Methods Guide
for Comparative Effectiveness Reviews

Grading the Strength of a Body of Evidence When Assessing Health Care Interventions for the Effective Health Care Program of the Agency for Healthcare Research and Quality: An Update

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

This report is based on research conducted by the RTI International-University of North Carolina Evidence-based Practice Center under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-2007-10056-I). The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

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

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

This document is in the public domain and may be used and reprinted without special permission. Citation of the source is appreciated.

Persons using assistive technology may not be able to fully access information in this report. For assistance contact info@ahrq.gov.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

Suggested citation: Berkman ND, Lohr KN, Ansari M, McDonagh M, Balk E, Whitlock E, Reston J, Bass E, Butler M, Gartlehner G, Hartling L, Kane R, McPheeters M, Morgan L, Morton SC, Viswanathan M, Sista P, Chang S. Grading the Strength of a Body of Evidence When Assessing Health Care Interventions for the Effective Health Care Program of the Agency for Healthcare Research and Quality: An Update. Methods Guide for Comparative Effectiveness Reviews (Prepared by the RTI-UNC Evidence-based Practice Center under Contract No. 290-2007-10056-I). AHRQ Publication No. 13(14)-EHC130-EF. Rockville, MD: Agency for Healthcare Research and Quality. November 2013. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

Authors:
Nancy D. Berkman, Ph.D., MLIR[a]
Kathleen N. Lohr, Ph.D., M.Phil., M.A. [a]
Mohammed Ansari**,** M.D., M.Med.Sc., M.Phil.[b]
Marian McDonagh, Pharm.D.[c]
Ethan Balk, M.D.[d]
Evelyn Whitlock, M.D., M.P.H.[e]
James Reston, Ph.D.[f]
Eric Bass, M.D., M.P.H.[g]
Mary Butler, Ph.D.[h]
Gerald Gartlehner, M.D., M.P.H.[i]
Lisa Hartling, Ph.D.[j]
Robert Kane, M.D., M.P.H.[h]
Melissa McPheeters, Ph.D.[k]
Laura Morgan, M.A.[a]
Sally C. Morton, Ph.D.[l]
Meera Viswanathan, Ph.D.[a]
Priyanka Sista, B.A.[a]
Stephanie Chang, M.D., M.P.H.[m]

[a]RTI International, Research Triangle Park, NC

[b]University of Ottowa, Ottawa, Ontario, Canada

[c]Oregon Health and Science University, Portland, OR

[d]Tufts University School of Medicine, Boston, MA

[e]Kaiser Permanente Center for Health Research, Portland, OR

[f]ECRI Institute, Plymouth Meeting, PA

[g]Johns Hopkins School of Medicine, Baltimore, MD

[h]University of Minnesota, Minneapolis, MN

[i]Danube University, Krems, Austria

[j]University of Alberta, Edmonton, Alberta, Canada

[k]Vanderbilt University, Nashville, TN

[l]University of Pittsburgh, Pittsburgh, PA

[m]Agency for Healthcare Research and Quality, Rockville, MD

Preface

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

Strong methodological approaches to systematic review improve the transparency, consistency, and scientific rigor of these reports. Through a collaborative effort of the Effective Health Care (EHC) Program, AHRQ, the EHC Program Scientific Resource Center, and the AHRQ Evidence-based Practice Centers have developed a Methods Guide for Comparative Effectiveness Reviews. This Guide presents issues key to the development of Systematic Reviews and describes recommended approaches for addressing difficult, frequently encountered methodological issues.

The Methods Guide for Comparative Effectiveness Reviews is a living document, and will be updated as further empiric evidence develops and our understanding of better methods improves. We welcome comments on this Methods Guide paper. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by email to epc@ahrq.hhs.gov.

Richard G. Kronick, Ph.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 and Task Order Officer
Evidence-based Practice Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Acknowledgments

The authors thank Gillian Sanders, Ph.D., of the Duke Evidence-based Practice Center for organizing an EPC retreat in 2012 that was a crucial step in the development of this guidance. We also express our appreciation to members of the GRADE working group, and in particular Holger Schünemann, M.D., MSc, Ph.D., for offering their guidance both in person and through thoughtful comments on the draft document. Lastly, we acknowledge with gratitude the expert document preparation provided by Loraine Monroe at RTI International.

This research was funded through the Agency for Healthcare Research and Quality Effective Health Care 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

Naomi Aronson, Ph.D.
Blue Cross and Blue Shield Association
Chicago, IL

Suzanne Belinson, Ph.D., M.P.H.
Blue Cross and Blue Shield Association
Chicago, IL

Kevin Bozic, M.D., M.B.A.
University of California at San Francisco
San Francisco, CA

Timothy Carey, M.D., M.P.H.
Cecil G. Sheps Center for Health Services Research, University of North Carolina
Chapel Hill, NC

David Moher, Ph.D.
Ottawa Hospital Research Institute
Ottawa, Quebec, Canada

Diana Petitti, M.D., M.P.H.
Arizona State University
Phoenix, AZ

Gregory Samsa, Ph.D.
Duke University
Durham, NC

David Samson, M.S.
Blue Cross and Blue Shield Association
Chicago, IL

Harold C. Sox, M.D.
Dartmouth Institute for Health Policy and Clinical Practice and Geisel School of Medicine
Dartmouth, NH

Jonathan Treadwell, Ph.D.
ECRI
Plymouth Meeting, PA

C. Michael White, Pharm.D.
University of Connecticut
Storrs, CT

*We did not formally respond to comments from peer reviewers that were received after the closing date (August 8, 2012)

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 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 synthesize a body of literature; the ultimate goal is to help clinicians, policymakers, and patients make well-considered decisions about health care. The goal of strength of evidence assessments is to provide clearly explained, well-reasoned judgments about reviewers' confidence in their systematic review conclusions so that decisionmakers can 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 chapter updates the original EPC strength of evidence approach,[5] presenting findings and recommendations of a work group with experience in applying previous guidance; it should be considered current guidance for EPCs. The guidance applies primarily to systematic reviews of drugs, devices, and other preventive and therapeutic interventions; it may apply to exposures (characteristics or risk factors that are determinants of health outcomes) and broader health services research questions. It does not address reviews of medical tests.

EPC reports support the work of many decisionmakers, but EPCs do not themselves develop recommendations or practice guidelines. In particular, we limit our grading strength of evidence approach to individual outcomes. Unlike grading systems that were designed to be used more directly by specific decisionmakers,[6-8] we do not develop global summary judgments of the relative benefits and harms of treatment comparisons.

We briefly explore the rationale for grading strength of evidence, define domains of concern, and describe our recommended grading system for systematic reviews. The aims of this guidance are twofold: (1) to foster appropriate consistency and transparency in the methods that different EPCs use to grade strength of evidence and (2) to facilitate users' interpretations of those grades for guideline development or other decisionmaking tasks. Because this field is rapidly evolving, future revisions are anticipated; they will reflect our increasing understanding and experience with the methodology.

Aims and Key Considerations for Grading Strength of Evidence

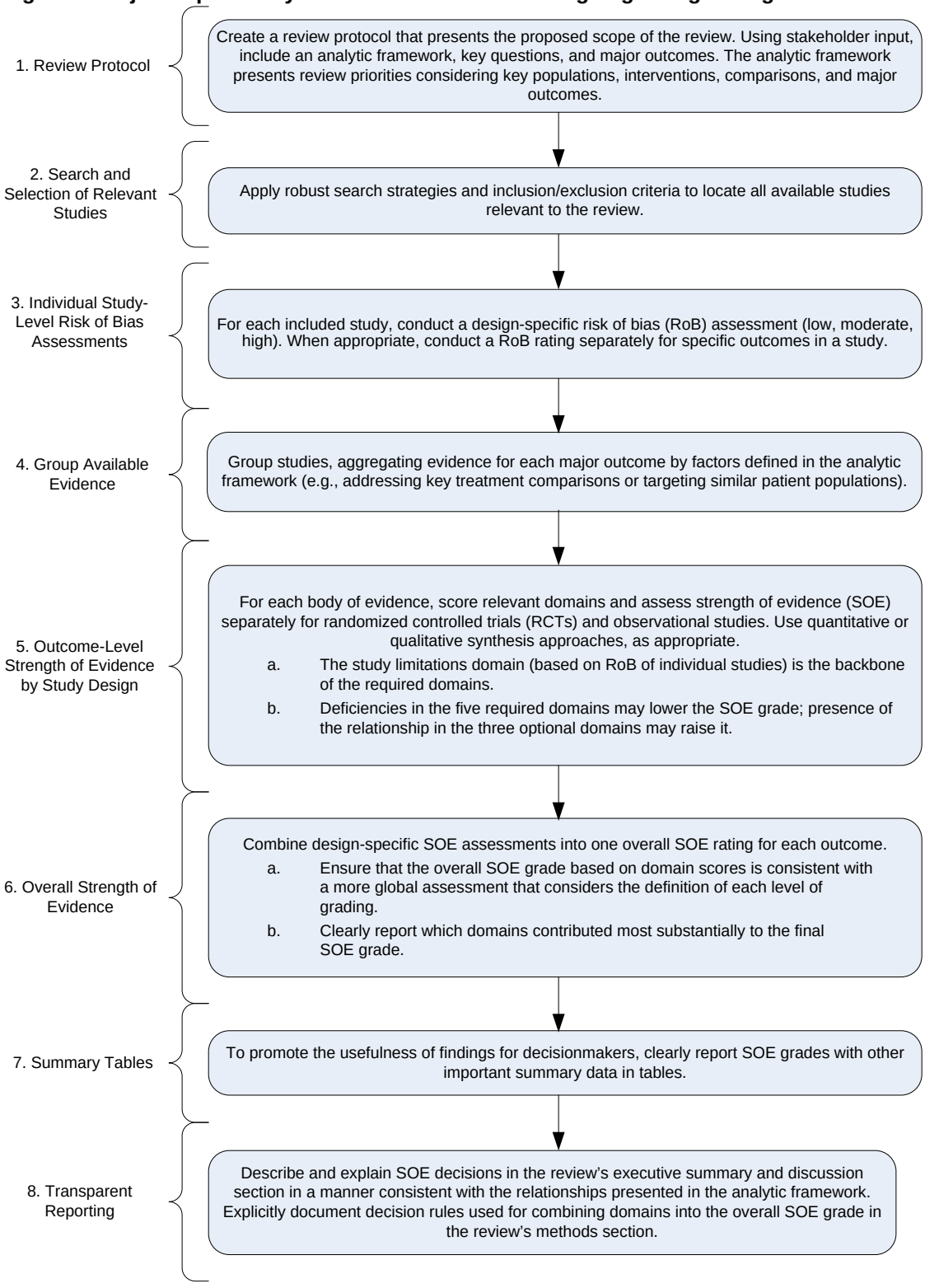
The primary purposes of a systematic review are to synthesize evidence for use by clinicians and patients and to facilitate the work of organizations that develop practice guidelines or make coverage decisions. Systematic reviewers examine all available evidence, summarize the findings, and communicate to end-users the reviewers' confidence in those findings. In some cases, reviewers may be able to conduct a meta-analysis to provide a quantitative estimate of effect (or no difference in effect) and related statistical inferences via a confidence interval (CI) or hypothesis test. In other cases, however, they may be able to speak only to the direction of effect through a qualitative (narrative) synthesis. The strength of evidence grade summarizes the reviewers' confidence in the findings based on either approach to evidence synthesis.

Grading the strength of evidence requires assessment of specific domains, including study limitations, directness, consistency, precision, and reporting bias. To assess the consistency and precision of a body of evidence, reviewers need to decide whether they are rating these domains with respect to estimating either an effect size or only the general direction of effect. The precision domain assesses possible random error; all other required domains assess possible sources of systematic bias that may distort true effects. Additional domains that may be considered for some bodies of evidence and increase confidence in the findings include increasing dose-response associations, plausible confounding that decreases the observed effect, and large magnitudes of effect.

Attaining the goals of consistency, transparency, and usability rests in part on uniformity and predictability in how EPC reviewers interpret these domains. 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 the broad range of potential audiences that AHRQ's work aims to reach.

Figure 1 illustrates the major steps of strength of evidence assessments, using hypothetical information. Some decisions must be made a priori and are documented during the stage in which review protocol are developed. Then, according to these decisions rules and procedures documented in protocols, EPCs assess individual domain scores and establish overall strength of evidence grades.

Figure 1. Major steps in a systematic review culminating in grading strength of evidence



Note: Adapted from © G.H. Guyatt, et al. Figure 1. Schematic view of GRADE's process for developing recommendations. J Clin Epidemiol 64 (2001) 385. Used with permission.

EPC and GRADE Approaches to Evaluating Evidence

The EPCs' strength of evidence approach is based in large measure on the approach developed by the GRADE working group for assessing evidence.[9-25] Although numerous grading systems have been available over the years,[26-28] the GRADE system has been widely used. EPCs recommend, consistent with GRADE, relying on ratings of specific domains and aggregating domain information into a single overall grade.[29] This update incorporates advice from members of the GRADE working group, information from their explanatory series of articles, and EPCs' experience in applying the original EPC guidance and recommendations.[13-25]

Differences in the specific guidance to EPCs and GRADE users involve some terminology, purposes of grading evidence, and characteristics of domains. As to the lexicon, EPCs refer to the assessment of *strength* of evidence, whereas GRADE refers to *quality* of evidence. Historically, EPCs referred to the evaluation of individual studies as quality assessment; EPCs have generally shifted in practice and terminology to assessing risk of bias.[30] In either case, EPC terminology was intended to distinguish rating specific studies from assessing a body of evidence. GRADE refers to risk of bias at the individual study level and in relation to a body of evidence. Finally, EPCs refer to three of the domains as directness, consistency, and precision; GRADE uses the terms indirectness, inconsistency, and imprecision.

The GRADE approach for systematic reviewers who are assessing the quality of evidence is often intended to complement activities of guideline developers who are also using a GRADE approach to look across outcomes to assess the strength of their recommendations; it assumes a close partnership between the two efforts.[13] In contrast, EPCs grade the strength of evidence only for individual outcomes and not across outcomes; EPCs do not themselves make or grade clinical recommendations. On any given systematic review, EPCs may work with a quite diverse body of end-users (policymakers, administrators, health professionals, advocacy groups, and patients)—even audiences of which they may not be aware at the start of a given review. They expect that end-users can and will make their own global summary judgments of relative benefits and harms across treatment comparisons.

EPCs consider applicability of the evidence explicitly but separately from strength of evidence in their reviews, so as to provide clear, direct descriptions to disparate sets of potential users.[31] The GRADE approach considers applicability as a part of the indirectness domain; reviewers using the GRADE approach typically have an identified target audience and can assess evidence against a specific target situation.[11]

Consistent with the Cochrane Risk of Bias tool for individual trial reports, the GRADE guidance recommends assessing outcome reporting bias within the domain of study limitations; it assesses publication bias as a separate domain.[20,32] EPC guidance newly directs EPCs to assess selective outcome reporting and publication bias within the single domain of reporting bias. No matter what the precise origin of the components of reporting bias, the risk of such bias lowers confidence that the evidence in the review reflects the true effect of a given intervention on an outcome of interest.

Overall, EPC and GRADE guidance both emphasize applying a structured, transparent method. The GRADE working group has developed detailed guidance in many areas, created software to conduct this task, and offer numerous examples on how to conduct the assessment, including when to upgrade or downgrade to reach a final quality of evidence rating.[9-25] A complete listing of the GRADE guidance series can be found at www.gradeworkinggroup.org/publications/index.htm.

Similarly, an EPC's final grades should reflect a reasoned weighting of domain ratings. Within that framework, this updated guidance addresses some particular challenges that commonly arise in EPC reviews. EPCs often need to assess evidence from both trials and observational studies in evaluating a single outcome. They frequently encounter substantial heterogeneity in populations, interventions, or outcomes that may preclude conducting meta-analyses. The approach to synthesis in such circumstances is necessarily qualitative (i.e., narrative, based on reasoned judgment, rather than based on statistical inference).

A Priori Determinations Required in the EPC Approach for Grading Strength of Evidence

Selecting Outcomes

Systematic reviews can be broad in scope, encompassing multiple patient populations, interventions, and outcomes. Because assessing strength of evidence can be labor intensive, especially when the combinations of comparisons and outcomes are numerous, 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 considerable importance to most users of the report. This decision contrasts with the Institute of Medicine recommendation in favor of assessing each outcome for strength of evidence,[33] but it is consistent with the GRADE approach.

We recommend that EPC authors identify a priori (in protocols) the major outcomes that they intend to grade and specify these core elements in analytic frameworks accompanied by an explanation for their choices in text. 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 and policy context, and the purpose of the report.

EPCs make these choices considering the input of key informants, including patients, during the topic refinement phase of the project[34] and subsequently through input from members of a technical expert panel (TEP). The final choices of outcomes rests on several considerations: the important needs that key informants, TEP members, and other end-users have expressed; the ultimate scope of the review (as reflected, for instance, in key questions); and the reliability, validity, responsiveness and other attributes of the outcome measures under consideration.

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."[35] 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."[36, p.15] Patient-centered health outcomes may include reductions in mortality or disease severity and improvements in patient-reported outcomes such as health-related quality of life; they may also involve known or potential harms, such as occurrences of serious and troubling adverse events and inconveniences.

An analytic framework can help to distinguish between these patient-centered, clinically important outcomes and intermediate outcomes. In some cases, EPCs may decide to grade intermediate outcomes that have clear, strong associations with health outcomes or that are, in

and of themselves, important to patients or other target users of the report (e.g., blood pressure control, cholesterol levels, adherence to treatment, or knowledge about an illness).

Specifying Study Eligibility

EPCs establish which studies will be eligible to answer the review questions.[37] Eligibility criteria will reflect the scope of the review but may take account of study design considerations. Sometimes EPCs may determine a priori (in protocols) that, even if a study might have met other inclusion criteria, some aspect of the study's design or execution was so flawed that it could not contribute meaningfully to the body of evidence. For instance, such studies may have very high attrition or high differential attrition, or they may use invalid or unreliable measures for a major outcome. When EPCs make such judgments, they may exclude such studies from the strength of evidence assessment and the review overall. Taking this stance is more likely for evaluating benefits than examining harms. Regardless of the types of decisions that EPCs might make about study eligibility, they should establish a priori criteria to identify studies with particular design elements that would constitute an unacceptably high risk of bias; they must also clearly state their rationale for these decisions.[30]

Specifying Procedures and Decision Rules

EPCs should decide a priori how they will ensure the accuracy and consistency of evidence assessments. For example, they should plan for specific steps to promote reliability and transparency in the whole process (i.e., in scoring individual domains and in using each domain to derive an overall strength of evidence grade). They should devise ways to identify and deal with disagreements among reviewers within a given review team. 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.[38]

We suggest that at least two reviewers with training in these methods independently score domains and determine final grades; in reaching final grades, at least two of the reviewers should be senior authors. Approaches to resolving disagreements in domain scores or final grades include invoking a third, senior author and consensus discussions that include senior authors or the EPC's leadership.

Finally, integrating individual domains into an overall strength of evidence grade is a considerable challenge. EPCs should describe their process for determining their overall strength of evidence assessment; steps include adopting a starting point and applying each domain score in upgrading or downgrading from that starting point. They should note how they will combine randomized controlled trial (RCT) and observational study bodies of evidence.

Major Steps in Grading Strength of Evidence

Scoring Domains: General Considerations

EPCs must 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 required domains are those in the EPC Program's original guidance: study limitations (previously named risk of bias), directness, consistency, and precision. The fifth required domain is reporting bias; it was previously an "additional" domain, limited to publication bias; now it also includes outcome reporting and analysis reporting bias. A set of three additional, but not required, domains are most relevant to bodies of evidence consisting of observational studies: dose-response association, plausible confounding, and strength of association (i.e., magnitude of effect). All are discussed in more detail below.

To score the first four required domains, EPCs evaluate the body of evidence that reports each outcome of interest. EPCs assess the fifth domain, reporting bias, when strength of evidence is high, moderate, or low based on the first four domains. In other words, evidence deemed insufficient is not scored on this domain. To score this fifth domain, EPCs need to identify whether additional evidence has *not* been reported either because entire studies have not been published or because included studies have not reported planned outcomes. Another Methods Guide chapter provides further direction on assessing reporting bias.[39]

For each outcome and intervention (or intervention comparison) of interest, EPCs should develop domain scores and strength of evidence grades *separately* for RCT evidence and observational study evidence when both contribute to evidence synthesis. We discuss considerations about when and how best to combine these separate bodies of evidence into one overall strength of evidence grade below.

The set of five required domains comprises the main constructs that EPCs should use for all major outcomes and comparisons of interest. As briefly defined in Table 1, these domains represent related but separate concepts, and each is scored independently. The concepts are explained in more detail in text.

Table 1. Required domains: definitions and scores

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: Whether RCTs or other designs such as nonexperimental or observational studies. • Study conduct. Aggregation of ratings of risk of bias of the individual studies under consideration.	Score as one of three levels, separately by type of study design: • **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 a health outcome of specific importance for the review, and (b) for comparative studies, whether the comparisons are based on 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: • The outcome being graded is considered intermediate (such as laboratory tests) in a review that is focused on clinical health outcomes (such as morbidity, mortality). • Data do not come from head-to-head comparisons but rather 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 direct comparisons of A vs. B. • 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 are capable of self-reporting and self-report is more reliable. Indirectness always implies that more than one body of evidence is required to link interventions to the most important health outcome.	Score as one of two levels: • **Direct** • **Indirect** If the domain score is indirect, EPCs should specify what type of indirectness accounts for the rating.

Table 1. Required domains and their definitions (continued)

Domain	Definition and Elements	Score and Application
Consistency	Consistency is the degree to which included studies find either the same direction or similar magnitude of effect. EPCs can assess this 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 minimally important difference [MID]) • Magnitude of effect: The range of effect sizes is similar. EPCs may consider the overlap of CIs when making this evaluation. The importance of direction vs. magnitude of effect will depend on the key question and EPC judgments.	Score as one of three levels: • **Consistent** • **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)."
Precision	Precision is the degree of certainty surrounding an effect estimate with respect to a given outcome, based on the sufficiency of sample size and number of events. • A body of evidence will generally be imprecise if the optimal information size (OIS) is not met. OIS refers to the minimum number of patients (and events when assessing dichotomous outcomes) needed for an evidence base to be considered adequately powered. • If EPCs performed a meta-analysis, then EPCs may also consider whether the CI crossed a threshold for an MID. • If a meta-analysis is infeasible or inappropriate, EPCs may consider the narrowness of the range of CIs or the significance level of p-values in the individual studies in the evidence base.	Score as one of two levels: • **Precise** • **Imprecise** A precise estimate is one that would allow users to reach a clinically useful conclusion (e.g., treatment A is more effective than treatment B).
Reporting Bias	Reporting bias results from selectively publishing or reporting research findings based on the favorability of direction or magnitude of effect. It includes: • Study publication bias, i.e., nonreporting of the full study. • Selective outcome reporting bias, i.e., nonreporting (or incomplete reporting) of planned outcomes or reporting of unplanned outcomes. • Selective analysis reporting bias, i.e., reporting of one or more favorable analyses for a given outcome while not reporting other, less favorable analyses. Assessment of reporting bias for individual studies depends on many factors–e.g. availability of study protocols, unpublished study documents, and patient-level data. Detecting such bias is likely with access to all relevant documentation and data pertaining to a journal publication, but such access is rarely available. Because methods to detect reporting bias in observational studies are less certain, this guidance does not require EPCs to assess it for such studies.	Score as one of two levels: • **Suspected** • **Undetected** Reporting bias is suspected when: • Testing for funnel plot asymmetry demonstrates a substantial likelihood of bias, And/or • A qualitative assessment suggests the likelihood of missing studies, analyses, or outcomes data that may alter the conclusions from the reported evidence. Undetected reporting bias includes all alternative scenarios.

CI = confidence internal; EPC = Evidence-based Practice Center; MID = minimally important difference; OIS = optimal information size

Study Limitations Domain

Definition

Scoring the study limitations domain is the essential starting place for grading strength of the body of evidence. It refers to the judgment that the findings from included studies of a treatment (or treatment comparison) for a given outcome are adequately protected against bias (i.e., have good internal validity), based on the design and conduct of those studies. That is, EPCs assess the ability of the evidence to yield an accurate estimate of the true effect without bias (nonrandom error).

Scoring

EPCs derive the score for the study limitations domain from their assessment of the risk of bias for each individual study (rated high, moderate, or low) based on guidance in another Methods Guide chapter.[30] EPCs consider differences in concerns about risk of bias that are based on study design by separately scoring bodies of evidence for two main designs (i.e., RCTs and observational studies). Then, for a particular outcome or comparison *within* each study design group, EPCs assign one of three levels of aggregate risk of study limitations based on study conduct; the scores are low, medium, or high.

Combining evidence from studies with a high risk of bias and those with less risk can be problematic. In particular, 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 the consistency in findings between the bodies of evidence. If results are inconsistent, EPCs should assess whether differing levels of risk of bias explain this inconsistency they should then determine whether combining these bodies of evidence may obscure the findings from evidence rated either low or moderate risk of bias. For example, a body of observational studies in an evidence base may have a high risk of bias; thus, combining them with a body of RCTs of low or moderate risk of bias could inappropriately lower the strength of evidence assessment and obscure the findings for a major outcome.

To determine which groups of studies to include in the domain score and the final strength of evidence assessment, EPCs can conduct sensitivity analyses involving the high risk-of-bias studies. They can explore whether meta-analytic findings with this subset of studies are systematically different from the findings limited to less biased studies, i.e., whether heterogeneity in study design or conduct can explain inconsistencies. If EPCs conclude that the findings do differ in material ways (with proper documentation of methods, explanation and justification), then they can give greater weight to the lower risk-of-bias studies or limit their final synthesis to these studies.[37] EPCs should describe clearly how they derived the score for this domain when some individual studies have high risk of bias but others have low or moderate risk of bias. They should also be sure to discuss in results the reasons that they assigned high risk-of-bias ratings to these studies and how they decided whether these studies did (or did not) contribute to the domain score, overall findings, and strength of evidence. Such high-risk-of-bias studies are still counted as part of the overall evidence base and cited in references.

Directness Domain

Definition

Directness of evidence expresses how closely available evidence measures an outcome of interest. Assessing directness has two parts: directness of outcomes and directness of comparisons. Applicability of evidence (external validity) is considered explicitly but separately from strength of evidence.[31]

Scoring

Directness (of outcomes or of comparisons) is scored as either direct or indirect. Generally, direct evidence for outcomes reflects a single link between an intervention and a patient-centered or clinically important ultimate health outcome, and direct evidence for comparisons requires head-to-head comparisons of interventions. EPCs score outcomes as indirect chiefly when an outcome is intermediate to or a proxy of an ultimate health outcome or when bodies of evidence lack head-to-head comparisons. EPCs should discuss considerations in determining directness in their synthesis of evidence, particularly links between intermediate and final health outcomes.

Directness of Outcomes

The focus of the review determines the evidence that EPCs should consider to be direct. As described earlier, insofar as possible EPCs should identify a priori which outcomes they will grade. In most cases those should be patient-centered or clinically important outcomes. For instance, for a review about treatment for heart disease, myocardial infarction (MI) or quality of life following an MI would be patient-centered outcomes (i.e., direct), whereas low density lipoprotein (LDL cholesterol) level would be considered an intermediate outcome, and in this illustrative review, thus, is indirect.

EPCs may consider some intermediate outcomes important enough to grade the strength of evidence. For example, in the heart disease example, if one key question concerns changes in risk factors for heart disease, EPCs can score the LDL outcomes on directness and consider this evidence direct. If, however, all key questions are limited to ultimate health outcomes of treatment for heart disease, EPCs would view LDL only as an intermediate outcome and consider the LDL evidence only as indirect. If EPCs have no direct evidence whatsoever to answer a key question regarding an ultimate outcome, then they may want to consider use of surrogate markers or intermediate outcomes and score them for this domain; such evidence would be considered indirect.

Evidence may also be considered indirect because investigators used proxy respondents to stand in for certain kinds of patients or subjects in measuring the outcome of interest. For instance, investigators may use surrogates (e.g., family members or nurses) to obtain patients' perceptions of their states of health, such as quality of life or measures of symptom improvement. However, when patient self-report is truly not possible, such as from infants or the cognitively impaired, EPCs may consider such data from proxy respondents to be direct.

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 a "usual care" (or similar) group but not specifically with a comparator intervention of interest, then the evidence is indirect.

EPCs can use separate bodies of evidence (e.g., A versus placebo, B versus placebo, and C versus placebo) to estimate indirectly the comparative effectiveness of the interventions. As an example, 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.[40] This evidence is considered indirect for making comparisons of one antipsychotic with another. Mixed treatment comparisons should be considered indirect (i.e., when the model combines direct and indirect evidence). Detailed guidance on indirect comparisons for EPCs has been reported previously.[41,42]

Consistency Domain

Definition

Consistency refers to the degree of similarity in the direction of effects or the degree of similarity in the effect sizes (magnitudes of effect) across individual studies within an evidence base. EPCs may choose which of these two notions of consistency (direction or magnitude) they are scoring; they should be explicit about this choice.

Scoring

Categories

The consistency of a body of evidence is scored using one of three categories: consistent, inconsistent, and consistency unknown. These categories apply for both direction of effect or magnitude of effect.

Some bodies of evidence may show consistency in the direction of effect but inconsistency in the magnitude of effect sizes. In such cases, EPCs would judge the evidence as consistent or inconsistent based on the choice they have made about grading direction or magnitude of effect in answering a key question.

Judging Direction of Effect (or Equivalence)

EPCs are most often judging consistency in evidence of superiority of one treatment over another. This is appropriate when comparing two interventions or an intervention with placebo or usual care. They look for consistency in direction of effect estimates in relation to the line that distinguishes superiority from inferiority (odds ratio [OR] or risk ratio [RR] =1.0 or absolute difference = 0). CIs may provide additional information on the consistency of the direction of effect in the body of evidence. For example, if all studies except one show estimates of effect in the same direction, but the CI for that one study overlaps the CIs for the estimates of effect in the other studies, then this body of evidence may still be considered consistent.

In contrast to superiority, EPCs may look for evidence to support noninferiority or equivalence when comparing two different interventions with each other. In distinguishing between superiority and equivalence, the EPC must define a line of difference in relation to a threshold; this is referred to as the minimally important difference (MID).[34] The MID is a clearly defined and justified clinical threshold below which EPCs would consider the evidence (effect estimates and corresponding CIs) to show no meaningful difference, and above which EPCs

would consider the evidence to show a benefit or harm of one treatment over another treatment or placebo. For example, EPCs can judge studies as consistent and find no meaningful difference between treatments when all estimates are between thresholds of an explicitly defined MID (e.g., between −0.75 and +1.25 for dichotomous outcomes).

Optimally, MID thresholds are based on empirical evidence or published guidelines. When such evidence is not available, then EPCs can use the consensus of the review team with input from clinical experts. Ideally, MIDs are determined a priori, but they may be established post hoc if necessary. In either case, EPCs should explicitly define meaningful clinical thresholds (and the rationale for them) in the methods section of the review.

Determining MIDs is not always possible. For example, studies in a review may use a variety of scales to measure the same outcome, and those scale scores may not have been calibrated or cross-walked against each other. Moreover, some or all of such scales may not have been subjected to reliability or validity testing. Thus, EPCs may not be able to determine a meaningful threshold across scales with different measurement properties. EPCs can find additional discussion concerning MIDs in the EPC guidance chapter on assessing equivalence and noninferiority.[43]

Judging Magnitude of Effect (and Heterogeneity)

EPCs judge consistency in the magnitude of effect by determining the degree to which point estimates are similar across studies. EPCs can consider studies to be consistent when the CIs of individual studies overlap a summary effect estimate calculated from a meta-analysis. When meta-analysis results are unavailable, EPCs will need to rely on the reviewers' judgment.

Substantial unexplained differences (heterogeneity) across studies may suggest the need for caution in estimating a summary magnitude-of-treatment effect. When EPCs can explain heterogeneity (e.g., a priori determined differences attributable to populations, intervention characteristics, comparators, study design, or conduct); they may not need to score the evidence as inconsistent. This may be the case when they can either stratify the evidence by meaningful subgroups, and separately score the magnitude of effect of outcomes for these subgroups; it may also be possible when they can select the most believable effect estimate from among the studies being considered and then adequately explain the difference between it and the results from the remaining studies.[44]

When EPCs cannot explain heterogeneity ahead of time but meta-analysis is appropriate, they can evaluate consistency in magnitude of effect both qualitatively and through statistical tests for heterogeneity (e.g., Cochran's Q test) or the magnitude of heterogeneity (e.g., I^2 statistic[3]). EPCs should not use results from statistical tests as the sole determinant of the presence of inconsistency because of potential problems in their interpretation and lack of statistical power.[45,46] No single measure is ideal, so EPCs need to explore heterogeneity by considering several statistical approaches, differences in effect estimates, and degree of overlap in CIs in individual studies. EPCs can find more detail about evaluating heterogeneity in GRADE guidance on inconsistency.[22]

Judging a Single-Study Evidence Base

Scoring consistency ideally requires an evidence base with independent investigations of the same treatment/outcome comparison in more than one study. EPCs cannot be certain that a single study, no matter how large or well designed, presents the definitive picture of any particular clinical benefit or harm for a given treatment.[47-49] Accordingly, we recommend that EPCs judge the consistency of a single-study evidence base as unknown.

Precision Domain

Definition

Precision is the degree of certainty surrounding an estimate of effect with respect to an outcome. It is based on the potential for random error evaluated through the sufficiency of sample size and, in the case of dichotomous outcomes, the number of events. A precise body of evidence should enable decisionmakers to draw conclusions about whether one treatment is inferior, equivalent, or superior to another.[50,51]

Scoring

Categories

The assessment of the precision of a body of evidence has two categories: precise and imprecise.

Judging Precision

When EPCs have conducted a quantitative synthesis and calculated a pooled estimate through meta-analysis, they can evaluate precision based on the CI from the meta-analysis. If the CI is wide, EPCs must judge whether it is caused by heterogeneity (which may be attributed to inconsistency) or imprecision. If a wide CI can be attributed to unexplained inconsistency in results, EPCs should not score evidence as imprecise as well. For greater details, see the later section on assigning an overall strength of evidence grade.

When a quantitative synthesis is not possible, EPCs must judge precision based on the constituent parts that would have contributed to the CI for the pooled estimate—i.e., the sample size and the assessment of variance within individual studies. EPCs can evaluate sufficiency of sample size relative to the optimal information size (OIS). OIS concerns the minimum number of patients (for continuous outcomes) and events (for dichotomous outcomes) that would be needed to regard a body of evidence as having adequate power. For a given effect size (such as an OR, a RR, or a weighted mean difference), the optimal number of patients derives from standard sample size calculations for a single, sufficiently powered trial. More detail on OIS is available in the GRADE guidance on imprecision.[21] If the OIS criteria are not met, EPCs may score the evidence as imprecise.

After assessing the adequacy of the sample size or events, EPCs must consider whether the potential for random error in individual studies would decrease their confidence in the study findings. In ideal circumstances, EPCs will have measures of variance for the outcomes of interest in the individual studies (e.g., standard deviation, CI), but in some cases they may have only p values. If more precise measures of variance in studies are not reported but the OIS is met, then EPCs may consider the evidence to be precise when studies report significance level of differences between treatments as p values of less than 0.05.

Reporting Bias

Definition

Reporting bias occurs when authors, journals, or both decide to publish or report research findings based on their direction or magnitude of effect.[52,53] Table 2 defines the three main types

of reporting bias that either authors or journals can introduce: publication bias and outcome and analysis reporting bias.

Table 2. Definitions and descriptions of reporting bias

Types of Reporting Bias	Definition	Examples and Implications
Publication	The whole study has been concealed from public access (nonregistration and/or nonpublication) or it will be made accessible only after an initial delay; this is the "file drawer phenomenon" and the "reporting lag time bias," respectively. A variant is purposeful publication of some or all of the study data in obscure platforms or journals.	Data included in the review are more likely to reflect favorable findings than unfavorable findings. For example, significant differences favoring an intervention for efficacy outcomes or nonsignificant differences for harms outcomes are likelier to be reported in study articles than other results.
Selective Outcome Reporting	The study is reported, but one or more of the planned outcomes are not reported and investigators do not provide a reasonable justification.	Data included in the review are more likely to reflect favorable findings than unfavorable findings. For example, significant differences favoring an intervention for efficacy outcomes or nonsignificant differences for harms outcomes are likelier to be reported in study articles than other results.
	Outcome data are reported but the specific outcome itself or the way it was measured was not as planned.	This phenomenon reflects data mining and increased risk for type I error when significant differences may be a chance occurrence rather than a true effect.
Selective Analysis Reporting	Outcome data are reported but they are based on the most favorable of several analyses undertaken; other analyses are suppressed.	This phenomenon includes presenting selective post hoc subgroup analyses, dichotomizing continuous data using a cut-point that gives the most favorable results, reporting more favorable adjusted versus unadjusted analyses, cherry-picking statistical assumptions, and reporting selective time-point analyses from among multiple follow-up points that had been planned.
	Precision of outcome data estimates is incompletely or not reported.	This problem includes presenting a point estimate without measures of dispersion or giving inexact, nonsignificant p-values (e.g., p>0.05)
	The same outcome data are ambiguously reported in multiple study reports.	Authors do not make the copublication status transparent, which may lead to double counting of outcomes data.

Methods to assess reporting bias exist only for RCTs. Further details on approaches to detecting reporting bias may be found in another paper in progress.[39] Observational studies may also be susceptible to reporting bias,[54-57] particularly because studies are generally not registered and lack a priori protocols. No comparable methods exist for assessing reporting bias for these study designs.

Scoring

Categories

The risk of reporting bias is scored as suspected or undetected.

Judging Reporting Bias

To judge the risk of reporting bias in a body of evidence, EPCs may be able to use a quantitative assessment that investigates the "missingness" of outcomes data from small studies

when those findings, if reported, would be either not statistically significant or unfavorable in direction.[58-64] EPCs can test for the impact of unreported data through, for instance, tests for funnel plot asymmetry, a trim and fill method, and selection modeling. When EPCs cannot do quantitative assessments, or in addition to quantitative assessments, they can conduct a qualitative assessment of reporting bias for the body of evidence. A proposed, but untested, decision aid to evaluate the risk of reporting bias provides guidance on taking a cautious approach for testing funnel plot asymmetry and conducting a qualitative assessment of the risk of reporting bias (see Appendix A).

Additional Domains

The second set of domains, which supplement the five required domains, has three components: dose-response association; uncontrolled confounding that would diminish an observed effect (which is referred to here as "plausible confounding"); and strength of association (i.e., large 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 body of evidence. Although these additional domains apply to RCTs, when they are present they can increase the strength of evidence and are, therefore, especially relevant for observational studies.

Table 3 defines these additional domains and ways to score and apply them. EPCs should explain which additional domains they have used in arriving at any overall strength of evidence grade and how they have altered a judgment that had otherwise been 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. Score as one of two levels: • **Present:** Dose-response pattern observed • **Undetected**: No dose-response pattern observed (dose-response relationship not present or could not be determined)
Plausible confounding that would decrease observed effect	Occasionally, in an observational study, plausible confounding 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 when plausible confounding exists that would decrease the observed effect. Score as one of two levels: • **Present**: Confounding factors that would decrease the observed effect may be present and have not been controlled for. • **Absent:** Confounding factors that would decrease the observed effect are not likely to be present or have been controlled for.
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 when the effect size is particularly large. Score as 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

EPCs define applicability as "the extent to which the effects observed in published studies are likely to reflect the expected results when a specific intervention is applied to the population of interest under "real-world' conditions." [31, p.2] Because of the broad target audiences of EPC reports, EPCs have chosen to make judgments about applicability explicit and separate from assessments of strength of evidence. The goal is to enable varied 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 describing applicability for the outcomes and comparisons for which they specify an overall strength of evidence grade. Separate guidance on applicability is available.[31]

Establishing an Overall Strength of Evidence Grade

Four Strength of Evidence Levels

The four levels of grades are intended to communicate to decisionmakers EPCs' confidence in a body of evidence for a single outcome of a single treatment comparison. Although assigning a grade requires judgment, having a common understanding of the interpretation will be useful for helping EPCs as they conduct their own global assessment and for improving consistency across reviewers and EPCs.

Table 4 summarizes the four levels of grades that EPCs use for the overall assessment of the body of evidence. Grades are denoted high, moderate, low, and insufficient. They are not designated by Roman numerals or other symbols. EPCs should apply discrete grades and should not use designations such as "low to moderate" strength of evidence.

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, i.e., another study would not change the conclusions.
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 reaching a conclusion.

Each level has two components. The first, principal definition concerns the level of confidence that EPCs place in the estimate of effect (direction or magnitude of effect) for the benefit or harm; this equates to their judgment as to how much the evidence reflects a 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, 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 for either the benefit or the harm. The designations of high, moderate, and low should convey how confident EPCs would be about decisions based on evidence of differing grades, which can be based on either quantitative or qualitative assessment.

For comparative effectiveness questions, the comparison is typically a choice of either direction (A>B, A=B, A<B) or magnitude (difference between A and B). In some instances assigning different grades regarding the direction and the magnitude of an effect may be appropriate. An example of this situation is when studies consistently find that an intervention improves an outcome (e.g., apnea-hypopnea index is reduced by a statistically significant amount or beyond a minimally important difference), but the degree of heterogeneity about the estimate is high (e.g., range -10 to -46 events/minute; $I^2 = 86\%$).

The importance of the distinctions among high, moderate, and low levels (and the distinction with insufficient strength of evidence) can vary by the type of outcome, comparison, and decisionmaker. EPCs understand that some stakeholders may want to take action only when evidence is of high or moderate strength, whereas others may want to understand clearly the implications of low versus insufficient evidence. Even when strength of evidence is low or insufficient, consumers, clinicians, and policymakers may find themselves in the position of having to make choices and decisions, and they may consider factors other than the evidence from a specific systematic review, such as patient values and preferences, costs, or resources.

Evidence Grade of Insufficient

In some cases, EPCs cannot draw any evidence-based conclusions for a particular outcome, specific comparison, or other question of interest. In these situations, EPCs should assign a grade of insufficient but be specific in text or tables as to why the evidence does not permit 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 EPCs cannot draw any conclusion regarding the effect from the body of evidence.

The first reason that EPCs may conclude that evidence is insufficient is that *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 reason for a grade of insufficient is that evidence on the outcome is too weak, sparse, or inconsistent to permit EPCs to draw any defensible conclusion concerning the effect. 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).

A grade of insufficient may be appropriate when the CI around the estimated effect in a meta-analysis or across the preponderance of evidence in a qualitative assessment is so wide that it includes two incompatible conclusions: that one treatment is clinically significantly better than the other, and that it is worse. This should not be misunderstood to mean that all statistically nonsignificant effects should lead to a grade of insufficient. Instead, EPCs should use the grade of insufficient when the imprecision results in no confidence regarding whether the effect of one intervention is superior, inferior, or equivalent to another.

Evidence based on a single study often warrants a grade of insufficient. Because the evidence includes only one study, consistency is unknown. When combined with a study size too small to meet OIS criteria, the resulting lowering of the precision domain score further reduces the confidence in the finding of that study, often leading the EPC to be unable to estimate an effect, and thus a grade of insufficient.

Incorporating Domains into an Overall Grade

Overview

For each outcome to be graded, EPCs should first score domains and strength of evidence separately for RCTs and observational studies. 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. They then combine those design-specific 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.

The final judgment for combining domains into an overall strength of evidence must weigh the relative importance of each of the domains in relation to the most worrisome uncertainty in the body of evidence. EPCs must clearly describe how the major concerns in each domain did 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 as long as their rationale for grading strength of evidence is clear and adheres to the important general principles in this guidance. The critical requirement is that EPCs explain the rationale for their approach to

grading of strength of evidence and note which domains were important in reaching a final grade.

Starting Point for Grades for RCTs and Observational Studies

Based on study design, RCT bodies of evidence initially start with a provisional grade of high strength of evidence. EPCs might change such an assessment after evaluating study limitations based on how the RCTs actually were conducted and the other domains.

In contrast, evidence based on observational studies is generally assumed to pose a greater risk of having study limitations because of the typically higher risk of bias attributable to a lack of randomization (and inability of investigators to control for critical confounding factors). This usually corresponds to an initial provisional grade of low strength of evidence.

EPCs may move up the initial grade for strength of evidence based on observational studies to moderate when the body of evidence is scored as low or medium study limitations, based on controls for risk of bias through study conduct or analysis. Similarly, EPCs may initially grade the strength of evidence as moderate for certain outcomes such as harms or certain key questions, when observational study evidence is at less of a risk for study limitations because of a lower risk of bias related to potential confounding.

Also, EPCs may well decide that, after assessing the additional domains, the overall strength of evidence of a body of observational studies can be upgraded to moderate (although rarely high).

Focusing the Strength of Evidence Assessment on Subsets of Studies

Based on reasonable standards of evidence for the subject area, EPCs may adopt a "best evidence" approach. That is, they may focus their assessment of strength of evidence on the subset of studies that provide the least limited, most direct, and most reliable evidence for an outcome or comparison, after analysis of all the evidence. EPCs may want to specify a dichotomy to define the best evidence subset; examples include active-controlled versus placebo-controlled, randomized versus nonrandomized, prospective versus retrospective, or lower risk of bias versus high risk of bias. For example, when EPCs locate a reasonable number of studies of head-to-head comparisons of important alternatives (i.e., Drug A versus Drug B), they are likely to elect not to use placebo-controlled comparisons (Drug A versus placebo, Drug B versus placebo) in their summary estimate of effect. This means that they also would not use the placebo-controlled comparisons in developing their summary findings and their strength of evidence grading.

EPCs may choose to determine an appropriate subset of studies for presenting review findings and strength of evidence assessment by conducting an analysis with and without the problematic studies (such as with a sensitivity analysis)[37,65] and consider which results are most valid and informative. No matter the criteria they use, EPCs must clearly identify studies that met their inclusion criteria and included in the review but did not use in the strength of evidence assessment.

Special Considerations Incorporating Consistency and Precision Domains into Overall Grades

Consistency and precision can be particularly challenging domains to use in reaching an overall strength of evidence grade. When consistency is unknown, EPCs may appropriately lower the overall strength of evidence. Scoring consistency becomes more challenging when

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 in data precludes not only statistical testing of heterogeneity but also qualitative assessment of consistency based on an examination of CIs. Even when the effect sizes appear to be generally consistent across directions or estimates of effect, EPCs cannot determine whether all CIs from the individual studies are above a threshold of no difference. In this case consistency may be uncertain, and EPCs' reviewers must use their judgment to decide whether lowering the grade 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 differed enough to make reviewers 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, EPCs cannot always determine whether the evidence base is truly inconsistent and whether lowering the grade is appropriate. Although precision may also be unknown in this example, an EPC would lower the grade no more than once (i.e., downgrade for unknown consistency or imprecision, but not both).

In many instances, in a body of evidence with estimates of effect that appear imprecise, EPCs may find it difficult to distinguish whether the evidence is inconsistent as well. The main reasons are that (a) the same measures are often used to assess both precision and consistency and (b) the underlying statistical model used in a meta-analysis may have incorporated measurement of both random error and heterogeneity. In meta-analyses with wide CIs, EPCs can examine whether most of the uncertainty can likely be attributed to inadequate sample size and random error (the OIS may be an indicator) or whether it arises mostly because of the heterogeneity in results. We recommend that when a meta-analysis has wide CIs that permit different interpretations, EPCs attribute the uncertainty to either imprecision or inconsistency and lower the grade only once unless they can justify otherwise.

Transparency: Documenting and Reporting Strength of Evidence

Overview

In arriving at an overall strength of evidence grade, a 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. Being explicit and transparent about what steps and criteria are used to arrive at a final strength of evidence grade is the essential element.

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 and underlying reasoning that were employed. Important considerations include how EPCs incorporated different study designs and studies with high risk of bias into the strength of evidence grading, how they weighted each of the required domains in assigning the grade for each outcome, and which additional domain was assessed (if any). For the sake of consistency across reviews and EPCs, EPCs should define the domains using the terminology presented in this paper.

EPCs should present information about all comparisons of interest for the outcomes that are most important to patients and other decisionmakers. Obtaining complete and perfect information is not an achievable goal. For some treatments, data may be lacking about one or more major outcome. In other cases, available evidence comes from studies that have important

flaws or is imprecise. For these reasons, EPCs should present explanations of their findings that will help decisionmakers judge the influence of study limitations on the estimates of effect, taking imprecision and other factors into account.

We 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; they can place supporting details in appendices. However, when a decision is complex or may appear counterintuitive, EPCs should explain it in the text. The placement and presentation of information should emphasize usability and readability of the document overall.

Tables

Much of the information (domain scores and overall strength of evidence) is presented in tables. Table 5 illustrates the suggested approach to providing actionable information to decisionmakers. We recommend that Table 5 or a comparable table—or a suite of tables, depending on the complexity of the review—summarizing key findings and strength of evidence grades be included in the main report. 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[a]

Outcome	Study Design[b]: No. Studies (N)	Findings and Direction [Magnitude] of Effect	Strength of Evidence
Major outcomes			
Mortality	RCT: 1 (56)	A single small RCT with medium study limitations and poor precision found no significant difference in mortality at 1 year.	Insufficient
Severity of [disease]	RCT: 3 (110)	Studies with medium-level study limitations found consistent but imprecise effects on disease severity measured through a range of specific outcomes. RRs ranged from 1.1 (0.75, 1.8) to 3.2 (1.8, 5.7). Outcome assessments were conducted at 1 month to 5 years. Overall, intervention A reduced the severity of [disease] more than intervention B	Low (improved Severity of [disease])
Other patient-centered outcomes			
Pain	RCT: 6 (160)	RCTs with medium study limitations all found that X reduced pain more than Y, between 3 months and 2 years. Summary SMD was 0.5 (0.2, 0.8), but inconsistency in the magnitude of effect was considerable. SMD estimates ranged from 0.13 to 0.94.	Moderate (reduced pain) Low (0.5 difference in pain reduction)
Sexual dysfunction	RCT: 3 (85)	Few studies, only in men. Results were consistent that treatment improves sexual dysfunction at 3 months, but imprecise.	Low (improved dysfunction)
Intermediate outcomes			
LDL cholesterol	RCT: 8 (212)	Small studies yielded a summary net change of -2.1% (95% CI -4, -0.1) with a wide (imprecise) CI.	Low (decreased cholesterol by 2.1%)
Radiology test	RCT: 0	No eligible studies	Insufficient
Adverse events			
Intestinal perforation	RCT: 1 (42)	Only a single event was reported in one small RCT.	Insufficient
Weight gain	Observational: 4 (600)	Observational studies with medium study limitations, including controls for some critical confounders, reported consistent effects on weight gain in 3 of 4 studies at 3 months (range: 0.2 to 13.8 kg)	Low (weight gain)

CI = confidence interval; LDL = low-density lipoprotein; kg = kilogram; RCT = randomized controlled trial; RR = risk ratio; SMD = standardized mean difference

[a] See Tables B-1 and B-2 in Appendix B for the full findings and strength of evidence profile.
[b] Other ways of categorizing the study designs may be appropriate, including active-controlled or placebo-controlled, prospective or retrospective.

The important components of Table 5 or a comparable strength of evidence summary table include the following: (a) the outcome (benefit or harm) of interest; (b) the number of contributing studies (in major study design categories) and number of participants; (c) a summary of the scored domains that were most influential in determining the grade; (d) a description of the length of followup; and (e) to avoid undue length in the table, a succinct description of the findings (e.g., direction or magnitude of effect), including summary estimates from meta-analyses, if calculated. Variations on the table design could further emphasize the findings from the comparison, while making clear the major weaknesses found in the evidence as well as the strength of evidence grade. The goal of the summary table is to assist readers in more easily understanding the available evidence for any given outcome or comparison. Tables should not describe findings from individual studies; a strength of evidence grade should always be accompanied by an overall estimate of effect (direction or magnitude).

If EPCs grade evidence for a given outcome or comparison as insufficient for drawing any conclusions, they can streamline the strength of evidence table by omitting that outcome or comparison and describe the insufficient evidence only in the text. This choice may be particularly preferable when the evidence includes a large number of findings that were graded as insufficient (because of how cumbersome the table would then become).

Additional tables that complement Table 5 may be useful to provide additional detail. Appendix B provides examples of two different approaches to providing more detail. Appendix B also presents examples of text that EPCs might use in the body of the report or an appendix to describe how they reached a strength of evidence grade.

We recommend that the title of each table state the intervention comparison being summarized. Based on the best presentation for each review, tables can either include whole topics or be specific to key questions or treatment or intervention comparisons. We believe that readability is enhanced when EPCs divide table outcomes into the following main categories: major, other patient-centered, intermediate, and adverse events. Major outcomes are those that are deemed most important for decisionmaking about the interventions reviewed. These four types of outcomes may overlap to some degree; however, EPCs should determine the outcome category into which they will place all included outcomes, based on discussions with their key informants and TEP members. The exact definitions of the categories and the determination of which outcomes belong in which category will vary for clinical topics and research questions.

Descriptive Explanatory Text

Transparency regarding strength of evidence grades requires EPCs to communicate clearly the finding that is being graded and the confidence they have in the finding. They should emphasize the criteria used to assign a strength of evidence grade; just stating such phrases as "per AHRQ guidance" or "standard practice" is considered inadequate. We recommend that the Methods section of the report include details about how EPCs handled the following steps: risk-of-bias ratings for individual studies; domain scores (e.g., how EPCs evaluated factors such as direction and magnitude of effect, thresholds, statistical heterogeneity, and overlapping CIs), and strength of evidence grades (i.e., approach to grading and what situations would result in one grade versus another, such as low versus insufficient).

We further recommend that EPCs marshal appropriate support for each conclusion they reach. Reviewers need to state clearly what the strength of evidence grade conveys—e.g., low evidence to determine the effect of X on Y—and the rationale for the grade. If EPCs considered one or more factors particularly salient, they should note this point directly. EPCs may present any needed commentary concerning the information in the strength of evidence tables in text or in the table itself (as footnotes). Lastly, when EPCs use evidence from both RCTs and observational studies in developing a final strength of evidence grade, they need to state explicitly in the Methods section the reasons for including both study designs and how they weighted conclusions from the two bodies of evidence.

Clearly articulating other available evidence that EPCs did not grade for strength of evidence and noting its location in the report will allow users to access findings according to their different priorities.

Finally, nothing about this grading chapter implies that EPCs should rely solely on a reductive, single grade of the evidence for explaining their findings and implications of those findings. Rather, in all systematic reviews, EPCs will present "narrative," qualitative synthesis, and that synthesis and the strength of evidence grades should be done in ways that make reviews as accessible and readable for the relevant stakeholder audiences as possible.

Discussion

The EPC program's approach to grading strength of evidence to assess and describe confidence in the review findings is based on an evaluation of a required group of domains that include aggregate study limitations, directness, consistency, precision, and reporting bias. We suggest that when EPCs are making their final determinations, they also consider the interaction among the domains and the unique concerns of the particular body of evidence. In relation to some findings, their confidence may be increased after also considering additional optional domains; magnitude of effect, a dose-response relationship, or uncontrolled confounding that is likely to be decreasing the observed effect.

This guidance to EPCs has drawn extensively from the GRADE approach—i.e., both during the initial conceptual development and subsequently, through incorporation of GRADE guidance and advice and discussion with members of the GRADE working group. Our guidance addresses application of this conceptually similar approach to grading to specific circumstances and experience of the EPC program. Our hope is that the EPCs and GRADE will continue to learn from each other's experiences and explore challenges in applying strength of evidence assessments.

The EPC program produces systematic reviews, but it is not involved directly in development of recommendations or practice 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.

This guidance does not extend to the idea of "combining" strength of evidence grades into a summary judgment that would take multiple outcomes into account simultaneously or that would reflect the tradeoffs between benefits and harms. We recognize that patients, clinicians, or others may wish to see such unitary judgments, but on balance we believe that different users may have distinctive views about how to combine or weight outcomes. With sufficient clarity about what they have done, EPCs can provide the full range of stakeholders with information that they, in turn, can apply in making treatment or other choices.

EPC systematic reviews have often focused on pharmaceutical therapies, for which both efficacy and effectiveness trials[66] are a major source of information. The strength of evidence domains discussed are directly relevant to studies of most drugs, procedures, and other therapeutic interventions.

By contrast, as EPCs increasingly assess diagnostic tests, screening strategies, and health services interventions such as quality improvement and patient safety studies, RCTs may not be a source of much relevant information; studies that are available may have some different methodologic concerns and be challenging to grade. With these types of nontherapeutic intervention questions, the challenge to EPCs is to determine the study design(s) that would be most appropriate to keep scores for the study limitations domain as robust as possible. For example, EPCs may find that particular types of studies, such as interrupted time series, have fewer study limitations than do other types of observational studies. Nevertheless, we caution that changing the criteria used in assessment of the study limitations domain for observational studies be done judiciously. EPCs should consult the separate AHRQ EPC methods guidance for instructions on grading strength of evidence for reviews on medical tests,[67] and future guidance may be necessary for other topics.

This guidance update did not consider or revise the additional optional domains, dose-response relationships, effect of confounding, or magnitude of effect. Of particular note, recent

approaches to evaluating the risk of bias from confounding in individual observational study evidence incorporate assessments of confounding across the body of evidence.[68,69] Experience with these approaches in evaluating risk of bias are likely to provide additional insights about evaluating confounding in bodies of evidence and may lead to future guidance revisions.

Conclusions

A consistent approach for grading the strength of evidence—one that decisionmakers can readily recognize and interpret—is highly desirable. To that end, EPCs will continue to refine and improve grading systems to be most applicable and useful for different types of reviews. 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. The key points include:

1. Assessing the strength of evidence is meant to communicate to end-users of systematic reviews EPCs' confidence in specific outcome findings of a given review.
2. EPCs should be clear what finding the strength of evidence grade is associated with—i.e., either a direction of effect or a summary estimate of effect.
3. Figure 1 defines the eight steps in assessing a body of evidence. This guidance focuses primarily on steps 5 through 8, which concern developing findings and reporting on individual outcomes. Tasks include scoring component domains (study limitations, directness, consistency, precision, and reporting bias, plus three additional optional domains that are more likely to be relevant when assessing observational studies) and combining the scores into an overall strength of evidence grade.
4. EPCs should strive to be transparent in their assessments and judgments at each stage of the process—from assessing individual domains to combining the domains into an overall strength of evidence grade.
5. EPCs score and initially grade RCT bodies of evidence separately from nonrandomized bodies of evidence. The final strength of evidence grade combines the two bodies of evidence.
6. When combining bodies of evidence with differing levels of study limitations, EPCs should consider all evidence, but they may ultimately choose to weight studies with lower risk of bias more heavily in the final analysis. They should describe clearly how all evidence was considered, but they may focus their presentation on the evidence that contributed most to the findings and on their confidence in those findings.

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. 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 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. Centre for Evidence Based Medicine. Oxford Centre for Evidence-based Medicine - Levels of Evidence (March 2009). Oxford, UK: University of Oxford; 2012. www.cebm.net/?o=1025. Accessed November 8, 2012.
7. The National Health and Medical Research Council (NHMRC) in Australia. NHMRC additional levels of evidence and grades for recommendations for developers of guidelines. Australia: National Institute of clinical Studies Officers of the NHMRC. www.nhmrc.gov.au/_files_nhmrc/file/guidelines/developers/nhmrc_levels_grades_evidence_120423.pdf. Accessed November 8, 2012.
8. Scottish Intercollegiate Guidelines Network. Implementing Grade. Scotland: Healthcare Improvement Scotland. www.sign.ac.uk/methodology/index.html. Accessed November 8, 2012.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.
14. 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.
15. Guyatt G, Thorlund K, Oxman AD, et al. GRADE guidelines 13. Preparing Summary of Findings (SoF) Tables and Evidence Profiles – continuous outcomes. In process.
16. 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.
17. 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.

18. 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.

19. 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.

20. 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.

21. 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.

22. 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.

23. 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.

24. 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.

25. 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.

26. 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.

27. Moroni A, Olmi R, Vicenzi G. [PMMA reinforced with pulverized hydroxyapatite crystals. Biomechanical characteristics of articular prosthetic implants. Preliminary results]. Chir Organi Mov. 1981 May-Jun;67(3):321-7. PMID: 7052350.

28. Carande-Kulis VG, Maciosek MV, Briss PA, et al. Methods for systematic reviews of economic evaluations for the Guide to Community Preventive Services. Task Force on Community Preventive Services. Am J Prev Med. 2000 Jan;18(1 Suppl):75-91. PMID: 10806980.

29. Falck-Ytter Y, Schunemann H, Guyatt G. AHRQ series commentary 1: rating the evidence in comparative effectiveness reviews. J Clin Epidemiol. 2010 May;63(5):474-5. PMID: 20189352.

30. 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/.

31. 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.

32. Meerpohl JJ, Langer G, Perleth M, et al. [GRADE guidelines: 4. Rating the quality of evidence - limitations of clinical trials (risk of bias)]. Z Evid Fortbild Qual Gesundhwes. 2012;106(6):457-69. PMID: 22857734.

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

34. Whitlock EP, Lopez SA, Chang S, et al. AHRQ series paper 3: identifying, selecting, and refining topics for comparative effectiveness systematic reviews: AHRQ and the Effective Health-Care Program. J Clin Epidemiol. 2010 May;63(5):491-501. PMID: 19540721.

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

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

37. Treadwell JR, Singh S, Talati R, et al. A framework for best evidence approaches can improve the transparency of systematic reviews. J Clin Epidemiol. 2012 Nov;65(11):1159-62. PMID: 23017634.

38. 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.

39. Finding Evidence and Assessing for Reporting Biases when Comparing Medical Interventions: AHRQ and the Effective Health Care Program. Draft Methods Guidance (Prepared by the University of Ottawa and the Oregon Health and Science University Evidence-based Practice Centers). Rockville, MD: Agency for Healthcare Research and Quality; August 2012. http://effectivehealthcare.ahrq.gov/ehc/products/486/1305/Reporting-Bias_DraftReport_20121023.pdf.

40. 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. 290-2007-10062- 1.) Comparative Effectiveness Review No. 43. Rockville, MD: Agency for Healthcare Research and Quality; September 2011. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

41. 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. Rockville, MD: Agency for Healthcare Research and Quality (US); Oct 25 2010.

42. 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.

43. Treadwell J, Uhl S, Tipton K, et al. Assessing Equivalence and Noninferiority. Methods Research Report. (Prepared by the EPC Workgroup under Contract No. 290-2007-10063.) AHRQ Publication No. 12-EHC045-EF. Rockville, MD. Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov.

44. Gartlehner G, West SL, Mansfield AJ, et al. Clinical heterogeneity in systematic reviews and health technology assessments: synthesis of guidance documents and the literature. Int J Technol Assess Health Care. 2012 Jan 5:1-8. PMID: 22217016.

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

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

47. Ioannidis JP. Mega-trials for blockbusters. JAMA. 2013 Jan 16;309(3):239-40. PMID: 23321760.

48. Shrier I, Platt RW, Steele RJ. Mega-trials vs. meta-analysis: precision vs. heterogeneity? Contemp Clin Trials. 2007 May;28(3):324-8. PMID: 17188025.

49. Charlton BG. Mega-trials: methodological issues and clinical implications. J R Coll Physicians Lond. 1995 Mar-Apr;29(2):96-100. PMID: 7595900.

50. 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.

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

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

53. 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.

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

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

56. 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.

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

58. 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.

59. 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.

60. 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.

61. 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.

62. 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.

63. 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.

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

65. Norris SL, Atkins D, Bruening W, et al. Observational studies in systemic reviews of comparative effectiveness: AHRQ and the Effective Health Care Program. J Clin Epidemiol. 2011 Nov;64(11):1178-86. PMID: 21636246.

66. 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.

67. 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.

68. Viswanathan M, Berkman ND, Dryden DM, et al. Assessing Risk of Bias and Confounding in Observational Studies of Interventions or Exposures: Further Development of the RTI Item Bank.. Methods Research Report. (Prepared by RTI–UNC Evidence-based Practice Center under Contract No. 290-2007-10056-I). AHRQ Publication No. 13-EHC106-EF. Rockville, MD: Agency for Healthcare Research and Quality; August 2013. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

69. Sterne J. Conversations concerning development of a new Cochrane Collaboration observational study risk of bias instrument. Personal communications with Berkman N; 2013.

Appendix A. A Tool for Evaluating the Risk of Reporting Bias

This appendix presents a conceptual framework and flow diagram (Figure A-1) that Evidence-based Practice Centers (EPCs) might use to assess the risk of reporting bias for a body of evidence for an outcome of interest. This is the fifth of the five required domains that EPCs are likely to need to score in grading strength of evidence. EPCs rate this domain as either "undetected" or "suspected."

Reporting bias, in this case, encompasses publication bias (i.e., not publishing a study whatsoever), outcome reporting bias (i.e., selectively reporting some but not all planned outcomes), and selective analysis reporting (i.e., selectively reporting only more favorable analyses from among all planned analyses). Reporting bias is defined and described in greater detail in Table 2 of the main text.

The framework considers both quantitative and qualitative assessments of reporting bias. Its use is intended to assist EPCs in reaching judgments, enhance standardization across EPCs, and promote transparency of their work, such that readers can see how EPCs reached judgments about reporting bias. The algorithm (in the figure) has not yet been tested in the context of conducting a systematic review; we would expect it to be modified based on EPC experience and feedback in the future.

This tool is intended to apply chiefly to evidence bases consisting of randomized controlled trials (RCTs). It is less relevant to nonexperimental or observational studies because of the difficulties of determining reporting bias for such studies. Methods for detecting such bias are (as of this writing) uncertain and unproven, particularly because such studies typically are not based on published or registered protocols. Although EPCs may assess the risk of reporting bias for observational evidence, the guidance offered in the main chapter does not require it.

Conceptual Framework and Steps in Using the Tool

Quantitative Assessments

As shown in Figure A-1, for each outcome of interest, EPCs begin assessing risk of reporting bias by determining whether the evidence lends itself to a quantitative assessment. We posit four main criteria for making this decision: at least 10 studies contribute data for the outcome in question; these studies are of unequal size; smaller and larger studies do not differ substantially in clinical factors or methods; and estimates of effect are accompanied by measures of dispersion.

If these criteria are met, such that a quantitative evaluation is permissible, the flow diagram takes EPCs down the left-hand column. If one or more of these criteria are not met, then EPCs would forego a quantitative evaluation and attempt only a qualitative evaluation instead (moving down the right-hand column of the figure). Because this effort is done for each outcome independently, one result of this first step is that, for some systematic reviews and bodies of evidence in them, EPCs may need to do both quantitative and qualitative assessments of reporting bias.

Assuming that the number of available studies is adequate and that smaller studies (just by visual inspection of findings) show more favorable results than larger studies, then EPCs can proceed with a quantitative evaluation. Specifically, they can test whether funnel plots reflect asymmetry and whether effect estimates from meta-analyses (direction or magnitude of effect)

differ in a meaningful way between smaller and larger studies, depending on whether analyses used a random effects or a fixed effects model.[1,2]

Because larger studies are more likely to be reported than smaller studies irrespective of their findings, nonpublication of less favorable results from smaller studies will result in a fixed effects estimate that is more conservative (i.e., closer to the null) than a random effects estimate. The reason is that a fixed effects model will reflect the estimates from the larger studies more than the smaller studies. If neither clinical nor methodological diversity is associated with study size, the likely explanations for any difference between the two models are study nonpublication or selective outcome reporting. EPCs would assign a rating of "suspected reporting bias" to such a difference.

Funnel plots have relatively serious limitations, however, in detecting reporting bias. On the one hand, when only a few studies constitute a body of evidence, then funnel plot tests may be underpowered. On the other hand, when the number of available trials is large, then the test becomes overly sensitive.[1] Furthermore, a statistically significant finding from a funnel plot test can imply one (or more) of several issues: reporting bias; clinical diversity, methodological diversity, or both, related to study size; or simply chance. Because of these multiple explanations,[2] minimizing alternative explanations is critical. Thus, we recommend that this test be used judiciously with bodies of evidence that meet the criteria specified in Figure A-1 concerning size, clinical and methodological heterogeneity, and estimated effects across studies.[2]

Qualitative Assessments

When a quantitative assessment is not possible or when it does not support a definitive conclusion, EPCs might undertake a qualitative assessment. The right hand column of Figure A-1, plus the seven items in the box at the bottom right, provides the guidance that EPCs can follow, considering the number and risk of bias of studies, the consistency in results, and confidence in the search process.

Timing of Reporting Bias Assessments

A body of evidence that includes many studies of a large number of patients, that reflects few study limitations in the design and conduct of the trials, and that yields relatively consistent effect estimates increases our confidence that a qualitatively or quantitatively synthesized summary estimate of effect is close to the truth. To be certain of that provisional conclusion, however, EPCs should evaluate the domain for risk of reporting bias last, i.e., after consideration of study limitations, consistency, directness, and precision. Rating this domain also assumes that EPCs have already done a reasonably diligent search for unpublished data to supplement published findings.

Scoring Reporting Bias

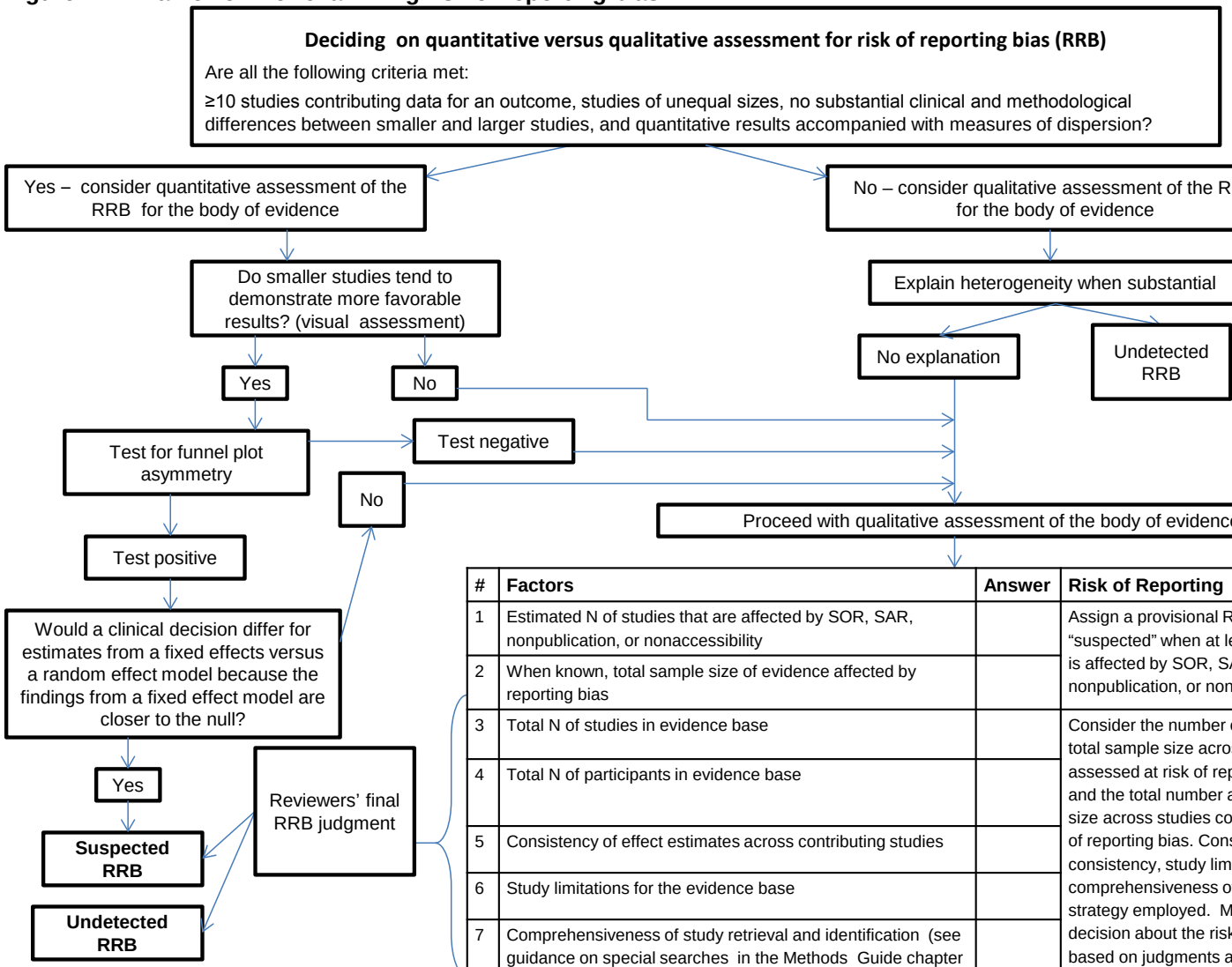
Generally, EPCs could decide that reporting bias is undetected because, in fact, they cannot find any evidence to support suspicions that it exists. In addition, EPCs may initially arrive at a provisional rating of "suspected" reporting bias in a body of evidence for a given outcome based on finding reporting bias in a small number of studies that include only a small proportion of the total patients across studies. They may conclude that this is not important enough to question the validity of the synthesized estimate. In such cases, reviewers may reasonably decide to judge the overall risk of reporting bias for the body of evidence as undetected.

In all other scenarios, EPCs can rate the risk of reporting bias as suspected.

Summary

In summary, EPCs make a provisional assessment of suspected when they identify selective outcome reporting bias, analysis reporting bias, or publication bias for individual studies. In light of the total size of the body of evidence, its internal validity (study limitations), consistency, directness, and precision, as well as the comprehensiveness of the search strategy for the review (see AHRQ's guidance on special searches and reporting bias[3]), reviewers judge the impact of their provisional risk assessment on the outcome results or conclusions associated with the available evidence base. They then develop a final rating for this domain as either suspected or undetected to inform their confidence on outcome results or conclusions.

Figure A-1. Framework for examining risk of reporting bias



#	Factors	Answer	Risk of Reporting Bias
1	Estimated N of studies that are affected by SOR, SAR, nonpublication, or nonaccessibility		Assign a provisional RRB rating of "suspected" when at least one study is affected by SOR, SAR, nonpublication, or nonaccessibility.
2	When known, total sample size of evidence affected by reporting bias		
3	Total N of studies in evidence base		Consider the number of studies and total sample size across studies assessed at risk of reporting bias and the total number and sample size across studies considered free of reporting bias. Consider consistency, study limitations , and comprehensiveness of search strategy employed. Make a final decision about the risk assessment based on judgments about the effect of reporting bias on findings .
4	Total N of participants in evidence base		
5	Consistency of effect estimates across contributing studies		
6	Study limitations for the evidence base		
7	Comprehensiveness of study retrieval and identification (see guidance on special searches in the Methods Guide chapter on individual study assessment of reporting bias)		

Abbreviations: N = number; RRB = risk of reporting bias; SAR = selective analysis reporting; SOR = selective outcome reporting.

Appendix A References

1. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. J Clin Epidemiol. 2000 Nov;53(11):1119-29. PMID: 11106885.

2. Sterne JA, Sutton AJ, Ioannidis JP, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ. 2011;343:d4002. PMID: 21784880.

3. Finding Evidence and Assessing for Reporting Biases when Comparing Medical Interventions: AHRQ and the Effective Health Care Program. Draft Methods Guidance. (Prepared by the University of Ottawa and the Oregon Health and Science University Evidence-based Practice Centers). Rockville, MD: Agency for Healthcare Research and Quality; August 2012. http://effectivehealthcare.ahrq.gov/ehc/products/486/1305/Reporting-Bias_DraftReport_20121023.pdf

Appendix B. Grading Strength of Evidence: Decisionmaking Examples

In this appendix, we present examples of detailed explanatory tables and text that EPCs may include in their systematic reviews. This material illustrates in particular the practice of using a "best evidence" approach in analyzing and synthesizing included studies. The examples are intended to supplement the presentation in Table 5 of the main guidance. We use "Severity of [Disease],"as presented in Table 5, as the outcome example for these tables and text. Tables B-1 and B-2 provide different format options for transparently reporting the score for each domain, the overall findings, and strength of evidence grade. We also present examples of text describing the results and analysis that led to the final conclusions and strength of evidence determination. EPCs can include similar text in either the main body of the report or an appendix.

Tables B-1 or B-2 and Illustrative Text

The footnotes included in the approach presented in Table B-1 are optional. In general, EPCs should use footnotes only when they are short and few. If footnotes would not clearly convey information or would be too numerous, then we recommend that EPCs use a version of Table B-2 instead.

Both tables B-1 and B-2 show a column documenting the size of the evidence used in the strength of evidence assessment: the number of studies of various study designs (e.g., randomized controlled trials [RCTs] and the total sample size (N). When using a best evidence approach, EPCs may use a footnote to document whether they included any studies in the report that did not contribute to the findings and strength of evidence. When documenting the study limitations of the body of evidence, EPCs should record the distribution of studies contributing to the findings and strength of evidence by the number receiving one of the three risk-of-bias assessments for individual studies.[1] Those scores are low, medium, or high.

Table B-1. [Intervention A] vs. [Intervention B] for the treatment of [Disease]: Strength of evidence domains

Outcome Strength of Evidence Grade	Study Design: No. Studies[a] (N)	Study Limitations	Directness	Consistency	Precision	Reporting Bias	Other Issues	Finding
Major outcomes								
Severity of [Disease] Low	RCT: 3 (110)	Medium[b]	Direct	Consistent	Imprecise[c]	Suspected[d]	None	Intervention A reduced the severity of [disease] more than intervention B.;

[a] Five high-risk-of-bias studies did not contribute to the final evidence assessment.
[b] Study limitations: risk-of-bias ratings for individual studies were medium (2 studies) or low (1 study); in general, lack of outcome assessor blinding and high attrition rates were the main concerns.
[c] Precision: evidence sample size did not meet OIS; CI surrounding the risk ratio for one of the three studies crossed 1.0
[d] Outcome reporting bias: inconsistent analyses of single and composite (multiple endpoints combined) outcomes raised concern about biased outcome reporting.
Abbreviations: RCT = randomized controlled trial.

Table B-2. [Intervention A] vs. [Intervention B] for the treatment of [Disease]: Details regarding strength of evidence domains

Outcome Strength of Evidence Grade	Study Design No. Studies[a] (N)	Risk of Bias of Individual Studies	Rating and Reasons for Domain Scores Descriptions of Other Issues Comments About Derivation of Overall Strength of Evidence Finding and Strength of Evidence
Severity of [Disease] Low	RCT: 3 (110)	1 Low 2 Medium	Study limitations: Medium. Unclear assessor blinding in one study; high attrition rates in two studies. Consistency: Consistent. Precision: Imprecise, confidence interval surrounding the risk ratio for one of the studies crossed 1.0. Reporting bias: Suspected. Inconsistent analyses of both single and composite (multiple endpoints combined) outcomes raises concerns. Other concerns: None Intervention A reduced the severity of [disease] more than intervention B.

[a] Five high-risk-of-bias studies did not contribute to the final strength of evidence assessment.

Abbreviations: RCT = randomized controlled trial.

Possible text to accompany Table B-1 and B-2 appears below. Note that this text reflects a best evidence approach that (for this hypothetical example) removed five trials rated as high risk of bias. Taking this approach may cause confusion for some end-users because of differences between either of these tables (on the one hand) and the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram in the main report (on the other hand). EPCs can mitigate the problem by documenting the studies that did and did not contribute to the findings and clearly describing their analyses in the main report.

Strength of Evidence for Severity of Disease

Of eight trials initially addressing the comparison of Intervention A with Intervention B for severity of [disease], three trials provide low strength of evidence that Intervention A reduced severity of [disease] more than Intervention B measured from 1 month to up to 5 years. Of the original eight trials, we considered five studies to be of high risk of bias. They did not contribute to the final conclusions and strength of evidence because including them obscured the conclusions from the three trials of low or moderate risk of bias.

We graded the strength of evidence for this conclusion as low, using the following rationale. Because the evidence consists of RCTs, of direct evidence but medium study limitations, we started with a grade of moderate strength of evidence. We further lowered the grade because of imprecision and the potential for outcome reporting bias, which is important enough to reduce the strength of evidence grade below moderate to low.

Appendix B References

1. 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/