," and "not inferior to." Overall, the conclusion statements of 19 reviews used direct terms of EQ/NI. Examples include the following:

- "Currently available evidence suggests a similar effectiveness of GNRH antagonists and agonists in the context of oocyte donation."
- "Current evidence does not support the notion that chest compression first prior to defibrillation improves the outcome of patients in out-of-hospital cardiac arrest. It appears that both treatments [chest compressions first vs. defibrillation first] are equivalent."
- "Arthroscopic and open acromioplasty have equivalent ultimate clinical outcomes, operative times, and low complication rates."

- "Intravenous immunoglobulin started within two weeks from onset hastens recovery as much as plasma exchange."
- "From the evidence discussed, it is evident that oral topotecan has similar efficacy to IV topotecan [direct comparison] and CAV [indirect comparison]."

As demonstrated in the first example listed above, the conclusion statements of many of the reviews that used direct terms of EQ/NI contained hedging language such as "seems," "suggests," and "may." In the remaining 31 reviews, the authors used less direct language to express what could be interpreted as a conclusion of equivalence. For example, in one review the authors concluded that "there was no difference in the hazard of death or recurrent myocardial infarction between patients treated with drug-eluting stents versus bare metal stents."[16]

In another review the authors concluded that "there is no evidence that VATS [video assisted thoracoscopic surgery] is more effective than fibrinolytic treatment."[17]

In only two reviews the authors indicated that their conclusions were supported by aspects of the strength of the evidence. In one review the authors reported that the conclusions drawn in their review were based on moderate quality evidence.[18] The authors of the other review reported that their conclusion of equivalence between chest compressions-first versus defibrillation-first for out-of-hospital cardiac arrest was based on "no treatment effect with fairly narrow confidence intervals and with very little heterogeneity."[6] None of the reviews assessed in this document provided a rating of the strength of evidence.

Summary

This document assessed current practices used within published systematic reviews that contain conclusions that could be interpreted as conclusions of equivalence or non-inferiority (EQ/NI) between two or more treatments. The information was intended to aid in the development of guidance for EPCs to follow when drawing conclusions of EQ/NI in the context of CERs. The focus of the document was on current practices within systematic reviews related to the following areas: assessing risk of bias, defining the minimum important difference (MID, or margin of EQ/NI), analytical basis for drawing conclusions of EQ/NI, and wording of conclusions of EQ/NI.

From a pool of 235 systematic reviews that met the inclusion criteria for this project, we randomly selected 50 for further review and data abstraction. We abstracted data related to the four areas listed above (See Table 2 for the full list of data abstracted from each review). Overall, none of the authors reported if any of the studies included in their review were designed to be EQ/NI trials. The majority of the authors reported using some method to assess the risk of bias or quality of the included studies. Many used a recognized method, instrument, or checklist to assess risk of bias, such as the Jadad scale, the method described in the Cochrane Handbook, and the Newcastle-Ottawa Quality Scale. A number of the factors assessed in the systematic reviews are likely to be important when assessing the risk of bias of studies that call themselves EQ/NI studies. These factors include: baseline comparability, blinding outcome assessors; completeness of data reporting, documentation of dropouts/lost to follow-up, and use of intention-to-treat analysis.

In only three systematic reviews did the authors define an MID for some of the outcomes considered in their review. In all three reviews, the definition of MID used was pre-specified and based on previously published studies or reports. The authors of the reviews did not indicate that the MID defined in their review was to be used as a margin for determining if the treatments being evaluated were equivalent or not inferior to each other. However, in one review, the

authors concluded that the treatments being evaluated appeared to be equivalent. This conclusion was based on the meta-analytic results of the review, which indicated no statistically or clinically significant difference between the treatments being compared.

In the majority of the reviews, the authors based their conclusions of EQ/NI on a statistically non-significant finding from meta-analyses. According to many of the regulatory documents reviewed in Methods Project 1, it is inappropriate to draw a conclusion of EQ/NI solely on the basis of a non-significant difference. Overall, 38 percent of the conclusion statements of the reviews assessed used direct terms of EQ/NI, such as "equivalent to," "similar to," "comparable to," and "not inferior to." In the remaining reviews, the authors used less direct language (e.g., "There is no difference [in outcomes] between patients treated with drug-eluting stents versus bare metal stents.") to express what could be interpreted as a conclusion of equivalence. Finally, in only two reviews the authors indicated that their conclusions were supported by aspects of the strength of the evidence.

Conclusion

This review underscores the need to develop guidance for EPCs to follow when drawing conclusions of equivalence or non-inferiority in the context of comparative effectiveness reviews. The results of our assessment suggest that authors of recently published systematic reviews do not always assess factors of risk of bias that may be important when drawing conclusions of EQ/NI, do not commonly define or use an MID (or margin of EQ/NI), and do not typically pre-specify how they will handle findings of no difference or similarity. In many of the reviews assessed in this document, meta-analytic findings of no statistically significant difference were interpreted as equivalence between two treatments. Thus, developing guidance on how and when it is appropriate to draw conclusions of equivalence or non-inferiority in the context of comparative effectiveness reviews would avoid confusion and misinterpretation of reports generated by EPCs.

References

1. Yang Q, Xie DR, Jiang ZM, et al. Efficacy and adverse effects of transdermal fentanyl and sustained-release oral morphine in treating moderate-severe cancer pain in Chinese population: a systematic review and meta-analysis. J Exp Clin Cancer Res 2010;29:67. PMID: 20529380

2. Kalil AC, Murthy MH, Hermsen ED, et al. Linezolid versus vancomycin or teicoplanin for nosocomial pneumonia: a systematic review and meta-analysis. Crit Care Med 2010 Sep;38(9):1802-8. PMID: 20639754

3. Kesselheim AS, Stedman MR, Bubrick EJ, et al. Seizure outcomes following the use of generic versus brand-name antiepileptic drugs: a systematic review and meta-analysis. Drugs 2010 Mar 26;70(5):605-21. PMID: 20329806

4. Tang Y, Li X, Yin S. Outcomes of MTA as root-end filling in endodontic surgery: a systematic review. Quintessence Int 2010 Jul-Aug;41(7):557-66. PMID: 20614042

5. Beauchamp MK, Nonoyama M, Goldstein RS, et al. Interval versus continuous training in individuals with chronic obstructive pulmonary disease--a systematic review. Thorax 2010 Feb;65(2):157-64. PMID: 19996334

6. Meier P, Baker P, Jost D, et al. Chest compressions before defibrillation for out-of-hospital cardiac arrest: a meta-analysis of randomized controlled clinical trials. BMC Med 2010;8:52. PMID: 20828395

7. Tamayo JM, Zarate CA Jr, Vieta E, et al. Level of response and safety of pharmacological monotherapy in the treatment of acute bipolar I disorder phases: a systematic review and meta-analysis. Int J Neuropsychopharmacol 2010 Jul;13(6):813-32. PMID: 20128953

8. Riemsma R, Simons JP, Bashir Z, et al. Systematic Review of topotecan (Hycamtin) in relapsed small cell lung cancer. BMC Cancer 2010;10:436. PMID: 20716361

9. Lanitis S, Tekkis PP, Sgourakis G, et al. Comparison of skin-sparing mastectomy versus non-skin-sparing mastectomy for breast cancer: a meta-analysis of observational studies. Ann Surg 2010 Apr;251(4):632-9. PMID: 20224371

10. Vermeulen H, Westerbos SJ, Ubbink DT. Benefit and harm of iodine in wound care: a systematic review. J Hosp Infect 2010 Nov;76(3):191-9. PMID: 20619933

11. Groeneveld AB, Navickis RJ, Wilkes MM. Update on the comparative safety of colloids: a systematic review of clinical studies. Ann Surg 2011 Mar;253(3):470-83. PMID: 21217516

12. Squizzato A, Manfredi E, Bozzato S, et al. Antithrombotic and fibrinolytic drugs for retinal vein occlusion: a systematic review and a call for action. Thromb Haemost 2010 Feb;103(2):271-6. PMID: 20126837

13. Loveman E, Jones J, Hartwell D, et al. The clinical effectiveness and cost-effectiveness of topotecan for small cell lung cancer: a systematic review and economic evaluation. Health Technol Assess 2010 Mar;14(19):1-204. PMID: 20356561

14. Chua TC, Liauw W, Saxena A, et al. Systematic review of neoadjuvant transarterial chemoembolization for resectable hepatocellular carcinoma. Liver Int 2010 Feb;30(2):166-74. PMID: 19912531

15. Vasiliadis HS, Wasiak J, Salanti G. Autologous chondrocyte implantation for the treatment of cartilage lesions of the knee: a systematic review of randomized studies. Knee Surg Sports Traumatol Arthrosc 2010 Dec;18(12):1645-55. PMID: 20127071

16. Dibra A, Tiroch K, Schulz S, et al. Drug-eluting stents in acute myocardial infarction: updated meta-analysis of randomized trials. Clin Res Cardiol 2010 Jun;99(6):345-57. PMID: 20221617

17. Krenke K, Peradzynska J, Lange J, et al. Local treatment of empyema in children: a systematic review of randomized controlled trials. Acta Paediatr 2010 Oct;99(10):1449-53. PMID: 20456264

18. Dubicka B, Elvins R, Roberts C, et al. Combined treatment with cognitive-behavioural therapy in adolescent depression: meta-analysis. Br J Psychiatry 2010 Dec;197:433-40. PMID: 21119148

19. Bodri D, Sunkara SK, Coomarasamy A. Gonadotropin-releasing hormone agonists versus antagonists for controlled ovarian hyperstimulation in oocyte donors: a systematic review and meta-analysis. Fertil Steril 2011 Jan;95(1):164-9. PMID: 20684954

20. Greenhalgh J, Hockenhull J, Rao N, et al. Drug-eluting stents versus bare metal stents for angina or acute coronary syndromes. Cochrane Database Syst Rev 2010;(5):CD004587. PMID: 20464732

21. Seitz DP, Adunuri N, Gill SS, et al. Antidepressants for agitation and psychosis in dementia. Cochrane Database Syst Rev 2011;2:CD008191. PMID: 21328305

22. Davis AD, Kakar S, Moros C, et al. Arthroscopic versus open acromioplasty: a meta-analysis. Am J Sports Med 2010 Mar;38(3):613-8. PMID: 19188562

23. Dong L, Zhang F, Shu X. Early administration of small-molecule glycoprotein IIb/IIIa inhibitors before primary percutaneous coronary intervention for ST-elevation myocardial infarction: insights from randomized clinical trials. J Cardiovasc Pharmacol Ther 2010 Jun;15(2):135-44. PMID: 20435991

24. Dong L, Zhang F, Shu X. Upstream vs deferred administration of small-molecule glycoprotein IIb/IIIa inhibitors in primary percutaneous coronary intervention for ST-segment elevation myocardial infarction: insights from randomized clinical trials. Circ J 2010 Aug;74(8):1617-24. PMID: 20571247

25. Fuentes JP, Armijo Olivo S, Magee DJ, et al. Effectiveness of interferential current therapy in the management of musculoskeletal pain: a systematic review and meta-analysis. Phys Ther 2010 Sep;90(9):1219-38. PMID: 20651012

26. Wu H, Xu MJ, Zou DJ, et al. Intensive glycemic control and macrovascular events in type 2 diabetes mellitus: a meta-analysis of randomized controlled trials. Chin Med J (Engl) 2010 Oct;123(20):2908-13. PMID: 21034605

27. Hughes RA, Swan AV, van Doorn PA. Intravenous immunoglobulin for Guillain-Barre syndrome. Cochrane Database Syst Rev 2010;(6):CD002063. PMID: 20556755

28. Myers J, Chan V, Jarvis V, et al. Intraspinal techniques for pain management in cancer patients: a systematic review. Support Care Cancer 2010 Feb;18(2):137-49. PMID: 19943068

29. Lee YH, Woo JH, Choi SJ, et al. Induction and maintenance therapy for lupus nephritis: a systematic review and meta-analysis. Lupus 2010 May;19(6):703-10. PMID: 20064907

30. Liu JG, Wang YJ, Du Z. Radiofrequency ablation in the treatment of small hepatocellular carcinoma: a meta analysis. World J Gastroenterol 2010 Jul 21;16(27):3450-6. PMID: 20632451

31. Liu Z, Zhang P, Ma Y, et al. Laparoscopy or not: a meta-analysis of the surgical effects of laparoscopic versus open appendectomy. Surg Laparosc Endosc Percutan Tech 2010 Dec;20(6):362-70. PMID: 21150411

32. Zhang Y, Zhang P, Mu Y, et al. The role of renin-angiotensin system blockade therapy in the prevention of atrial fibrillation: a meta-analysis of randomized controlled trials. Clin Pharmacol Ther 2010 Oct;88(4):521-31. PMID: 20811347

33. Wang X, Liu R, Ma B, et al. High dose rate versus low dose rate intracavity brachytherapy for locally advanced uterine cervix cancer. Cochrane Database Syst Rev 2010;(7):CD007563. PMID: 20614461

34. Macedo LG, Smeets RJ, Maher CG, et al. Graded activity and graded exposure for persistent nonspecific low back pain: a systematic review. Phys Ther 2010 Jun;90(6):860-79. PMID: 20395306

35. Machado M, Einarson TR. Comparison of SSRIs and SNRIs in major depressive disorder: a meta-analysis of head-to-head

randomized clinical trials. J Clin Pharm Ther 2010 Apr;35(2):177-88. PMID: 20456736

36. Milito G, Cadeddu F, Muzi MG, et al. Haemorrhoidectomy with Ligasure vs conventional excisional techniques: meta-analysis of randomized controlled trials. Colorectal Dis 2010 Feb;12(2):85-93. PMID: 19220374

37. Murphy JD, Yan D, Hanna MN, et al. Comparison of the postoperative analgesic efficacy of intravenous patient-controlled analgesia with tramadol to intravenous patient-controlled analgesia with opioids. J Opioid Manag 2010 Mar-Apr;6(2):141-7. PMID: 20481179

38. Pan XH, Chen YX, Xiang MX, et al. A meta-analysis of randomized trials on clinical outcomes of paclitaxel-eluting stents versus bare-metal stents in ST-segment elevation myocardial infarction patients. J Zhejiang Univ Sci B 2010 Oct;11(10):754-61. PMID: 20872982

39. Sbruzzi G, Ribeiro RA, Schaan BD, et al. Functional electrical stimulation in the treatment of patients with chronic heart failure: a meta-analysis of randomized controlled trials. Eur J Cardiovasc Prev Rehabil 2010 Jun;17(3):254-60. PMID: 20560163

40. Sgourakis G, Gockel I, Radtke A, et al. The use of self-expanding stents in esophageal and gastroesophageal junction cancer palliation: a meta-analysis and meta-regression analysis of outcomes. Dig Dis Sci 2010 Nov;55(11):3018-30. PMID: 20440646

41. Simpson PM, Goodger MS, Bendall JC. Delayed versus immediate defibrillation for out-of-hospital cardiac arrest due to ventricular fibrillation: a systematic review and meta-analysis of randomised controlled trials. Resuscitation 2010 Aug;81(8):925-31. PMID: 20483525

42. Sunkara SK, Siozos A, Bolton VN, et al. The influence of delayed blastocyst formation on the outcome of frozen-thawed blastocyst transfer: a systematic review and meta-analysis. Hum Reprod 2010 Aug;25(8):1906-15. PMID: 20542896

43. Tamhane UU, Chetcuti S, Hameed I, et al. Safety and efficacy of thrombectomy in patients undergoing primary percutaneous coronary intervention for acute ST elevation MI: a meta-analysis of randomized controlled trials. BMC Cardiovasc Disord 2010;10:10. PMID: 20187958

44. Testa L, Agostoni P, Vermeersch P, et al. Drug eluting stents versus bare metal stents in the treatment of saphenous vein graft disease: a systematic review and meta-analysis. EuroIntervention 2010 Sep;6(4):527-36. PMID: 20884442

45. Valachis A, Mauri D, Polyzos NP, et al. Partial breast irradiation or whole breast radiotherapy for early breast cancer: a meta-analysis of randomized controlled trials. Breast J 2010 May-Jun;16(3):245-51. PMID: 20210799

46. Xie X, Dendukuri N, McGregor M. Percutaneous radiofrequency ablation for the treatment of early stage hepatocellular carcinoma: a health technology assessment. Int J Technol Assess Health Care 2010 Oct;26(4):390-7. PMID: 20923590

47. Agarwal S, Tuzcu EM, Desai MY, et al. Updated meta-analysis of septal alcohol ablation versus myectomy for hypertrophic cardiomyopathy. J Am Coll Cardiol 2010 Feb 23;55(8):823-34. PMID: 20170823

48. Avouac J, Vicaut E, Bardin T, et al. Efficacy of joint lavage in knee osteoarthritis: meta-analysis of randomized controlled studies. Rheumatology (Oxford) 2010 Feb;49(2):334-40. PMID: 19955221

49. Devaiah AK, Andreoli S. Postmaneuver restrictions in benign paroxysmal positional vertigo: an individual patient data meta-analysis. Otolaryngol Head Neck Surg 2010 Feb;142(2):155-9. PMID: 20115966

50. Valachis A, Mauri D, Polyzos NP, et al. Fulvestrant in the treatment of advanced breast cancer: a systematic review and meta-analysis of randomized controlled trials. Crit Rev Oncol Hematol 2010 Mar;73(3):220-7. PMID: 19369092

Appendix A

Table 2. Characteristics of systematic reviews

Reference	Purpose/ Objective of Review	Medical Condition	Intervention(s)	Number of Trials Included	Total Sample Size	Original Trial Design[1]	Number of Trials using EQ/NI Design[2]
Bodri et al. 2011[19]	"To compare GnRH agonists vs. antagonists in oocyte-donation IVF treatment cycles by a systematic review and meta-analysis of trials."	Controlled ovarian hyperstimulation in oocyte donors	GnRH agonists vs. antagonists	8	1,024 oocyte donors	RCTs of head-to-head comparisons of GnRH agonists vs. antagonists	Not reported
Groeneveld et al. 2011[11]	"To provide an update on the comparative safety of colloids based upon the extensive clinical data accumulated since 2002."	Acutely ill patients	Colloids: HES gelatin, dextran, or albumin	69	10,382 RCTs: 6,106 cohort; 26,318 non-RCTs	42 RCTs; 8 cohort studies; 7 non-RCTs; 7 meta-analyses, 4 systematic reviews; and 1 pharmaco-surveillance study	Not reported
Hockenhull et al. 2011[20]	"To examine evidence from RCTs to assess the impact of DES compared to BMS in the reduction of cardiac events."	Stable angina or acute coronary syndrome	DES vs. BMS (standard treatment)	47 RCTs	14,500 patients	RCTs of head-to-head comparisons of DES vs. BMS	Not reported
Seitz et al. 2011[21]	"To assess the safety and efficacy of antidepressants in treating psychosis and agitation in older adults with AD, vascular, or mixed dementia."	Agitation and/or psychosis in dementia (patients diagnosed with AD, vascular, or mixed AD and vascular dementia, DLB, dementia not otherwise specified)	Antidepressants (SSRIs, TCAs, trazodone, and other antidepressants) vs. other psychotropic medications (benzodiazepines, antipsychotics, anticonvulsants) or placebo	9	692	RCTs that compared antidepressants vs. other psychotropic medications or placebo	Not reported

Reference	Purpose/ Objective of Review	Medical Condition	Intervention(s)	Number of Trials Included	Total Sample Size	Original Trial Design[1]	Number of Trials using EQ/NI Design[2]
Beauchamp et al. 2010[5]	"To compare the effects of interval versus continuous training on peak oxygen uptake, peak power, six minute walk test distance and HRQoL in individuals with COPD."	COPD	Interval exercise training vs. Continuous exercise training	8	388	RCTs of head-to-head comparisons of Interval exercise training vs. continuous exercise training	Not reported
Davis et al. 2010[22]	To perform a systematic review of the literature that compares the efficacy arthroscopic compared to open acromioplasty.	Subacromial impingement syndrome	Arthroscopic decompression vs. open acromioplasty (standard treatment)	9 comparative studies	775 (432 arthroscopic and 343 open acromioplasty)	Studies that directly compared arthroscopic to open acromioplasty	Not reported
Dibra et al. 2010[16]	"To consolidate and extend current knowledge on the safety and efficacy of DES after primary percutaneous coronary intervention."	Acute myocardial infarction	DES vs. BMS (standard treatment)	14 RCTs	7,781	RCTs of head-to-head comparisons of DES and BMS	Not reported
Dong et al. 2010[23]	"To compare the clinical safety and efficacy of early smGPIs with abciximab before primary PCI in patients with STEMI."	STEMI	Early smGPIs vs. abciximab	4	2,040	RCTs of head-to-head comparisons of early smGPIs vs. abciximab	Not reported

Reference	Purpose/ Objective of Review	Medical Condition	Intervention(s)	Number of Trials Included	Total Sample Size	Original Trial Design[1]	Number of Trials using EQ/NI Design[2]
Dong et al. 2010[24]	"To exclusively evaluate the relative safety and efficacy of upstream vs. deferred administration of smGPIs in STEMI patients scheduled for mechanical reperfusion."	Myocardial infarction	Upstream administration of smGPIs (tirofiban or eptifibatide) vs. deferred periprocedural administration of smGPIs (tirofiban or eptifibatide)	10	2,724	RCTs of head-to-head comparisons of upstream administration of smGPIs vs. deferred periprocedural administration of smGPIs	Not reported
Dubicka et al. 2010[18]	To assess if CBT provides additional benefit to antidepressant treatment in adolescents with unipolar depression for depressive symptoms, suicidality, impairment, and depression.	Unipolar depression in adolescents	CBT plus antidepressant therapy vs. antidepressant therapy alone (standard treatment)	5 RCTs	1,206 adolescents	RCTs of head-to-head comparisons of CBT plus antidepressants vs. antidepressants alone	Not reported
Fuentes et al. 2010[25]	"To analyze the available information regarding the efficacy of IFC in the management of musculoskeletal pain."	Musculoskeletal pain	Isolated or Coadjuvant IFC vs. placebo control, another physical therapy intervention, or another type of intervention	20	1,757	RCTs that compared isolated or coadjuvant IFC vs. placebo control, another physical therapy intervention, or another type of intervention	Not reported

Reference	Purpose/ Objective of Review	Medical Condition	Intervention(s)	Number of Trials Included	Total Sample Size	Original Trial Design[1]	Number of Trials using EQ/NI Design[2]
Hong et al. 2010[26]	"To evaluate the efficacy of intensive glucose control in the prevention of cardiovascular events when compared with standard glucose controls, including all RCTs and, when appropriate, to explore the sources of heterogeneity of results."	Type II diabetes mellitus	Intensive glucose control strategies including oral agents, insulin and multiple cardiovascular intervention (<7% target level for HbA_1C) vs. Conventional glucose control strategies including oral agents, insulin and multiple cardiovascular intervention (no strict target level)	6	28,065	RCTs of head-to-head comparisons of intensive glucose strategies vs. conventional glucose strategies	Not reported
Hughes et al. 2010[27]	"To determine the efficacy of intravenous immunoglobin for Guillain-Barre syndrome."	Guillian-Barre syndrome	Immunoglobin (IVIg, newer treatment) vs. plasma exchange (standard treatment), other immunomodulatory treatment, or placebo. For this project, we only consider the evidence of IVIg vs. other active treatment.	9 comparative trials (5 trials comparing IVIg to plasma exchange, 3 comparing IVIg compared to supportive care, and 1 comparing plasma exchange followed by IVIg compared to plasma exchange alone)	860 patients	RCTs and quasi-RCTs comparing treatments	Not reported
Kalil et al. 2010[2]	"To test the hypothesis that linezolid may be superior to glycopeptide."	Nosocomial pneumonia	Linezolid (newer treatment) vs. vancomycin or teicoplanin (standard care)	9 RCTs (7 comparing linezolid to vancomycin and 2 comparing linezolid to teicoplanin)	2,329 patients	RCTs of head-to-head comparisons of linezolid to vancomycin or teicoplanin	Not reported

Reference	Purpose/ Objective of Review	Medical Condition	Intervention(s)	Number of Trials Included	Total Sample Size	Original Trial Design[1]	Number of Trials using EQ/NI Design[2]
Kesselheim et al. 2010[3]	"To evaluate studies comparing brand-name and generic AEDs, and to determine whether evidence exists of superiority of the brand-name version in maintaining seizure control."	Seizure disorders	AEDs	16 studies (9 RCTs, 1 prospective non-randomized trial, and 6 observational studies)	249 in RCTs	Studies that reported a comparative evaluation of one brand-name drug and at least one alternate version produced by a distinct manufacturer.	Not reported
Krenke et al. 2010[17]	"To evaluate data from RCTs on the efficacy of using intrapleural fibrinolytic agents in the treatment of complicated parapneumonic effusions or empyema in children."	Complicated parapneumonic effusions or empyema	Intrapleural fibrinolytic agents (streptokinase, urokinase, alteplase)	4 RCTs (2 that compared fibrinolysis to VAT	96 (48 fibrinolysis and 48 VATs)	RCTs that compared intrapleural fibrinolytic agents to surgical interventions, chest tube drainage, thoracocentesis, placebo, or any other intervention	Not reported
Lanitis et al. 2010[28]	"To evaluate differences in outcomes of breast cancer patients undergoing either conventional mastectomy without reconstruction or skin-sparing mastectomy with immediate reconstruction."	Breast cancer	Conventional mastectomy without reconstruction (NSSM, standard) vs. skin-sparring mastectomy with immediate reconstruction (SSM)	9 retrospective, non-randomized comparative trials	3,739 (1,104 SSM and 2,635 NSSM)	Comparative studies of SSM and NSSM	Not reported

Reference	Purpose/ Objective of Review	Medical Condition	Intervention(s)	Number of Trials Included	Total Sample Size	Original Trial Design[1]	Number of Trials using EQ/NI Design[2]
Lee et al. 2010[29]	"To assess the efficacies and toxicities of immunosuppressive treatments for lupus nephritis versus CYC."	Lupus nephritis	MMF (standard treatment) vs. CYC induction therapy; MMF vs. AZA as maintenance therapy; and low-dose IV CYC and high-dose IV CYC therapy	10 RCTs (6 studies compared MMF to CYC, 2 compared MMF to AZA, and 2 compared high dose CYC to low dose CYC	891 (450 patients in experimental group and 441 in control group)	RCTs of head-to-head comparisons of treatments	Not reported
Liu et al. 2010[30]	"To evaluate survival and recurrence after radiofrequency ablation for the treatment of small HCC using meta-analysis."	HCC	Surgical resection (standard care) vs. RFA (newer treatment)	10 comparative trials (7 retrospective studies and 3 cohort studies)	1,522 patients (787 RFA and 735 surgical resection)	Studies comparing RFA to surgical resection	Not reported
Liu et al. 2010[31]	"To compare the surgical effects of laparoscopic versus open appendectomy."	Appendicitis	OA (standard) vs. LA	16 RCTs	3,261 (1,587 LA and 1,674 OA)	RCTs of head-to-head comparisons of treatments	Not reported
Liu et al. 2010[32]	"To retrospectively evaluate the long-term results of [RFA] as compared with hepatic resection for the treatment of [HCC]."	HCC	RFA vs. hepatic resection (standard treatment)	8 RCTs	1,188 (534 hepatectomy group and 654 RFA	RCTs of head-to-head comparisons of treatments	Not reported
Liu et al. 2010[33]	"To assess the efficacy and safety of HDR vs. LDR ICBT."	Locally advanced uterine cervix cancer	HDR vs. LDR ICBT with whole pelvic EBRT	4	1,265	3 RCTs; 1 quasi-RCT of head-to-head comparisons of HDR vs. LDR ICBT with whole pelvic EBRT	Not reported

Reference	Purpose/ Objective of Review	Medical Condition	Intervention(s)	Number of Trials Included	Total Sample Size	Original Trial Design[1]	Number of Trials using EQ/NI Design[2]
Macedo et al. 2010[34]	"To systematically review RCTs that evaluated the effectiveness of graded activity and graded exposure interventions for the treatment of persistent (>6 weeks in duration or recurrent) non-specific LBP at short, intermediate, and long-term follow-up evaluations."	Non-specific LBP	Graded activity or graded exposure vs. placebo, no treatment, or another active treatment, or when graded activity of graded exposure was added as a supplement to other interventions	15	1,654	RCTs that compared graded activity or graded exposure vs. placebo, no treatment, or another active treatment or when graded activity or graded exposure was added as a supplement to other interventions	Not reported
Machado et al. 2010[35]	"To compare clinical outcomes of adults treated with SSRIs vs. SNRIs for MDD under ideal clinical condition, research design, and outcome measure."	MDD	SSRIs (citalopram, escitalopram, fluoxetine, fluvoxamine, paroxetine or sertraline) vs. SNRIs (venlafaxine, duloxetine or milnacipran)	15	3,094	RCTs	Not reported
Meier et al. 2010[6]	"To evaluate the beneficial effect of chest compression-first versus defibrillation-first on survival in patients with out-of-hospital cardiac arrest."	Out-of-hospital cardiac arrest	Chest compression-first vs. defibrillation-first (standard) CPR	4 RCTs	1,503 patients	RCTs of head-to-head comparison of treatments	Not reported
Milito et al. 2010[36]	"To compare the use of LigaSure devices with conventional excisional techniques, circular stapling and use of Harmonic Scalpel in patients with symptomatic haemorrhoides."	Symptomatic haemorrhoides	Instruments used to perform haemorrhoidectomy: LigaSure vs. other techniques of haemorrhoidectomy	11 RCTs	850 patients	RCTs of head-to-head comparison of treatments	Not reported

Reference	Purpose/ Objective of Review	Medical Condition	Intervention(s)	Number of Trials Included	Total Sample Size	Original Trial Design[1]	Number of Trials using EQ/NI Design[2]
Murphy et al. 2010[37]	"To compare the analgesic efficacy of IV PCA tramadol with that of IV PCA with opioids."	Adult patients	IV PCA with tramadol vs. IV PCA with opioids	12	782	RCTs of head-to-head comparisons of IV PCA with tramadol vs. IV PCA with opioids	Not reported
Myers et al. 2010[9]	To systematically review the evidence regarding the effectiveness of intraspinal techniques for cancer pain.	Pain related to any type of cancer	Intraspinal techniques alone or in combination vs. medical management, or different intraspinal techniques, external pumps vs. internal pumps, or timing of intraspinal techniques	3 previous systematic reviews, 12 RCTs, and 3 consensus statements; Of the 12 RCTs, 6 compared intraspinal techniques alone or in combination with other techniques alone or in combination, 4 compared different intraspinal medications, and 2 compared different intraspinal techniques	1,234 patients enrolled in RCTs	RCTs of head-to-head comparisons of treatments	Not reported
Pan et al. 2010[38]	"To address the efficacy and safety of paclitaxel-eluting stent in STEMI patients."	STEMI	Paclitaxel-eluting stent (DES) vs. BMS (standard)	6 RCTs	4,190 (1,344 for BMS and 2,846 DES)	RCTs of head-to-head comparisons of DES vs. BMS	Not reported
Riemsa et al. 2010[8]	To systematically review available data for oral and intravenous topotecan in adults with relapsed SCLC for whom retreatment with the first line regimen is not appropriate.	SCLC	Oral topotecan plus BSC vs. BSC alone; IV topotecan vs. CAV; and Oral topotecan vs. IV topotecan;	7 RCTs only 4 RCTs included in analysis (1 comparing oral topotecan plus BSC to BSC alone; 1 comparing IV topotecan to CAV, and 2 comparing oral topotecan to IV topotecan	762 (141 doe study of topotecan vs. BSC; 211 for study of topotecan vs. CAV; and 410 for 2 studies comparing oral vs. IV topotecan)	RCTs of head-to-head comparisons of treatments	Not reported

Reference	Purpose/ Objective of Review	Medical Condition	Intervention(s)	Number of Trials Included	Total Sample Size	Original Trial Design[1]	Number of Trials using EQ/NI Design[2]
Sbruzzi et al. 2010[39]	"To systematically review the effect of treatment with [FES] compared to conventional aerobic exercise training or control group in patients with chronic heart failure."	Chronic heart failure	FES vs. conventional aerobic exercise or control (the same regimen as the FES group, except that the intensity of stimulation did to lead to visible or palpable contractions)	7 RCTs (5 compared FES to exercise and 2 compared FES to control)	224 patients	RCTs comparing FES to conventional exercise or control	Not reported
Sgourakis et al. 2010[40]	"To examine the impact of self-expanding stents versus locoregional treatment modalities in the setting of esophageal cancer palliation."	Esophageal and gastroesophageal junction cancer	Self-expanding stents vs. other treatment modalities	16 RCTs	1,027	RCTs of head-to-head comparisons of treatments	Not reported
Simpson et al. 2010[41]	"To conduct a systematic review and meta-analysis of RCTs comparing the effect of delayed defibrillation preceded by CPR with intermediate defibrillation on survival to hospital discharge."	Cardiac arrest due to ventricular fibrillation	Delayed vs. immediate defibrillation	3	658	RCTs of head-to head comparisons of delayed vs. immediate defibrillation	Not reported
Squizzato et al. 2010[12]	"To systematically summarize the best available evidence on the acute treatment and secondary prevention of RVO with antithrombotic and fibrinolytic drugs."	RVO	Anti-thrombotic drugs vs. fibrinolytic drugs	6	384	RCTs of head-to-head comparisons of anti-thrombotic drugs vs. fibrinolytic drugs	Not reported

Reference	Purpose/ Objective of Review	Medical Condition	Intervention(s)	Number of Trials Included	Total Sample Size	Original Trial Design[1]	Number of Trials using EQ/NI Design[2]
Sunkara et al. 2010[42]	"To compare treatment outcomes following transfer of frozen-thawed blastocysts that had developed on Day 5 or Day 6 after fertilization, and attempted to explore the reasons for the inconsistencies currently present in the literature."	Women having frozen-thawed blastocyst transfer cycles	Transfer of thawed blastocysts frozen on Day 5 vs. Day 6 following fertilization in vitro	15	2,502	Observational studies that compared transfer of thawed blastocysts frozen of Day 5 vs. Day 6 following fertilization in vitro	Not reported
Tamayo et al. 2010[7]	"Evaluated comparative data of commonly prescribed MDRs for both manic and depressive phases of bipolar disorder type I (BP 1)."	Manic and depressive phases of BP 1	MDR vs. placebo or active treatment	40	11,585	31 RCTs that compared MDR vs. placebo or active treatment for acute mania	Not reported
Tamhane et al. 2010[43]	"To systematically evaluate currently available data comparing thrombectomy followed by primary PCI with conventional PCI alone in patients with acute STEMI and to assess for differences if any between the various types of thrombectomy devices."	Patients with acute STEMI	Thrombectomy followed by PCI vs. conventional PCI	17	3,909	RCTs of head-to-head comparisons of thrombectomy followed by PCI vs. conventional PCI	Not reported

Reference	Purpose/ Objective of Review	Medical Condition	Intervention(s)	Number of Trials Included	Total Sample Size	Original Trial Design[1]	Number of Trials using EQ/NI Design[2]
Tang et al. 2010[4]	"To compare the clinical outcomes of mineral trioxide aggregate used as root-end filling with other materials in endodontic surgery to determine which modality offers more favorable outcomes."	Tooth to be treated had a dental history of a root canal treatment and demonstrated a periradicular lesion of strictly endodontic origin with or without clinical signs or symptoms	MTA (newer treatment) versus other (more standard) root end filling materials, such as amalgam, composite resin, glass ionomer, Super-ethoxy benzoic acid, or intermediate restorative.	5 comparative studies (2 RCTs comparing MTA to IRM, 2 studies comparing MTA to amalgam, and 1 RCT comparing MTA to gutta-percha)	414 teeth	RCTs and quasi-RCTs comparing treatments of interest.	Not reported
Testa et al. 2010[44]	"To assess, by means of a meta-analytic approach, the risk/benefit profile of DES vs. BMS in the treatment of SVG disease."	SVG disease	DES vs. BMS	18	3,294	3 RCTs; 15 registries of head-to-head comparisons of DES vs. BMS	Not reported
Valachis et al. 2010[45]	To compare treatment outcomes of patients with breast cancer treated with partial breast irradiation versus whole breast radiation.	Women with breast cancer who have undergone breast conserving therapy	PBI or limited field versus WBRT (standard)	3 RCTs comparing PBI to WBRT	1,140 (575 WBRT and 565 to limited field or PBI)	RCTs of head-to-head comparisons of PBI vs. WBRT	Not reported
Vasiliadis et al. 2010[15]	"To assess the effectiveness and safety of ACI compared to other treatment options (either conservative or surgical) for patients who require repair of clinically significant, symptomatic defects of the knee joint."	Symptomatic cartilage defects of the femur or patella	ACI vs. conservative or surgical treatments	9	626	RCTs or quasi-RCTs of head-to-head comparisons of ACI vs. conservation or surgical treatments	Not reported

Reference	Purpose/ Objective of Review	Medical Condition	Intervention(s)	Number of Trials Included	Total Sample Size	Original Trial Design[1]	Number of Trials using EQ/NI Design[2]
Vermeulan et al. 2010[10]	"To investigate the possible beneficial and harmful clinical effects of iodine in the treatment of all kinds of contaminated wounds."	Various wounds, including chronic, acute, burn wounds, pressure ulcers, and skin grafts	Local wound care product containing iodine (standard) vs. other antiseptic dressings or agents and non-antiseptic dressings or agents	29 RCTs	4,495 patients	RCTs comparing iodine containing wound care products to other antiseptic or non-antiseptic wound care products.	Not reported
Xie et al. 2010[46]	"To compare the clinical effectiveness and cost of PRFA and SRS for the management of early stage HCC."	Early stage HCC	PRFA vs. SRS	6	1,014	1 RCTs; 5 comparative cohort studies of head-to-head comparisons of PRFA vs. SRS	Not reported
Yang et al. 2010[1]	To update and perform a systematic review of a previous meta-analysis that suggested that "transdermal fentanyl was not inferior to sustained-release oral morphine in treating moderate to severe cancer pain."	Moderate to severe cancer pain	Transdermal fentanyl vs. sustained-release oral morphine (standard)	32 comparative trials	2,651 (1,296 fentanyl and 1355 morphine)	Prospective cohort study matched for age, gender, performance status, and type of tumor	Not reported
Agarwal et al. 2010[47]	To perform a "meta-analysis of comparative studies to compare outcomes of septal ablation with septal myectomy for treatment of hypertrophic obstructive cardiomyopathy."	Hypertrophic obstructive cardiomyopathy	SA vs. SM (standard)	12 retrospective cohort studies	706 (380 SA and 326 SM)	Included all comparative studies, including prospective/retrospective cohort and case control studies.	Not reported

Reference	Purpose/ Objective of Review	Medical Condition	Intervention(s)	Number of Trials Included	Total Sample Size	Original Trial Design[1]	Number of Trials using EQ/NI Design[2]
Avouac et al. 2010[48]	"To assess the efficacy of joint lavage alone or joint lavage combined with intra-articular steroid injection to alleviate pain and improve function in knee osteoarthritis."	Knee osteoarthritis	Joint lavage alone vs. joint lavage plus intra-articular steroid injections (standard).	6 RCTs; 3 of which compared joint lavage with steroid injection to joint lavage alone; the remaining studies included a placebo control group	415 (299 patients received joint lavage plus corticoid injection and 116 received joint lavage alone)	2 RCTs of head-to-head comparison of joint lavage with steroid injection vs. joint lavage alone.	Not reported
Chua et al. 2010[14]	"To provide an extensive overview of all published studies of neoadjuvant TACE in the setting of resectable HCC as a prelude towards establishing an evidence-based practice of this procedure in the management of resectable HCC."	Resectable HCC	Neoadjuvant TACE vs. non-TACE	18	3297	Observational studies, non-randomized case-control analytics studies, 3 RCTs of head-to-head comparisons of Neoadjuvant TACE vs. non-TACE	Not reported
Devaiah et al. 2010[49]	To evaluate the efficacy of Epley and Semont maneuvers with or without postural restrictions for treating vertigo.	Benign paroxysmal positional vertigo	Epley and Semont repositioning maneuvers with post-maneuver restrictions vs. no restrictions	6 controlled trials	523 patients	Prospective or retrospective controlled trials	Not reported
Loveman et al. 2010[13]	"To assess the clinical and cost-effectiveness of topotecan as second-line treatment for SCLC."	SCLC	Intravenous or oral topotecan vs. intravenous and oral compared with each other; BSC including radiotherapy; CAV; other chemotherapy regimens	5	826	RCTs of head-to-head comparisons of intravenous or oral topotecan vs. BSC, CAV, intravenous, oral, other chemotherapy agents	Not reported

Reference	Purpose/ Objective of Review	Medical Condition	Intervention(s)	Number of Trials Included	Total Sample Size	Original Trial Design[1]	Number of Trials using EQ/NI Design[2]
Valachis et al. 2010[50]	"To compare efficacy and tolerability of fulvestrant with aromatase inhibitors and tamoxifen that actually represents the standard of care in hormone-sensitive breast cancer."	Advanced breast cancer	Fulvestrant vs. AIs and tamoxifen (standard)	4 RCTs (2 compared fulvestrant to anastrozole, 1 compared fulvestrant to exemestane, and 1 compared fulvestrant to tamoxifen.	2,125 patients	RCTs of head-to-head comparisons of fulvestrant vs. any other hormonal therapy	Not reported

[1] Number of RCTs, Non-RCT comparative trials, other trial designs
[2] Number of included trials using a non-inferiority or equivalence research design

ACI	Autologous chondrocyte implantation
AD	Alzheimer's disease
AED	Antiepileptic drugs
AI	Aromatase inhibitors
AZA	Azathioprine
BMS	Bare metal stents
BPPV	Benign paroxysmal positional vertigo
BSC	Best supportive care
CAV	Cyclophosphamide, Adriamycin, and vincristine
CBT	Cognitive behavioral therapy
CHS	Chronic heart failure
COPD	Chronic obstructive pulmonary disease
CPR	Cardiopulmonary resuscitation
CYC	Cyclophosphamide
DES	Drug eluting stents
DLB	Dementia with lewy bodies
EBRT	External beam radiation therapy
FES	Functional electrical stimulation
GnRH	Gonadotropin-releasing hormone
HCC	Hepatocellular carcinoma
HDR	High-dose rate
HES	Hydroxyethyl starch
HOCM	Hypertrophic obstructive cardiomyopathy
ICBT	Intracavity brachytherapy
IFC	Interferential current therapy
IV	Intravenous
IVF	In vitro fertilization
LA	Laparoscopic appendectomy
LBP	Low back pain
LDR	Low-dose rate
MDD	Major depressive disorder
MDR	Monotherapy drug regimen
MMF	Mycophenolate mofetil
MTA	Mineral trioxide aggregate
NSSM	Mastectomy without reconstruction
OA	Open appendectomy
PBI	Partial breast irradiation
PCA	Patient-controlled analgesia
PCI	Percutaneous intervention
PPCI	Percutaneous coronary intervention
PRFA	Percutaneous radiofrequency ablation
RCT	Randomized controlled trial
RFA	Radiofrequency ablation
RVO	Retinal vein occlusion
SA	Septal ablation
SCLC	Small cell lung cancer
SM	Septal myectomy
smGPIs	Small-molecule glycoprotein IIb/IIIa receptor inhibitors
SNRI	Serotonin-norepinephrine reuptake inhibitor
SRS	Surgical resection
SSM	Skin-sparing mastectomy
SSRI	Selective serotonin reuptake inhibitor
STEMI	ST-segment elevation myocardial infarction
SVG	Saphenous vein graft
TACE	Transarterial chemoembolization
TCA	Tricyclic antidepressants
WBRT	Whole breast irradiation

Appendix B

Table 3. Assessment of internal validity within systematic reviews

Reference	Primary Outcomes	Secondary Outcomes	Method/Instrument Used to Assess Internal Validity	Overall Internal Validity Rating
Bodri et al. 2011[19]	Recipient ongoing pregnancy rate per randomized donor	Number of retrieved oocytes, duration stimulation, total gonadotropin consumption, and OHSS) incidence per randomized oocyte donor.	Using components of study design that are related to internal validity (Center for Reviews and Dissemination, 2001): randomization, allocation concealment, blinding, ITT, follow-up.	Quality of included trails was generally good
Groeneveld et al. 2011[11]	Safety: mortality, morbidity, bleeding and impaired coagulation, acute kidney injury, edema, hypoalbuminemia, pruritus, and anaphylactoid reactions	Not reported	Study quality and reliability for assessment of safety were judged according to several factors: randomized study design, size of patient population and statistical power to evaluate safety endpoints, colloid dose and demonstration of a dose-response relationship, adequacy of follow-up period, sensitivity of employed diagnostic methods for detecting complications, type of control fluid used as a comparator for safety, co-morbidities and severity of illness as indicators for the likelihood of observing complications, adherence to an a priori analysis plan, and multivariate analysis with adjustment for potential confounding factors.	Not reported
Hockenhull et al. 2011[20]	Death, major adverse cardiac and cerebrovascular events, or other major adverse events	Acute myocardial infarction, target vessel revascularization, target lesion revascularization, repeat treatment, and thrombosis	Used methods proposed by the Heart Collaborative Review Group and grading similar to that used in Villanueva, which considers the following: adequacy of randomization, allocation concealment, potential for selection bias, and adequacy of masking.	The quality of most of the included studies was rated as B due to lack of adequacy of allocation concealment and masking/blinding. According to the authors, the reporting of the use of intention-to-treat analysis was very good across all trials.

Reference	Primary Outcomes	Secondary Outcomes	Method/Instrument Used to Assess Internal Validity	Overall Internal Validity Rating
Seitz et al. 2011[21]	Change in symptoms of agitation and psychosis in dementia as measured on the various dementia NPS scales, drop-up due to adverse events	Changes in total scores for dementia NPS scales, changes on the CGI scale, changes in cognitive impairment scores, and caregiver stress; falls, headache, gastrointestinal upset, worsening of dementia, anxiety, headache, bleeding extrapyramidal symptoms, and hyponatremia	Risk of bias assessment provided by the Cochrane Collaboration: sequence generation, allocation concealment, blinding of participants, personnel and outcome assessors, incomplete outcome data, selective outcome reporting, other sources of bias.	Not reported
Beauchamp et al. 2010[5]	Peak power and peak oxygen uptake measured during an incremental exercise test on cycle ergometer or treadmill; endurance time measured from a constant power test; functional exercise capacity measured by 6MWT or 12MWT; HRQoL measured by CRQ; anxiety and depression measured by HAD	Lactate threshold, isotime ventilation, heart rate, breathing frequency and symptoms	Jadad (0 to 5 scale): randomization, blinding, withdrawals; PEDro 10 point scale: blinding, randomization, withdrawals, comparability of baseline characteristics, data reporting	Jadad – 2/5 (range 1–3)
Davis et al. 2010[22]	Clinical outcomes (e.g., scales that measure functional improvement), patient satisfaction, complications, length of hospital stay, and return to work	Not reported	Not reported	Not reported
Dibra et al. 2010[16]	All-cause death, recurrent myocardial infarction, reintervention, and stent-thrombosis	Not reported	Trials were evaluated for adequacy of allocation of concealment, intent-to-treat analysis, and blind assessment of outcomes. The authors reported using the criteria of Altman et al. and Juni et al. to assess adequacy of allocation of concealment.	The authors indicate that the main limitation was the absence of blinding of outcome assessors.

Reference	Primary Outcomes	Secondary Outcomes	Method/Instrument Used to Assess Internal Validity	Overall Internal Validity Rating
Dong et al. 2010[23]	Angiographic: Rates of TIMI grade 3 flow before and after PCI, Myocardial perfusion evaluated by cumulative ST-segment resolution in postprocedural electrocardiograms Clinical: 30-day and 8-month rates of mortality and reinfarction Safety: Major and minor bleeding complications according to the criteria of the TIMI trial	Not reported	Not reported	Not reported
Dong et al. 2010[24]	Angiographic: combined TIMI grade 2 and 3 on the initial angiogram; preprocedural and post-procedural TIMI grade 3 flow were also assessed; pre and post procedural myocardial perfusion evaluated by TMBG 3; Clinical: mortality at 30 day follow-up; incidence of reinfarction evaluated as another clinical outcome of interest Safety: major bleeding complications	Not reported	QUOROM guidelines for meta-analysis. Evaluated studies for the adequacy of allocation concealment, performance of the analysis according to the intention-to-treat principle, and blind assessment of the outcomes of interest. We used the criteria recommended by Altman and Schulz and Juni et al. to decide whether treatment allocation was adequately concealed.	No summary score used
Dubicka et al. 2010[18]	Depression and impairment scores, overall improvement, suicidality, and adverse events	NR	Used a method based on the authors of other systematic reviews on similar topics. Nine features of study quality were rated on a scale of 0 to 3, with a maximum score of 27. The 9 features included: method of randomization, intent-to-treat analysis, blinding of outcome assessors, blinding of patients, description of improvement, use of multiple outcome assessors, description of treatment dosage, use of manualized therapy, assessment of therapy adherence, and assessment of adherence to medication.	Mean quality score was 21, range 18 to 24.
Fuentes et al. 2010[25]	Pain measured by the VAS or numeric pain rating scale	Not reported	7 scales used: Delphi List, PEDro, Masstricht, Maastricht-Amsterdam List, Bizzini, van Tulder, and Jadad compiled in a set of 39 items. Categories included: patient selection, blinding, intervention, outcomes, statistics.	Not reported

Reference	Primary Outcomes	Secondary Outcomes	Method/Instrument Used to Assess Internal Validity	Overall Internal Validity Rating
Hong et al. 2010[26]	Microvascular events, mortality	Not reported	Quality of each selected trial evaluated by means of empirical evidence: randomization, allocation concealment, blinding.	Not reported
Hughes et al. 2010[27]	Disability rating	Recovery of unaided walking, recovery of walking with aid, discontinuation of ventilation, mortality, death or disability, treatment related fluctuation, and adverse events	Used methods described in the Cochrane Handbook, which considers sequence generation, allocation concealment, blinding, completeness of follow-up, freedom for selective reporting and other sources of bias.	Moderate quality evidence
Kalil et al. 2010[2]	Clinical cure	Microbiological eradication, methicillin-resistant, adverse events	Used the Jadad scale QUOROM guidelines to assess study quality.	Mean quality score according to the Jadad scale was 3.3, range 3 to 4.
Kesselheim et al. 2010[3]	Number or severity of seizures	NR	Used the Jadad scale to assess quality of RCTs and the Newcastle-Ottawa scale for assessing non-randomized trials.	Mean quality score 2.7, range 2 to 4
Krenke et al. 2010[17]	Length of hospital stay and failure rate	Duration of chest tube drainage, duration of fever, duration of respiratory distress, and volume of pleural fluid drainage	The quality of studies was assessed for the following strategies: allocation concealment, blinding, intention-to-treat analysis, and completeness of follow-up Authors did not report using a specific assessment scale or checklist.	The two studies that made up the comparative treatment evidence base used methods to conceal allocation, did not report blinding, used an intent-to-treat analysis, and adequate follow-up.
Lanitis et al. 2010[28]	Local recurrence, distant relapse, disease-free interval, severe postoperative complications, other outcomes (quality of life and cosmetic satisfaction)	Adverse events, complications	Since the studies used in the review were all non-randomized, the Newcastle-Ottawa Scale was used to assess study quality. Studies achieving 6 or more stars were considered to be of higher quality.	7 studies scored 6 or more stars on the modified Newcastle-Ottawa scale.
Lee et al. 2010[29]	Number of patients that: experienced (1) full remission; (2) partial remission; (3) overall remission; (4) relapse; (5) treatment failure; (6) end stage renal disease; or (7) died.	Number of patients who experienced a side effect	Used the Jadad scale to assess study quality.	Studies of induction therapy (MMF vs. CYC) Jadad scores ranged from 1 to 3; Studies of maintenance therapy (MMF vs. AZA) Jadad scores were 1 and 2; and for studies of high-dose vs. low-dose CYC scores were 2 and 2.

Reference	Primary Outcomes	Secondary Outcomes	Method/Instrument Used to Assess Internal Validity	Overall Internal Validity Rating
Liu et al. 2010[30]	Survival (during 1-year follow-up, during 3-year follow-up, and up to the end of the follow-up period) and recurrence (during 1-year follow-up, during 3-year follow-up, and up to the end of the follow-up period)	Not reported	No specific instrument reported. The authors indicate that study quality was measured according to the non-randomized controlled clinical trial quality evaluation standard.	Quality scores for the included studies ranged from 7 to 9.
Liu et al. 2010[31]	Operating time, complications, death rates, time of hospital stay, time to return to normal activities, and time to return to normal diet	Overall cost	Used the Jadad scale to assess study quality.	The mean Jadad score of the included studies was 4.25. The main limitations included: sample size, allocation concealment, and double blinding.
Liu et al. 2010[32]	Complications, death rates, survival rates, recurrence-free survival rates, and recurrence	Not reported	Used the Jadad scale to assess study quality.	The mean Jadad score of the included studies was 3. The quality of the studies was limited in terms of sample size, allocation concealment, and double blinding.
Liu et al. 2010[33]	OS, RFS, pelvic control rate	Rates of local and distant recurrence, complications	Juni quality assessment criteria for RCTs: randomization method, blinded assessment of outcomes, allocation concealment, losses to follow-up, ITT.	Not reported
Macedo et al. 2010[34]	Pain, disability, global perceived effect, and return to work	Not reported	PEDro scale (0-10): masking, baseline comparability, allocation concealment, ITT, adequate follow-up	Median=6 (range 3 to 9)
Machado et al. 2010[35]	Remission rate defined as scores $\leq$7 or 8 and $\leq$10 or 12 for the HAM-D and MADRS scales, respectively, and measured at 8-12 weeks of treatment.	Not reported	Downs-Black, 27-item quality assessment checklist. Categories included: study design, sample selection, data presentation, statistical analysis and statistical power.	All studies reported quality above 80%
Meier et al. 2010[6]	Survival to hospital discharge,	Return to spontaneous circulation, favorable neurologic outcome at discharge, and long-term outcome (survival at 1-year)	Used the Jadad scale to assess study quality	The mean Jadad score of the included studies was 4. The quality of the studies was limited in terms of lack of double blinding.

Reference	Primary Outcomes	Secondary Outcomes	Method/Instrument Used to Assess Internal Validity	Overall Internal Validity Rating
Milito et al. 2010[36]	Healing	Operative time and hospitalization, pain, analgesic requirements, blood loss, wound healing, convalescence period, postoperative continence impairment, anal stenosis and relapse, and cost effectiveness	No specific method reported; discuss the following aspects of quality of RCTs: allocation of concealment, blinding, mean outcome measures, statistical methods, and length of follow-up.	Not reported
Murphy et al. 2010[37]	VAS for pain, postoperative nausea and vomiting, pruritus	Not reported	Not reported	Not reported
Myers et al. 2010[9]	Pain measured using a validated scale	Not reported	No specific method reported; discuss the following aspects of quality of RCTs: publication status (full publication or meeting abstract), allocation of concealment, blinding, statement of statistical power or sample size calculation, intention-to-treat analysis, and statement of funding/sponsorship.	Not reported
Pan et al. 2010[38]	Mortality, recurrent myocardial infarction, repeat revascularization, and stent thrombosis	Not reported	Not reported	Not reported
Riemsa et al. 2010[8]	Overall survival, progression free survival, time to progression, response rate, complete response, response duration, stable disease, and quality of life	Adverse events	Used the Cochrane Collaboration quality assessment checklist.	The authors state that overall all 4 studies included in analysis had a low risk of bias.
Sbruzzi et al. 2010[39]	Functional capacity (measured by peak oxygen uptake	Distance of 6-min walk test and muscle strength	Used the Jadad and PEDro scales and specifically considered concealment of allocation, intention to treat analysis, baseline comparability, blinding of outcome assessors, and description of losses and exclusions	The quality of the studies was poor, all studies received a Jadad score ≥ 3 and 5 studies received a PEDro score of ≥ 5 (out of 10).
Sgourakis et al. 2010[40]	Patients requiring re-intervention, reflux, complications, procedural death, and overall survival.	Not reported	Used the Jadad scale to assess study quality.	The mean Jadad score of included studies was 2.7, range 1 to 4.

Reference	Primary Outcomes	Secondary Outcomes	Method/Instrument Used to Assess Internal Validity	Overall Internal Validity Rating
Simpson et al. 2010[41]	Survival to hospital discharge (overall), survival to hospital discharge (response time ≤ 5 minutes), survival to hospital discharge (response time > 5 minutes)	Not reported	Internal validity assessment performed using methodology recommended by Cochrane Collaboration: risk of bias across the following domains – sequence generation, allocation concealment, blinding of participants, personnel and outcome assessors, incomplete outcome data, selective outcome reporting, other potential threats to validity	Not reported
Squizzato et al. 2010[12]	Visual acuity, neovascular complications, recurrent events, bleeding complications	Not reported	Jadad scale: randomization, blinding, follow-up	Not reported
Sunkara et al. 2010[42]	Ongoing pregnancy/live birth rate	Post-thaw blastocyst survival rate, clinical pregnancy rate, and miscarriage rate	Newcastle-Ottawa Quality Assessment Scales: selection of cases and controls, study group comparability, exposure to intervention and treatment outcome.	Not reported
Tamayo et al. 2010[7]	Response rates, remission rates, discontinuation rates due to adverse events, lack of efficacy, or discontinuation due to any cause, NNT or NNH	Not reported	Jadad scale	Not reported
Tamhane et al. 2010[43]	Clinical: Death, stroke, TVR, reinfarction. Myocardial perfusion Angiographic: post procedural rates of TIMI grade 3 flow and TMGB	Not reported	Allocation concealment, study design, ITT, blinding assessment of outcome measures	Did not use a quality score
Tang et al. 2010[4]	Success (complete healing or incomplete healing) and failure (uncertain healing or unsatisfactory healing)	Not reported	Assessed quality based on the following factors: RCT, control, double-blinding, allocation of concealment, description of withdrawals and dropouts, sample size predetermined, intent-to-treat, operator experience reported, treatment procedures described, measurements standardized, and evaluation methods clearly described. Studies with low risk of bias graded A, moderate risk of bias B, and high risk of bias graded C.	RCTs rated as A-low risk of bias

Reference	Primary Outcomes	Secondary Outcomes	Method/Instrument Used to Assess Internal Validity	Overall Internal Validity Rating
Testa et al. 2010[44]	Combined rate of MAE, defined as cumulative risk of all cause death and nonfatal acute myocardial infarction, TVR and TLR, rate of stent thrombosis	Not reported	Cochrane Collaboration Newcastle-Ottawa scale for assessing quality of cohort study: sequence generation, allocation concealment, blinding, selective reporting, and incomplete data.	Registries overall low quality RCTs overall good quality
Valachis et al. 2010[45]	OS	Number of local recurrences, true and elsewhere breast recurrences, axillary recurrences, supraclavicular recurrences, and distant recurrences	No specific method reported; discuss the following aspects of quality: method of randomization, allocation concealment, intent-to-treat analysis, and patient withdrawal	The authors reported that 2 trials described the model of randomization and the model of allocation of concealment.
Vasiliadis et al. 2010[15]	Lysholm score, Tegner score, Modified Cincinnati score, VAS, Meyers score, Stanmore scores, SF-36 scores, repair tissue evaluation and histological assessment, complications, post-operative clinical improvement	Not reported	Cochrane Handbook for Systematic Reviews of Interventions: selective reporting, baseline comparability.	Overall the quality of evidence can be rated as average to low
Vermeulan et al. 2010[10]	Bacterial load and wound healing	Adverse events, cost, and length of hospital stay	No specific method reported	Authors indicate that study quality was limited, mainly because of lack of concealment of randomization, use of quasi-randomization procedures, not reporting if performed an intent-to-treat analysis, not reporting if outcome assessor were blinded, not using independent outcome assessors, and not reporting funding source.
Xie et al. 2010[46]	Survival, recurrence, DFS, additional treatment, complications, hospitalization, patients' attitudes toward treatment options, cost analysis	Not reported	Not reported	Not reported

Reference	Primary Outcomes	Secondary Outcomes	Method/Instrument Used to Assess Internal Validity	Overall Internal Validity Rating
Yang et al. 2010[1]	Remission rate of pain and incidence of opioid-related adverse effects	Quality of life	Used a quality checklist adapted from MOOSE standard, which includes the following 6 measures: prospective study design, group comparability on confounding factors, blinding of outcome assessors, length of follow-up, relation between outcome and exposure appropriately measured, and used appropriate statistical analysis	Not reported
Agarwal et al. 2010[47]	30 day all-cause mortality	Functional status, reinterventions, pacemaker insertion, ventricular arrhythmias, cardiac dimensions, mitral regurgitation, systolic anterior motion of mitral valve, length of hospital stay, and exercise tolerance	Not reported	Not reported
Avouac et al. 2010[48]	Pain and functional status	Not reported	Used the Jadad scale to assess study quality.	Jadad scores ranged from 1 to 5.
Chua et al. 2010[14]	Improved DFS	Overall survival, morbidity, mortality, pathological response, pattern of recurrence	Not reported	Not reported
Devaiah et al. 2010[49]	Relapse of subjective vertigo at follow-up	A negative Dix-Hallpike maneuver	Not reported	Not reported
Loveman et al. 2010[13]	Time to disease progression, progression-free survival, response rate, response duration, overall survival, symptom control, health-related quality of life, cost-effectiveness, adverse effects	Not reported	Criteria recommended by the Center for Reviews and Dissemination (CRD): randomization, allocation concealment, baseline characteristics, eligibility, blinding assessors, blinding care provider, patient blinding, reporting outcomes, ITT, withdrawals explained.	Not reported

Reference	Primary Outcomes	Secondary Outcomes	Method/Instrument Used to Assess Internal Validity	Overall Internal Validity Rating
Valachis et al. 2010[50]	OS, time to tumor progression, proportion of patients with complete or partial response after treatment (objective response), and proportion of patients with an objective response or stable disease lasting ≥ 24 weeks (clinical benefit)	Adverse events	Recorded the following methodological quality items: mode of randomization, allocation concealment, subject withdrawals, blinding, if interim analysis was planned or performed, and if intent-to-treat analysis performed	No overall rating reported; 3 of 4 trials were double blind, 1 study reported mode of randomization and methods for ensuring allocation concealment, and no study was stopped early because of statistically significant differences in an interim analysis

AKI	Acute kidney injury
AZA	Azathioprine
CGI	Clinical global impression
CRD	Center for reviews and dissemination
CRQ	Chronic Respiratory Questionnaire
CYC	Cyclophosphamide
DFS	Disease-free survival
HAD	Hospital Anxiety and Depression scale
HAM-D	Hamilton Depression rating scale
HRQoL	Health-related Quality of life
ITT	Intention-to-treat
MAE	Major adverse events
NNH	Number needed to harm
NNT	Number needed to treat
NPS	Neuropsychiatric symptoms
NRS	Numeric pain rating scale
OHSS	Ovarian hyperstimulation syndrome
OS	Overall survival
PEDro	Physiotherapy evidence base database
RFS	Relapse-free survival
TIMI	Thrombolysis in Myocardial Infarction
TLR	Target lesion revascularization
TMBG	TIMI myocardial blush grade
TVR	Target vessel revascularization
VAS	Visual analog scale

Appendix C

Table 4. Method of defining MID within systematic reviews

Reference	MID Defined by Systematic Reviewers	Pre-specified by Reviewers	Range of MID Defined within Original Trials
Bodri et al. 2011[19]	No MID was defined or used	No	Not reported
Groeneveld et al. 2011[11]	No MID was defined or used	No	Not reported
Hockenhull et al. 2011[20]	No MID was defined or used	No	Not reported
Seitz et al. 2011[21]	No MID was defined or used	No	Not reported
Beauchamp et al. 2010[5]	0.5 minimal important difference for the CRQ domains of dyspnea and total score, and 54 m for the 6 minute walk test (6MWT)	Yes	Not reported
Davis et al. 2010[22]	No MID was defined or used	No	Not reported
Dibra et al. 2010[16]	No MID was defined or used	No	Not reported
Dong et al. 2010[23]	No MID was defined or used	No	Not reported
Dong et al. 2010[24]	No MID was defined or used	No	Not reported
Dubicka et al. 2010[18]	No MID was defined or used	No	Not reported
Fuentes et al. 2010[25]	No MID was defined or used	No	Not reported
Hong et al. 2010[26]	No MID was defined or used	No	Not reported
Hughes et al. 2010[27]	No MID was defined or used	No	Not reported
Kalil et al. 2010[2]	No MID was defined or used	No	Not reported
Kesselheim et al. 2010[3]	No MID was defined or used	No	Not reported
Krenke et al. 2010[17]	No MID was defined or used	No	Not reported
Lanitis et al. 2010[28]	No MID was defined or used	No	Not reported
Lee et al. 2010[29]	No MID was defined or used	No	Not reported

Reference	MID Defined by Systematic Reviewers	Pre-specified by Reviewers	Range of MID Defined within Original Trials
Liu et al. 2010[30]	No MID was defined or used	No	Not reported
Liu et al. 2010[31]	No MID was defined or used	No	Not reported
Liu et al. 2010[32]	No MID was defined or used	No	Not reported
Liu et al. 2010[33]	No MID was defined or used	No	Not reported
Macedo et al. 2010[34]	No MID was defined or used	No	Not reported
Machado et al. 2010[35]	No MID was defined or used	No	Not reported
Meier et al. 2010[6]	The authors defined a "clinically relevant change for the primary outcome-survival to hospital discharge as a relative change of 20% to 25%."	Yes	Not reported
Milito et al. 2010[36]	No MID was defined or used	No	Not reported
Murphy et al. 2010[37]	No MID was defined or used	No	Not reported
Myers et al. 2010[9]	No MID was defined or used	No	Not reported
Pan et al. 2010[38]	No MID was defined or used	No	Not reported
Riemsa et al. 2010[8]	No MID was defined or used	No	Not reported
Sbruzzi et al. 2010[39]	No MID was defined or used	No	Not reported
Sgourakis et al. 2010[40]	No MID was defined or used	No	Not reported
Simpson et al. 2010[41]	No MID was defined or used	No	Not reported
Squizzato et al. 2010[12]	No MID was defined or used	No	Not reported
Sunkara et al. 2010[42]	No MID was defined or used	No	Not reported
Tamayo et al. 2010[7]	Only NNTs or NNHs <10 are considered clinically meaningful	Yes	Not reported
Tamhane et al. 2010[43]	No MID was defined or used	No	Not reported
Tang et al. 2010[4]	No MID was defined or used	No	Not reported
Testa et al. 2010[44]	No MID was defined or used	No	Not reported
Valachis et al. 2010[45]	No MID was defined or used	No	Not reported
Vasiliadis et al. 2010[15]	No MID was defined or used	No	Not reported

Reference	MID Defined by Systematic Reviewers	Pre-specified by Reviewers	Range of MID Defined within Original Trials
Vermeulan et al. 2010[10]	No MID was defined or used	No	Not reported
Xie et al. 2010[46]	No MID was defined or used	No	Not reported
Yang et al. 2010[1]	No MID was defined or used	No	Not reported
Agarwal et al. 2010[47]	No MID was defined or used	No	Not reported
Avouac et al. 2010[48]	No MID was defined or used	No	Not reported
Chua et al. 2010[14]	No MID was defined or used	No	Not reported
Devaiah et al. 2010[49]	No MID was defined or used	No	Not reported
Loveman et al. 2010[13]	No MID was defined or used	No	Not reported
Valachis et al. 2010[50]	No MID was defined or used	No	Not reported

MID Minimal important difference
NNH Number needed to harm
NNT Number needed to treat

Appendix D

Table 5. Methods used to draw conclusions within systematic review

Reference	Performed Meta-analysis	Meta-analytic Approach[1]	Statistical Method(s) Used[2]	Meta-analytic Results	Results of Heterogeneity Testing
Bodri et al. 2011[19]	Yes	Fixed effects models if significant heterogeneity was absent. Ongoing pregnancies were meta-analyzed for studies in which one donor donated eggs to one recipient only; the reason for this approach is that studies that contain multiple recipients from one donor have multiple nonindependent outcome data which invalidate the assumptions made in standard meta-analytical approaches.	Binary data risk ratios (RR), 95% CI; WMD was calculated for continuous variables	Ongoing Pregnancy Rate RR 1.15, 95% CI 0.97 to 1.36, p = 0.12 Oocytes retrieved WMD -0.60, 95% CI -2.26 to 1.07, p = 0.48 Duration of stimulation WMD -0.90 days, 95% CI -1.61 to -0.20, days p = 0.001 Gonadotropin consumption WMD -264 IU, 95% CI -682 to 154 IU, p = -.22 OHSS incidence RR 0.61, 95% CI 0.18 to 2.15, p = 0.45	Ongoing Pregnancy Rate I^2 = 0% Oocytes retrieved I^2 = 71% Duration of stimulation I^2 = 91% Gonadotropin consumption I^2 = 95% OHSS incidence I^2 = 0%
Groeneveld et al. 2011[11]	No	Due to the presence of clinical heterogeneity, study results summarized narratively.	Not reported	Not reported	Not reported
Hockenhull et al. 2011[20]	Yes	Pooled estimates using fixed or random effects model depending on the presence of heterogeneity.	Odds ratio (OR) with 95% confidence intervals (CIs)	"There were no statistically significant differences in death, [acute myocardial infarction], or thrombosis between [drug eluting stents] and [bare metal stents]. For composite events, [target lesion revascularization] and target vessel revascularization reductions were evident with use of sirolimus, paclitaxel, everolimus, dexamethasone, zotarolimus, and tacrolimus-eluting stents."	I^2 ranged from 0.0% to 30% across outcomes.

Reference	Performed Meta-analysis	Meta-analytic Approach[1]	Statistical Method(s) Used[2]	Meta-analytic Results	Results of Heterogeneity Testing
Seitz et al. 2011[21]	Yes	Random effect models for studies reporting continuous outcomes using same rating scale. Random effect Mantel-Haenszel risk ratios and CI for meta-analysis of binary outcomes for studies reporting the same outcome.	Weighted mean differences (WMD), risk ratios (RR) standard deviations	SSRIs vs. Typical Antipsychotics (Haloperidol) Change in Cohen Mansfield Agitation Inventory (CMAI) WMD 4.66, 95% CI -3.58 to 12.90 Trazodone vs. Typical Antipsychotics (Haloperidol) Change in CMAI total scores WMD 3.28, 95% CI -3.28 to 15.74 Clinical Global Impression (CGI) RR 1.25, 95% CI 0.82 to 2.34	SSRIs vs. Typical Antipsychotics (Haloperidol) Change CMAI $I^2 = 0\%$ Trazodone vs. Typical Antipsychotics (Haloperidol) Change in CMAI total scores $I^2 = 0\%$ CGI $I^2 = 0\%$
Beauchamp et al. 2010[5]	Yes	Random effects model used for all analyses	Weighted mean difference (WMD) selected when estimating the total effect of combined data. Point estimates and CIs for the difference between groups assessed to see if they exceeded minimal important difference	Peak power WMD 1 W, 95% CI -1 to 3 Peak oxygen uptake WMD -0.04 l/min, 95% CI -0.13 to 0.05 Chronic Respiratory Questionnaire (CRQ) dyspnea score WMD -0.2 units, 95% CI -0.5 to 0.0 6MWT WMD 4 m, 95% CI -15 to 23	Tests of heterogeneity on all measures of exercise capacity and QoL were not significant and I^2 ranged from 0% to 13%
Davis et al. 2010[22]	Yes	Pooled mean difference using methods described by Basu for continuous measures and rate difference.	Mean difference (MD) with 95% CIs	"No significant differences were found in clinical outcomes or complications for the 2 groups. However, open acromioplasty was associated with longer hospital stays and a greater length I time until return to work compared with the arthroscopic technique."	Not reported

Reference	Performed Meta-analysis	Meta-analytic Approach[1]	Statistical Method(s) Used[2]	Meta-analytic Results	Results of Heterogeneity Testing
Dibra et al. 2010[16]	Yes	Pooled data in direct comparison using DerSimonian and Laird's random effects model.	Hazard ratio (HR) with 95% CIs	Hazard of death or recurrent myocardial infarction: HR = 0.91, 95% CI 0.75 to 1.09; Reduction in hazard of reintervention: HR = 0.41, 95% CI 0.32 to 0.52, favors drug-eluting stents; Stent thrombosis: HR = 0.84, 95% CI 0.61 to 1.17	Death or recurrent myocardial infarction: I^2 = 0.0%, reintervention: I^2: = 29.9%, stent thrombosis: I^2 = 0.0%
Dong et al. 2010[23]	Yes	Mantel-Haenszel method according to fixed effect model or DerSimonian and Laird method according to random effect model when heterogeneity across trials was found	Odds Ratio (OR) and 95% CI for dichotomous variables	Angiographic Initial TIMI OR 1.12, 95% CI 0.90 to 1.39, p = 0.31 Clinical 30 day mortality OR 0.83, 95% CI 0.46-1.51, p = 0.54; 8 month mortality OR 0.78, 95% CI 0.41 to 1.46, p = 0.43 Safety Major bleeding complications OR 1.32, 95% CI 0.66 to 2.62; Minor bleeding complications OR 0.82, 95% CI 0.53 to 1.27	Angiographic Initial TIMI I^2 = 0% Final TIMI 3 flow I^2 = 0.70% Clinical 30 day mortality I^2 = 0% 8 month mortality I^2 = 0% Safety Major bleeding complications I^2 = 0% Minor bleeding complications I^2 = 0%
Dong et al. 2010[24]	Yes	Mantel-Haenszel fixed effect method; DerSimonian and Laird random effects method	Intention to treat; Odds ratios (OR) with 95% CI	Angiographic & Electrocardiographic endpoints TIMI grade 2 or 3 flow rate OR 1.40, 95% CI 1.20-1.64, p <0.001 Final TIMI 3 flow OR 0.87, 95% CI 0.70-1.10, p = 0.25 Clinical 30 day mortality OR 1.04, 95% CI 0.67 to 1.61, p = 0.85 Safety Major bleeding complications OR 1.25, 95% CI 0.76 to 2, p = 0.38	Angiographic & Electrocardiographic endpoints TIMI grade 2 or 3 flow rate I^2 = 11.2% Final TIMI 3 flow I^2 = 0% Clinical 30 day mortality I^2 = 2.2% Safety Major bleeding complications I^2 = 0%

Reference	Performed Meta-analysis	Meta-analytic Approach[1]	Statistical Method(s) Used[2]	Meta-analytic Results	Results of Heterogeneity Testing
Dubicka et al. 2010[18]	Yes	Pooled estimates using fixed or random effects model depending on the presence of heterogeneity.	Weighted mean difference (WMD) for continuous data and odds ratio (OR) for dichotomous data, all with 95% CIs	12 weeks follow-up Depression outcomes: WMD = 0.04, 95% CI -0.09 to 0.17; Impairment outcomes: WMD = 0.20, 95% CI 0.08 to 0.33, favors combined treatment. These results were based on data from one scale in favor of combined treatment; Suicidality: 0.00, 95% CI -0.14 to 0.15; Adverse events: no statistically significant difference on other adverse events; Later follow-up No change, except evidence of improved impairment no longer significant	Depression: I^2 = 0.0% at 12 weeks and 32% at later follow-up; Impairment: I^2 = 0.0% at both follow-ups; Suicidality: 0.0% at 12 weeks and 19.3% at later follow-ups.
Fuentes et al. 2010[25]	Yes	DerSimonian and Laird random-effects model. If there was relative homogeneity, a fixed-effects model was used	Mean difference, 95% CI; chi-square	IFC vs. Comparison group: Pain Intensity at Discharge MD 0.16 95% CI -0.62 to 0.31 IFC as a supplement to another treatment vs. Control group: Pain intensity at discharge MD 2.45 95% CI 1.69 to 3.22 IFC as supplement to another treatment vs. comparison: Pain intensity at discharge MD 0.55, 95% CI -0.33 to 1.44	IFC vs. Comparison group: Pain Intensity at Discharge I^2 = 0% IFC as a supplement to another treatment vs. Control group: Pain intensity at discharge I^2 = 70% IFC as supplement to another treatment vs. comparison Pain intensity at discharge I^2 = 81%
Hong et al. 2010[26]	Yes	Random effects model	Relative risk (RR), 95% CI	Macrovascular events RR 0.92, 95% CI 0.87 to 0.98, p = 0.005 Mortality RR 0.95, 95% CI 0.80 to 1.12, p = 0.55 Cardiovascular deaths RR 1.10, 95% CI 0.79 to 1.53, p = 0.57	Macrovascular events I^2 = 0% Mortality I^2 = 66.7% Cardiovascular deaths I^2 = 65.2%

Reference	Performed Meta-analysis	Meta-analytic Approach[1]	Statistical Method(s) Used[2]	Meta-analytic Results	Results of Heterogeneity Testing
Hughes et al. 2010[27]	Yes	Pooled estimates using a fixed effects model.	Relative risk (RR) for dichotomous outcomes and weighted mean difference (WMD), all with 95% CIs	Comparison of immunoglobulin to plasma exchange (5 studies; main comparison) Improvement of disability scale: WMD = -0.02, 95% CI -0.25 to 0.20. No statistically significant different results found between groups for any secondary outcomes (e.g., recovery of walking, adverse events)	Improvement of disability scale: I^2 = 41.0%
Kalil et al. 2010[2]	Yes	Pooled estimates using fixed or random effects model depending on the presence of heterogeneity.	Relative risk (RR) with 95% CIs	Clinical cure: RR = 1.01, 95% CI 0.93 to 1.10, p = 0.83; Microbiological eradication: RR = 1.10, 95% CI 0.98 to 1.22, p = 0.10; The risk of thrombocytopenia and gastrointestinal events is higher with linezolid, but no difference for renal dysfunction or all-cause mortality.	Clinical cure: I^2 = 0.0%, microbiological eradication: I^2: 16.0%
Kesselheim et al. 2010[3]	Yes, included 7 of 9 RCTs	Compared percentage of controlled and uncontrolled patients for brand-name and generic drugs. The summary estimate was a weighted average of odds ratios (ORs) from the studies were the inverse of the variance of the ORs.	Odds ratio (OR) with 95% CIs	Controlled seizures: OR = 1.1; 95% CI 0.9 to 1.2	No evidence of statistically significant heterogeneity.
Krenke et al. 2010[17]	Yes	Not reported	Computed risk ratio and 95% CIs for binary outcomes and mean difference and 95% CIs for continuous outcomes	The findings of the RCTs that compared fibrinolysis to video-assisted thoracoscopic surgery (VATs) demonstrated no statistically significant benefit of VATs duration of hospital stay, failure rates, or other outcomes.	Not reported

Reference	Performed Meta-analysis	Meta-analytic Approach[1]	Statistical Method(s) Used[2]	Meta-analytic Results	Results of Heterogeneity Testing
Lanitis et al. 2010[28]	Yes	Pooled estimates using fixed or random effects model depending on the presence of heterogeneity.	Odds ratio (OR) with 95% CIs	Local recurrence: OR = 1.22, 95% CI 0.85 to 1.74; Distant relapse: OR = 0.67, 95% CI: 0.48 to 0.94, favors SSM group; Disease-free interval: OR = No meta-analysis performed due to lack of reliable data; Postoperative severe complication: OR = 0.81, 95% CI 0.57 to 1.16	Local recurrence: I^2 = 25.9%; distant relapse: I^2 = 0.0%; No statistically significant evidence of heterogeneity between studies reporting on severe complications
Lee et al. 2010[29]	Yes	Pooled estimates using random or fixed effect model depending on the presence of heterogeneity.	Risk ratio (RR) with 95% CIs	MMF vs. CYC: Complete remission: RR = 1.613, 95% CI 0.908 to 2.863, p = 0.013 Partial remission: RR = 1.031, 95% CI 0.678 to 1.567, p = 0.887 The two groups did not differ significantly in terms of the risk of developing ESRD or mortality MMF vs. AZA: Developing ESRD: RR = 1.360, 95% CI 0.236 to 7.828, p = 0.730 No other MA performed; Results of 1 study indicated that survival rates and renal failure was higher in MMF group than AZA group. Low-dose IV CYC vs. High-dose CYC: Relapse rate: RR = 0.465, 95% CI 0.261 to 0.830, p = 0.010, favors low-dose Treatment failure: RR = 0.451, 95% CI 0.202 to 1.09, p = 0.053 Infection risk: RR = 0.688, 95% CI 0.523 to 0.905, p = 0.008	Between-study heterogeneity was found during analysis of complete remission, response rate, and amenorrhea associated with induction therapy.

Reference	Performed Meta-analysis	Meta-analytic Approach[1]	Statistical Method(s) Used[2]	Meta-analytic Results	Results of Heterogeneity Testing
Liu et al. 2010[30]	Yes	Pooled estimates using random or fixed effects model depending on the presence of heterogeneity.	Odds ratio (OR) with 95% CIs	Survival at 1 year: OR = 0.94, 95% CI 0.65 to 1.36, p = 0.75; Survival at 3 year: OR = 0.92, 95% CI 0.56 to 1.51, p = 0.73 Survival at end of follow-up: OR = 0.82, 95% CI 0.48 to 1.39, p = 0.46; Recurrence at 1 year: OR = 0.96, 95% CI 0.69 to 1.33, p = 0.80; Recurrence at 3 year: OR = 1.19, 95% CI: 0.63 to 2.27, p = 0.59; Recurrence at end of follow-up: OR = 1.73, 95% CI 1.04 to 2.87, p = 0.03; Compared to surgical resection, RFA did not cause liver function damage, had a lower rate of complications, and was more affordable.	Survival at 1 year: I^2 = 0.0%, Survival at 3 year: I^2 = 75.5%, Survival at end of follow-up: I^2 = 78.5%; for recurrence outcomes authors indicated the presence of significant between-study heterogeneity, but did not present I^2 values.
Liu et al. 2010[31]	Yes	Pooled estimates using random or fixed effects model depending on the presence of heterogeneity.	Odds ratio (OR) for dichotomous variables and weighted mean difference (WMD) for continuous variables, all with 95% CIs.	Operating time: WMD = 7.60 min, 95% CI 6.03 to 9.17 min, p <0.001, longer for laparoscopic (LA); Hospital stay: WMD = -0.82, 95% CI -0.93 to -0.70 days, favors LA; Normal activities: WMD = -6.85, 95% CI -7.62 to -6.09, favors LA; Normal diet: WMD = -0.61, 95% CI: -0.86 to -0.36 days, favors LA; Complications: OR = 0.97, 95% CI 0.29 to 3.25, p = 0.96	Operating time: I^2 = 65.3%, complications: I^2 = 22.0%; hospital stay: I^2 = 80.9%; normal activities: I^2 = 97.3%

Reference	Performed Meta-analysis	Meta-analytic Approach[1]	Statistical Method(s) Used[2]	Meta-analytic Results	Results of Heterogeneity Testing
Liu et al. 2010[32]	Yes	Pooled estimates using random or fixed effects model depending on the presence of heterogeneity.	Odds ratio (OR) for dichotomous variables and weighted mean difference (WMD) for continuous variables, all with 95% CIs.	Complications: OR = 1.41, 95% CI 0.55 to 3.65, p = 0.48; Death rates: OR = 3.28, 95% CI 0.86 to 12.56, p = 0.08; Survival rates: 1 year OR = 1.03, 95% CI 0.65 to 1.62, p = 0.91; 5 year OR = 1.38 95% CI, 1.01 to 1.87, p = 0.04; Recurrence-free survival rates: 1 year OR = 1.54, 95% CI 1.07 to 1.21, p = 0.02; 5 year OR = 1.52, 95% CI 1.12 to 2.06, p = <0.007; Recurrence in previous site: OR = 0.28, 95% CI 0.12 to 0.65, p = 0.03	Complications: I^2 = 11.4%; Death rates: I^2 = 25.6%; 1 year survival: I^2 = 22.0%; 5 year survival: I^2 = 40.6%; 1 year recurrence free survival: I^2 = 61.9%; 5 year survival: I^2 = 59.9%; Recurrence in same site: I^2 = 24.7;
Liu et al. 2010[33]	Yes	Fixed-effect model used to pool all results if there was no evidence of heterogeneity between studies. Random effects model used if there was evidence of heterogeneity	If published and unpublished data were contrary, the results based on intent-to-treat (ITT) analysis were used. IF ITT analysis was not available the more conservative results were used. Binary outcomes, risk ratios (RR), 95% CI. For continuous variables, weighted mean difference (WMD), 95% CIs	5 year OS Overall (stage I, II, III cervical carcinoma) RR 0.93, 95% CI 0.84 to 1.04 5 year Disease Specific survival (DSS) Pooled RR 0.95, 95% CI 0.84 to 1.07 Locoregional recurrence rate RR 1.09, 95% CI 0.83 to 1.43 Para-aortic lymph node metastasis RR 0.79, 95% CI 0.40 to 1.53 Combined local and distant metastasis RR 2.23, 95% CI 0.78 to 6.34 Complications severe bladder complications pooled RR 1.33 95% CI 0.53 to 3.34, p = 0.54 Rectosigmoid complications RR 1.00, 95% CI 0.52 to 1.91, p = 0.00 small bowel complications RR 3.37 95% CI 1.06 to 10.72, p = 0.04	P values reported for heterogeneity for 5 year OS and DSS: 5 year OS P = 0.22 5 year DSS overall P = 0.38 Locoregional recurrence rate I^2 = 0% Para-aortic lymph node metastasis I^2 = 0% Combined local and distant metastasis I^2 = 0% Small bowel complications not reported

Reference	Performed Meta-analysis	Meta-analytic Approach[1]	Statistical Method(s) Used[2]	Meta-analytic Results	Results of Heterogeneity Testing
Macedo et al. 2010[34]	Yes	Random and fixed effect models	Weighted mean differences (WMD), 95% CI	Graded activity vs. minimal intervention or no treatment: Pain short term -6.2 95% CI -9.4 to -3.0 Pain Intermediate term: -5.5 95% CI -9.9 to -1.0 Disability short term: -6.5 95% CI -10.1 to -3.0 Disability intermediate term: -3.9 95% CI -7.4 to -0.4	Not reported
Machado et al. 2010[35]	Yes	Random effects model and inverse variance method	Odds ratio (OR), differences in success rates, 95% CI, ITT and per-protocol (PP) data analyzed in which dropouts due to adverse drug reactions (ADRs) or lack efficacy (LoE)/effectiveness were included as failures	Remission rates: ITT OR 1.27, 95% CI 1.06 to 1.52 PP OR 1.56, 95% CI 1.19 to 2.04 Studies produced meta-analytic differences favoring SNRIs of 5.7%, $p = 0.007$ and 11.6%, $p < 0.001$ using ITT and PP models, respectively Meta-analytic rates for total dropouts and those due to ADRs were higher but non-significant ($p = 0.086$) in SNRI group vs. rates in SSRI group Rates of dropout due to LOE were non-significant between studied groups. SSRIs reported statistically significant lower dropout rates due to ADRs when compared with SNRIs	Remission Rates: ITT $I^2 = 29\%$ PP $I^2 = 52\%$

Reference	Performed Meta-analysis	Meta-analytic Approach[1]	Statistical Method(s) Used[2]	Meta-analytic Results	Results of Heterogeneity Testing
Meier et al. 2010[6]	Yes	Pooled estimates using the DerSimonian and Laird's random effects model.	Odds ratio for each outcome with 95% CIs	Survival to hospital discharge: OR = 1.10, 95% CI 0.70 to 1.70, p = 0.686; Return to spontaneous circulation: OR = 1.10, 95% CI 0.82 to 1.26, p = 0.979; Favorable neurologic outcomes: OR = 1.02, 95% CI 0.31 to 3.38, p = 0.979 1-year survival: OR = 1.38, 95% CI 0.95 to 2.02, p = 0.092, trend toward favoring chest compression	Survival to hospital discharge: I^2 = 34.4%, return to spontaneous circulation: OR = I^2 = 0.0%, neurologic outcomes: I^2 = 74.9%, 1 year survival: I^2 = 0.0%
Milito et al. 2010[36]	Yes	Pooled estimates using a random effects model	Standardized mean difference (SMD) with 95% CIs for most outcomes	Authors state that there was no significant difference in the proportion of patients cured after LigaSure compared to other excisional techniques; Operative time (mins): SMD = -0.859, 95% CI -1.142 to -0.575, p <0.01; Bleeding: SMD = 0.450. 95% CI 0.199 to 1.019, p = 0.056; Pain: SMD = -0.545, 95% CI -0.836 to -0.255, p <0.01 Stenosis/relapse: SMD = 0.344, 95% 0.136 to 0.868, p = 0.024	Authors indicate statistically significant heterogeneity for operation time and pain.
Murphy et al. 2010[37]	Yes	Random effects	WMD, OR, 95% CI	VAS pain scores at 48 hours postoperatively WMD 0.04, 95% CI 0.10 to 0.18, p = 0.46 Postoperative nausea and vomiting OR 1.52, 95% CI 1.07 to 2.14 Postoperative pruritus OR 0.43, 95% CI 0.19 to 0.98 Fatigue or sedation OR 0.82, 95% CI 0.31 to 2.21, p = 0.70	VAS pain scores at 48 hours postoperatively I^2 = 17.8% Postoperative nausea and vomiting I^2 = 0% Postoperative pruritus I^2 = 0% Postoperative fatigue or sedation I^2 = 33.6%

Reference	Performed Meta-analysis	Meta-analytic Approach[1]	Statistical Method(s) Used[2]	Meta-analytic Results	Results of Heterogeneity Testing
Myers et al. 2010[9]	No	Narratively summarized the findings of RCTs.	Not applicable	Narrative findings: No statistically significant differences were observed in pain relief between intraspinal techniques and conventional medical treatment (oral or subcutaneous administration of morphine); both treatments achieved similar pain relief.	Not reported
Pan et al. 2010[38]	Yes	Pooled estimates using DerSimonian and Laird random-effects model.	Odds ratio with 95% CIs	Mortality: OR = 0.88, 95% CI 0.63 to 1.21, p = 0.42; Recurrent myocardial infarction: OR = 0.80, 95% CI 0.56 to 1.13, p = 0.21; Repeat revascularization: OR = 0.56, 95% CI 0.44 to 0.72, p <0.001, favors drug eluting stents; Stent thrombosis rate: OR = 0.81, 95% CI 0.52 to 1.26, p = 0.35	I^2 <10% for all outcomes

Reference	Performed Meta-analysis	Meta-analytic Approach[1]	Statistical Method(s) Used[2]	Meta-analytic Results	Results of Heterogeneity Testing
Riemsa et al. 2010[8]	Yes	Pooled data in direct comparison using DerSimonian and Laird's random effects model and, when appropriate, in an indirect comparison using the method of Butcher et al.	Relative risk (RR) for dichotomous data; weighted mean difference (WMD) for continuous data, and hazard ratio (HR) for survival data, all with 95% CIs	Direct comparisons: Oral topotecan plus BSC vs. BSC alone: 1 study concluded that topotecan was more effective than BSC alone; IV topotecan vs. CAV: 1 study concluded that IV topotecan was at least as effective as CAV for symptom improvement; Oral vs. IV topotecan: Overall survival at median survival time: HR = 0.98, 95% CI, 0.77 to 1.25 Symptom control and toxicity: similar between oral and IV topotecan Indirect comparison; IV topotecan vs. CAV: Survival time: HR = 1.02, 95% CI 0.70 to 1.49 Response rate: RR = 1.29, 95% CI 0.67 to 2.50	Oral vs. IV topotecan: I^2 = 42% IV topotecan vs. CAV: I^2 = 42%
Sbruzzi et al. 2010[39]	Yes	Pooled estimates using a fixed effects model. Separate analysis done for studies comparing functional electrical stimulation to conventional exercise and control.	Weight mean difference (WMD) with 95% CIs from baseline to end of study.	Peak VO_2: WMD = -0.74, 95% CI -1.38 to -0.10; Distance of 6-min walk: WMD = 2.73, 95% CI -15.39 to 20.85; Muscle strength: WMD = -0.33, 95% CI -4.56 to 3.90	Peak VO_2: I^2 = 0.0%; distance of 6-min walk: I^2 = 0.0%; muscle strength: I^2 = 41.0%

Reference	Performed Meta-analysis	Meta-analytic Approach[1]	Statistical Method(s) Used[2]	Meta-analytic Results	Results of Heterogeneity Testing
Sgourakis et al. 2010[40]	Yes	Pooled estimates (with 95% CIs) using random or fixed effects model depending on the presence of heterogeneity.	Odds ratio (OR) or Relative Risk (RR), or Risk Difference (RD) with 95% CI for dichotomous data and weighted mean difference (WMD) with 95% CIs for continuous data.	Number of patients requiring re-intervention: OR = -1.0903, 95% CI -1.7197 to -0.4135, favors conventional stent; Overall 1-year survival: RD = -0.1628, 95% CI -0.2195 to -0.0952, favors locoregional modality treatments, no longer significant upon sensitivity analysis; No difference was observed between the use of the anti-reflux stents and conventional stents in relieving reflux. No difference in overall complications.	Re-intervention: I^2 = 82%; No evidence of significant between-study heterogeneity for other outcomes.
Simpson et al. 2010[41]	Yes	Random effects model for primary outcome of survival to discharge	ORs, 95% CI	Overall survival to hospital discharge OR 0.94, 95% CI 0.46 to 1.94 Survival to hospital discharge (<5 minutes) OR 0.70, 95% CI 0.24 to 2.08 Survival to hospital discharge (>5 minutes) OR, 1.35 95% CI 0.30 to 1.65	Overall survival to hospital discharge I^2 = 49% Survival to hospital discharge (<5 minutes) I^2 = 30% Survival to hospital discharge (>5 minutes) I^2 = 75%
Squizzato et al. 2010[12]	No	Narratively summarized study findings.	Data for qualitative variables were presented as incidence rates. Data from continuous variables were summarized using measures of central tendency and dispersion	Not reported	Not reported

Reference	Performed Meta-analysis	Meta-analytic Approach[1]	Statistical Method(s) Used[2]	Meta-analytic Results	Results of Heterogeneity Testing
Sunkara et al. 2010[42]	Yes	Fixed and random effects	RR, 95% CI	Ongoing pregnancy/live birth rate RR 1.15, 95% CI 1.01 to 1.30, p = 0.03 Blastocyst Survival Rate RR 1.05, 95% CI 0.96 to 1.14, p = 0.27 Clinical Pregnancy Rate RR 1.14, 95% CI 1.03 to 1.26, p = 0.01 Miscarriage Rate RR 0.82, 95% CI 0.43 to 1.58, p = 0.56	Ongoing pregnancy/live birth rate I^2 = 0% Blastocyst Survival Rate I^2 = 94.2% Clinical Pregnancy Rate I^2 = 42.6% Miscarriage Rate I^2 = 57.4%
Tamayo et al. 2010[7]	Yes	Random effects	RR, 95% CI, effect sizes such as number needed to treat (NNT) and number needed to harm (NNH)	Lithium vs. other MDRs Response RR 0.90, 0.81 to 1.00 Olanzapine vs. other MDRs remission RR 1.17, 1.06 to 1.30 Valproate vs. other MDRs had similar chance of response, but lower risk of discontinuation due to AEs Aripiprazole vs. other MDRs had similar chance of response and remission Risperidone vs. other MDRs had similar chance of response, similar risk of discontinuation due to AEs, higher risk of discontinuation due to any cause Comparisons between active compound and another MDR (lithium, valproate, haloperidol) no significant differences were found for response, remission, discontinuation due to AEs, lack of efficacy, or discontinuation due to any cause.	Lithium vs. other MDRs I^2 = 0% Olanzapine vs. other MDRs remission I^2 = 0% Haloperidol vs. other MDRs response I^2 = 62% Haloperidol vs. other MDRs remission I^2 = 51% Quetiapine vs. other MDRs response I^2 = 69% Quetiapine vs. other MDRs remission I^2 = 69%

Reference	Performed Meta-analysis	Meta-analytic Approach[1]	Statistical Method(s) Used[2]	Meta-analytic Results	Results of Heterogeneity Testing
Tamhane et al. 2010[43]	Yes	Fixed effects and random effects models. Since there was significant heterogeneity for some endpoints, random-effects models are preferentially reported for entire group although fixed effects gave similar results. Data on results were collected on an "intention-to-treat" basis.	Summary odds ratios (OR); 95% CI	30-day mortality OR 0.84, 95% CI 0.54 to 1.29, p = 0.42 Stroke OR 2.88, 95% CI 1.06 to 7.85, p = 0.04 TVR OR 0.92, 95% CI 0.57 to 1.49, p = .073 Post procedural MBG 3 OR 2.42, 95% CI 1.63 to 3.61, p<0.001 Post procedural TIMI 3 flow OR 1.41, 95% CI 1.10 to 1.81, p = 0.007 Electrocardiographic end point: OR 2.30, 95% CI 1.64 to 3.23, p <0.001	30-day mortality: I(2) = 0%, Stroke: heterogeneity = 0.97 Post procedural MBG 3 I(2) = 81.5% Post procedural TIMI 3 flow: I(2) = 22% Electrocardiographic end point: I(2) = 73.8%
Tang et al. 2010[4]	Yes	Pooled estimates using random or fixed effects model depending on the presence of heterogeneity.	Relative risk (RR) with 95% CIs	MTA vs. intermediate restorative material Success rate: RR = 0.62, 95% CI 0.34 to 1.16 MTA vs. amalgam Success rate: RR = 0.35, 95% CI 0.13 to 0.94, favors MTA MTA vs. gutta-percha Success rate (based on 1 study): RR = 0.08, 95% CI 0.01 to 0.57	MTA vs. intermediate restorative material Success rate: I^2 = 0.0% MTA vs. amalgam Success rate: I^2 = 0.0%
Testa et al. 2010[44]	Yes	Binary outcomes from individual studies were combined with both DerSimonian and Laird random effect model and fixed effect model, according to intention to treat analysis	OR, 95% CI, calculated the number needed to treat (NNT) to prevent a MAE as the inverse of absolute risk reduction (ARR)	Overall death OR 1.32, 95% CI 1.00 to 1.74, p = 0.05 (NNT = 55) TVR OR 1.86, 95% CI 1.33 to 2.61, p = 0.0003 (NNT = 16) TLR OR 1.77, 95% CI 1.27 to 2.48, P <0.0001 (NNT = 25)	Overall death I^2 = 0 TLR I^2 = 0% TVR I^2 = 59%

Reference	Performed Meta-analysis	Meta-analytic Approach[1]	Statistical Method(s) Used[2]	Meta-analytic Results	Results of Heterogeneity Testing
Valachis et al. 2010[45]	Yes	Pooled estimates using random or fixed effects model depending on the presence of heterogeneity.	Odds ratio (OR) with 95% CIs for each outcome	Overall survival: OR = 0.912, 95% CI 0.674 to 1.234, p = 0.550; No differences were observed for supraclavicular (OR = 1.415, 95% CIs 0.278 to 7.202, p = 0.560) or distant (OR = 0.740, 95% CI 0.506 to 1.082, p = 0.120) recurrence. Partial breast irradiation resulted in significantly higher risk of developing local and axillary recurrences.	Overall survival: heterogeneity not significant (p = 0.575) No statistically significant heterogeneity observed for any of the secondary outcomes.
Vasiliadis et al. 2010[15]	No	Narratively summarized study findings.	Not reported	Not reported	Not reported
Vermeulan et al. 2010[10]	No, authors indicated the presence of clinical heterogeneity.	Studies were analyzed using vote-counting methods.	Individual study effect size estimates calculated as mean difference (MD with 95% CI) for continuous data and risk difference (RD with 95% CIs) for dichotomous outcomes.	Reported results of individual studies: “Iodine was significantly superior to other antiseptic agents (such as silver sulfadiazine cream) and non-antiseptic dressings, but seemed inferior to a local antibiotic and, when combined with alcohol, to crude honey in reducing bacterial count and/or wound size.” Adverse effects did not occur more frequently with iodine.	Clinical heterogeneity was present.
Xie et al. 2010[46]	Yes	Carried out separate meta-analyses to estimate the pooled 1, 3, and 5 year survival rates of each treatment strategy within two subgroups of studies – those with Child-Pugh A patients and those with a mix of Child-Pugh A and B patients.	Exact likelihood approach based on binomial distribution	Not reported	Not reported

Reference	Performed Meta-analysis	Meta-analytic Approach[1]	Statistical Method(s) Used[2]	Meta-analytic Results	Results of Heterogeneity Testing
Yang et al. 2010[1]	Yes	Pooled estimates using random or fixed effects model depending on the presence of heterogeneity.	Risk ratio (RR) with 95% CI	Remission rate: RR = 1.13, 95% CI 0.92 to 1.38; Constipation: RR = 0.35, 95% CI 0.27 to 0.45; Nausea/vomiting: RR = 0.57, 95% CI 0.49 to 0.67; Vertigo: RR = 0.59, 95% CI 0.51 to 0.68 No meta-analysis performed for quality of life due to differences in how the outcome was defined.	No significant heterogeneity observed for remission rate, but heterogeneity was observed for adverse events
Agarwal et al. 2010[47]	Yes	Pooled estimates using random or fixed effects model depending on the presence of heterogeneity.	Odds ratio (OR), risk difference (RD), or standardized mean difference (SMD) depending on type of data, all with 95% CIs.	Short-term mortality: OR = 0.01, 95% CI -0.01 to 0.03, p = 0.35; Long-term mortality: OR = 0.02, 95% CI -0.05 to 0.09, p = 0.55; Risk of right bundle branch block (RBBB): OR = 56.3, 95% CI 11.6 to 273.9, p <0.001, increase in septal ablation; Pacemaker implantation: OR = 2.57, 95% CI 1.68 to 3.98, p <0.001, increase in septal; No statistically significant difference found in functional class, ventricular arrhythmia, re-interventions, and post-procedural mitral regurgitation.	Short-term mortality: I^2 = 0.0% long-term mortality: I^2 = 75%, post-intervention pacemaker: I^2 = 40.6%
Avouac et al. 2010[48]	Yes	Pooled estimates using a random-effects model.	Mean difference with 95% CIs	Joint lavage plus steroid injection vs. joint lavage alone Pain: Effect size (ES) = -0.82, 95% CI -2.47 to 0.82, not significant Physical function: ES = 0.09, 95% CI -0.28 to 0.45, not significant	Pain: I^2 = 97%; no I^2 reported for physical function
Chua et al. 2010[14]	No	Studies were analyzed using vote-counting methods	Not reported	Not reported	Not reported

Reference	Performed Meta-analysis	Meta-analytic Approach[1]	Statistical Method(s) Used[2]	Meta-analytic Results	Results of Heterogeneity Testing
Devaiah et al. 2010[49]	Yes	Authors reported using individual patient analyses in a one-stage comparison to perform meta-analysis. They pooled individual patient data for all positional restrictions.	Calculated Chi-square and p-values	Results of pooled analyses indicated that all restriction types showed no advantage over no restrictions. No restriction type was found to be statistically significant.	Not reported
Loveman et al. 2010[13]	No	Narratively summarized study findings.	Not reported	Not reported	Not reported
Valachis et al. 2010[50]	Yes	Pooled estimates using Peto method for log transformed hazard ratios or Mantel-Haenszel method for odds ratios. DerSimonian-Laird random effects model was used in cases of heterogeneity.	Hazard ratio (HR) for overall survival and time to progression; Odd ratios (OR) for clinical benefit, objective response, and adverse events, all with 95% CIs	Overall survival: HR = 1.047, 95% CI 0.688 to 1.592, p = 0.830 Time to tumor progression: HR = 0.994, 95% CI 0.691 to 1.431, p = 0.975 Objective response: OR = 1.044, 95% CI 0.828 to 1.315, p = 0.716 Clinical benefit: OR = 0.949, 95% CI 0.736 to 1.224, p = 0.687 Deaths: OR = 0.754. 95% CI 0.267 to 2.127, p = 0.594	No statistically significant heterogeneity was observed for primary outcomes

[1] Did systematic reviewer use a frequentist or Bayesian approach? Were direct or indirect analyses performed?
[2] Statistical method(s) used to derive summary effect size(s) (e.g., standardized mean difference, risk ratio, etc.)

6MWT	Six minute walk test
ADR	Adverse drug reaction
ARR	Absolute risk reduction
CGI	Clinical global impression
CI	Confidence interval
CMAI	Cohen Mansfield agitation inventory
CRQ	Chronic respiratory questionnaire
ES	Effect size
HR	Hazard ratio
ITT	Intention-to-treat
IV	Intravenous
LoE	Lack of efficacy
MD	Mean difference
MDR	Monotherapy drug regimen
MTA	Mineral trioxide aggregate
NNT	Number needed to treat
OHSS	Ovarian hyperstimulation syndrome
OR	Odds ratio
PP	Per protocol
QoL	Quality of life
RD	Risk difference
RFA	Radiofrequency ablation
SMD	Standardized mean difference
TIMI	Thrombolysis in Myocardial Infarction
TLR	Target lesion revascularization
TVR	Target vessel revascularization
VAT	Video-assisted thoracoscopic surgery
VO_2	Peak oxygen
WMD	Weighted mean difference

Appendix E

Table 6. Wording of conclusion

Reference	Authors' Conclusion Statement Reviewers	Direct Use of EQ/NI Terms	Strength of Evidence Rating	Conclusions Paired with Strength of Evidence or Internal Validity Rating of Evidence
Bodri et al. 2011[19]	"Currently available evidence suggests a similar effectiveness of GnRH antagonists and agonists in the context of oocyte donation. Owing to its increased safety potential, the GnRH antagonist protocol combined with GnRH agonist triggering could be advocated as the treatment of first choice for oocyte donors."	Yes	Not reported	Not reported
Groeneveld et al. 2011[11]	"Albumin displayed a more favorable safety profile than HES. Available evidence does not support the existence of consistent safety differences between HES solutions."	No	Not reported	Not reported
Hockenhull et al. 2011[20]	"The [drug eluting stents] in this analysis performed no better than [bare metal stents] on the outcomes of mortality, [acute myocardial infarction], or thrombosis. Drug eluting stents releasing sirolimus, paclitaxel, dexamethasone and zotarolimus reduce composite cardiac events. However, this reduction is due largely to reductions in repeat revascularization rates as there is no evidence of a significant effect on rates of death, myocardial infarction, or thrombosis. The increased cost of drug eluting stents and lack of evidence of their cost-effectiveness means that various health fund agencies are having to limit or regulate their used in relation to price premium."	No	Not reported	Not reported
Seitz et al. 2011[21]	"Currently there are relatively few studies of antidepressants for the treatment of agitation and psychosis in dementia. The SSRIs sertraline and citalopram were associated with a reduction in symptoms of agitation when compared to placebo in two studies. Both SSRIs and Trazodone appear to be tolerated reasonably well when compared to placebo, typical antipsychotics, and atypical antipsychotics. Future studies involving more subjects are required to determine if SSRIs, trazadone, or other antidepressants are safe and effective treatments for agitation and psychosis in dementia."	No	Not reported	Not reported
Beauchamp et al. 2010[5]	"Interval and continuous training modalities did not differ in their effect on measures of exercise capacity or HRQoL. Interval training may be considered as an alternative to continuous training in patients with varying degrees of COPD severity."	No	Not reported	Not reported
Davis et al. 2010[22]	"Arthroscopic and open acromioplasty have equivalent ultimate clinical outcomes, operative times, and low complication rates. However, arthroscopic acromioplasty results in faster return to work and fewer hospital inpatient days compared with the arthroscopic technique."	Yes	Not reported	Not reported

Reference	Authors' Conclusion Statement Reviewers	Direct Use of EQ/NI Terms	Strength of Evidence Rating	Conclusions Paired with Strength of Evidence or Internal Validity Rating of Evidence
Dibra et al. 2010[16]	"There was no difference in the hazard of death or recurrent myocardial infarction between patients treated with drug-eluting stents versus bare metal stents. Treatment with drug-eluting stents resulted in a significant reduction in the hazard of reintervention. The hazards of death, myocardial infarction, and stent thrombosis were not significantly different between [patient groups]. Use of drug eluting stents in patients with acute myocardial infarction is safe and markedly reduces the need for reintervention."	No	Not reported	Not reported
Dong et al. 2010[23]	"This meta-analysis shows that early administration of smGPIs is as effective as abciximab in the setting of PPCI for STEMI, without an increase in bleeding complications."	Yes	Not reported	Not reported
Dong et al. 2010[24]	"In STEMI patients scheduled for primary PCI, although early smGPIs treatment improved initial epicardial patency, no beneficial effect on post-procedural angiographic or 30-day clinical outcome was found. Thus, the current available data do not support the routine utilization of upstream smGPIs in STEMI patients treated with primary PCI."	No	Not reported	Not reported
Dubicka et al. 2010[18]	"There was no evidence of a statistically significant benefit of combined treatment over antidepressants for depressive symptoms, suicidality and global improvement after acute treatment or follow-up. There was a statistically significant advantage of combined treatment for impairment in the short-term. Adding CBT to antidepressants confers limited advantage for the treatment of an episode of depression in adolescents."	No	Not reported	Not reported
Fuentes et al. 2010[25]	"IFC as a supplement to another intervention seems to be more effective for reducing pain than a control treatment at discharge and more effective than a placebo treatment at the 3-month follow-up. However, it is unknown whether the analgesic effect of IFC is superior to that of the concomitant interventions. IFC alone was not significantly better than placebo or other therapy at discharge or follow-up. Results may be considered with caution due to the low number of studies that used IFC alone. In addition, the heterogeneity across studies and methodological limitations prevents conclusive statements regarding analgesic efficacy."	Yes	Not reported	Not reported

Reference	Authors' Conclusion Statement Reviewers	Direct Use of EQ/NI Terms	Strength of Evidence Rating	Conclusions Paired with Strength of Evidence or Internal Validity Rating of Evidence
Hong et al. 2010[26]	"Control of glycemia to normal (or near normal levels) in type II diabetes appears to be effective in reducing the incidence of major macrovascular events, but there were no significant differences of either the mortality from any cause or from cardiovascular death between the two glycemia-control strategies."	No	Not reported	Not reported
Hughes et al. 2010[27]	"Intravenous immunoglobulin started within two weeks from onset hastens recovery as much as plasma exchange. Adverse events were not significantly more frequent with either treatment but intravenous immunoglobulin is significantly much more likely to be completed than plasma exchange."	No	Not reported	Authors indicate that conclusions are based on moderate quality evidence.
Kalil et al. 2010[2]	"Our study does not demonstrate clinical superiority of linezolid vs. glycopeptides for the treatment of nosocomial pneumonia despite statistical power of 95%. Linezolid shows a two-fold increase in the risk of thrombocytopenia and gastrointestinal events. Vancomycin and teicoplanin are not associated with more renal dysfunction than linezolid."	No	Not reported	Not reported
Kesselheim et al. 2010[3]	"Although most RCTs were short-term evaluations, the available evidence does not suggest an association between loss of seizure control and generic substitution of at least three types of [antiepileptic drugs] AEDs. The observational study may be explained by factors such as undue concern from patients of physicians about the effectiveness of generic AEDs after a recent switch."	No	Not reported	Not reported
Krenke et al. 2010[17]	"There is no evidence that [video assisted thoracoscopic surgery] VATs is more effective than fibrinolytic treatment."	No	Not reported	Not reported
Lanitis et al. 2010[28]	"Our results suggest that in breast cancer patients, SSM was not significantly different from NSSM in terms of rates of local reoccurrence."	No	Not reported	Not reported
Lee et al. 2010[29]	Findings from meta-analyses suggest that "MMF and CYC are comparable in terms of efficacy; that MMF induction therapy tends to have a more favorable safety profile; and that MMF and AZA are comparable maintenance therapies in terms of efficacy and toxicity." Findings also suggest that low-dose IV CYC is more efficacious and safer than high-dose IV CYC for the treatment of severe lupus nephritis.	Yes	Not reported	Not reported
Liu et al. 2010[30]	"Radiofrequency ablation did not decrease the number of overall recurrences, and had no effect on survival when compared with surgical resection in a selected group of patients. For patients who do not have the opportunity or are unwilling to accept surgical treatment, RFA is an acceptable means of palliative care."	No	Not reported	Not reported

Reference	Authors' Conclusion Statement Reviewers	Direct Use of EQ/NI Terms	Strength of Evidence Rating	Conclusions Paired with Strength of Evidence or Internal Validity Rating of Evidence
Liu et al. 2010[31]	"In conclusion, laparoscopic appendectomy may have advantages over open appendectomy in hospital stay and postoperative recovery. There is no convincing difference in complications and death rates."	No	Not reported	Not reported
Liu et al. 2010[32]	"There is no significant difference in death rates of the treatment of HCC [hepatocellular carcinoma] in the groups of hepatectomy and RFA [Radio-frequency ablation], although recurrence of HCC may be lower in hepatectomy group. "RFA may have comparable results with surgical resection in patients in the therapeutic effect of ablation for the treatment of HCC, if recurrence of HCC after RFA could be timely detected and effectively treated."	Yes	Not reported	Not reported
Liu et al. 2010[33]	"This review showed no significant differences between HDR and LDR-ICBT when considering OS< DSS, RFS, local control rate, recurrence, metastasis and treatment related complications for women with cervical carcinoma. Due to some potential advantages of HDR-ICBT (rigid immobilization, outpatient treatment, patient convenience, accuracy of source and applicator positioning, individualized treatment) we recommend the use of HDR-ICBT for all clinical stages of cervix cancer."	No	Not reported	Not reported
Macedo et al. 2010[34]	"The available evidence suggests that graded activity in the short term and intermediate term is slightly more effective than a minimal intervention but not more effective than other forms of exercise for persistent low back pain. The limited evidence suggests that graded exposure is as effective as minimal treatment or graded activity for persistent low back pain."	No	Not reported	Not reported
Machado et al. 2010[35]	"We conclude that the 5.7% (OR = 1.27) represents the difference in efficacy (i.e. remission rates) between SNRIs and SSRIs in treating patients with MDD. However, such efficacy difference was not considered clinically relevant. This result was based on some clinical assumptions, such as ideal clinical conditions (i.e. 8 to 12 week treatment duration for MDD), research design (i.e. head to head trials), and outcome measure (i.e. remission rates). As well, the present meta-analysis examines a wide variety of pharmacological treatments, including the newest drugs in the classes of interest, duloxetine and escitalopram. Other meta-analyses confirmed the results found here."	No	Not reported	Not reported
Meier et al. 2010[6]	"Current evidence does not support the notion that chest compression first prior to defibrillation improves the outcome of patients in out-of-hospital cardiac arrest. It appears that both treatments are equivalent. However, subgroup analyses indicate that chest compression first may be beneficial for cardiac arrests with a prolonged response time."	Yes	Not reported	The authors report that their conclusions were based on no treatment effect with fairly narrow confidence intervals (precision) and with very little heterogeneity.

Reference	Authors' Conclusion Statement Reviewers	Direct Use of EQ/NI Terms	Strength of Evidence Rating	Conclusions Paired with Strength of Evidence or Internal Validity Rating of Evidence
Milito et al. 2010[36]	"There was no significant difference in the proportion of patients cured after LigaSure haemorrhoidectomy or other excisional techniques. Patients treated with LigaSure had a significantly shorter operative time, postoperative pain, wound healing and time off work than the patients submitted to excisional techniques. Postoperative bleeding did not significantly differ between the two groups." "Our meta-analysis shows that LigaSure haemorrhoidectomy is a fast procedure characterized by limited postoperative pain, short hospitalization, fast wound healing and convalescence."	No	Not reported	Not reported
Murphy et al. 2010[37]	"In summary, we have undertaken a meta-analysis to examine RCTs comparing IV PCA tramadol to IV PCA opioids in the postoperative period and found no difference in pain scores between the two regimens. We did find differences in the profile of side effects of the two drugs, which should be taken into consideration if both analgesic regimens are available for use. Further studies may be needed to assess the relative safety and analgesic efficacy, particularly for severe pain, of the two regimens."	No	Not reported	Not reported
Myers et al. 2010[9]	"Reports indicate intraspinal analgesia is equally or more effective than conventional medical management and often associated with fewer side effects." "Intraspinal techniques monitored by an interprofessional health care team should be included as part of a comprehensive cancer pain management program."	Yes	Not reported	Not reported
Pan et al. 2010[38]	Pooled analyses indicated no significant difference between bare metal and drug eluting stents for the following outcomes: 1-year mortality, recurrent myocardial infarction, and rate of stent thrombosis. Analysis did show that drug eluting stents resulted in significantly lower rate of revascularization than bare metal stents. The authors concluded that "the results show that [drug eluting stents] improved clinical outcomes in [patients with ST-segment elevation myocardial infarction] with a decreased need for repeat revascularization and no concerns for safety."	No	Not reported	Not reported
Riemsa et al. 2010[8]	"From the evidence discussed, it is evident that oral topotecan has similar efficacy to IV topotecan (direct comparison) and CAV (indirect comparison). There is no further evidence base of direct or possible indirect comparisons for other comparators than CAV or either oral or IV topotecan."	Yes	Not reported	Not reported
Sbruzzi et al. 2010[39]	"Treatment with [functional electrical stimulation] provides a similar gain in the distance of the 6-min walk test and in the muscle strength when compared to [conventional exercise, CA], but a small gain in the peak VO_2. Thus, FES may be an alternative in relation with CA for patients with [chronic heart failure]."	Yes	Not reported	Not reported

Reference	Authors' Conclusion Statement Reviewers	Direct Use of EQ/NI Terms	Strength of Evidence Rating	Conclusions Paired with Strength of Evidence or Internal Validity Rating of Evidence
Sgourakis et al. 2010[40]	"Conventional self-expanding stents and anti-reflux stents are equally effective. Although the risk difference for 1-year survival favored locoregional palliative treatment modalities, the latter were associated with a higher number of patients requiring reintervention."	Yes	Not reported	Not reported
Simpson et al. 2010[41]	"Delaying initial defibrillation to allow a short period of CPR in out-of-hospital cardiac arrest due to VF demonstrated no benefit over immediate defibrillation for survival to hospital discharge irrespective of response time. There is no evidence that CPR before defibrillation is harmful. Based on the existing evidence, EMS jurisdictions are justified continuing with current practice using either defibrillation strategy."	No	Not reported	Not reported
Squizzato et al. 2010[12]	"Antithrombotic drugs may play a role in the treatment of the acute phase RVO, at least in some patient's categories. However, given the complexity of this condition, a multidisciplinary approach, concomitantly including ophthalmologic and antithrombotic treatment strategies, should be assessed to improve the management of what we can call a still 'orphan' disease."	No	Not reported	Not reported
Sunkara et al. 2010[42]	"Slower developing blastocysts cryopreserved on Day 6 but at the same stage of development as those developing to the blastocyst stage on Day 5 have similar clinical pregnancy and ongoing pregnancy/live birth rates following frozen-thawed blastocyst transfers."	Yes	Not reported	Not reported
Tamayo et al. 2010[7]	"Although there are patients who are unresponsive to acute treatment with monotherapy, these results suggest that MDRs should be considered as a first therapeutic option for the treatment of nonrefractory manic episodes. This approach may result in the reduction of direct costs of medications, the number and magnitude of AEs and may improve treatment adherence and patient compliance. For depressive episodes, the new data with SGAs (quetiapine and olanzapine) suggest that these MDR, especially quetiapine, are efficacious and well tolerated."	No	Not reported	Not reported
Tamhane et al. 2010[43]	"Thrombectomy devices appear to improve markers of myocardial perfusion in patients undergoing primary PCI, with no difference in overall 30-day mortality but an increased likelihood of stroke. The clinical benefits of thrombectomy appear to be influenced by the device type with a trend towards survival benefit with MAT and worsening outcome with mechanical devices."	No	Not reported	Not reported
Tang et al. 2010[4]	"The results of this systematic review indicate that the outcomes of MTA as root-end filling are significantly better than amalgam and pure gutta-percha, but are similar to IRM."	Yes	Not reported	Not reported

Reference	Authors' Conclusion Statement Reviewers	Direct Use of EQ/NI Terms	Strength of Evidence Rating	Conclusions Paired with Strength of Evidence or Internal Validity Rating of Evidence
Testa et al. 2010[44]	"Use of DES in SVG substantially reduces both TVR and TLR. These data also demonstrates that using DES in SVG is safe and contradicts previous reports of potential risks."	No	Not reported	Not reported
Valachis et al. 2010[45]	"Partial breast irradiation [PBI] does not seem to jeopardize survival and may be used as an alternative to whole breast-radiation [WBRT]. Risk for death is comparable among both treatment modalities."	Yes	Not reported	Not reported
Vasiliadis et al. 2010[15]	"Complications rates were comparable between interventions except from an increase rate of graft hypertrophies after ACI with periosteum. There is insufficient data to say whether ACI is superior to other treatment strategies. More high quality studies and harmonization in the reported outcomes are needed before specific suggestions for practice can be made."	Yes	Not reported	Not reported
Vermeulan et al. 2010[10]	"Based on the available evidence from clinical trials, iodine is an effective antiseptic agent that shows neither the purported harmful effects nor a delay of wound-healing process, particularly in chronic and burn wounds. The antiseptic effect of iodine is not inferior to that of other (antiseptic) agents and does not impair wound healing."	Yes	Not reported	Not reported
Xie et al. 2010[46]	"Continuing doubts on this issue can only be resolved by a substantial RCT. Meanwhile, for early stage HCC patients classified as Child-Pugh A, who despite a possibly higher recurrence rate, prefer the less invasive PRFA to open surgery with its attendant risks, there is sufficient evidence to justify such a choice. For those classified as Child-Pugh (B) it is possible that overall survival is equally good with PRFA, but the evidence is less certain."	No	Not reported	Not reported
Yang et al. 2010[1]	"Our study showed again that both transdermal fentanyl and oral morphine had the same efficacy in the treatment of moderate-severe cancer pain in Chinese population, but the former might have less adverse effects and better quality of life."	Yes	Not reported	Not reported
Agarwal et al. 2010[47]	"[Septal ablation] does seem to show promise in treatment of [hypertrophic obstructive cardiomyopathy] owing to similar mortality rates as well as functional status compared with septal myectomy; however, the caveat is increased conduction abnormalities and a higher post-intervention [left ventricular outflow tract gradient."	Yes	Not reported	Not reported
Avouac et al. 2010[48]	"This meta-analysis of RCTs investigating joint lavage for knee OA suggests that at three months, (1) joint lavage alone [vs. placebo] does not provide significant improvement in pain or function and (2) the combination of joint lavage and IA steroid injection is no more efficacious that lavage alone."	No	Not reported	Not reported

Reference	Authors' Conclusion Statement Reviewers	Direct Use of EQ/NI Terms	Strength of Evidence Rating	Conclusions Paired with Strength of Evidence or Internal Validity Rating of Evidence
Chua et al. 2010[14]	"This systematic review presents the role of TACE in the setting of a resectable HCC. Current evidence indicates that there appears to be no DFS advantage despite its safety and feasibility. A well-designed prospective multi-institutional RCT, with a clearly defined protocol for concealed allocation, eligibility criteria, TACE intervention regimen and endpoints will be potentially meaningful."	No	Not reported	Not reported
Devaiah et al. 2010[49]	"The restrictions examined in controlled trials did not differ significantly in clinical outcomes, which suggest that restrictions do not appear to significantly affect the efficacy of benign paroxysmal positional vertigo."	No	Not reported	Not reported
Loveman et al. 2010[13]	"Topotecan appears to improve survival in people with SCLC when compared to BSC alone, is as effective as CAV but less effective than amrubicin in terms of response rates, and shows comparable rates of treatment toxicities and adverse events with CAV and amrubicin based on the data available. Oral and IV topotecan were not seen to be different from one another on survival or measures of response."	Yes	Not reported	Not reported
Valachis et al. 2010[50]	"Fulvestrant was similar to other hormonal agents with respect to efficacy measures, with good tolerability profile."	Yes	Not reported	Not reported

AED Antiepileptic drugs
BMS Bare metal stents
BSC Best supportive care
CAV Cyclophosphamide, Adriamycin, and vincristine
CBT Cognitive behavioral therapy
COPD Chronic obstructive pulmonary disease
CYC Cyclophosphamide
DES Drug eluting stents
FES Functional electrical stimulation
GnRH Gonadotropin-releasing hormone
HCC Hepatocellular carcinoma
HDR High-dose rate
HES Hydroxyethyl starch
HRQoL Health-related quality of life
ICBT Intracavity brachytherapy
IFC Interferential current therapy
IV Intravenous
LDR Low-dose rate
MDD Major depressive disorder
MDR Monotherapy drug regimen
MMF Mycophenolate mofetil
MTA Mineral trioxide aggregate
OS Overall survival
PBI Partial breast irradiation
PCA Patient-controlled analgesia
PCI Percutaneous intervention
PRFA Percutaneous radiofrequency ablation
RCT Randomized controlled trial
RFA Radiofrequency ablation
RFS Relapse-free survival
RVO Retinal vein occlusion
SCLC Small cell lung cancer
smGPIs small-molecule glycoprotein IIb/IIIa receptor inhibitors
SNRI Serotonin norepinephrine reuptake inhibitor
SSRI Selective serotonin reuptake inhibitor
STEMI ST-segment elevation myocardial infarction
SVG Saphenous vein graft
TACE Transarterial chemoembolization
TLR Target lesion revascularization
TVR Target vessel revascularization
VAT Video-assisted thoracoscopic surgery
VO_2 Peak oxygen
WBRT Whole breast irradiation