Grading the Strength of a Body of Evidence When Assessing Health Care Interventions – AHRQ and the Effective Health Care Program: An Update

Draft Report

Submitted to:

Agency for Healthcare Research and Quality
540 Gaither Road
Rockville, Maryland 20850

Submitted by:

Contract No. XXX

Project No. XXX

Date

Methods Research Report

Number xx

Grading the Strength of a Body of Evidence When Assessing Health Care Interventions – AHRQ and the Effective Health Care Program: An Update

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.

Prepared by:

Investigators:

AHRQ Publication No. xx-EHCxxx

<Month Year>

Suggested Citation:

<Authors> Grading the Strength of a Body of Evidence When Assessing Health Care Interventions – AHRQ and the Effective Health Care Program: An Update <EPC> Evidence-based Practice Center, Contract No. <Contract Number>. (AHRQ to provide the remainder)

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

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. 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 bring the broadest range of experts into the development of evidence reports and health technology assessments, AHRQ encourages the EPCs to form partnerships and enter into collaborations with other medical and research organizations. The EPCs work with these partner organizations to ensure that the evidence reports and technology assessments they produce will become building blocks for health care quality improvement projects throughout the Nation. The reports undergo peer review prior to their release.

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.

We welcome comments on this evidence report. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by e-mail to epc@ahrq.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
Evidence-based Practice Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Task Order Officer
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Acknowledgments

This research was funded through the Agency for Healthcare Research and Quality Effective Healthcare Program The opinions expressed here are those of the authors and do not necessarily represent the views of the Agency for Healthcare Research and Quality, the Department of Health and Human Services, or the Department of Veterans Affairs.

Peer Reviewers

<Name>

<Place>

<City>, <ST>

<Name>

<Place>

<City>, <ST>

Grading the Strength of a Body of Evidence When Assessing Health Care Interventions – AHRQ and the Effective Health Care Program: An Update

Structured Abstract

Objective. To revise guidance on grading strength of evidence for systematic reviews and similar products from the Evidence-based Practice Center (EPC) program of the US Agency for Healthcare Research and Quality.

Study Design and Setting. Authors reviewed authoritative systems for grading strength of evidence, revised domains and methods for grading bodies of evidence in systematic reviews through discussions based on their experience with the current system, methods expertise, and discussions with representatives of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group.

Results. The EPC approach is conceptually similar to the GRADE system of evidence rating. It requires assessment of five domains: study limitations (risk of bias), consistency, directness, precision, and reporting bias (publication, outcome and selective analysis reporting bias). Additional domains to be used when appropriate include dose-response association, presence of confounders that would diminish an observed effect, and strength (magnitude) of association. Strength of evidence receives a single grade: high, moderate, low, or insufficient. We give definitions, examples, mechanisms for scoring domains, and an approach for assigning strength of evidence.

Conclusion. EPCs should grade strength of evidence separately for each major outcome and each major comparison. We will continue to work with the GRADE group to address ongoing challenges in assessing the strength of evidence.

Contents

Figure

Tables

Appendixes

Introduction

Systematic reviews are essential tools for summarizing information to help users make well-informed decisions about health care options.[1] The Evidence-based Practice Center (EPC) program, supported by the US Agency for Healthcare Research and Quality (AHRQ), produces substantial numbers of such reviews, including those that explicitly compare two or more clinical interventions (sometimes termed comparative effectiveness reviews). These reports summarize, accurately and transparently, a body of literature; the primary goal is to help clinicians, policymakers, and patients make well-informed decisions about health care. Reviews should provide clearly explained, well-reasoned judgments about the strength of the evidence that underlies conclusions to enable decisionmakers to use them effectively.[2]

Beginning in 2007, AHRQ supported a cross-EPC set of work groups to develop guidance on major elements of designing, conducting, and reporting systematic reviews.[3] Together the materials form the EPC Methods Guide for Effectiveness and Comparative Effectiveness Reviews;[4] one chapter focused on grading the strength of evidence.[5] This paper reports updated findings and recommendations of a cross-EPC work group based on 5 years of experience in applying previous guidance. The guidance is developed for systematic reviews of drugs, devices and other preventive and therapeutic intervention. It does not address particular issues for reviews of medical tests, disease epidemiology, and broader health services research.

EPC authors prepare reports that many decisionmakers use, but EPCs do not themselves develop recommendations. Separating those who grade strength of evidence from the activities of various decisionmakers (e.g., patients, caregivers, clinicians, guideline developers, policymakers, and consumer groups) led us to develop guidance that differs in some ways from other rating systems that are designed to be used more directly by specific decisionmakers. In particular, we limit our grading strength of evidence approach to individual outcomes; we do not develop more global summary judgments of the relative benefits and harms of treatment comparisons.

The EPC's strength of evidence approach was based in large measure on the GRADE (Grading of Recommendations Assessment, Development and Evaluation) working group approach (which refers to the "quality" of evidence).[6-8] Although a wide variety of grading systems has been available for some time;[9] the GRADE system for assessing the quality of evidence, based on eight domains, has been widely used. We have continued communication with the GRADE working group, so this update includes insights and expertise gained from their direct input and the GRADE guidance series.[10-22] We will continue to explore and address particular challenges in applying the GRADE principles to EPC systematic reviews with the GRADE working group. This paper presents an update of the original EPC approach[23] and should be considered current guidance for EPCs. We briefly explore the rationale for grading strength of evidence, define domains of concern for evidence strength, and describe our recommended grading system for systematic reviews. Because this field is rapidly evolving, future revisions are anticipated and will reflect our increasing understanding and experience with the methodology.

Rationale and Approach

A systematic approach to making judgments about the strength of a body of evidence is needed to inform the decisions of individual clinicians and patients and to facilitate the work of organizations that develop practice guidelines or make coverage decisions. Assessment of the strength of evidence relies heavily on the assessment of the risk of bias of the individual studies included in the body of evidence. In addition, the grade of the overall body of evidence includes assessments across the body of evidence for several domains.

Evidence hierarchies are not equivalent to strength of evidence systems. Evidence hierarchies categorize our confidence in a causal inference in principal by focusing on only select elements of study design, such as randomization. By contrast, the more commonly used strength of evidence systems consider other elements of study design implementation that may reduce the risk of bias, as well as factors that may increase or decrease confidence when looking across all studies within the body of evidence, such as directness (or indirectness) of evidence and comparisons, consistency of the evidence and precision of the estimates. By including these additional components in grading the strength of evidence, we give decisionmakers a more comprehensive and fair evaluation of the evidence than could ever be done with simple study hierarchies.

The aims of this work are twofold: (1) to ensure appropriate consistency and transparency in the methods that different EPCs use to grade the strength of evidence and (2) to facilitate users' interpretations of those grades for their use in guideline development or other decisionmaking tasks. Attaining these goals rests in part on uniformity and predictability in the domains that EPCs use in this effort. Although no single approach for reporting results and grading the related strength of evidence is likely to suit all users, documentation and a consistent approach in reporting of the most important summary information about a body of literature —the general concept of transparency—will make reviews more useful to a broader range of potential audiences that AHRQ's work is intended to reach.

Figure 1 presents the major steps in conducting a strength of evidence assessment. Some decisions must be made *a priori*, and documented during the protocol development stage. According to the decision rules and procedures documented in the protocol, the EPC will assess individual domain scores and an overall strength of evidence grade.

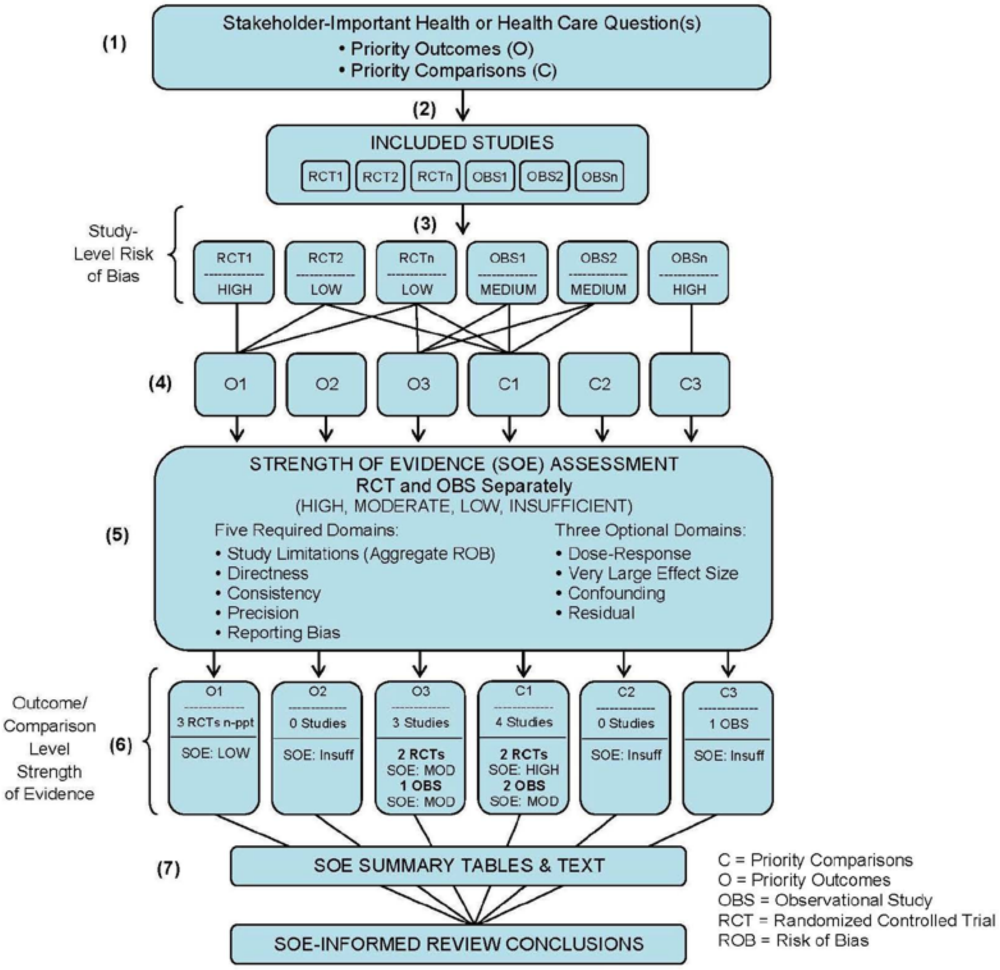
A priori Determinations

Selection of Outcomes

EPCs will not likely grade all outcomes for all treatment comparisons in their review protocols or key questions. Because assessing strength of evidence can be labor intensive, especially when the combinations of comparisons and outcomes are numerous, EPCs may restrict this step to outcomes of major salience to the end users of the review. We note that this decision contrasts with the Institute of Medicine recommendation in favor of assessing each outcome for strength of evidence.[24]

We recommend that EPC authors identify *a priori* the major outcomes they intend to grade in the review protocol and specify these core elements in the analytic framework. Also, we recommend that major outcomes include both benefits and harms. Determining which outcomes and comparisons are most important to decisionmakers in clinical practice and health policy depends heavily on the key questions and their specified outcomes or comparisons, the clinical or policy context, and the purpose of the report. EPCs can make these choices considering the input of key informants, including patients, during the topic refinement phase of the project (Whitlock, 2011) and subsequently through input from Technical Expert Panel (TEP) members. The final choices should reflect the scope of the review, the needs that key informants, TEP members and other end users express (as reflected in the protocol and final key questions), and the reliability, validity, and usefulness of the outcomes under consideration.

Figure 1. Fictional Example Illustrating Major Steps in a Systematic Review Related to Rating of Strength of Evidence (SOE)



© 2012, Kaiser Permanente, Center for Health Research, Evelyn P. Whitlock, MD, MPH

Major steps in Systematic Review Related to Rating of Strength of Evidence

1) Define clinical or policy questions, priority outcomes, and comparisons that are important to stakeholders (patients, clinicians, others) within a relevant, sensible overall scope for systematic review.
(2) Define search strategies and inclusion/exclusion criteria to locate and include all available studies relevant to the review.
(3) Conduct study design-specific risk of bias (ROB) assessment for included study, documenting whether overall study ROB assessment applies to all important outcomes and comparisons or not.
(4) Aggregate all studies that report each priority outcome or comparison into separate groups for strength of evidence (SOE) assessment.
(5) Within the group of studies reporting each priority outcome or comparison, conduct SOE assessment within study design (RCTs and observational studies separately) based on five required domains (and three optional domains when relevant).
(6) Combine study-design-specific SOE assessments into one overall SOE rating for each outcome and comparison.
(7) For each outcome and comparison, clearly report SOE domain assessments (study-design-specific and overall) with other relevant summary information in SOE Summary Tables and reflect these appropriately in report text, executive summary, and review conclusions.

Ideally, outcomes that EPC authors elect to grade will be patient-centered. The Patient-Centered Outcomes Research Institute (PCORI) has defined patient-centered outcomes as those that "*people notice and care about.*"[25] They can also be considered to reflect "*an event that is perceptible to the patient and is of sufficient value that changing its frequency would be of value* *to the patient.*"[26] Patient-centered care has been defined as "providing care that is respectful of and responsive to individual patient preferences, needs, and values and ensuring that patient values guide all clinical decisions."[27] Other clinically important health outcomes may include reductions in mortality or disease severity and improvements in health-related quality of life (patient-reported outcomes); they may also involve known or potential harms such as occurrences of serious and troubling adverse events and inconveniences.

The analytic framework can help in distinguishing between these patient centered, clinically important outcomes from intermediate outcomes. In rare cases, the EPC may decide to grade intermediate outcomes that have clear and strong associations with health outcomes or that are, in and of themselves, important to the target population are preferred over those without such links. Intermediate outcomes may include blood pressure control, cholesterol levels, adherence to treatment, or knowledge. Systematic reviews can be broad in scope, encompassing multiple patient populations, interventions, and outcomes. EPCs are *not* expected to grade every possible comparison for every outcome. Rather, reviewers should specify their priorities in the review protocol for those combinations (patients-interventions-outcomes) that are likely to be of greatest interest to most users of the report.

Selection of Studies

EPCs establish, up front, which studies will be eligible to answer the review questions. These criteria may be determined by the scope of the study, but may also consider the study design. In some cases, the EPC may determine that, given the body of literature or the question being asked, some study design characteristics would be so flawed that they could not contribute meaningfully to the body of evidence and thus should be excluded from the beginning. In these cases the EPC should establish *a priori* criteria (in the review protocol) to identify studies with particular design elements that would constitute an unacceptably high risk of bias.[28] For instance, such studies may have very high attrition or high differential attrition or studies may use invalid or unreliable measures for a major outcome. The rationale for excluding these studies from the review must be clearly stated *a priori*. When not explicitly excluded *a priori*, EPCs may, after reviewing the entire body of literature conduct an analysis with and without these problematic studies (such as with a sensitivity analysis), and consider which results are most valid and informative.

Decision Rules for Assessing the Overall Strength of Evidence

EPCs should decide *a priori* (to the extent possible) how they will incorporate each domain into an overall strength of evidence grade and how they will ensure the accuracy and consistency of evidence ratings. They should develop an explicit procedure for ensuring a high degree of inter-rater reliability for rating individual domains. This assumes that at least two reviewers who have received training specific to the concerns of the review will rate each domain, with recourse to a third, senior rater in instances of important disagreement. EPCs should pay close attention to the extent of disagreement, because recent empirical work documents that inter-rater reliability for domain scoring can be problematic when studies have markedly different strengths and weaknesses, use different or incompatible outcome measures, or do not report all their findings clearly.[29]

They should take specific steps to promote reliability and transparency when incorporating domains into an overall grade. Initially, they should be explicit about whether the evidence grade will be determined by an algorithmic point system for combining ratings of the domains, by a qualitative consideration, or by some combination of these approaches (expanded on further in later steps). In contrast to scoring domains, which may be done by more junior staff, strength of evidence grading should be done by senior reviewers. EPCs should use at least two senior reviewers with clinical or methods expertise and invoke a third, experienced author in cases of significant disagreement.

Assessing Strength of Evidence Domains

EPCs consistently assess a set of agreed upon (required) domains when grading the strength of evidence for each major outcome and comparison (Table 1). Four of these domains are the same as in those required in the original guidance: study limitations (previously named risk of bias), directness, consistency, precision, and reporting bias. The fifth domain, reporting bias (previously an “additional” domain, limited to publication bias, now includes outcome reporting bias) is a “required” domain and should be assessed when there is a high or moderate strength of evidence based on the first four domains. A second set of “additional” domains are most relevant to observational study bodies of evidence (Table 3).

In order to score the initial four required domains, EPCs should first identify the studies that address the outcomes of interest. When appropriate to score the fifth domain of reporting bias, the EPC may also need to identify studies that measured but did not publish or report on the outcome because of the direction of effect or lack of effect. Further information on this can be found in another Methods guide paper on Reporting bias (in process).

For each outcome and comparison of interest, EPCs should develop domain scores and strength of evidence grades *separately* for RCT evidence and observational study evidence when both contributed to evidence synthesis. Considerations when combining these separate bodies of evidence into one final strength of evidence grade can be found in a later section.

EPCs should have two or more reviewers with the appropriate clinical and methods expertise separately assess each required domain (or each optional domain, as relevant) for each major outcome (whether benefit or harm) and comparison. Those reviewers should resolve any differences in scores by either consensus discussion or adjudication by an additional expert reviewer.

The set of five “required domains” comprises the main constructs that EPCs should use for all major outcomes and comparison(s) of interest. As defined in Table 1, these represent related but separate concepts, and each is scored independently, although considering other domains using an appropriate scale. In some cases, concerns in the body of evidence may be attributable to more than one domain. When this happens, the EPC may decide where to attribute the concern and note it clearly. Each of these domains can individually and as a group, decrease the overall strength of the body of evidence. We discuss each of the five required domains (i.e., study limitations, directness, consistency, precision, and reporting bias) in more detail below.

Table 1. Required domains and their definitions

Domain	Definition and Elements	Score and Application
Study Limitations	Study limitations is the degree to which the included studies for a given outcome have a high likelihood of adequate protection against bias (i.e., good internal validity), assessed through two main elements: • Study design (e.g., RCTs or observational studies) • Aggregate risk of bias of the studies under consideration, assessed separately for RCTs and observational studies. Information for this determination comes from rating of risk of bias (high, medium, low) for individual studies.	Score one of three levels of aggregate study limitations: • **Low** level of study limitations • **Medium** level of study limitations • **High** level of study limitations
Directness	Directness relates to (a) whether evidence links interventions directly to health outcomes of specific importance for the review, and (b) for comparative studies, whether the comparisons have been done in head-to-head studies. The EPC should specify the comparison and outcome for which the SOE grade applies. Evidence may be indirect in several situations such as: • Data are available on only intermediate outcomes (such as laboratory tests) when the review is focused on clinical health outcomes • Data are available only for proxy respondents (e.g., obtained from family members or nurses) instead of directly from patients for situations in which patients self-report can be thought capable and more reliable, even if, at the time of the review, self-report evidence is graded as insufficient. • Data come from two or more bodies of evidence to compare interventions A and B -- e.g., studies of A vs. placebo and B vs. placebo, or studies of A vs. C and B vs. C but not A vs. B. Indirectness always implies that more than one body of evidence is required to link interventions to the most important health outcomes.	Score dichotomously as one of two levels • **Direct** • **Indirect** If the domain score is indirect, EPCs should specify what type of indirectness accounts for the rating
Consistency	Consistency is the degree to which included studies appear to have the same direction of effect or the same magnitude of effect. This can be assessed through two main elements: • Direction of effect: Effect sizes have the same sign (that is, are on the same side of “no effect” or a “minimum important difference”) • Magnitude of effect: The range of effect sizes is similar. When a meta-analysis is conducted, this may consider the overlap of confidence intervals. The importance of direction versus magnitude of effect will depend on the key question and EPC author	Score one of three levels of consistency: • **Consistent** (i.e., no inconsistency) • **Inconsistent** • **Unknown** (e.g., single study) Single-study evidence bases (including mega-trials) cannot be judged with respect to consistency. In that instance, use “Consistency unknown (single study).”

	judgments, but the EPC should make any threshold of “minimum important difference” explicit and how this determination was made.	

Table 1. Required domains and their definitions (continued)

Domain	Definition and Elements	Score and Application
Precision	Precision is the degree of certainty surrounding an effect estimate with respect to a given outcome (i.e., for each outcome separately), based on the sufficiency of sample size and number of events. In some cases, the actual precision (as measured in meta-analyses by the confidence interval) may also incorporate elements related to consistency. EPCs should clearly delineate whether uncertainty about an effect estimate or direction of effect is due to inconsistency or imprecision. A body of evidence will be imprecise if the optimal information size (OIS) is not met. If the OIS is met and the EPC performed a meta-analysis for the outcome, precision may also consider whether the confidence interval crossed a threshold for a “minimum important difference.” If a meta-analysis is infeasible or inappropriate, the EPC may consider the narrowness of the range of effect size estimates in the evidence base.	Score one of three levels of precision: • **Precise** • **Imprecise** • **Unknown** A precise estimate is one that would allow users to reach a clinically useful conclusion. An imprecise estimate is one for which the effect estimates in the evidence is wide enough to include clinically distinct conclusions. For example, results may be statistically compatible with both clinically important superiority and inferiority (i.e., the direction of effect is unknown), a circumstance that will preclude a valid conclusion. Precision is unknown when the precision of the evidence base cannot be determined (e.g., when studies do not report measures of dispersion for effect estimates).
Reporting Bias	Reporting bias results from selection of publication or reporting of research findings based on their direction or magnitude of effect. It includes: • *study publication bias*, i.e., nonreporting of results that are not “newsworthy” (the file drawer phenomenon), • *selective outcome reporting bias,* i.e., nonreporting (or incomplete reporting) of planned outcomes or reporting of unplanned outcomes, and • *selective analysis reporting bias*, i.e., reporting of the most favourable analyses conducted for a given outcome. Reporting bias is extremely difficult to detect. Registration and posting of protocols can help detect reporting bias for RCT evidence, but the effort may only be worthwhile when there is sufficient evidence for a potential strength of evidence grade of high or moderate. For observational study evidence, reporting bias is even more difficult to determine and methods for detection are uncertain at this time. Further recommendations on approaches may be found in another paper in progress and observational studies may be scored but it is not required.	Score one of two levels of reporting bias: • **Suspected** • **Undetected** Suspected reporting bias may include: a substantial difference in the pooled fixed effect estimate between small and large studies, such that small study effect reflects an exaggerated benefit or harm, or a qualitative assessment of the risk based on reviewers’ consensual judgment of the likely impact of reporting bias on the included evidence. Undetected reporting bias includes all alternative scenarios.

Study Limitations Domain

The study limitations domain, based on the design and conduct of the available studies, is an essential component of strength of evidence; rating this domain is the starting place for grading the strength of the overall body of evidence. The overall rating for the study limitations domain is a judgment of the limitations due to risk of bias in all of the individual studies, aggregated separately for RCTs and observational studies. This reflects the author's assessment of the ability of the evidence, given the design and conduct of individual studies, to accurately estimate effect the truth without bias (nonrandom error).

EPCs derive the overall study limitations domain score for an evidence base from their assessment of the risk of bias for each individual study;[28] with each study rated low, medium, or high risk of bias.

RCTs will generally be assessed to have a low risk of bias (a score of low on study limitations); this rating typically correlates with a grade of high strength of evidence, but such an assessment may be changed after evaluation of other domains. Evidence based on observational studies is generally assumed to have a higher risk of bias, which would correlate with a lower strength of evidence, but EPCs may well decide that, after actually assessing study limitations and evaluating other domains, the overall strength of evidence of a body of observational studies can be graded moderate (although rarely high).

EPCs may act on the judgment that, for certain outcomes such as harms, observational studies have less risk of bias than do RCTs or that the available RCTs have a substantial risk of bias. In such instances, the EPC may move up the initial grade for strength of evidence based on observational studies to moderate or move down the initial rating based on RCTs to moderate or high.

If the evidence for a given outcome or comparison of interest comes from a single or a small number of high risk-of-bias studies (negating the value of other domain ratings), and the EPC has determined that this small body of evidence is insufficient to draw any conclusions, EPCs may choose not to complete other domain scores.

If studies included in a body of evidence differ substantially in risk of bias, based on study design, study conduct, or both, EPCs may consider whether including high risk-of-bias studies will obscure the findings from the studies rated either low or medium risk of bias. If that is their conclusion, with proper documentation, EPCs may elect to give greater weight to the latter two sets of studies or, in fact, to limit their final synthesis to the studies with a lower risk of bias. For example, observational studies typically have higher risk of bias and may downgrade the strength of evidence assessment for a set of studies addressing an important outcome that also consists of many RCTs. Reviewers may reasonably focus first on studies with low or moderate risk of bias in their initial grading of summary of evidence across other required domains. They may do this in formal meta-analyses involving only studies of low or medium risk of bias, although they may consider conducting sensitivity analyses involving the less desirable studies. When quantitative analysis is not possible and results rest on qualitative analysis, EPCs should evaluate how consistent findings from studies with high risk of bias are with findings from the other, more desirable studies. If EPCs elect not to include studies that are individually rated high risk of bias, with clear communication of methods used and rationale, they may omit them from strength of evidence grading and from tables and text, in order to focus on data from better studies. Such studies are, however, counted as part of the overall evidence base and are included in references.

Although the study limitations domain is important, it is typically only the starting point in most instances. The EPC will generally go on to incorporate assessments of the other required domains in addition to study limitations to determine an overall strength of evidence grade.

For rating the study limitations domain, EPCs can assign one of three levels of aggregate risk of study limitations (low, medium, or high level of study limitations) to a body of evidence for a particular outcome or comparison. Because of unique issues in study designs, we recommend scoring the body of evidence from randomized controlled studies separately from the body of evidence from non-randomized studies, and combining the two bodies of evidence as a later stage, as described below.

Directness

Directness of evidence expresses the closeness of the available evidence to measuring the ultimate health concern. Directness is scored as direct or indirect. Assessing directness has two parts: directness of outcomes and directness of comparisons. Applicability of the evidence is considered explicitly but separately in EPC systematic reviews.[30] This practice contrasts with the GRADE approach which considers it as part of directness.[8]

The situations for which EPCs might score important outcomes as indirect will be limited to bodies of evidence with indirect comparisons or, in some cases to situations in which intermediate outcomes or proxy respondents are used to measure an important outcome, as described below. EPCs should discuss any issues of directness in the review, particularly links between intermediate and ultimate health outcomes in their synthesis of the evidence.

Directness of Outcomes

The focus of the review itself determines what type of evidence should be considered "directness." As described earlier, the EPC should identify *a priori* which outcomes will be graded. In most cases those outcomes should be patient important or clinically important outcomes, although there may be rare cases where intermediate outcomes are considered important to be graded. In either case, if there is no direct evidence on the named outcome, some reviewers may then consider use of surrogate markers or intermediate outcomes. This is rarely done in EPC reports, but if done, such evidence may be considered indirect. Other examples where a body of evidence may be considered indirect because investigators have used a proxy to stand in for or to measure the outcome of interest. An example of a proxy measure is when a surrogate (e.g., family member or nurse) is used to obtain patients' perceptions of their states of health, such as quality of life or measures of symptom improvement.

Directness of Comparisons

Comparisons are considered direct when the evidence derives from studies that compare interventions specifically with each other; that is, the studies are head-to-head comparisons. For the directness domain, this is the most desirable situation. In many circumstances, such head-to-head evidence is not available. When studies compare an intervention group with a placebo control or "usual care" (or similar) group but not specifically with the comparator intervention of interest, then the evidence is indirect. EPC can use separate bodies of evidence (e.g., A vs. placebo, B vs. placebo, and C vs. placebo) to estimate indirectly the comparative effectiveness of the interventions. As a case in point: in a review of off-label use of atypical antipsychotic drugs, only placebo-controlled trials evaluated changes in depression scores in patients with major depressive disorder who had been treated with olanzapine, quetiapine, or risperidone as adjunct therapy to antidepressants.[31] This evidence is considered indirect for making comparisons of one antipsychotic with another. Detailed guidance on indirect comparisons for EPCs has been reported previously.[32, 33]

Consistency

Main Considerations

Consistency refers to the degree of similarity in the effect sizes (sometimes termed magnitude of effect) or the degree of similarity in the direction of effects across different studies within an evidence base. Assessment of the consistency of an evidence base divides into three categories: consistent, inconsistent, and consistency unknown.

For most comparisons, the direction of the effect (a benefit or harm of one intervention over another or no difference between the interventions) is paramount when rating consistency. In order to determine the difference between benefit or harm and no difference, it may be necessary to identify a minimum important difference below which the EPC considers there to be no meaningful difference. This threshold should be explicitly and clearly defined. After determining the minimum important difference threshold, EPCs can then assess whether outcome effects across studies are consistent in direction of effect. For example, if equal numbers of study effect sizes all on opposite sides of a line of no difference (e.g., 0) but between thresholds for minimally important differences (e.g., -1 to +1), the effect sizes and CIs indicate no important difference between interventions and could be judged as consistent in direction of effect. These considerations apply for both qualitative and quantitative analyses. Studies with non-overlapping CIs and effects that go in different directions with respect to thresholds are clearly inconsistent.

When examining the consistency of the magnitude of effect, EPCs should determine the degree to which confidence internals (CIs) for those outcomes in the individual studies overlap; greater overlap suggests greater consistency. However, studies with nonoverlapping CIs that are all above a threshold, coupled with effects in the same direction, are at least qualitatively consistent.

If meta-analysis is appropriate, EPCs can evaluate consistency both qualitatively and using statistical tests and measures of heterogeneity (such as Cochran's Q test or I^2 statistics[3]). If the heterogeneity can be explained *a priori*, EPCs can stratify the evidence into subgroups whose outcomes are given separate strength-of-evidence ratings. If the heterogeneity cannot be explained, statistical significance of I^2 statistics and other statistical tests for heterogeneity should not be the sole determinant of the presence of inconsistency because of potential problems in their interpretation.[34, 35] Because no single measure is ideal, EPCs need to explore heterogeneity based on consideration of several factors, including I^2, τ^2, p-values, differences in point estimates, and degree of overlap in CIs of individual study effect sizes.

Some bodies of evidence may show heterogeneity in effect sizes but consistency in the direction of effect. Even if EPCs cannot explain the former heterogeneity satisfactorily, they can still judge the evidence base to be consistent in direction of effect. With substantial unexplained heterogeneity, however, EPCs need to be appropriately cautious about estimating treatment effects.

Evaluation of a Single-Study Evidence Base

Evaluation of consistency ideally requires an evidence base with independent replication of findings. EPCs cannot be certain that a single trial, no matter how large or well-designed, presents the definitive picture of any particular clinical benefit or harm for a given treatment. Accordingly, we recommend that EPCs judge the consistency of a single-study evidence base as unknown, which may decrease the strength of evidence grade, especially if the optimal information size is not met.

Precision

Precision is the degree of certainty surrounding an estimate of effect with respect to an outcome, based on the sufficiency of sample size and number of events. This domain should be scored as precise, imprecise, or precision unknown separately for each important outcome and comparison. A precise estimate should enable decisionmakers to draw conclusions about whether one treatment is, clinically speaking, inferior, equivalent (neither inferior nor superior), or superior to another.[36, 37] Precision is unknown when, for various reasons, a reviewer cannot determine the precision of the evidence base (e.g., when studies do not report measures of dispersion for effect sizes).

When rating precision, EPCs need to consider two main factors: the optimal information size[18] (OIS, a threshold for establishing the minimum number of patients and events) and the 95% CI around the summary effect estimate. This assessment evaluates assesses the likelihood that random error may lead to exaggerated intervention effects.[38] For example, studies of small or moderate sample size and with low numbers of events can generate precise effect sizes, but switching a few events between groups can dramatically change the effect size.

Guidance for assessing the OIS has been previously published by the GRADE working group.[18] If OIS is not met, then in most instances EPCs should consider the evidence to be imprecise.[10, 38] However, when the total sample size across the body of evidence is reasonably large (e.g., 4000 patients), EPCs can consider the estimate to be precise because even with a low number of total events, prognostic factors are likely to be evenly distributed.[18]

Despite meeting the OIS criteria, EPCs may still not be able draw a definitive conclusion due to imprecision in the effect estimates. This is most obvious when a meta-analysis is conducted and confidence intervals cross a minimum important difference threshold. Overlap between the CI and the threshold may indicate imprecision, but it is important to distinguish between wide confidence intervals due to heterogeneity (which may be attributed to inconsistency) and due to imprecision.

Assessment of effect sizes and variation across minimum important difference thresholds should be used rather than assessments of statistical significance. If the threshold for precision is defined as the boundary of statistical significance, in cases where the CI around an effect size overlaps with the possibility of no effect, even a tight CI will be considered imprecise. Also it biases the review away from concluding no difference. To account for such situations, EPCs should attempt to determine thresholds for MIDs (i.e., the minimum effect size that identifies a meaningful difference between groups for benefits or harms). This step is essential to identify interventions that are equivalent or noninferior to each other; that is, the effect size and CI falls below the level of a meaningful between-group difference. For superiority comparisons, use of MIDs is recommended but optional; the decision may depend on the outcome being evaluated and the degree of evidence or expert consensus supporting a threshold. Choice of MID thresholds should be based preferably on empirical evidence, but if this is not possible then EPCs should use the consensus of the review team with input from key informants and technical experts. They should ideally be determined a priori (and be included in review protocols), but may be established post hoc, after OIS criteria are first met. In either case, EPCs should explicitly define thresholds in the methods section of the review.

Determining a MID is not always possible. Studies included in a review may create a separate scale for each outcome measure or use a variety of scales to measure the same outcome, and these scales may not have been subjected to reliability or validity testing. Reviewers may not be able determine a clinically meaningful threshold across scales with different measurement properties. In other instances, studies may not provide a measure of effect and a CI but present only statistical significance tests. Furthermore, clinically important differences are much harder to determine for surrogate or intermediate outcomes and may not be appropriate.

When studies cannot be pooled in a meta-analysis, precision is more difficult or may even be impossible to judge. A common reason that meta-analysis is not feasible is that one or more studies do not report measures of dispersion around effect sizes, and data are not available to perform independent calculations. In such scenarios, EPCs may score the body of evidence as imprecise if the total number of patients or number of events is below the OIS. If the OIS is met but measures of dispersion are not reported, EPCs may score the body of evidence as unknown precision.

Reporting Bias

Reporting bias results from selection of publication or reporting of research findings based on their magnitude or direction of effect.[39, 40] The risk of reporting bias is scored as suspected or not detected. An assessment of risk of reporting bias would only contribute to lowering the strength of evidence, and thus would only be helpful for outcomes that have been graded as moderate or high strength of evidence, based on all other relevant domain scores. As such, risk of reporting bias must be evaluated last.

Empiric evidence guiding assessment of the risk of reporting bias in observational studies is lacking. Currently, methods and infrastructure to assess reporting bias is really only possible for RCT bodies of evidence. Observational studies may also be susceptible to reporting bias,[41-44] particularly because studies are generally not registered and lack *a priori* protocols, but no empiric evidence or mechanism currently exists for assessing reporting bias for observational studies. Further guidance on assessing reporting bias as a whole is in development.[4]

For a given outcome of interest, reporting bias can occur through publication bias and outcome reporting bias, as summarized in Table 2.

Table 2. Definitions of reporting bias

TYPES OF REPORTING BIAS		
Outcome reporting bias	**Publication bias**	**Examples and implications of reporting bias.**
Outcomes data are missing; Results for the outcomes of interest are not reported, when the study is reported	The whole study has been concealed from public access (nonregistration and/or nonpublication) or will be made accessible later after an initial delay – “file drawer phenomenon” and “reporting lag time bias,” respectively. A variant is when the study is published in obscure platforms or journals.	Results included in review are more likely to have positive findings. Findings of no effect are less likely to be published or reported.
Outcome data are reported but the outcome, or the way it was measured was not planned to be investigated		Reflects data dredging and likelier to be a chance finding
Outcome data are reported but originate in the most favorable of the several analyses undertaken		E.g., selective post hoc subgroup analyses, selective cut-offs to dichotomize continuous outcomes, cheery picking statistical assumptions, etc.
Outcome data are incompletely reported		E.g., effect estimate without measures of dispersion or exact p-value
Outcome data are reported in multiple study reports		Co-publication status is not transparently reported leading to double counting of outcomes data

Where applicable, a quantitative assessment of reporting bias, testing for the impact of missing data (e.g., tests for funnel plot asymmetry, trim and fill method and selection modeling) can be used to inform the risk of reporting bias for a body of evidence[45-51] that originates in “missingness” of small study outcomes data that are either nonsignificant or unfavorable in direction. When a quantitative assessment is precluded, a qualitative assessment of reporting bias can be conducted. A proposed, but untested algorithmic approach to evaluate the risk of reporting bias, including guidance on an approach for testing funnel plot asymmetry and a qualitative assessment of the risk of reporting bias, is presented in Figure A-1, Appendix A.

Additional Domains

The second set of domains, which supplement the five required domains, include dose-response association, existence of confounding that would diminish an observed effect (which is referred to in this document as “plausible confounding”), and strength of association (i.e., magnitude of effect). EPCs should consider the additional domains when appropriate; they need not report on those domains when they regard them as irrelevant to the review in question. The additional domains may increase strength of evidence and are especially relevant for observational studies where one may begin with a lower overall strength of evidence grade based on study limitations. Table 3 provides their definitions and ways to rate and apply them. Presence of a clear dose-response association or a very strong association would justify raising a strength of evidence grade. If the confounding that may exist in studies would *decrease* the observed effect, but an effect is observed despite this possible confounding, the EPC may wish to upgrade the strength of evidence. EPCs should explain in their reviews the degree to which additional domains that are used in arriving at any overall strength of evidence grade have altered a judgment based on only the required domains.

Table 3. Additional domains and their definitions

Domain	Definition and Elements	Score and Application
Dose-response association	This association, either across or within studies, refers to a pattern of a larger effect with greater exposure (dose, duration, adherence)	This domain should be considered when studies in the evidence base have noted levels of exposure. Use one of two levels: • **Present:** Dose-response pattern observed • **Not present**: No dose-response pattern observed (dose-response relationship not present)
Plausible confounding that would decrease observed effect	Occasionally, in an observational study, plausible confounding factors would work in the direction opposite that of the observed effect. Had these confounders not been present, the observed effect would have been even larger than the one observed.	This additional domain should be considered if plausible confounding exists that would decrease the observed effect. Use one of two levels: • **Present**: Confounding factors that would decrease the observed effect may be present. • **Absent:** Confounding factors that would decrease the observed effect are not likely to be present.
Strength of association (magnitude of effect)	Strength of association refers to the likelihood that the observed effect is large enough that it cannot have occurred solely as a result of bias from potential confounding factors.	This additional domain should be considered if the effect size is particularly large. Use one of two levels: • **Strong:** large effect size that is unlikely to have occurred in the absence of a true effect of the intervention • **Weak:** small enough effect size that it could have occurred solely as a result of bias from confounding factors

Applicability

A wide array of groups use EPC reviews and other products. Not surprisingly, the populations and contexts these users consider relevant may differ. Thus, evidence that one group may regard as applicable for making clinical or policy decisions, for its population of interest and circumstances, may not be similarly applicable for another decisionmaker. This situation may arise even though key informants, technical experts, or other partners have specified which comparisons, outcomes, or constituencies are very important for the review. EPCs have chosen to make our judgments about applicability explicit and separate from assessments of strength of evidence; separate guidance on applicability is available.[30] Our goal in assessing applicability separately is to enable decisionmakers to take into account how well the evidence maps to the patient populations, diseases or conditions, interventions, comparators, outcomes, and settings that are most relevant to their decisions. EPCs should record information about applicability for the outcomes and comparisons for which they specify an overall strength of evidence grade.

Assessing an Overall Strength of Evidence Grade

Incorporating Multiple Domains into an Overall Grade

For each outcome, EPCs should score domains and strength of evidence separately for RCTs and observational studies. They may then combine those domain scores and strength of evidence grades into one overall strength of evidence grade or they may choose to rely on one study design if it clearly provides stronger evidence. EPCs should describe whether evidence from observational studies complements or conflicts with evidence from RCTs, give plausible reasons for any differences, and note pertinent limitations in both bodies of evidence.

Similarly, based on reasonable standards of evidence for the subject area, EPCs may focus their assessment of strength of evidence on the set of studies that provide the least limited, most direct and reliable evidence for an outcome or comparison. For example, when EPCs locate a reasonable number of studies of head-to-head comparison of important alternatives (i.e., Drug A vs. Drug B), they may elect not to utilize placebo-controlled comparisons (Drug A vs placebo, Drug B vs. placebo) in their summary estimate of effect and therefore in the strength of evidence grading. As stated above, evidence may also focus on studies that do not have a high risk of bias based on their study design.

In some systems, such as that of the GRADE working group,[6, 8, 10-22, 52] the overall grade for strength of evidence (which GRADE calls quality of evidence) is calculated primarily from the ratings for each domain using an approach that provides guidance on how to upgrade or downgrade to reach the overall strength of evidence grade. GRADE uses such an algorithm to help reviewers (and readers) be clear about how they considered domains in producing their final grade. Such a system has the advantage of transparency, documenting how that upgrading or downgrading has been done (e.g., adding a point, subtracting a point); delineating the path from the evidence to its grade.

Although a system that uses such an algorithmic method may offer advantages in terms of transparency, as yet no empirical evidence supports the superiority of a particular point system compared with a more qualitative approach. Furthermore, some evidence suggests no difference in accuracy between quantitative and qualitative systems.[9] Members of the GRADE working group acknowledge that their more arithmetic method should not hold dominance over the sensible “gestalt” that fits the overall body of evidence.[10] This is particularly important when considering the potential for overlap and “double jeopardy” between domains.

Consistency and precision can be particularly challenging domains. When consistency is unknown, downgrading the overall strength of evidence may be appropriate. Scoring consistency becomes more challenging if some studies in the evidence base do not report (or reviewers cannot independently calculate) measures of dispersion around between-group differences in effect. This gap precludes not only statistical testing of heterogeneity but also qualitative assessment of consistency based on CIs. Even if the effect sizes appear to be in the same direction, an EPC cannot determine whether all CIs from the individual studies are above (or fall between) the threshold(s). In this case consistency is unknown, and an EPC must use its judgment to decide whether a downgrade is appropriate.

Another example of a challenging consistency scenario is an evidence base consisting of studies that all measured roughly the same construct (e.g., functional limitation) but used instruments that differ enough to make an EPC doubt the wisdom of converting to a standardized measurement for conducting any meta-analysis. Because differences in effect sizes may reflect differences in measurement instruments, reviewers cannot always determine whether the evidence base is truly inconsistent. The consistency is unknown, and the EPC must decide whether a downgrade is appropriate. Although precision will also be unknown in this example, an EPC would downgrade no more than once (i.e., downgrade for unknown consistency or unknown precision, but not both).

When a meta-analysis cannot be performed and the precision of the body of evidence is unknown, the EPC can downgrade the strength of evidence unless the reviewer has a strong reason for not doing so. For example, if all individual studies have effect sizes that are relatively close, there may be no need to downgrade. Conversely, if studies have precise effect estimates and meet the OIS, but the effect sizes are scattered around threshold(s) then downgrading may be justified, because if a meta-analysis could be performed the summary effect size would likely be imprecise.

In many instances, evidence bases with outcomes that are imprecise will be inconsistent as well. Likewise, if precision cannot be determined then consistency may also be unknown. The question arises as to whether EPCs should downgrade once or twice in these circumstances. We recommend that a single downgrade is usually sufficient in such instances. This means that inconsistency in direction of effect usually precludes the need to rate precision. However, if studies are inconsistent and imprecision includes the possibility of benefit and harm, EPCs should downgrade twice.

The final judgment for combining domains into an overall strength of evidence cannot always simply be reduced to an algorithm. Reviewers must weigh the relative importance of each of the domains in relation to the most concerning uncertainty in the body of evidence and clearly describe how the major concerns in each domain contributed (or did not contribute) to the overall strength of evidence.

Thus, EPCs may use different approaches to incorporate multiple domains into an overall strength of evidence grade. EPCs may use the GRADE system or their own weighting system, or they may elect to use a qualitative approach, so long as the rationale for ratings of strength of evidence is clear and adheres to the following important general principles. The critical requirement is that they explain the rationale for their approach to grading of strength of evidence and note which domains were important in reaching a final grade.

Four Strength of Evidence Levels

The four levels of grades are intended to communicate to decision-makers the confidence in a body of evidence. Although judgment is required, having a common understanding of the interpretation will be useful for systematic reviewers conducting their own global assessment as well as for improving consistency across reviewers.

Table 4 summarizes the four levels of grades that EPCs for the overall assessment of the body of evidence. Overall grades are denoted high, moderate, low, and insufficient. They are not designated by Roman numerals or other symbols.

Table 4. Strength of evidence grades and definitions

Grade	Definition
High	**We are very confident that the estimate of effect lies close to the true effect for this outcome.** The body of evidence has few or no deficiencies. We believe that the findings are stable.
Moderate	**We are moderately confident that the estimate of effect lies close to the true effect for this outcome.** The body of evidence has some deficiencies. We believe that the findings are likely to be stable, but some doubt remains.
Low	**We have limited confidence that the estimate of effect lies close to the true effect for this outcome.** The body of evidence has major or numerous deficiencies (or both). We believe that additional evidence is needed before concluding either that the findings are stable or that the estimate of effect is close to the true effect.
Insufficient	**We have no evidence, we are unable to estimate an effect, or we have no confidence in the estimate of effect for this outcome**. No evidence is available or the body of evidence has unacceptable deficiencies, precluding judgment.

Each level has two components. The first, principal definition concerns the level of confidence the authors place in the estimate of effect for the benefit or harm (i.e., their judgment that the evidence reflects the true effect). The second, subsidiary definition involves an assessment of the level of deficiencies in the body of evidence and belief in the stability of the findings, based on domain scores and a more holistic and summary appreciation of the possibly complex interaction among the individual domains.

Assigning a grade of high, moderate, or low implies that an evidence base is available from which to estimate an effect. The designations of high, moderate, and low should convey how confident EPCs would be about decisions based on evidence of differing grades. EPCs should apply discrete grades and avoid designations such as “low to moderate” strength of evidence.

The importance of the distinctions between these levels (and the distinction with insufficient strength of evidence) can vary by the type of outcome or comparison and the decisionmaker. EPCs understand that some stakeholders may have interest in taking action only when evidence is of high or moderate strength, whereas others may want to understand clearly the implications of low vs. insufficient evidence. Even when strength of evidence is low, consumers, clinicians, and policymakers may find themselves in the position of having to make choices and decisions.

In some cases, EPCs cannot draw any evidence-based conclusions for a particular outcome, specific comparison, or other question of interest. In these situations, the EPC should assign a grade of insufficient but be specific in text or tables as to why they were unable to reach a conclusion. EPCs need to take particular care not to conflate “low” strength of evidence with “insufficient.” If a body of evidence is truly “insufficient,” that should mean that no conclusion can be drawn that is associated with that body of evidence.

The first reason an EPC may conclude there is insufficient evidence is when *no* evidence is available from the included studies. This case includes the absence of any relevant studies whatsoever. In some systematic reviews, for example, certain drug comparisons may never have been studied (or published) in head-to-head trials *and* placebo-controlled trials of the multiple drugs of interest may not provide adequate indirect evidence for any comparisons.

Another common example when EPCs may conclude a grade of insufficient is when evidence on the outcome is too weak, sparse, or inconsistent to permit EPCs to draw any defensible conclusion concerning the effect—that is, either a benefit or a harm (or a finding of no difference). This situation can reflect one or more of several complicated conditions, such as unacceptably high study limitations or a major unexplained inconsistency (e.g., two studies with the same risk of bias that found opposite results, with no clear reason for the discrepancy). Imprecise data can also lead to a grade of insufficient, specifically when the confidence interval is so wide that it includes two incompatible conclusions: that one treatment is clinically significantly better than the other, and that it is worse. In addition, evidence based on a single study or comparison (particularly if it does not meet OIS criteria) also usually warrants a grade of insufficient.

Even when several studies are available, the strength of evidence could be considered insufficient if a single quantitative estimate is desired and the EPC cannot calculate any effect size from reported information or cannot explain heterogeneity. If, however, just the general direction of the effect is needed, this same evidence base might be considered sufficient to permit a conclusion.

Transparency: Documenting and Reporting Strength of Evidence

EPCs should carefully document procedures used to grade strength of evidence (in the review’s Methods section) and provide enough detail to assure that users can grasp the methods that were employed. For example, important considerations may include how different study designs and studies with high risk of bias were incorporated into the strength of evidence grading and how each of the domains was weighted in assigning the grade for each outcome. For the sake of consistency across reviews, the domains should be defined using the terminology presented in this paper.

As noted above, EPCs’ systematic reviews should present information about all comparisons of interest for the outcomes that are most important to patients and other decisionmakers. Complete and perfect information is rarely available. For some treatments, data may be lacking about one or more of the outcomes. In other cases, the available evidence comes from studies that have important flaws, is imprecise, or is not applicable to some populations of interest. For these reasons, EPCs should also present information that will help decisionmakers judge study limitations that would increase the risk of bias in the estimates of effect, take imprecision and other factors into account, and assess the applicability of the evidence to populations of interest.

We acknowledge and emphasize the need to balance transparency with readability of reviews. Transparency does not mean that EPCs must provide all details about all decisions in the body of the report; supporting details may have been recorded in review protocols and can be provided in appendices. The placement and presentation of information should emphasize usability and readability of the document overall.

Much of the information (domain scores and overall strength of evidence) is presented in tables. Tables 5 through 7 illustrate one approach to providing actionable information to decisionmakers. (Table 5 is below and Tables 6 and 7 are in Appendix B.) We recommend that Table 5 or a comparable table—or a suite of tables, depending on the complexity of the review--presenting a summary of key findings and strength of evidence grades be included in the main report, typically in the discussion section. All or most of this table could also be presented in the Executive Summary.

Table 5. Summary of key outcomes, findings, and strength of evidence. See Tables B-1 and B-2 in Appendix B.

Outcome	Study Design: No. Studies (N)	Findings and Magnitude of Effect	Strength of Evidence
Critical outcomes			
Mortality	RCT: 1 (56)	A single study with poor precision and directness of outcome assessment found no significant difference in mortality at 1 year.	Insufficient
Severity of [Disease]	RCT: 8 (250)	High risk of bias studies found inconstant and imprecise effects on a range of specific outcomes. RRs ranged from 0.45 (0.11, 1.8) to 3.2 (1.8, 5.7), Outcome assessments were done at 1 month to 5 years. Overall, the effect on severity of [disease] is unclear.	Low
Patient-reported outcomes			
Pain	RCT: 6 (160)	Mostly moderate risk of bias studies consistently found X reduced pain more than Y between 3 months and 2 years. Summary SMD = 0.5 (0.2, 0.8); however, there was large statistical heterogeneity (χ^2 P=0.003; I^2=0.97) and SMD estimates ranged from 0.13 to 0.94).	Moderate (direction) Low (magnitude)
Sexual dysfunction	RCT: 3 (85)	Text on sexual dysfunction… e.g., few studies; measured only in men; results consistent but imprecise	Low
Intermediate outcomes			
Hb A1c	RCT: 13 (845)	Numerous studies (NN) reported on reduction of Z. Summary net change was -2.1% (95% CI -4, -0.2)	Low
Radiology test	RCT: 0	No eligible studies	Insufficient
Adverse Events			
Intestinal perforation	RCT: 1 (42)	One study (NN) reported on event; frequency very rare ranging from PP to PPP …;	Low
Weight gain	Observational: 5 (1100)	Numerous studies (NN) reported on event; measured by clinician (or BMI, or …); Results consistent; some effect size (2.8 kg; 95% CI, 1.5, 3.5)	Low

The important components of Table 5 or a comparable strength of evidence summary table include: the number of contributing studies and number of participants, a summary of study limitations and other scored domains, a description of the direction of effect, a description of the length of followup, and, to avoid undue length in the table, a succinct description of the findings and magnitude of effect (including summary estimates from meta-analyses, if appropriate). In this way, readers can better understand the available evidence for any given outcome or comparison.

However, if the evidence for a given outcome or comparison is from a single study or a small number of studies and strength of evidence was graded as insufficient to permit drawing any conclusion, EPCs can highlight this finding if it is a principal finding or the report or summarize it in a paragraph that includes all such studies.

Tables B-1 and B-2 present additional detail and rationale and are examples of supporting tables that would be included in an appendix and called out in the body of the report (Appendix B). Table B-1 includes domain scores and strength of evidence grades for RCT and observational study evidence for each graded outcome and the final strength of evidence grade. Table B-2 summarizes the reasons and logic that was used in developing the grades.

We recommend that the title of each table state the intervention comparison being summarized. Based on the best presentation for each review, tables can include whole topics, or be specific to key questions or treatment/intervention comparisons. We believe that readability is enhanced by tables dividing outcomes into the following categories: clinical, patient-related, quality-of-life, intermediate, and adverse events.

Transparency regarding strength of evidence grades should emphasize how important decisions were made; just stating such phrases as "per AHRQ guidance" or "standard practice" are considered inadequate. We recommend that the methods section of the report include details about how the following were operationalized: individual study risk of bias ratings (i.e., what is meant by good, fair, or poor); strength of evidence grades (i.e., approach to grading and what situations would result in one grade versus another, such as low vs. insufficient); consistency (i.e., how factors such as direction, magnitude of effect, thresholds, statistical heterogeneity, and overlapping CIs were evaluated, and prioritized).

We further recommend that the report contains support for each conclusion. Reviewers need to clearly state what the strength of evidence grade conveys (e.g., insufficient evidence to determine the effect of X on Y) and rationale for the grade. If one or more factors were considered particularly salient, they should be noted. EPCs may present any needed commentary concerning the information in the strength of evidence tables in text or in the table itself. Lastly, when RCT and observational study evidence were used in developing a final strength of evidence grade, the EPC needs to explicitly state the reasons for including both study designs and how the bodies of evidence were combined.

Clearly articulating other available evidence that has not been graded and its location in the report will allow users to access findings according to different priorities. For example, an evidence grade might apply to a link in an analytic framework, or to a specific intervention, for a specific set of outcomes in a particular population.

Discussion

The EPC approach to grading strength of evidence draws heavily on the international GRADE system; both conceptually and substantively, it is similar to GRADE. (Table 6 compares the EPC, GRADE and United States Preventive Services Task Force (USPSTF) approaches to the task.) Our recommendations address specific circumstances of the EPC program, which differ from those of some groups that use GRADE. The EPC program produces systematic reviews, but it is not involved directly in development of recommendations or guidelines. Rather, a wide spectrum of government agencies, professional societies, patient advocacy groups, and other stakeholders use EPC reports. Our approach for grading strength of evidence aims to facilitate use of the EPC reports by these diverse groups.

We recommend that EPCs grade strength of evidence based on a core group of domains that include aggregate study limitations, directness, consistency, precision, and reporting bias. We suggest that EPCs also consider the interaction among the domains and the unique concerns of the particular body of evidence, when making their final determination.

We recognize that some types of evidence, such as evidence about public health interventions, quality improvement or patient safety studies, and studies of diagnostic tests may be challenging to grade. With these types of nontherapeutic intervention questions, the challenge to the EPCs is to determine the study design(s) that would be most appropriate to reduce risk of bias. For example, EPCs may find that particular types of studies, such as interrupted time series, lower the risk of bias more than do other types of observational studies. Nevertheless, we caution that changing the assessment of observational studies for risk of bias should be done judiciously.

AHRQ systematic reviews have often focused on pharmaceutical therapies, for which both efficacy and effectiveness trials are a major source of information. The domains discussed above are directly relevant to studies of most drugs, procedures, and other therapeutic interventions. In the future, however, EPCs may increasingly assess diagnostic tests or screening strategies. For these technologies, RCTs may not be the source of much relevant information, and the studies that are available may have special methodologic features. EPCs should consult the separate methods guidance for instructions on grading strength of evidence for such reviews.[54]

In arriving at an overall strength of evidence grade, the crucial requirement is transparency. EPCs should make a global assessment of the overall strength of evidence with explicit consideration for how the scores for each domain contribute to that overall grade, although may not follow a standard algorithm. EPCs should make judgments for individual domains as a first step and to be especially sensitive to the effects of any “borderline” scores for those domains and their impact on the overall score. A case in point involves decisions about grading evidence about an important outcome as “low” versus “insufficient.” Being explicit and transparent about what steps and criteria are used to arrive at a final strength of evidence grade is the essential element.

Table 6. Comparison of terms and levels within different widely used strength of evidence systems

Evidence System	**USPSTF**	**GRADE**	**EHC/EPC Program** **(Owens & in-process update)**
Systematic Review (SR) Approach	A Clinical Preventive Service Analytic Framework (AF) and Key Questions (KQ)	Clinical Questions about Alternative Health Management Strategies (PICO)	Health and Health Care Questions (Prevention, Diagnosis, Treatment, Prognosis) (AF) (PICO)
By	USPSTF and EPC—public posting	Clinicians, Experts, GL developers	EPC, nominators, stakeholders---public posting
Individual Study Level **(Rating Scale)**	***Quality*** (Internal Validity) (**Good, Fair, Poor**) Within Type of Study Design (and by outcome if appropriate)	RCTs: start as high-quality evidence Down-graded: for study limitations Observational studies: start as low quality evidence (for estimates of intervention effects) Down-graded: for study limitations, upgraded for large effects, dose-response, or residual confounding projected to minimize effect GRADE is “…..less comprehensive than many other systems”	***Risk of Bias*** (**Low, Medium, High, Unclear**) For selection bias, confounding, attrition, performance, detection, reporting, and other biases (such as selective reporting, outcome measurement biases) Within study design and by type of outcome as appropriate using
By	EPC	SR / GL developers	EPC
1st Summary Level	Summary of Evidence (SOE) Tables Display of no of studies and participants, designs, major limitations, consistency, applicability, overall study quality, quantitative/qualitative summary of findings **Not rated**	***Quality of Evidence: Evidence Profile*** Depending on starting point (based on study design), body of evidence per outcome is downgraded for study limitations, imprecision, inconsistency of results, indirectness, publication bias and upgraded for dose-response, large effects, direction of effect of possible confounding	***Strength of Evidence*** Five Required Domains: study limitations, directness, consistency, precision, reporting bias **(High, Moderate, Low, Insufficient)**

Evidence System	USPSTF	GRADE	EHC/EPC Program (Owens & in-process update)
	Within KQ, stratified by intermediate (benefits & harms) or health outcomes (benefits & harms), and by population, intervention, as appropriate	(**High, Moderate, Low, Very Low**) Within patient & treatment groups and with effect estimates for each critical or important patient-important outcome	Only for major outcomes and comparisons, by each outcome and comparison separately and by type of study design separately (RCT versus observational) Also displays magnitude of effect
By	EPC	SR/GL developers	EPC

Table 6. Comparison of terms and levels within different widely used strength of evidence systems (continued)

Evidence System	**USPSTF**	**GRADE**	**EHC/EPC Program** **(Owens & in-process update)**
2nd Summary Level	Critical Appraisal Grid for each link in AF (**Convincing, Adequate, Inadequate**) 6 critical appraisal questions for ***overall adequacy of evidence*** by population subgroups as appropriate	***Quality of Evidence: Summary of Findings*** (with PICOS) with effect estimates and Quality of Evidence (confidence in effect estimates) ratings for each outcome (**High, Moderate, Low, Very Low**)	***Overall Strength of Evidence*** (**High, Moderate, Low, Insufficient**) Considering how RCTs and observational evidence combine for an overall strength of evidence for a major outcome or comparison, specifying purpose of observational evidence
By	USPSTF	SR/GL developers	EPC
3rd Summary Level	Body of evidence for CPS ***Certainty of Net Benefit*** (**High, Moderate, Low**) By population/service subgroups	Overall Quality of Evidence Confidence in Estimate of Effect across All Critical Outcomes (**High, Moderate, Low, Very Low**)	
By	USPSTF	G/L developers only	
4th Summary Level	Magnitude of effect ***Magnitude of net benefit*** (**Substantial, Moderate, Small, Zero/Negative**) By population/service subgroups with at least adequate evidence	Recommended Treatment Which option is preferable **Strength of Recommendation** **(Strong, Weak)**	

By	USPSTF (EPC may do outcomes table)	GL developers only	
5th Summary Level	Recommendation statements (**A,B,C, D, I statement**) Recommendation grid of ***certainty and magnitude of net benefit*** for each populations/subgroup as appropriate		
By	USPSTF		
Applicability	Displayed in SOE table by EPC, considered by USPSTF in summary levels 2 and 3	Part of Directness consideration in Quality of Evidence Assessment at the Outcome Level	Separate consideration in addition to strength of evidence; in text

Conclusions

A consistent approach for grading the strength of evidence—one that decisionmakers can readily recognize and interpret—is highly desirable. To that end, the EPCs and the GRADE working group will continue to collaborate to facilitate consistency across grading systems. Meanwhile, this paper codifies the guidance that EPCs can follow now to strengthen the consistency, clarity, and usefulness of the reviews and other products from AHRQ's EPC program.

This paper recommends key points which will help improve consistency and improved understanding and use of systematic reviews by decision-makers.

1. Document decisions in the protocol
    a. Identify major outcomes and comparisons to be graded
    b. Identify criteria for excluding studies based on risk of bias
    c. Describe decision rules and procedures that will be used to ensure consistency and transparency of scoring of individual domains and the overall strength of evidence
2. Assess each domain in an iterative manner
    a. Study limitations considers the risk of bias of individual studies and is the backbone of the required domains.
    b. There may be specific situations in which assessment of other required domains (directness, consistency, precision, reporting bias) is unlikely to change the overall strength of evidence.
    c. Three additional domains may increase the strength of evidence for a body of non-randomized studies.
3. Assess overall strength of evidence for each comparison and outcome
    a. EPCs should grade the RCT body of evidence separately from non-randomized studies and then combine them into an overall body of evidence.
    b. EPCs should "check" their overall strength of evidence grade based on domains with a global assessment that considers the definitions of each level of grading.
4. Document clearly in report
    a. EPCs should be explicit in the methods section about their decision rules for combining across domains for an overall strength of evidence grade.
    b. EPCs should describe clearly in summary tables which domains contributed significantly to the final strength of evidence rating

References

1. Helfand M. Using evidence reports: progress and challenges in evidence-based decision making. Health Aff (Millwood). 2005;24(1):123-7.

2. Atkins D, Fink K, Slutsky J. Better information for better health care: the Evidence-based Practice Center program and the Agency for Healthcare Research and Quality. Ann Intern Med. 2005 Jun 21;142(12 Pt 2):1035-41. PMID: 15968027.

3. Agency for Healthcare Research and Quality. Methods Reference Guide for Effectiveness and Comparative Effectiveness Reviews, Version 1.0. [Draft posted Oct. 2007]. Rockville, MD. Available at: http://effectivehealthcare.ahrq.gov/repFiles/2007_10DraftMethodsGuide.pdf; 2007.

4. Agency for Healthcare Research and Quality. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. 2008. http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318. Accessed on June 22, 2011.

5. Owens DK, Lohr KN, Atkins D, et al. AHRQ series paper 5: 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.

6. Atkins D, Eccles M, Flottorp S, et al. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches The GRADE Working Group. BMC Health Serv Res. 2004 Dec 22;4(1):38. PMID: 15615589.

7. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. Bmj. 2008 Apr 26;336(7650):924-6. PMID: 18436948.

8. Guyatt GH, Oxman AD, Kunz R, et al. What is "quality of evidence" and why is it important to clinicians? BMJ. 2008 May 3;336(7651):995-8. PMID: 18456631.

9. West S, King V, Carey TS, et al. Systems to Rate the Strength of Scientific Evidence. Evidence Report/Technology Assessment No. 47 (Prepared by the Research Triangle Institute-University of North Carolina Evidence-based Practice Center under Contract No. 290-97-0011). AHRQ Publication No. 02-E016. Rockville, MD: Agency for Healthcare Research and Quality; 2002.

10. Guyatt G, Oxman AD, Sultan S, et al. GRADE guidelines 11-making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. J Clin Epidemiol. 2012 Apr 27PMID: 22542023.

11. Guyatt GH, Oxman AD, Santesso N, et al. GRADE guidelines 12. Preparing Summary of Findings tables-binary outcomes. J Clin Epidemiol. 2012 May 18PMID: 22609141.

12. Guyatt G, Thorlund K, Oxman AD, et al. GRADE guidelines 13. Preparing Summary of Findings (SoF) Tables and Evidence Profiles – continuous outcomes. in process.

13. Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011 Apr;64(4):383-94. PMID: 21195583.

14. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol. 2011 Apr;64(4):395-400. PMID: 21194891.

15. Balshem H, Helfand M, Schunemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011 Apr;64(4):401-6. PMID: 21208779.

16. Guyatt GH, Oxman AD, Vist G, et al. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). J Clin Epidemiol. 2011 Apr;64(4):407-15. PMID: 21247734.

17. Guyatt GH, Oxman AD, Montori V, et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. J Clin Epidemiol. 2011 Dec;64(12):1277-82. PMID: 21802904.

18. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 6. Rating the quality of evidence--imprecision. J Clin Epidemiol. 2011 Dec;64(12):1283-93. PMID: 21839614.

19. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. J Clin Epidemiol. 2011 Dec;64(12):1294-302. PMID: 21803546.

20. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 8. Rating the quality of evidence--indirectness. J Clin Epidemiol. 2011 Dec;64(12):1303-10. PMID: 21802903.

21. Guyatt GH, Oxman AD, Sultan S, et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol. 2011 Dec;64(12):1311-6. PMID: 21802902.

22. Brunetti M, Ian Shemilt I, Pregno S, et al. GRADE guidelines: 10. Considering resource use and rating the quality of economic evidence J Clin Epidemiol. in press.

23. Guyatt GH, Oxman AD, Schunemann HJ, et al. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol. 2011 Apr;64(4):380-2. PMID: 21185693.

24. Institute of Medicine (IOM). Finding what works in health care: standards for systematic reviews. Washington, DC: The National Academies Press; 2011.

25. Patient-Centered Outcomes Research Institute. 1 of 7 – Rationale: Working Definition of Patient-Centered Outcomes Research. Washington, DC: Patient-Centered Outcomes Research Institute. http://www.pcori.org/images/PCOR_Rationale.pdf. Accessed on March 12, 2012.

26. Crowther MA. Introduction to surrogates and evidence-based mini-reviews. Hematology Am Soc Hematol Educ Program. 2009:15-6. PMID: 20008177.

27. National Research Council. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, DC: The National Academies Press; 2001.

28. Viswanathan M, Ansari MT, Berkman ND, et al. Assessing the Risk of Bias of Individual Studies in Systematic Reviews of Health Care Interventions Methods Guide for Comparative Effectiveness Reviews. AHRQ Publication No. 12-EHC047-EF. Agency for Healthcare Research and Quality; March 2012. www.effectivehealthcare.ahrq.gov/.

29. Berkman ND, Lohr KN, Morgan LC, et al. Reliability Testing of the AHRQ EPC Approach to Grading the Strength of Evidence in Comparative Effectiveness Reviews. Methods Research Report. (Prepared by RTI International–University of North Carolina Evidence-based Practice Center under Contract No. 290-2007-10056-I.). AHRQ Publication No. 12-EHC067-EF. Rockville, MD: Agency for Healthcare Research and Quality; May 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

30. Atkins D, Chang SM, Gartlehner G, et al. Assessing applicability when comparing medical interventions: AHRQ and the Effective Health Care Program. J Clin Epidemiol. 2011 Nov;64(11):1198-207. PMID: 21463926.

31. Maglione M, Ruelaz Maher A, Hu J, et al. Off-Label Use of Atypical Antipsychotics: An Update (Prepared by the Southern California Evidence-based Practice Center under Contract No. HHSA290-2007-10062- 1.). Comparative Effectiveness Review No. 43. Rockville, MD: Agency for Healthcare Research and Quality; September 2011. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

32. Fu R, Gartlehner G, Grant M, et al. Conducting Quantitative Synthesis When Comparing Medical Interventions: AHRQ and the Effective Health Care Program Methods Guide for Effectiveness and Comparative Effectiveness Reviews [Internet]. 2008-.AHRQ Methods for Effective Health Care. Rockville, MD: Agency for Healthcare Research and Quality (US); Oct 25 2010.

33. 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. PMID: 21477993.

34. Higgins JP. Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. Int J Epidemiol. 2008 Oct;37(5):1158-60. PMID: 18832388.

35. Patsopoulos NA, Evangelou E, Ioannidis JP. Heterogeneous views on heterogeneity. Int J Epidemiol. 2009 Dec;38(6):1740-2. PMID: 18940836.

36. Sackett DL. Superiority trials, noninferiority trials, and prisoners of the 2-sided null hypothesis. ACP J Club. 2004 Mar-Apr;140(2):A11. PMID: 15122874.

37. Sackett D. The principles behind the tactics of performing therapeutic trials. In: Haynes RBS, Guyatt DL, Gordon H, Tugwell P, eds. Clinical Epidemiology: How to Do Clinical Practice Research. New York: Lippincott Williams & Wilkins; 2005.

38. Thorlund K, Imberger G, Walsh M, et al. The number of patients and events required to limit the risk of overestimation of intervention effects in meta-analysis--a simulation study. PLoS One. 2011;6(10):e25491. PMID: 22028777.

39. Dickersin K. The existence of publication bias and risk factors for its occurrence. Jama. 1990 Mar 9;263(10):1385-9. PMID: 2406472.

40. Man-Son-Hing M, Wells G, Lau A. Quinine for nocturnal leg cramps: a meta-analysis including unpublished data. J Gen Intern Med. 1998 Sep;13(9):600-6. PMID: 9754515.

41. Loder E, Groves T, Macauley D. Registration of observational studies. BMJ. 2010;340:c950. PMID: 20167643.

42. Chanock SJ, Manolio T, Boehnke M, et al. Replicating genotype-phenotype associations. Nature. 2007 Jun 7;447(7145):655-60. PMID: 17554299.

43. Bekkering GE, Harris RJ, Thomas S, et al. How much of the data published in observational studies of the association between diet and prostate or bladder cancer is usable for meta-analysis? Am J Epidemiol. 2008 May 1;167(9):1017-26. PMID: 18403406.

44. Kavvoura FK, Liberopoulos G, Ioannidis JP. Selection in reported epidemiological risks: an empirical assessment. PLoS Med. 2007 Mar;4(3):e79. PMID: 17341129.

45. Kirkham JJ, Dwan KM, Altman DG, et al. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. Bmj. 2010;340:c365. PMID: 20156912.

46. Egger M, Davey Smith G, Schneider M, et al. Bias in meta-analysis detected by a simple, graphical test. Bmj. 1997 Sep 13;315(7109):629-34. PMID: 9310563.

47. Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. Stat Med. 2006 Oct 30;25(20):3443-57. PMID: 16345038.

48. Rucker G, Schwarzer G, Carpenter J. Arcsine test for publication bias in meta-analyses with binary outcomes. Stat Med. 2008 Feb 28;27(5):746-63. PMID: 17592831.

49. Peters JL, Sutton AJ, Jones DR, et al. Comparison of two methods to detect publication bias in meta-analysis. Jama. 2006 Feb 8;295(6):676-80. PMID: 16467236.

50. Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics. 2000 Jun;56(2):455-63. PMID: 10877304.

51. Copas J, Shi JQ. Meta-analysis, funnel plots and sensitivity analysis. Biostatistics. 2000 Sep;1(3):247-62. PMID: 12933507.

52. Atkins D, Best D, Briss PA, et al. Grading quality of evidence and strength of recommendations. BMJ. 2004 Jun 19;328(7454):1490. PMID: 15205295.

53. van der Heijde D, Klareskog L, Boers M, et al. Comparison of different definitions to classify remission and sustained remission: 1 year TEMPO results. Ann Rheum Dis. 2005 Nov;64(11):1582-7. PMID: 15860509.

54. Singh S, Chang S, Matchar DB, et al. Grading a body of evidence on diagnostic tests Chapter 7 of Methods Guide for Medical Test Reviews. AHRQ Publication No. 12-EHC079-EF. Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/ final.cfm. Also published as a special supplement to the Journal of General Internal Medicine, July 2012.