Methods Guide for
Comparative Effectiveness Reviews

Grading the Strength of a Body of Evidence When Comparing Medical Interventions

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

Comparative Effectiveness Reviews are systematic reviews of existing research on the effectiveness, comparative effectiveness, and harms of different health care interventions. They provide syntheses of relevant evidence to inform real-world health care decisions for patients, providers, and policymakers. Strong methodologic 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, the Agency for Healthcare Research and Quality (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 Comparative Effectiveness 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. Comments and suggestions on the *Methods Guide for Comparative Effectiveness Reviews* and the Effective Health Care Program can be made at www.effectivehealthcare.ahrq.gov.

This research was funded through contracts from the Agency for Healthcare Research and Quality to the following Evidence-based Practice Centers: ECRI Institute (290-02-0019), Johns Hopkins University (290-02-0018), Oregon Health & Science University (290-02-0009), RTI International and the University of North Carolina (290-02-0016), and Stanford University (290-02-0017).

None of the authors has a financial interest in any of the products discussed in this document.

Suggested citation: Owens DK, Lohr KN, Atkins D, et al. Grading the strength of a body of evidence when comparing medical interventions. In: Agency for Healthcare Research and Quality. Methods Guide for Comparative Effectiveness Reviews [posted July 2009]. Rockville, MD. Available at: http://effectivehealthcare.ahrq.gov/healthInfo.cfm?infotype=rr&ProcessID=60.

Grading the Strength of a Body of Evidence When Comparing Medical Interventions

Authors:

Douglas K Owens, M.D., M.S.[a]

Kathleen N. Lohr, Ph.D.[b]

David Atkins, M.D, M.P.H.[c]

Jonathan R. Treadwell, Ph.D.[d]

James T. Reston, Ph.D., M.P.H.[e]

Eric B. Bass, M.D., M.P.H.[f]

Stephanie Chang, M.D., M.P.H.[g]

Mark Helfand, M.D.[h]

[a]VA Palo Alto Healthcare System; Stanford-University of California San Francisco Evidence-based Practice Center; Center for Primary Care and Outcomes Research, Stanford University, Palo Alto, CA

[b]RTI International, Research Triangle Park, NC

[c]Health Services Research & Development Service, Department of Veterans Affairs, Washington, DC

[d]ECRI Institute Evidence-based Practice Center, Plymouth Meeting, PA

[e]ECRI Institute, Plymouth Meeting, PA

[f]Johns Hopkins University Evidence-based Practice Center, Baltimore, MD

[g]Center for Outcomes and Evidence, Agency for Healthcare Research and Quality, Rockville, MD

[h]Oregon Health & Science University Evidence-based Practice Center, Portland VA Medical Center, Portland, OR

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 view of AHRQ or the Veterans Health Administration. Therefore, no statement in this report should be construed as an official position of these entities, the U.S. Department of Health and Human Services, or the U.S. Department of Veterans Affairs.

Grading the Strength of a Body of Evidence When Comparing Medical Interventions

Key Points

- The EPC (Evidence-based Practice Center) approach is conceptually similar to the GRADE (Grading of Recommendations Assessment, Development and Evaluation) system of evidence rating.
- It requires assessment of four domains: risk of bias, consistency, directness, and precision.
- Additional domains to be used when appropriate include dose-response association, presence of confounders that would diminish an observed effect, strength of association, and publication bias.
- Strength of evidence receives a single grade: high, moderate, low, or insufficient.
- EPCs should grade strength of evidence separately for each major outcome and, for Comparative Effectiveness Reviews, all major comparisons.
- EPCs will collaborate with the GRADE group to address ongoing challenges in assessing the strength of evidence.

Introduction

Comparative Effectiveness Reviews (CERs), like systematic reviews in general, are essential tools for summarizing information to help make well-informed decisions about health care options.[1] CERs explicitly compare two or more screening or diagnostic strategies or therapeutic interventions. The Evidence-based Practice Center (EPC) program, supported by the U.S. Agency for Healthcare Research and Quality (AHRQ), produces substantial numbers of evidence reports and CERs. These reports are designed to accurately and transparently summarize a body of literature with the goal of helping clinicians, policymakers, and patients make well-informed decisions about health care. Reviews should provide clear judgments about the strength of the evidence that underlies conclusions to enable decisionmakers to use them effectively.[2]

In 2007, AHRQ supported a cross-EPC set of workgroups to develop guidance on major elements of designing, conducting, and reporting CERs.[3] This paper reports the outcomes of the EPC workgroup on grading strength of evidence. We briefly explore the rationale for grading strength of evidence, define the domains of concern for evidence strength, and describe our recommended grading system for such reviews. Our main objective was to give guidance to EPCs for grading strength of evidence in CERs, but this guidance may also apply to other systematic reviews.

The EPCs prepare reports that are used by a variety of decisionmakers, but they do not themselves develop recommendations. Therefore, the goal of our evidence rating system is to facilitate use of the reports by decisionmakers who may have differing perspectives. This separation of the raters of the strength of evidence from the decisionmakers led to some differences in the system we propose relative to other rating systems that are designed to be used directly by decisionmakers.

The EPC approach is based in large measure on the GRADE (Grading of Recommendations Assessment, Development and Evaluation) working group approach.[4-6] We briefly discuss the differences in emphasis between the two systems. EPC and GRADE experts will explore ways to harmonize the two methods and to offer reviewers and decisionmakers a coordinated model for grading strength of evidence. This paper presents the approach that EPCs are expected to implement for CERs in the meantime.

Strength of Evidence: Rationale

Among organizations that make practice guidelines or coverage decisions and among experts who develop systematic reviews, assessment of the strength of a body of evidence is widely accepted. In drawing conclusions about strength of evidence, a growing number of organizations adopt systematic approaches to making judgments about the strength of evidence. A wide variety of grading systems is available for this purpose,[7] and different organizations may weigh features, or domains, of a body of evidence differently. Consequently, discrepant, contradictory, or variable ratings may arise, and results may not be of practical help to some organizations.

We note the important distinction between strength-of-evidence systems and evidence hierarchies. Evidence hierarchies traditionally focus only on study design, with systematic reviews of randomized controlled trials (RCTs) and individual RCTs at the highest levels. By contrast, strength-of-evidence systems incorporate not only study design but also many other facets of the evidence, including study conduct, presence or absence of bias, quantity of evidence, directness (or indirectness) of evidence, consistency of evidence, and precision of estimates. By including these additional components in our approach, we have attempted to give decisionmakers a more comprehensive evaluation of the evidence.

The aims of this work are to ensure appropriate methodologic consistency in how different EPCs grade the strength of evidence and to facilitate users' interpretations of those grades and how they apply them in guideline development or other decisionmaking tasks. Attaining these goals rests in part on consistency and predictability in the domains that EPCs use in this effort. Although no one system for reporting results and grading the related strength of evidence is likely to suit all users, documentation and consistent reporting of the most important summary information about a body of literature will make reviews more useful to a broader range of potential audiences.

Strength of Evidence: Domains

The EPC approach to grading evidence begins with assessments of a set of agreed-upon domains pertaining to entire bodies of evidence about major outcomes (benefits and harms) and comparisons—i.e., outcomes and comparisons that are most important to decisionmakers in clinical practice and health policy. A determination of which outcomes and comparisons the EPCs consider important enough to warrant formal grading of the strength of the evidence will depend on the key questions, the clinical or policy context, and the purpose of the report. Major outcomes may include mortality, health-related quality of life, costs, potential harms, and for some reviews, intermediate end points or surrogate markers (for example, blood pressure control or cholesterol levels).

The four major domains are risk of bias, consistency, directness, and precision of the evidence. In selecting these domains, we reviewed work by the U.S. Preventive Services Task Force,[8] the GRADE Working Group[4] (http://www.gradeworkinggroup.org/), and other research by EPCs.[7,9] EPC reviewers aggregate judgments about the strength of evidence with respect to

the domains into an overall evidence grade (explained below) for each major outcome. Tables 1 and 2 present two sets of domains: "required" and "additional," respectively. Because the strength of evidence may vary between key questions in a systematic review and among comparisons within a key question, the EPC should evaluate the strength of evidence separately for each important comparison for each key question.

Required Domains

The first set, "required domains," comprises four major constructs that EPCs should use for all major outcomes and comparison(s) of interest: risk of bias, consistency, directness, and precision. Table 1 defines these and indicates how to assess and apply them. These four domains are discussed in more detail below.

Before assessing the required domains, EPCs should first identify the studies that address the outcomes and comparisons of interest. When no study is available for an outcome or comparison of interest, the evidence should be graded simply as insufficient.

For the remaining major outcomes and comparisons of interest, the strength-of-evidence grade will depend on the required domains. EPCs have decided that focusing on consistency, directness, and precision is more informative than emphasizing just the number of studies. Nevertheless, for CERs, EPCs should record the numbers of studies both in total and for specific comparisons. They should also indicate the numbers of studies that form the basis of given findings or conclusions. In this way, readers can better understand the available evidence for any given outcome or comparison.

Risk of Bias

As noted in Table 1, the risk of bias for an evidence base will be derived from assessment of the risk of bias in individual studies. Risk of bias incorporates both study design and study conduct. For strength-of-evidence grading, this domain requires reviewers to assess the aggregate quality of studies within each major study design and integrate those assessments into an overall risk-of-bias score.

Scores are denoted high, medium, or low. High risk of bias lowers the strength-of-evidence grade; low risk of bias raises it. If studies included in a systematic review differ substantially in risk of bias, EPCs may give greater weight or emphasis to the studies with a lower risk of bias. In formal meta-analyses, EPCs may choose to evaluate the influence of studies with differing risk of bias to aid in their assessment of the overall strength of evidence.

Consistency

Main considerations. Consistency refers to the degree of similarity in the effect sizes of different studies within an evidence base. If effect sizes indicate the same direction of effect and if the range of effect sizes is narrow, an evidence base can be judged to be consistent. This assessment enhances the overall strength-of-evidence grade. Nonoverlapping confidence intervals, significant unexplained clinical or statistical heterogeneity, or similar problems may reflect inconsistency. The presence of inconsistency is the chief concern for grading strength of evidence in this domain, and it would lead EPCs to reduce the overall strength-of-evidence grade.

If meta-analysis is appropriate, EPCs can evaluate consistency using statistical tests and measures of heterogeneity (such as Cochran's Q test or I^2 statistics, as discussed in the

Quantitative Synthesis chapter of the *Methods Guide for Comparative Effectiveness Reviews* (http://effectivehealthcare.ahrq.gov/healthInfo.cfm?infotype=rr&ProcessID=60).

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

EPCs should designate an evidence base as inconsistent when different studies show statistically significant effect sizes in opposite directions. In the absence of statistical testing or measurement of heterogeneity, EPCs can assess consistency on the basis of similarity of populations, interventions, and outcome measures.

Evaluation of a single-study evidence base. Evaluation of consistency ideally requires an evidence base with independent replication of findings; therefore, EPCs cannot properly evaluate consistency in an evidence base with a single study. Even if the study is a large multicenter trial (i.e., a mega-trial), findings from different centers within such a study are rarely reported separately. If the results are reported separately for each center, EPCs may be able to evaluate consistency within the overall trial, but this is not truly independent replication. Any flaw (reported or not reported) in the trial design or conduct will likely be replicated at every center. Even pairs of mega-trials addressing the same clinical question (i.e., the same patient intervention-outcome combinations) may report discrepant results,[10] and the methodology of mega-trials has been further questioned.[11]

Thus, EPCs cannot be certain that a single trial, no matter how large or well designed, presents the definitive picture of any particular clinical benefit or harm for a given treatment. Accordingly, with respect to consistency, we recommend that EPCs judge single-study evidence bases "consistency unknown (single study)," which would generally decrease the strength-of-evidence grade.

Directness

Directness concerns whether the evidence being assessed reflects a single, direct link between the interventions of interest and the ultimate health outcome under consideration (whether a benefit or harm). If direct evidence linking an intervention to the most ultimate outcomes is lacking, then two or more bodies of evidence are needed to link the intervention to health outcomes. When several bodies of evidence are involved, the ultimate decision about using an intervention may depend on the strength of evidence for every link in the causal chain.

Some links in the causal chain will be more important than others. Thus, the final assessment of directness requires EPCs to consider the strength of evidence for each link as well as the importance of each link in the chain. Of particular salience is the extent to which evidence pertains to intermediate or surrogate outcomes rather than to ultimate patient-centered outcomes such as mortality, morbidity, and quality of life. More direct links enhance strength-of-evidence assessments (and vice versa).

In an example involving enteral feeding[12] used in this *Methods Guide* (see Principles for Developing Guidance for Comparing Medical Interventions),[3] a large body of well-conducted randomized trials might demonstrate that enteral supplementation improved nutritional status and delivery of nutrients to the area of the wound. However, evidence of an association between a richer nutritional milieu and the ultimate outcome of complete healing may be weak. If this is a critical link in the causal chain, then the EPC can decide to grade the overall body of evidence as

indirect, which would weaken the strength of evidence. As illustrated in the chapter on Principles for Developing Guidance for Comparing Medical Interventions of this *Methods Guide*,[3] use of an analytic framework is an important heuristic for determining how to evaluate evidence in a causal chain (e.g., in an overarching link or only in subsidiary linkages).

For CERs in particular, directness also applies to comparing interventions. For example, if there are three alternative interventions—A, B, and C—having evidence that compares them directly—A vs. B, A vs. C, and B vs. C—is desirable. In many circumstances, such head-to-head evidence is not available. Under these circumstances, reviewers must look to indirect evidence, such as evidence for A vs. C and B vs. C but not A vs. B. Grades for such indirect evidence will not be as strong as those obtained from truly direct evidence.

A single body of evidence is preferable to two bodies of evidence, particularly if the strengths of evidence for those two bodies of evidence differ in material ways. Assessing directness clarifies the degree to which evidence between the intervention and the ultimate health outcome does or does not meet the ideal set of studies addressing the overarching question.

Precision

Precision is the degree of certainty surrounding an estimate of effect with respect to a specific outcome. EPCs should assess the boundaries of the pooled confidence interval for that effect estimate in relation to a threshold that would allow CER users to make judgments about the treatments being compared. Relevant thresholds for precision include the boundary of statistical significance—that is, whether the estimate of an effect reaches accepted levels for statistical significance. A precise estimate should enable decisionmakers to draw conclusions about whether one treatment is, clinically speaking, inferior, equivalent (neither inferior nor superior), or superior to another.[13,14]

Judgments about precision may depend on the importance of the outcome being measured, other clinically important outcomes, and the context of decisionmaking. They may also be contingent on whether the central issue is harms or benefits and the relative impact or size of those harms or benefits. This domain should be rated as precise or imprecise separately for each important outcome.

Substantial variability does not necessarily render an estimate imprecise. A truly imprecise estimate is one with a confidence interval so wide that it does not rule out the superiority or inferiority of either treatment being compared—that is, an estimate whose confidence interval includes two incompatible possibilities: one treatment is clinically significantly better than the other, and the difference is in the opposite direction. In this case, no conclusion can be reached about the relative effectiveness of the two treatments.

Additional Domains

The second set of domains, which supplement the four required domains, consists of secondary constructs that EPCs should use and report if they are relevant to a particular CER. These domains are dose-response association, existence of confounders that would diminish an observed effect, strength of association (i.e., magnitude of effect), and publication bias. These domains also derive from our review of other rating systems, including GRADE. Table 2 provides their definitions and ways to rate and apply them. Generally, we expect three of these domains—dose-response association, existence of confounding factors that would diminish an observed effect, and strength of association—to be applied more often to evidence from observational studies (of all types) than to evidence from RCTs.

The EPCs will invoke publication bias concerns when they have reason to believe that relevant empirical findings have not been published or are not otherwise available. Three situations are particularly relevant: (1) when negative, no-difference, or other studies with results that are substantially different from published studies are unavailable; (2) when the results of completed studies (e.g., those noted in ClinicalTrials.gov as having been ended 3 or more years in the past) have clearly not been published (save, perhaps, in abstract form); and (3) when trial protocols specify certain secondary end points for which results have not been reported (even if other results have been published). EPCs should consider and report on publication bias insofar as it appears to influence scores for either required or other domains (e.g., consistency or precision).

Applicability

A wide array of groups use EPC reports; not surprisingly, the context and populations these users consider relevant may differ. Thus, evidence that one group may consider applicable to the population of interest may not be applicable to the population of interest of another group. For this reason, we have chosen to make our judgments about applicability explicit and separate from assessments of other domains of strength of evidence. In doing so, we aim to make it clear when our statements about the evidence are based on applicability rather than on other aspects of the evidence. Our goal in assessing applicability separately is to enable decisionmakers to take into account how well the evidence maps to the patient populations, settings, diseases or conditions, interventions, comparators, and outcomes that are most relevant to their decisions. Decisionmakers may determine that evidence is not readily applicable to their population of interest, and they should make recommendations accordingly.

Thus, we recommend that EPCs summarize characteristics that decisionmakers may need to consider in assessing the applicability of the evidence. In particular, EPCs should record information about applicability for the outcomes and comparisons for which they specify an overall strength-of-evidence rating. Summarizing such information in a separate table, which decisionmakers can review along with the strength-of-evidence table, may be helpful. Guidance for this process will be available in the Assessing Applicability paper of the *Methods Guide*, which was under review at the time of publication of this paper.

Procedures for Assessing Domains

EPCs should have two or more reviewers with the appropriate clinical and methodological expertise separately assess each required domain (or each optional domain, as relevant) for each major outcome (whether benefit or harm). Differences should be resolved by consensus or mediation by an additional expert reviewer. Although the consensus judgments will appear in tables in the reviews, EPCs should record and save each reviewer's individual judgments about domains as background documentation.

Overall Strength-of-Evidence Grade

Four Strength-of-Evidence Levels

The overall grade for strength of evidence reflects a global assessment that takes the required domains directly into account and, as needed, incorporates judgments about the additional domains as well. For each comparison of interest, EPCs should rate strength of evidence for each

major benefit (e.g., positive impact on health outcomes such as physical function or quality of life, or effects on laboratory measures or other surrogate variables) and each major harm (ranging from rare, serious, or life-threatening adverse events to common but bothersome effects). For both benefits and harms, EPCs should focus on the outcomes most relevant to patients, clinicians, and policymakers.

Systematic reviews and CERs can be broad in scope, encompassing multiple patient populations, interventions, and outcomes. EPCs are not expected to grade every possible comparison for every outcome. Rather, reviewers should set clear priorities, assigning grades to those combinations (patients-interventions-outcomes) that are likely to be of greatest interest to users of the report. EPCs should also state clearly which interventions, outcomes, and comparators they included for each strength-of-evidence grade. For example, an evidence grade might apply to a link in an analytic framework, or it might apply to a specific intervention for a specific set of outcomes in a particular population. EPCs should also make clear which of the comparators or interventions is favored for each strength-of-evidence grade.

Table 3 summarizes the four levels of grades that EPCs should use. Each level has two components. The first, principal definition concerns the level of confidence the authors place in the estimate of effect for the benefit or harm (i.e., their judgment that the evidence reflects the true effect). The second, subsidiary definition involves a subjective assessment of the likelihood that future research might affect the level of confidence in the estimate or actually change that estimate.

Grades are denoted high, moderate, low, and insufficient. They are not designated by Roman numerals or other symbols.

High, moderate, or low strength of evidence

Assigning a grade of high, moderate, or low implies that an evidence base is available from which to estimate an effect. EPCs understand that, even when evidence is low, consumers, clinicians, and policymakers may find themselves in the position of having to make choices and decisions. The designations of high, moderate, and low should convey how secure reviewers feel about decisions based on evidence of differing grades. EPCs should apply discrete grades and avoid designations such as “low to moderate” strength of evidence.

Insufficient

In some cases, the reviewers cannot draw conclusions for a particular outcome, specific comparison, or other question of interest. In these situations, the EPC should assign a grade of insufficient. Such situations arise in two main ways.

First, evidence for an outcome receives a grade of insufficient when no evidence is available from the included studies. This case includes the absence of any relevant studies whatsoever. In CERs, 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.

Second, a grade of insufficient is also appropriate when evidence on the outcome is too weak, sparse, or inconsistent to permit any conclusion to be drawn. This situation can reflect several complicated conditions, such as unacceptably high risk of bias or a major inconsistency that cannot be explained (e.g., two studies with the same risk of bias that found opposite results, with no clear explanation for the discrepancy). Imprecise data may also lead to a grade of insufficient, specifically when the confidence interval is so wide that it includes two

incompatible conclusions: that one treatment is clinically significantly better than the other and that it is worse. Indirect data based on only one study or comparison could also receive a grade of insufficient. If a single quantitative estimate is desired, the strength of evidence may be insufficient if an effect size cannot be calculated from reported information or if heterogeneity cannot be explained. This same evidence base may still be sufficient to permit a conclusion about the general direction of the effect, but EPCs need to take care not to conflate "low" strength of evidence with "insufficient."

Incorporating Multiple Domains into an Overall Grade

To assign an overall grade to the strength of a body of evidence, EPCs must decide how to incorporate multiple domains into that overall assessment. In some systems, such as that of the GRADE working group,[4-6] the overall grade for strength of evidence (which it calls quality of evidence) is calculated from the ratings for each domain using a method that provides guidance on how to upgrade or downgrade the rating of the evidence. Such a system has the advantage of transparency because it clearly delineates a direct path from the evidence to its grade.

Although a system that uses such a method may offer advantages in terms of transparency, as yet there is not empirical evidence to support the superiority of a particular point system compared with a more qualitative approach. Furthermore, some evidence suggests no difference in accuracy between quantitative and qualitative systems.[7] Research is needed to compare the performance of a point system approach with other grading systems before we can recommend that EPCs use any specific system. Thus, EPCs may use different approaches to incorporate multiple domains into an overall strength-of-evidence grade.

The EPCs should explain the rationale for their approach to rating of strength of evidence and note which domains were important in upgrading or downgrading the strength of evidence. GRADE uses an algorithm to help reviewers to be clear about how they consider domains to produce the grade. EPCs may use the GRADE system or their own weighting system, or they may elect to use a qualitative approach, so long as the rationale for ratings of strength of evidence is clear. Several general principles that all should follow are important.

First, the risk of bias based on the design and conduct of the available studies is an essential component to rating the overall body of evidence. In considering the risk-of-bias domain, EPCs should consider which study design is most appropriate to reduce bias for each question. For many of the traditional therapeutic interventions, evidence that is based on well-conducted randomized trials will have less risk of bias than does evidence based on observational studies. For these outcomes, if randomized trial data are available, EPCs may choose to start with a rating of low for the risk-of-bias domain and change the assessment of this domain if the RCTs have important flaws. For these traditional therapeutic intervention questions, observational data would generally start with a high risk of bias but may be altered depending on the conduct of the study. As with all questions, the overall strength of evidence must incorporate assessments of other domains in addition to risk of bias.

Second, EPCs should assess each of the major domains for rating the overall strength of evidence. Assessment of consistency, directness, and precision may reveal strengths or weaknesses with the entire body of evidence and lead to a strength of evidence that is either higher or lower than would be obtained by considering only risk of bias. EPCs should also consider the additional domains when appropriate; they need not report on those domains when they regard them as irrelevant to the review in question. The strength of the evidence would be weakened by concerns about publication bias. In contrast, several factors may increase strength

of evidence and are especially relevant for observational studies, where one may typically begin with a lower overall strength of evidence based on the risk of bias. Presence of a clear dose-response association or a very strong association would justify increasing strength of evidence. If the confounding that may exist in a study would decrease the observed effect, but an effect is observed despite this possible confounding, the EPC may wish to upgrade the strength of evidence.

Third, EPCs should decide a priori how to incorporate each domain into an overall strength of evidence and what measures they will use to ensure accuracy and consistency of evidence ratings. The degree to which the overall strength of evidence is altered by additional domains that are used is a judgment that EPCs should explain in the report.

Key Procedures

EPCs should also take specific steps to ensure reliability and transparency within their own work (both in individual reviews and across them) when incorporating domains into an overall grade. As a first step, they should be explicit about whether the evidence grade will be determined by a point system for combining ratings of the domains or by a qualitative consideration of the domains. They should carefully document procedures used to grade strength of evidence and provide enough detail within the report to assure that the users can grasp the methods that were employed. EPCs should, furthermore, keep records of their procedures and results for each review so that they may contribute to the overall EPC expertise and science of grading evidence.

Second, EPCs should identify the domains that are most important for the targeted body of evidence and decide how to weight the domains when assigning the evidence grade. For the sake of consistency across reviews, the domains should be defined using the terminology presented in this chapter. In the absence of evidence to support specific systems for weighting of the domains, both qualitative and quantitative approaches are acceptable. EPCs may also choose to follow GRADE guidance for downgrading and upgrading evidence based on assessments of each domain. In general, the first or highest priority should be given to the domain for risk of bias, as it is well established that evidence is strongest when the study design and conduct have the lowest risk of bias.

The third step is to develop an explicit procedure for ensuring a high degree of inter-rater reliability for rating individual domains. As mentioned earlier, this assumes that at least two reviewers with appropriate clinical and methodological expertise will rate each domain. In addition, EPCs should assess the resulting inter-rater reliability for each domain. Although EPCs generally will not include the details of the reliability assessment in their CERs, they should keep records of this information. By documenting this information, EPCs will be able to increase knowledge about the reliability of the grading system.

The fourth step is to use the ratings of the domains to assign an overall strength-of-evidence grade according to the decisions made in the first through third steps. If this action involves a qualitative approach with subjective weighting of the domains, EPCs should consider using at least two reviewers and assessing the inter-rater reliability of this step in the process. That will not be necessary if the approach involves a formulaic calculation or algorithm based on the ratings of the domains. However, the scoring system or algorithm should be specified in sufficient detail to permit readers to replicate it if desired.

The fifth step is to prepare a narrative explanation of the reasoning used to arrive at the overall grade for each body of evidence. This should include an explanation of what domains played important roles in the ultimate grades.

Reporting Strength of Evidence

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

Table 4 illustrates one approach to providing actionable information to decisionmakers that reflects strength of evidence. It presents information pertinent to assessing evidence strength from different types of studies—specifically on the four required domains—and it displays estimates of the magnitude of effect (right column).

For the outcome as a whole (e.g., mortality or quality of life), the table also gives the overall rating. It shows, for instance, that one fair-quality RCT reported mortality, which was lower by one patient per 100 treated (i.e., 1 percent), a difference that was not statistically significant (95-percent confidence interval [CI], -4 percent to +3 percent). For the same comparison, 14 retrospective cohort studies had a wide range of effect sizes (range -7 percent to +5 percent). Had these estimates been precise and consistent (e.g., narrower CI for the RCT, consistent cohort studies to allow a summary effect size), one might have been able to reach a conclusion. However, the evidence is insufficient to allow a conclusion for mortality.

Although Table 4 illustrates how EPCs might organize information about the strength of evidence and magnitude of effect in ways useful to decisionmakers, it is incomplete. First, the table does not convey any information about the applicability of the evidence, which would be presented through other means (text or table). Second, a narrative summary of the results is also essential for interpreting the results of a literature synthesis.

Discussion

The EPC approach to rating the strength of evidence draws heavily on the international GRADE system; both conceptually and substantively, it is similar to GRADE. Our recommendations address specific circumstances of the EPC program, which differ from those of some groups that use GRADE. The EPC program produces systematic reviews, but it is not involved directly in development of recommendations or guidelines. Rather, EPC reports are used by a spectrum of government agencies, professional societies, and other stakeholders. Our approach for grading strength of evidence and discussing applicability of the evidence is meant to facilitate use of the EPC reports by this broad group of users.

We recommend that EPCs rate strength of evidence based on a core group of domains that include risk of bias, consistency, directness, and precision. Randomized trials will generally be assessed to have a low risk of bias, which correlates with a high strength of evidence, but may be changed after evaluation of other domains. Evidence based on observational studies will generally have a high risk of bias, which correlates with a low strength of evidence, but may be

rated higher after evaluating other domains. When appropriate, the EPCs can also use additional domains of dose-response association, the impact of plausible confounding, strength of association, and publication bias to upgrade or downgrade the strength of evidence.

This overall approach is similar to the methods used in the GRADE system. In GRADE, evidence based on observational studies starts with a strength of low and can be upgraded based on several factors. In the approach we describe here, the EPC may believe that, for certain outcomes, such as harms, observational studies have less risk of bias than do randomized trials or that the available randomized trials have a substantial risk of bias. In such instances, the EPC may either move up the initial rating of strength of evidence based on observational studies to moderate or move down the initial rating based on randomized trials to moderate or low.

We recognize that some types of evidence, such as evidence about public health interventions, quality improvement studies, and studies of diagnostic tests, may be challenging to rate. With these nontherapeutic intervention questions, the challenge to the EPCs is to determine the study design that is most appropriate to minimize the risk of bias. For example, the EPCs may find that particular types of studies, such as interrupted time series, reduce the risk of bias more than do other types of observational studies. Although the EPCs can take into account criteria other than those specified expressly by GRADE in assessing the risk of bias of observational (nonrandomized) studies as moderate, we caution that changing the assessment of observational studies for risk of bias should be done judiciously.

AHRQ CERs have often focused on pharmaceutical therapies, for which both efficacy and effectiveness trials[15] are a major source of information. The domains discussed above are directly relevant to studies of most drugs. In the future, CERs may increasingly assess diagnostic tests or strategies. For these technologies, RCTs may not be the origin of much relevant information, and the studies that are available may have special methodologic features. Further conceptual or empirical work may be warranted to explore whether the EPC approach to grading strength of evidence described here remains appropriate for such interventions. EPCs are encouraged to keep careful records of the application of these methods to nonpharmacologic interventions.

In arriving at an overall strength-of-evidence grade, the crucial requirement is transparency. The EPC method implies that EPCs can, if they choose, make a global assessment of the overall quality of evidence rather than explicitly use scores for each domain and then combine them. Nevertheless, EPCs are encouraged to make judgments for individual domains as a first step and to be especially sensitive to the effects of any "borderline" scores for those domains and their impact on the overall score. Being explicit and transparent about what criteria are used to raise or lower grades is the essential element in this step.

As noted earlier, the EPC approach emphasizes assessment of applicability separately from strength of evidence. GRADE also addresses applicability, which is incorporated within the general concept of directness. The rationale for the EPC approach is that many stakeholders use EPC reviews for developing guidelines or making clinical or health policy decisions, and they may have quite different views on how much, or little, the evidence applies to populations of interest to them. Future EPC reports will have a discussion and information about applicability, and the intention is for the various users and audiences to read this section of the report and make their own judgments.

A consistent approach for grading the strength of evidence—one that decisionmakers can readily recognize and interpret—is highly desirable. To that end, the EPCs and the GRADE working group will continue to collaborate to facilitate consistency across grading systems. Refinements and modifications of the approach outlined here can be found at

www.effectivehealthcare.ahrq.gov as they become available. Meanwhile, this paper codifies the interim guidance that EPCs can follow to strengthen the consistency within the AHRQ program's current and coming reports and products.

Acknowledgments

The authors thank Valerie King, M.D., M.P.H., of the John M. Eisenberg Center at Oregon Health & Science University, for her insightful comments on an earlier draft and Loraine Monroe, of RTI International, for superior assistance with manuscript preparation. We thank Gordon Guyatt, Holger Schünemann, and the members of the GRADE working group for their work on rating quality of evidence, for very helpful discussions about our approach, and for comments on the manuscript.

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.

3 Helfand M, Balshem H. Principles for developing guidance: AHRQ and the Effective Health-Care Program. J Clin Epidemiol 2009; in press.

4 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.

5 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.

6 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.

7 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). Rockville, MD: Agency for Healthcare Research and Quality, 2002. AHRQ Publication No. 02-E016.

8 Harris RP, Helfand M, Woolf SH, et al. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med 2001 Apr;20(3 Suppl):21-35.

9 Treadwell JR, Tregear SJ, Reston JT, et al. A system for rating the stability and strength of medical evidence. BMC Med Res Methodol 2006;6:52.

10 Furukawa TA, Streiner DL, Hori S. Discrepancies among megatrials. J Clin Epidemiol 2000 Dec;53(12):1193-9.

11 Charlton BG. Fundamental deficiencies in the megatrial methodology. Curr Control Trials Cardiovasc Med 2001;2(1):2-7.

12 Langer G, Schloemer G, Knerr A, et al. Nutritional interventions for preventing and treating pressure ulcers. Cochrane Database Syst Rev 2003(4):CD003216.

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

14 Sackett D. The principles behind the tactics of performing therapeutic trials. In: Haynes RBS, Guyatt DL, Gordon H, Tugwell P, eds. Clinical epidemiology: how to do clinical practice research. New York: Lippincott Williams & Wilkins; 2005.

15 Gartlehner G, Hansen RA, Nissman D, et al. A simple and valid tool distinguished efficacy from effectiveness studies. J Clin Epidemiol 2006 Oct;59(10):1040-8.

Table 1. Required domains and their definitions

Domain	Definition and elements	Score and application
Risk of bias	Risk of bias is the degree to which the included studies for a given outcome or comparison have a high likelihood of adequate protection against bias (i.e., good internal validity), assessed through two main elements: • Study design (e.g., RCTs or observational studies) • Aggregate quality of the studies under consideration. Information for this determination comes from the rating of quality (good/fair/poor) done for individual studies	Use one of three levels of aggregate risk of bias: • **Low** risk of bias • **Medium** risk of bias • **High** risk of bias
Consistency	The principal definition of consistency is the degree to which reported effect sizes from included studies appear to have the same direction of effect. This can be assessed through two main elements: • Effect sizes have the same sign (that is, are on the same side of "no effect") • The range of effect sizes is narrow.	Use one of three levels of consistency: • **Consistent** (i.e., no inconsistency) • **Inconsistent** • **Unknown or not applicable** (e.g., single study) As noted in the text, single-study evidence bases (even mega-trials) cannot be judged with respect to consistency. In that instance, use "consistency unknown (single study)."
Directness	The rating of directness relates to whether the evidence links the interventions directly to health outcomes. For a comparison of two treatments, directness implies that head-to-head trials measure the most important health or ultimate outcomes. Two types of directness, which can coexist, may be of concern: Evidence is indirect if: • It uses intermediate or surrogate outcomes instead of health outcomes. In this case, one body of evidence links the intervention to intermediate outcomes and another body of evidence links the intermediate to most important (health or ultimate) outcomes. • It uses two or more bodies of evidence to compare interventions A and B— e.g., studies of A vs. placebo and B vs. placebo, or studies of A vs. C and B vs. C but not A vs. B. Indirectness always implies that more than one body of evidence is required to link interventions to the most important health outcomes. Directness may be contingent on the outcomes of interest. EPC authors are expected to make clear the outcomes involved when assessing this domain.	Score dichotomously as one of two levels of directness: • **Direct** • **Indirect** If indirect, specify which of the two types of indirectness accounts for the rating (or both, if that is the case)—namely, use of intermediate/surrogate outcomes rather than health outcomes and use of indirect comparisons. Comment on the potential weaknesses caused by, or inherent in, the indirect analysis. The EPC should note if both direct and indirect evidence was available, particularly when indirect evidence supports a small body of direct evidence.
Precision	Precision is the degree of certainty surrounding an effect estimate with respect to a given outcome (i.e., for each outcome separately). If a meta-analysis was performed, this will be the confidence interval around the summary effect size.	Score dichotomously as one of two levels of precision: • **Precise** • **Imprecise** A precise estimate is an estimate that would allow a clinically useful conclusion. An imprecise estimate is one for which the confidence interval is wide enough to include clinically distinct conclusions. For example, results may be statistically compatible with both clinically important superiority and inferiority (i.e., the direction of effect is unknown), a circumstance that will preclude a valid conclusion.

EPC, Evidence-based Practice Center; RCT, randomized controlled trial.

Table 2. 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 additional domain should be rated if studies in the evidence base have noted levels of exposure. Use one of three levels: • **Present:** Dose-response pattern observed • **Not present:** No dose-response pattern observed (dose-response relationship *not* present) • **NA (not applicable or not tested)**
Plausible confounding that would decrease observed effect	Occasionally, in an observational study, plausible confounding factors would work in the direction *opposite* that of the observed effect. Had these confounders not been present, the observed effect would have been even larger than the one observed. In such a case, an EPC may wish to upgrade the level of evidence.	This additional domain should be considered if plausible confounding exists that would decrease the observed effect. Use one of two levels: • **Present:** Confounding factors that would decrease the observed effect may be present • **Absent:** Confounding factors that would decrease the observed effect are not likely to be present
Strength of association (magnitude of effect)	Strength of association refers to the likelihood that the observed effect is large enough that it cannot have occurred solely as a result of bias from potential confounding factors.	This additional domain should be considered if the effect size is particularly large. Use one of two levels: • **Strong:** Large effect size that is unlikely to have occurred in the absence of a true effect of the intervention • **Weak:** Small enough effect size that it could have occurred solely as a result of bias from confounding factors
Publication bias	Publication bias indicates that studies may have been published selectively, with the result that the estimated effect of an intervention based on published studies does not reflect the true effect. The finding that only a small proportion of relevant trials (or other studies) has been published or reported in a results database may indicate a higher risk of publication bias, which in turn may undermine the overall robustness of a body of evidence.	Publication bias need not be formally scored. However, it can influence ratings of consistency, precision, magnitude of effect, and, to a lesser degree, risk of bias and directness. If EPCs identify unpublished trials and if the results differ from those of published studies, they can take these factors into account in their rating for consistency and in calculating a summary confidence interval for an effect. We encourage authors to comment on publication bias when circumstances suggest that relevant empirical findings, particularly negative or no-difference findings, have not been published or are not otherwise available.

EPC, Evidence-based Practice Center.

Table 3. Strength-of-evidence grades and definitions

Grade	Definition
High	**High confidence that the evidence reflects the true effect.** Further research is very unlikely to change our confidence in the estimate of effect.
Moderate	**Moderate confidence that the evidence reflects the true effect.** Further research may change our confidence in the estimate of effect and may change the estimate.
Low	**Low confidence that the evidence reflects the true effect.** Further research is likely to change the confidence in the estimate of effect and is likely to change the estimate.
Insufficient	Evidence either is unavailable or does not permit a conclusion.

Table 4. Treatment 1 vs. Treatment 2: Numbers of studies and subjects, strength-of-evidence domains, magnitude of effect, and strength of evidence for key outcomes

Number of studies; subjects	**Domains pertaining to strength of evidence**				**Magnitude of effect and strength of evidence**
	Risk of bias:	Consistency	Directness	Precision	Absolute risk difference per 100 patients
Mortality					**Insufficient SOE**
1;80	RCT/Medium	Unknown	Direct	Imprecise	-1 (95% CI -4 to +3)
14;384	Retrospective cohort/ Medium	Inconsistent	Direct	Imprecise	-7 to +5 (range)
Myocardial infarction					**Low SOE**
7: 625	Retrospective cohort/ High	Consistent	Direct	Precise	-3 (95% CI -5 to -1)
Severe diarrhea					**Moderate SOE**
4; 256	RCTs/ Medium	Consistent	Direct	Imprecise	-4 (95% CI -8 to +1)
14; 28,400	Cohort / Medium	Consistent	Direct	Precise	-5 (95% CI -8 to -2)
Improved quality of life					**High SOE**
6; 265	RCTs/ Low	Consistent	Direct	Precise	-5 (95% CI -1 to -7)
Ulcer healing					**High SOE**
6; 265	RCTs/ Low	Consistent	Direct	Precise	+12 (95% CI +4 to +27)
5; 684	Retrospective cohort / Low	Consistent	Direct	Precise	+17 (95% CI +12 to +22)

CI, confidence interval; RCT, randomized controlled trial; SOE, strength of evidence.