Grading the strength of a body of evidence when assessing health care interventions: an EPC update

Nancy D. Berkmana,a,*, Kathleen N. Lohra, Mohammed T. Ansarib, Ethan M. Balkc, Robert Kane, Marian McDonagh, Sally C. Morton, Meera Viswanathan, Eric B. Bass, Mary Butler, Gerald Gartlehner, Lisa Hartling, Melissa McPheeters, Laura C. Morgan, James Reston, Priyanka Sista, Evelyn Whitlock, Melissa McPheeters, Stephanie Chang

aDivision of Social Policy, Health & Economics Research, RTI International (Research Triangle Institute), PO Box 12194, 3040 Cornwallis Road, Research Triangle Park, NC 27709, USA
bClinical Epidemiology Program, Ottawa Hospital Research Institute, Box 201B, Ottawa Hospital - General campus, 501 Smyth Road, Ottawa, ON K1H 8L6, Canada
cCenter for Evidence-Based Medicine, Brown University, School of Public Health, Box G-S121-8, Providence, RI 02912, USA
dUniversity of Minnesota, School of Public Health, 420 Delaware St SE, MMC 197, Minneapolis, MN 55455, USA
ePacific Northwest Evidence-based Practice Center, Oregon Health Science University, 318 J SW Sam Jackson Park Road, Portland, OR 97239, USA
fDepartment of Biostatistics, Graduate School of Public Health, University of Pittsburgh, 130 DeSoto Street, Pittsburgh, PA 15261, USA
gDepartment of Medicine, and Department of Health Policy and Management, Johns Hopkins University, 624 North Broadway, Room 680A, Baltimore, MD, 21205, USA
hDepartment for Clinical Epidemiology and Evidence-based Medicine, Danube University, Dr. Karl Dorrek Strasse 30, 3500 Krems, Austria
iDepartment of Pediatrics, University of Alberta, ECHA 4-472, 11405-87 Avenue, Edmonton, Alberta, T6G 1C9, Canada
jDepartment of Health Policy, Vanderbilt University, Medical Center, 6th Floor, 2525 West End Avenue, Nashville, TN 37203, USA
kEvidence-Based Practice Center and Health Technology Assessment Group, ECRI Institute Headquarters, 5200 Butler Pike, Plymouth Meeting, PA 19462, USA
lUniversity of North Carolina School of Medicine, 1001 Bondurant Hall, CB# 9535, Chapel Hill, NC 27599, USA
mCenter for Health Research, Kaiser Permanente, Northwest, 3800 N. Interstate Avenue, Portland, OR 97227, USA
nCenter for Evidence and Practice Improvement, Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, USA

Accepted 12 November 2014; Published online 20 December 2014

Abstract

Objectives: To revise 2010 guidance on grading the strength of evidence (SOE) of the effectiveness of drugs, devices, and other preventive and therapeutic interventions in systematic reviews produced by the Evidence-based Practice Center (EPC) program, established by the US Agency for Healthcare Research and Quality (AHRQ).

Study Design and Setting: A cross-EPC working group reviewed authoritative systems for grading SOE [primarily the approach from the Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group] and conducted extensive discussions with GRADE and other experts.

Results: Updated guidance continues to be conceptually similar to GRADE. Reviewers are to evaluate SOE separately for each major treatment comparison for each major outcome. We added reporting bias as a required domain and retained study limitations (risk of bias), consistency, directness, and precision (and three optional domains). Additional guidance covers scoring consistency, precision, and reporting bias, grading bodies of evidence with randomized controlled trials and observational studies, evaluating single study bodies of evidence, using studies with high risk of bias, and presenting findings with greater clarity and transparency. SOE is graded high, moderate, low, or insufficient, reflecting reviewers’ confidence in the findings for a specific treatment comparison and outcome.

Conclusion: No single approach for grading SOE suits all reviews, but a more consistent and transparent approach to reporting summary information will make reviews more useful to the broad range of audiences that AHRQ’s work aims to reach. EPC working groups will consider ongoing challenges and modify guidance as needed, on issues such as combining trials and observational studies in bodies of evidence, weighting domains, and combining qualitative and quantitative syntheses. © 2015 Elsevier Inc. All rights reserved.

Conflict of interest: This article was supported by a contract award from the US Agency for Healthcare Research and Quality (AHRQ; Contract No. HHSA-290-2007-10056-I-EPC3, Task Order #5) to RTI International (the RTI-University of North Carolina Evidence-based Practice Center). The findings and conclusions expressed here are those of the authors and do not necessarily represent the views of AHRQ or the U.S. Department of Health and Human Services.

Funding: None.

* Corresponding author. Tel.: 919-541-8337; fax: (919) 990-8454.
E-mail address: berkman@rti.org (N.D. Berkman).

http://dx.doi.org/10.1016/j.jclinepi.2014.11.023
0895-4356/© 2015 Elsevier Inc. All rights reserved.
1. Introduction

Systematic reviews are essential tools for summarizing a body of literature to help users make well-informed decisions about health care options [1]. The Evidence-based Practice Center (EPC) program, supported by the US Agency for Healthcare Research and Quality (AHRQ), produces many such reviews. Strength of evidence (SOE) assessment is one of the final tasks in conducting a systematic review. The goal is to provide clearly explained, well-reasoned judgments about reviewers’ confidence in their conclusions about effects of interventions so that decisionmakers can use them effectively [2].

AHRQ supported a cross-EPC work group that updated and revised earlier guidance [3] on grading SOE (codified within the larger Methods Guide for Effectiveness and Comparative Effectiveness Reviews [4]) [5]. The recommendations apply primarily to systematic reviews of drugs, devices, and other preventive and therapeutic interventions; they may apply to reviews of other health services research questions such as the effects of exposures (characteristics or risk factors) on health outcomes. The updated guidance aims to (1) foster appropriate consistency and transparency in EPC methods and (2) facilitate users’ interpretations of SOE grades for guideline development or other decision-making tasks.

The EPCs’ SOE approach is based in large measure on widely used methods developed by the Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group [6–22]. GRADE provides helpful detailed guidance concerning many aspects of assessing evidence, offering numerous examples on how to conduct these assessments (see http://www.gradeworkinggroup.org/publications/index.htm).

This updated guidance expands and clarifies earlier guidance on grading SOE and examines challenges that commonly arise in EPCs’ reviews. The latter includes (1) assessing evidence from both randomized controlled trials (RCTs) and observational studies in evaluating a single outcome, (2) dealing with substantial heterogeneity in populations, interventions, or outcomes that may preclude conducting meta-analyses, and (3) presenting findings that can be readily used by a variety of audiences (eg, policymakers, administrators, health professionals, advocacy groups, and patients).

EPCs do not judge the balance between benefits and harms of interventions; they grade SOE only for individual outcomes and not across outcomes. End users can and will make their own global summary judgments of relative benefits and harms across treatment comparisons. EPCs consider applicability of evidence explicitly but separately from SOE grades. The aim is to enable the myriad groups of end users to consider how closely the evidence assists with their decisionmaking [23].

2. A priori determinations required in grading SOE

2.1. Selecting outcomes

EPC reviews can be broad in scope, encompassing multiple patient populations, interventions, outcomes, and study designs; they are thus often long and complex. Reviewers are expected to identify and grade outcomes that are of the greatest importance to most readers. This approach contrasts with the Institute of Medicine recommendation to assess all outcomes for SOE [24], but it is consistent with the GRADE approach.

We recommend that reviewers identify in review protocols both the major outcomes—benefits and harms—and the comparisons that they intend to grade [25]. They should give due consideration to the critical clinical and policy concerns that key informants, including patients, and technical experts suggest. Also important are the attributes of the outcome measures under consideration. Ideally, outcomes that receive SOE grades will be patient-centered—that is, those that “people notice and care about” [26] and that 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” [27] (p. 15).

2.2. Specifying study eligibility

In protocols, reviewers establish which studies will be eligible to answer review questions [28]. Even if a study might have met other inclusion criteria and addressed a question of interest, reviewers may decide that some aspect of the study’s execution (eg, high differential attrition, unreliable outcome measurement) may be sufficiently flawed that it will not contribute meaningfully to the body of evidence. Reviewers may exclude such studies from the entire review or from the SOE assessment even if they use them in some fashion in the overall review. Either way, they need to state clearly their rationale for these decisions [29].

2.3. Specifying procedures and decision rules

Reviewers should take steps to ensure the accuracy and consistency in their approach to SOE assessment. For each outcome and comparison that is to be graded, key tasks include scoring individual domains, using domain scores to derive an overall SOE grade. We advocate dual,
What is new?

This article updates and revises earlier guidance on grading the strength of evidence (SOE) for systematic reviewers from Evidence-based Practice Centers (EPCs) supported by the US Agency for Healthcare Research and Quality (AHRQ) and other review groups.

Key findings

- Reviewers should assess five domains to grade SOE: study limitations, consistency, directness, precision, and reporting bias. Additional domains that may be important for some topics or study designs are dose-response association, presence of confounders that would diminish an observed effect, and strength of association. SOE grades are denoted high, moderate, low, or insufficient.

What this adds to what was known?

- The former rigorous, transparent approach for grading SOE in systematic reviews has been improved through use of recent research on systematic review methods and considerations such as minimally important differences and optimal information size, additional discussions with representatives of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system, and more extensive examples and detailed description of ways to rate domains, combine evidence from randomized controlled trials and observational studies and present SOE grades.

What is the implication and what should change now?

- This newer, more explicit approach will help clinicians, patients, and policymakers interpret levels of evidence about health care interventions and their outcomes; this will in turn enhance their ability to create practice guidelines, engage in shared decision making, and make well-informed health care and policy choices. Systematic reviewers at EPCs should adopt this updated guidance for grading SOE.

3. Major steps in grading SOE: scoring domains

Reviewers must assess (score) a set of “required” domains when grading SOE (Table 1): study limitations (previously named risk of bias), consistency, directness, precision, and reporting bias. 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 (ie, magnitude of effect). Because of concerns about differences in risk of bias arising from study design, EPCs develop domain scores and SOE grades separately for bodies of evidence from RCTs and observational studies.

To score the first four required domains, EPCs evaluate the entire body of evidence that reports on an outcome for a distinct combination of patients and treatments. Because reporting bias can only lower a grade, reviewers need assess this fifth domain only when SOE is high, moderate, or low (but not insufficient) based on the first four domains. Another Methods Guide chapter provides further direction on reporting bias [31].

3.1. Study limitations

Study limitations, the essential starting point, refers to the summary judgment of whether the studies as a whole are adequately protected against bias (ie, have good internal validity) and can yield an accurate unbiased estimate of the true effect. The study limitation domain score is derived from the assessment of risk of bias for each individual study [29]. Within each study design group, this domain receives one of three designations: low, medium, or high.

3.2. Directness

Directness of evidence—how closely available evidence measures an outcome—has two parts: directness of outcomes and directness of comparisons (Table 1). Directness is scored as either direct or indirect.

A review’s key questions determine whether an outcome is considered direct. For instance, for a review about treating patients for heart disease, survival following a myocardial infarction is a direct patient-centered health outcome, whereas low-density lipoprotein (LDL cholesterol) is an intermediate outcome and, in this example, indirect. However, if a key question concerns changes in risk factors for heart disease, then evidence on LDL cholesterol is scored as direct.

Some indirect outcomes may be important enough to warrant grading SOE. Especially when direct evidence is lacking, reviewers may want to consider using surrogate markers or intermediate outcomes; such evidence would be considered indirect.

Outcomes may also be considered indirect when investigators used proxy respondents (eg, family members or nurses) to stand in for certain kinds of patients to obtain perceptions of the patient’s state of health. When patient self-report is truly not possible, such as from infants or
<table>
<thead>
<tr>
<th>Domain</th>
<th>Definition and elements</th>
<th>Score and application</th>
</tr>
</thead>
</table>
| Study limitations | Study limitations are the degree to which the included studies for a given outcome have a high likelihood of adequate protection against bias (ie, 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 results are based on head-to-head comparisons. Reviewers 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) for a key question that is focused on clinical health outcomes (such as morbidity, mortality).  
• Data are available only for proxy respondents (eg, 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.  
• Data do not come from head-to-head comparisons but rather from two or more bodies of evidence to compare interventions A and B—for example, 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.  
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, reviewers should specify what type of indirectness accounts for the rating. |
| Consistency     | Consistency is the degree to which included studies find either the same direction or similar magnitude of effect. Reviewers can assess this through two main elements:  
• Direction of effect: Effect sizes have the same sign (ie, are on the same side of no effect or a minimally important difference (MID)).  
• Magnitude of effect: The range of effect sizes is similar. Reviewers 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 reviewer judgments. | Score as one of three levels:  
• Consistent  
• Inconsistent  
• Unknown (eg, 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 reviewers performed a meta-analysis, then reviewers may also consider whether the CI crossed a threshold for an MID.  
• If a meta-analysis is infeasible or inappropriate, reviewers 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 (eg, treatment A is more effective than treatment B). |

(Continued)
the cognitively impaired, data from proxy respondents can be considered direct.

Comparisons are considered indirect when the evidence derives from studies that do not compare two interventions of interest specifically with each other; that is, the evidence is not from head-to-head studies. Mixed treatment comparisons are considered indirect (ie, when the model combines direct and indirect evidence). Detailed EPC guidance on indirect comparisons is available [32,33].

3.3. Consistency

Consistency refers to the degree of similarity in either the direction of effects or magnitude of effect (effect sizes) across individual studies within an evidence base (Table 1). EPCs choose whether to score direction, magnitude, or both; reviewers need to be explicit about the choice. Consistency (in either case) is scored as consistent, inconsistent, or consistency unknown.

3.4. Direction of effect

In judging the superiority of one treatment over another, reviewers 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]. Confidence intervals (CIs) may provide additional information on the consistency of the direction of effect.

In contrast to superiority, reviewers may look for evidence to support noninferiority or equivalence of two interventions. To do this, reviewers must define a line of difference in relation to a threshold; this is referred to as the minimally important difference (MID) [25,34,35]. The MID is a clearly, explicitly defined and justified clinical threshold; below it, reviewers would consider the evidence to show no meaningful difference and above it to show a benefit or harm of one treatment over another.

Optimally, MID thresholds are based on empirical evidence or guidelines. When such thresholds are not available, 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. EPC guidance on assessing equivalence and non-inferiority discusses MIDs further [34,35].

3.4.1. Magnitude of effect (and heterogeneity)

Consistency in the magnitude of effect reflects the degree to which point estimates are similar across studies. Studies can be considered consistent when the CIs of individual studies overlap a summary effect estimate calculated from a meta-analysis. When meta-analyses are unavailable, reviewers’ judgments are acceptable.

Substantial unexplained differences (heterogeneity [19]) across studies invoke a need for caution in estimating a summary magnitude of effect. When meta-analysis is appropriate, reviewers should evaluate consistency in magnitude of effect through statistical tests for heterogeneity (eg, Cochran’s Q test) or the magnitude of heterogeneity (eg, $I^2$ statistic [4]) and with qualitative evaluations. Reviewers 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 [36,37]. GRADE provides more detail about evaluating heterogeneity [19].

Table 1. Continued

<table>
<thead>
<tr>
<th>Domain</th>
<th>Definition and elements</th>
<th>Score and application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reporting bias</td>
<td>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, that is, nonreporting of the full study. • Selective outcome reporting bias, that is, nonreporting (or incomplete reporting) of planned outcomes or reporting of unplanned outcomes. • Selective analysis reporting bias, that is, 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—for example, 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 reviewers to assess it for such studies.</td>
<td>Score as one of two levels: • Suspected • Undetected Reporting bias is suspected when: • Testing for funnel plot asymmetry demonstrates a substantial likelihood of bias. • 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.</td>
</tr>
</tbody>
</table>

Abbreviations: RCT, randomized controlled trials; CI, confidence interval.
3.4.2. A single-study evidence base

Scoring consistency requires 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 [38–40]. Accordingly, we recommend that the consistency of a single-study evidence base be scored as unknown.

3.5. Precision

Precision is the degree of certainty surrounding an estimate of effect, that is, the potential for random error with respect to an outcome. A precise body of evidence should enable decisionmakers to draw conclusions about whether one treatment is inferior, equivalent, or superior to another [41,42]. Precision is scored as precise or imprecise.

Reviewers can evaluate sufficiency of sample size for determining precision relative to an optimal information size (OIS). Similar to the calculation for determining whether a single trial is sufficiently powered, 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 determining that an estimate is precise (see GRADE guidance [18]). If OIS criteria are not met, reviewers may score the evidence as imprecise.

When reviewers have conducted a meta-analysis, they can evaluate precision based on the CI calculated from the pooled estimate. For a qualitative synthesis, precision is assessed based on the constituent parts that would have contributed to the CI for the pooled estimate—that is, the sample size relative to OIS and the variance within individual studies. Ideally, reviewers will have measures of variance for individual studies (eg, standard deviation, CI); in some cases, they may have only P-values. If individual studies report only P-values but meet OIS standards, estimates might be regarded as precise when they report significance levels of differences between treatments as P-values of less than 0.05.

3.6. Reporting bias

Reporting bias occurs when authors, journals, or both publish or report research findings based on their direction or magnitude of effect [43,44]; it encompasses three main problems—publication bias, outcome reporting bias, and analysis reporting bias (Table 2). The risk of reporting bias is scored as suspected or undetected.

Methods to assess reporting bias exist only for RCTs (by comparing protocols and final methods; see related AHRQ

<table>
<thead>
<tr>
<th>Types of reporting bias</th>
<th>Definition</th>
<th>Implications and examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Publication</td>
<td>The whole study has been concealed from public access (nonregistration and/or nonpublication) or it is anticipated to be 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 the study data in obscure platforms or journals.</td>
<td>Data included in the review are more likely to reflect favorable than unfavorable findings. Example: significant differences favoring an intervention for efficacy outcomes or nonsignificant differences for harms outcomes are more likely to be reported in published studies.</td>
</tr>
<tr>
<td>Selective Outcome Reporting</td>
<td>The study is reported, but one or more of the planned outcomes are not reported and investigators do not provide a reasonable justification. Planned outcome data are reported, but unplanned outcome data are reported as well or the way outcome data were measured was not as planned.</td>
<td>Data included in the review are more likely to reflect favorable than unfavorable findings. Example: significant differences favoring an intervention for efficacy outcomes or nonsignificant differences for harms outcomes are more likely to be reported in published studies than other results. This reflects data mining and increased risk for type I error; significant differences may be a chance occurrence rather than a true effect. Example: changing outcome measure cut points.</td>
</tr>
<tr>
<td>Selective Analysis Reporting</td>
<td>Outcome data are reported, but they are based on the most favorable of several analyses undertaken; other analyses are suppressed. Precision of outcome data estimates is incompletely reported or not reported. The same outcome data are ambiguously reported in multiple study reports.</td>
<td>Examples: presenting selective post hoc subgroup analyses, dichotomizing continuous data using a cut point that gives the most favorable results, reporting adjusted vs. unadjusted analyses based on which appear more favorable, cherry-picking statistical assumptions, and reporting selective time-point analyses from among multiple planned follow-up points. Study data are presented as point estimates without measures of dispersion or with only inexact, nonsignificant P-values (eg, P &gt; 0.05). Authors do not make the copublication status transparent, which may lead to double counting of outcomes data.</td>
</tr>
</tbody>
</table>
report [31]). Observational studies may also be susceptible to reporting bias [45–48]. These studies generally lack protocols, so comparable methods are not available for evaluating reporting bias.

Reviewers can quantitatively test for the impact of unreported data through tests for funnel plot asymmetry, the trim and fill method, and selection modeling [49–55]. A qualitative assessment in addition to or in place of quantitative assessments (when the latter are not possible) is recommended. A proposed decision aid for evaluating the risk of reporting bias advocates taking a cautious approach when using both assessment approaches [5].

3.7. Additional domains

Three additional domains supplement the five required domains (Table 3): dose-response association, plausible uncontrolled confounding that would diminish an observed effect, and strength of association (ie, large magnitude of effect). Although these domains are more likely to be relevant for observational study bodies of evidence because they can increase SOE, they can also apply to RCTs. Reviewers should consider the additional domains but need not report on them when they regard them as irrelevant to the body of evidence.

4. Major steps in grading SOE: establishing an overall SOE grade

4.1. Four grades for SOE

4.1.1. Overview

Four levels of grades are intended to communicate reviewers’ conclusions of the strength of the overall assessment
of a body of evidence for a single outcome of a single treatment comparison. Grades are denoted as high, moderate, low, and insufficient (Table 4). These are discrete grades and should not be designated as ranges (eg, “low to moderate” SOE).

Each grade has two components. The first, principal definition concerns the level of confidence that reviewers place in the estimate for the benefit or harm; this equates to their judgment as to how closely the evidence is likely to reflect 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. The grade is based on domain scores as well as 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. Reviewers can base their designation on either a quantitative or a qualitative synthesis of the evidence.

For comparative effectiveness questions, reviewers typically consider either direction (A > B, A = B, A < B), magnitude (difference between A and B), or both. In some instances, assigning different grades for direction and magnitude of effect may be appropriate (eg, when a significant outcome exceeds an MID [direction] but heterogeneity about that estimate [magnitude] is quite high).

4.1.2. Grade of insufficient
   EPCs assign a grade of insufficient when they cannot draw any evidence-based conclusions for a single outcome for a single comparison. Particular care is needed so as not to confuse “low” SOE with “insufficient.” That is, low means the reviewers can draw a conclusion but, because of concerns about major or numerous deficiencies, place little faith in it. By contrast, insufficient means that reviewers cannot draw a conclusion at all (even with some findings).

Evidence is insufficient when literally no evidence is available from the included studies or when evidence about the outcome is too weak, sparse, or inconsistent to draw any defensible conclusions. This situation can reflect one or more conditions, such as unacceptably high study limitations. It can also reflect major unexplained inconsistency. Two studies with the same risk of bias may find opposite results with no clear reason for the discrepancy. Another example is when the CI around the estimated effect in a meta-analysis or across the preponderance of studies in a qualitative assessment is so wide that it includes two truly incompatible conclusions—that one treatment is clinically and significantly better than the other and that it is worse. Statistically nonsignificant effects per se should not prompt an automatic grade of insufficient; rather, insufficient is appropriate when the imprecision results in no confidence regarding whether the effect of one intervention is superior, inferior, or equivalent to another.

Finally, evidence based on a single study often, but not always, warrants a grade of insufficient. Because the evidence includes only one study, consistency is unknown. When the study is also too small to meet OIS criteria, the resulting lowering of the precision domain score further reduces the confidence in the finding of that study.

4.2. Incorporating domains into an overall grade
   In arriving at a final judgment for combining domains into a SOE grade, reviewers must weigh the relative importance and seriousness of concerns for each domain. EPCs may use different approaches to incorporate multiple domains into an overall SOE grade. The critical requirement is that they explain the rationale for their approach and note which domains were most important in reaching a final grade.

Consistency and precision can be particularly challenging domains for reaching an overall SOE grade. (See the updated methods guidance chapter for more detail [5].) When both inconsistency and imprecision may explain various hard-to-interpret findings, reviewers should generally attribute uncertainty to one or the other problem and lower an overall SOE grade only once. Only rarely would reviewers be justified in lowering a grade twice.

4.3. Developing grades for RCTs and observational studies
   SOE grades are first completed separately for RCT and observational study bodies of evidence. Then reviewers combine those design-specific SOE grades into one overall SOE grade. Describing whether evidence from observational studies complements or conflicts with evidence from RCTs is important for diverse end users. Giving plausible reasons for any differences and noting pertinent limitations in both bodies of evidence are helpful steps for all audiences.

Typically, RCT bodies of evidence initially start with a provisional grade of high SOE. Reviewers can lower this assessment after they evaluate each domain. In contrast, evidence about interventions based on observational studies is assumed to pose a greater risk of having study limitations; this usually corresponding to an initial provisional grade of low SOE. Reviewers may raise an initial SOE grade for an observational study body of evidence to moderate when the study limitation domain is scored as low or medium, most likely because risk of bias from confounding is low (eg, certain outcomes such as harms; study questions that focus on prognosis). Reviewers may also decide that after assessing the additional domains, the overall SOE of a body of observational studies can be upgraded to moderate (although rarely high).

4.4. Developing overall SOE grades from subsets of studies

4.4.1. Adopting best-evidence approaches
   Based on reasonable standards of evidence for the subject area, EPCs may adopt a best-evidence approach for
assessing SOE. They can rely on one study design when it clearly provides stronger evidence—that is, studies that provide the least limited, most direct, and most reliable evidence for an outcome or comparison. Criteria for this approach may be randomization or may be active-controlled vs. placebo-controlled, prospective vs. retrospective, and lower risk of bias vs. high risk of bias. Reviewers can determine an appropriate subset of studies for presenting review findings and SOE assessment through analyses with and without weaker studies (such as with a sensitivity analysis) [28,56].

4.4.2. Using studies with high risk of bias

Combining evidence from studies with a high risk of bias and those with less risk can be problematic [5]. Moreover, inconsistent results between two bodies of evidence calls for assessing whether different levels of risk of bias explain the inconsistency. Also important is examining whether combining bodies of evidence might obscure findings from studies rated either low or moderate risk of bias.

To determine which studies (within different study designs) to include in the final SOE assessment, reviewers can conduct sensitivity analyses or qualitative assessments involving the high-risk-of-bias studies. They can explore whether a finding including these studies is systematically different from one limited to less biased studies—that is, whether heterogeneity in study design or conduct can explain inconsistencies. If reviewers conclude that the findings do differ materially, then, with clear explanation, they can give greater weight to the lower risk-of-bias studies or limit final synthesis to these studies [28]. Such high-risk-of-bias studies are still counted as part of the overall evidence base and cited in references.

4.4.3. Accounting for heterogeneity

When reviewers can explain heterogeneity (eg, differences attributable to populations, interventions, comparators, study design, or conduct), then reporting findings

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Study design</th>
<th>Findings and direction (magnitude) of effect</th>
<th>Strength of evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major outcomes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mortality</td>
<td>RCT: 1 (56)</td>
<td>A single small RCT with medium study limitations and poor precision found no significant difference in mortality at 1 yr.</td>
<td>Insufficient</td>
</tr>
<tr>
<td>Severity of (disease)</td>
<td>RCT: 3 (110)</td>
<td>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 mo to 5 yr. Overall, intervention A reduced the severity of (disease) more than intervention B.</td>
<td>Low (improved severity of [disease])</td>
</tr>
<tr>
<td>Other patient-centered outcomes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain</td>
<td>RCT: 6 (160)</td>
<td>RCTs with medium study limitations all found that X reduced pain more than Y, between 3 mo and 2 yr. 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.</td>
<td>Moderate (reduced pain) Low (0.5 difference in pain reduction)</td>
</tr>
<tr>
<td>Sexual dysfunction</td>
<td>RCT: 3 (85)</td>
<td>Few studies, only in men. Results were consistent that treatment improves sexual dysfunction at 3 mo but imprecise.</td>
<td>Low (improved dysfunction)</td>
</tr>
<tr>
<td>Intermediate outcomes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDL cholesterol</td>
<td>RCT: 8 (212)</td>
<td>Small studies yielded a summary net change of −2.1% (95% CI: −4.0, −0.1) with a wide (imprecise) CI.</td>
<td>Low (decreased cholesterol by 2.1%)</td>
</tr>
<tr>
<td>Radiology test</td>
<td>RCT: 0</td>
<td>No eligible studies</td>
<td>Insufficient</td>
</tr>
<tr>
<td>Adverse events</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intestinal perforation</td>
<td>RCT: 1 (42)</td>
<td>Only a single event was reported in one small RCT. 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 mo (range 0.2 to 13.8 kg).</td>
<td>Insufficient</td>
</tr>
<tr>
<td>Weight gain</td>
<td>Observational: 4 (600)</td>
<td></td>
<td>Low (weight gain)</td>
</tr>
</tbody>
</table>

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

a See Tables A and B in the Supplementary Material Appendix at www.jclinepi.com 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.
by subgroups may be very helpful for decisionmakers. Doing this may entail stratifying evidence by meaningful subgroups and separately scoring domains for these subgroups or limiting reported findings to a subset of studies containing the most believable effect estimates. Adequately explaining differences between subgroup results and results from remaining studies is important [57].

5. Transparency: documenting and reporting SOE

In assessing overall SOE, EPCs should explicitly explain how scores for each domain contribute to the overall grade. Important considerations include how reviewers incorporated different study designs and studies with high risk of bias into the grade, how they weighted each required domain, and which additional domain (if any) they assessed. Reviewers should carefully document their approach with enough detail to ensure that diverse user groups can grasp the methods and underlying reasoning.

5.1. Presenting SOE grades in tables

Table 5 illustrates a suggested approach to providing actionable information to decisionmakers. Its important components include the following: (1) outcome (benefit or harm); (2) number of contributing studies (in major study design categories) and number of participants; (3) summary of the scored domains that influenced the overall grade; (4) length of follow-up; and (5) succinct description of findings (eg, direction or magnitude of effect), including summary estimates from meta-analyses, if calculated. An SOE grade should always be accompanied by an overall direction and/or magnitude of effect. If evidence was graded insufficient, omitting those outcomes or comparisons and describing insufficient evidence only in text will streamline this table.

Reviewers can enhance readability by dividing outcomes into key categories: major, other patient-centered, intermediate, and adverse events. Major outcomes are those deemed most important for decisionmaking about the interventions reviewed. These four categories likely overlap to some degree; giving exact definitions of categories and determining which outcomes belong where will vary for clinical topics and research questions.

5.2. Providing descriptive explanatory text

Transparency about SOE grades requires clearly communicating the findings and the confidence that reviewers have in those findings. Particularly salient factors include approaches to rating risk of bias for individual studies, scoring domains, and assigning final SOE grades (eg, why situations resulted in one grade vs. another, such as low vs. insufficient). Examples of more detailed complementary tables and explanatory text are presented in the Supplementary Material Appendix at www.jclinepi.com.

Reviewers cannot rely solely on a reductive, single grade for explaining their findings and implications of those findings. All systematic reviews need to present “narrative” syntheses of findings. The goal is making reviews accessible, readable, and interpretable for many types of stakeholders.

6. Discussion

The EPC program’s approach to grading SOE rests on evaluating a required group of domains (study limitations, directness, consistency, precision, and reporting bias) and, at times, additional domains (dose-response relationship, uncontrolled confounding that would diminish an observed effect, and strength of association). When reviewers determine their final SOE grades, they should consider the relative importance of each domain and interactions among them and the unique concerns that a particular body of evidence may engender.

The SOE domains discussed are directly relevant to studies of drugs, procedures, and other therapeutic interventions. However, EPCs increasingly assess diagnostic tests, screening strategies, and health services interventions such as quality improvement and patient safety studies. Separate guidance on grading SOE for reviews on medical tests is available [58].

Unlike GRADE, AHRQ’s EPC program is not directly involved in developing clinical recommendations or practice guidelines. Partly for that reason, this guidance does not extend to the idea of “combining” SOE grades into a summary judgment that would take multiple outcomes into account simultaneously or that would reflect trade-offs between benefits and harms. Rather, EPC systematic reviews are intended to provide a wide range of stakeholders with information that they, in turn, can apply in making their own treatment, policy, or administrative choices.

The importance of the distinctions among high, moderate, and low SOE (and the distinction with insufficient SOE) can vary by the type of outcome, comparison, and decision maker. Some stakeholders may want to act only when evidence is of high or moderate strength; others may want to understand clearly the implications of low vs. insufficient evidence. Even when SOE is low or insufficient, individuals, clinicians, and policymakers may need to make choices and decisions, and they may consider factors in addition to the evidence from a specific systematic review, such as patient values and preferences, costs, or resources.

Recent approaches to evaluating the risk of bias arising from confounding in individual observational studies incorporate assessments of potential confounding across the body of evidence from observational studies [59,60]. Experience with these ways to assess risk of bias will likely
provide additional insights that, in turn, enhance methods for SOE grading.

An understandable approach for grading SOE—one that decisionmakers can readily recognize and interpret—is highly desirable for all types of systematic reviews. This updated guidance drew extensively from the GRADE approach, applying it to the specific circumstances and experience of AHRQ’s EPC program. AHRQ’s EPC program (like GRADE) will continue to refine and improve grading systems to be most applicable and useful for different types of health care questions, reviews, and diverse stakeholders. Meanwhile, this article codifies the guidance that EPCs and other review groups can follow now to strengthen the consistency, clarity, and usefulness of their reviews and points to the special challenges that lie ahead. Even if no single approach for reporting results and grading the related SOE suits all reviews, using and documenting a consistent approach to reporting critical summary information—the general concept of transparency—makes all reviews more useful to the broad range of potential audiences that AHRQ’s work aims to reach.

Acknowledgments

The authors thank Gillian Saunders, PhD, 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, MD, MSc, PhD, for offering guidance both in person and through thoughtful comments on a draft of the full report from which this article is adapted. Finally, we are indebted as well to our RTI colleague Loraine Monroe for superlative assistance with preparation of this document and the original report from which it is adapted.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2014.11.023.

References

