Methods Guide
for Comparative Effectiveness Reviews

Assessing the Risk of Bias of Individual Studies in Systematic Reviews of Health Care 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 Effectiveness and 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 Effectiveness and Comparative Effectiveness Reviews and the Effective Health Care Program can be made at www.effectivehealthcare.ahrq.gov.

This document was written with support from the Effective Health Care Program at AHRQ.

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

Suggested citation: Viswanathan M, Ansari MT, Berkman ND, Chang S, Hartling L, McPheeters LM, Santaguida PL, Shamliyan T, Singh K, Tsertsvadze A, Treadwell JR. Assessing the Risk of Bias of Individual Studies in Systematic Reviews of Health Care Interventions. Agency for Healthcare Research and Quality Methods Guide for Comparative Effectiveness Reviews. March 2012. AHRQ Publication No. 12-EHC047-EF. Available at: www.effectivehealthcare.ahrq.gov/

Assessing the Risk of Bias of Individual Studies in Systematic Reviews of Health Care Interventions

Authors:

Meera Viswanathan, Ph.D.[a]

Mohammed T. Ansari, M.B.B.S., M.Med.Sc, M.Phil[b]

Nancy D. Berkman, Ph.D.[a]

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

Lisa Hartling, Ph.D.[d]

Melissa McPheeters, M.P.H., Ph.D.[e]

P. Lina Santaguida, P.T., Ph.D.[f]

Tatyana Shamliyan, M.D., M.S.[g]

Kavita Singh, M.P.H.[b]

Alexander Tsertsvadze, M.D., M.Sc.[b]

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

[a]RTI International–University of North Carolina at Chapel Hill Evidence-based Practice Center, Research Triangle Park, NC
[b]University of Ottawa Evidence-based Practice Center, Ottawa, Ontario, Canada
[c]Agency for Healthcare Research and Quality, Rockville, MD
[d]University of Alberta Evidence-based Practice Center, Edmonton, Alberta, Canada
[e]Vanderbilt University Evidence-based Practice Center, Nashville, TN
[f]McMaster University Evidence-based Practice Center, Hamilton, Ontario, Canada
[g]Minnesota University Evidence-based Practice Center, Minneapolis, MN
[h]ECRI Institute Evidence-based Practice Center, Plymouth Meeting, PA

Acknowledgments

The authors gratefully acknowledge the following individuals for their contributions to this project: Kathleen N. Lohr, Ph.D.; Mark Helfand, M.D., M.P.H.; Jeffrey C. Andrews, M.D.; and Loraine Monroe, EPC Publications Specialist. We also wish to acknowledge the thoughtful contributions of Susan Norris, M.D., M.Sc., M.P.H., our Associate Editor.

Assessing the Risk of Bias of Individual Studies in Systematic Reviews of Health Care Interventions

Key Points

- The task of assessing the risk of bias of individual studies is part of assessing the strength of a body of evidence. In preparation for evaluating the overall strength of evidence, reviewers should separate criteria for assessing risk of bias of individual studies from those that assess precision, directness, and applicability.
- EPCs may choose to use the terms "assessment of risk of bias" or "quality assessment." EPCs should define clearly the term used in their systematic review (SR) and comparative effectiveness review (CER) protocols and describe the constructs included as part of the assessment.
- We recommend that AHRQ reviews:
  - Opt for tools that are specifically designed for use in systematic reviews; have demonstrated acceptable validity and reliability; specifically address items related to methodological quality (internal validity) and preferably are based on empirical evidence of bias; where available, are specific to the study designs being evaluated; and avoid the presentation of risk of bias assessment as a composite score.
  - Do not use study design labels (e.g., RCT, cohort, case-control) as a proxy for assessment of risk of bias of individual studies.
  - Explicitly evaluate risk of selection, performance, attrition, detection, and selective outcome reporting biases.
  - Allow for separate risk of bias ratings by outcome to account for outcome-specific variations in detection bias and selective outcome reporting bias. Categories of outcomes, such as harms and benefits, may have different sources of bias.
  - Select items from recommended criteria for each included study design, as appropriate for the topics.
  - Evaluate validity and reliability of outcome measures as a component of detection bias and fidelity to the protocol as a component of performance bias.
  - Generally speaking, exclude precision and applicability when assessing the risk of bias because these are assessed in other domains when evaluating the strength of a body of evidence.
  - Assess risk of bias based on study design and conduct rather than reporting. Poorly reported studies may be judged as unclear risk of bias.
  - Not rely solely on poor reporting, industry funding, or disclosed conflict of interest, to rate a study as high risk of bias, although reviewers should report these issues transparently.
  - Conduct sensitivity analyses, when appropriate, for the body of evidence to evaluate whether poor reporting, industry funding. or disclosed conflict of interest may be associated with the studies' results. Industry funding or other conflict of interest may raise the risk of bias in design, analysis, and reporting. Reviewers suspecting high risk of bias because of industry funding should pay attention to the risk of selective outcome reporting.
  - Define decision rules for assessing the risk of bias category for each outcome from an individual study to improve transparency and reproducibility.

- Conduct dual assessment of risk of bias.

Introduction

This document updates the existing Agency for Healthcare Research and Quality (AHRQ) Evidence-based Practice Center (EPC) Methods Guide for Effectiveness and Comparative Effectiveness Reviews on assessing the risk of bias of individual studies. As with other AHRQ methodological guidance, our intent is to present standards that can be applied consistently across EPCs and topics, promote transparency in processes, and account for methodological changes in the systematic review process. These standards are based on available empirical evidence, theoretical principles, or workgroup consensus: as greater evidence accumulates in this methodological area, our standards will continue to evolve. When possible, our recommended standards offer flexibility to account for the wide range of AHRQ EPC review topics and included study designs.

Some EPC reviews may rely on an assessment of high risk of bias to serve as a threshold between included and excluded studies; in addition, EPC reviews use risk-of-bias assessments in grading the strength of the body of evidence. Assessment of risk of bias as unclear, high, medium, or low may also guide other steps in the review process, such as study inclusion for qualitative and quantitative synthesis, and interpretation of heterogeneous findings.

This guidance document begins by defining terms as appropriate for the EPC program, explores the potential overlap in various constructs used in different steps of the systematic review, and offers recommendations on the inclusion and exclusion of constructs that may apply to multiple steps of the systematic review process. We note that this guidance applies to reviews—such as AHRQ-funded reviews—that separately assess the risk of bias of outcomes from individual studies, the strength of the body of evidence, and applicability of the findings. This guidance applies to comparative effectiveness reviews that require interventions with comparators and systematic reviews that may include noncomparative studies. A key construct, however, is that risk-of-bias assessments judge whether the design and conduct of the study compromised the believability of the link between exposure and outcome. This guidance may not be relevant for reviews that combine evaluations of risk of bias or quality of individual studies with applicability.

Later sections of this guidance document provide guidance on the stages involved in assessing risk of bias and design-specific minimum criteria to evaluate risk of bias. We discuss and recommend tools and conclude with guidance on summarizing risk of bias.

Terminology and Constructs

Differences in Terminology

Risk of bias, defined as the risk of "a systematic error or deviation from the truth, in results or inferences,"[1] is interchangeable with internal validity, defined as "the extent to which the design and conduct of a study are likely to have prevented bias"[2] or "the extent to which the results of a study are correct for the circumstances being studied."[3] Despite the central role of the assessment of the believability of individual studies in conducting systematic reviews, the specific term used has varied considerably across review groups. A common alternative to "risk of bias" is "quality assessment," but the meaning of the term *quality* varies, depending on the source of the guidance. One source defines quality as "the extent to which all aspects of a study's

design and conduct can be shown to protect against systematic bias, nonsystematic bias, and inferential error."[4] The Grading of Recommendations Assessment, Development and Evaluation Working Group (GRADE) uses the term quality to refer to *an individual study* and judgments based about the strength of the *body of evidence* (quality of evidence).[5] The U.S. Preventive Services Task Force (USPSTF) equates quality with internal validity and classifies individual studies first according to a hierarchy of study design and then by individual criteria that vary by type of study.[6] In contrast, the Cochrane collaboration argues for wider use of the phrase "risk of bias" instead of "quality," reasoning that "an emphasis on risk of bias overcomes ambiguity between the quality of reporting and the quality of the underlying research (although does not overcome the problem of having to rely on reports to assess the underlying research)."[1]

Because of inconsistency and potential misunderstanding in the use of the tem "quality," this guidance will refer to risk of bias. We understand risk of bias to refer to the extent to which a single study's design and conduct protect against all bias in the estimate of effect using the more precise terminology "assessment of risk of bias." Thus, assessing the risk of bias of a study can be thought of as assessing the risk that the study results reflect bias in study design or execution in addition to the true effect of the intervention or exposure under study.

Guidance on Terminology

This guidance uses risk of bias as the preferred terminology. Nonetheless, we recognize the competing demands for flexibility across reviews to account for specific clinical contexts and consistency within review teams and across EPCs. We advocate transparency of planned methodological approach and documentation of decisions and therefore recommend that EPCs define the term selected in their SR and Comparative Effectiveness Review (CER) protocols and describe the constructs included in the assessment.

Differences in Constructs Included in Risk-of-Bias Assessment

Across prior guidance documents and instruments, the types of constructs included in risk of bias or quality assessments have included one or more of the following issues: (1) conduct of the study/internal validity, (2) random error, (3) external validity or applicability, (4) completeness of reporting, (5) selective outcome reporting, (6) choice of outcome measures, (7) study design, (8) fidelity of the intervention, and (9) conflict of interest in the conduct of the study.

The lack of agreement on what constructs to include in risk-of-bias assessment stems from two sources. First, no strong empirical evidence supports one approach over another; this gap leads to a proliferation of approaches based on the practices of different academic disciplines and the needs of different clinical topics. Second, in the absence of updated guidance on risk-of-bias assessment that accounts for how new guidance on related components of systematic reviews (such as selection of evidence,[7] assessment of applicability,[8] or grading the strength of evidence[5,9-17]) relate to, overlap with, or are distinct from risk-of-bias assessment of individual studies, some review groups continue to use quality practices that have served well in the past.

In the absence of strong empirical evidence, methodological decisions in this guidance document rely on epidemiological principles.[1] Thus, this guidance document presents a conservative path forward. Systematic reviewers have the responsibility to evaluate potential sources of bias and error if these concerns could plausibly influence study results; we include these concerns even if no empirical evidence exists that they influence study results.

Guidance on Constructs To Include or Exclude From Risk-of-Bias Assessment

The constructs selected in the assessment of risk of bias may differ because of the academic orientation of the reviewers, guidelines by sponsoring organizations, and clinical topic. In AHRQ-sponsored reviews, recent guidance and requirements for systematic reviews have reduced the variability in other related steps of the systematic review process and, therefore, allow for greater consistency in risk-of-bias assessment as well. Some constructs that EPCs may have considered part of risk of bias (or quality) assessment in the past now overlap with or fall within the domains of other systematic review tasks. Table 1 illustrates which constructs to include for each systematic review task when systematic reviews separately assess the risk of bias of individual studies, the strength of the body of evidence, and applicability of the findings for individual studies. We note that the GRADE approach to grading the strength of evidence incorporates applicability within strength of evidence assessments,[12] and the AHRQ-EPC approach does not, but the distinction between concepts relevant for risk of bias and applicability are relevant to both systems.[9]

Table 1. Inclusion and exclusion of constructs for risk-of-bias assessment, applicability, and strength of evidence

Construct	Included in appraisal of individual study risk of bias?	Included in assessing applicability of studies and the body of evidence?	Included in grading strength of the body of evidence?
Risk of bias (from selection bias and confounding, attrition, performance, detection, reporting, and other biases)	Yes	No	Yes (required domain of risk of bias)
Precision	Only when no quantitative pooling or presentation is possible	No	Yes (required domain of precision)
Applicability/external validity	Only when components of applicability influence risk of bias (e.g., duration of follow-up varies across intervention arms)	Yes	Depends on the SOE system. GRADE includes applicability as part of directness, AHRQ-EPC does not (with the exception of rating surrogate outcomes as indirect evidence)
Poor or inadequate reporting	Yes, studies may be rated as having unclear risk of bias	No	No
Selective outcome reporting	Yes, only when judgments can be made about the impact of differences between outcomes listed in full protocol and published materials	No	Yes
Outcome measures	Yes (potential for outcome measurement bias, specifically validity, reliability, variation across study arms)	Yes (applicability of outcomes measures)	Yes (directness of outcome measures)

Table 1. Inclusion and exclusion of constructs for risk-of-bias assessment, applicability, and strength of evidence (continued)

Construct	Included in appraisal of individual study risk of bias?	Included in assessing applicability of studies and the body of evidence?	Included in grading strength of the body of evidence?
Study design	Assessment should evaluate the relevant sources of risk of bias by study design rather than rate the study risk of bias by design labels alone	No	Yes (overall risk of bias is rated separately for randomized and nonrandomized studies)
Fidelity to protocol	Yes	Yes	No
Conflict of interest from sponsor bias	Indirectly (sponsor bias may influence one or more sources of bias)	Indirectly (sponsor bias may limit applicability)	Indirectly (sponsor bias may influence domains of risk of bias, directness, and publication bias)

Abbreviations: GRADE=Grading of Recommendations Assessment, Development and Evaluation; SOE=strength of evidence.

Types of Bias Included in Assessment of Risk of Bias

Numerous, often discipline-specific, taxonomies exist for classifying the different phenomena that introduce bias in studies.[18] For example, although some use the terms confounding and selection bias interchangeably, others see a very clear structural difference between the two and the manner in which they should be handled when detected.[19] What constitutes performance and detection bias in one scheme may be classified under the broader category of information bias in another[1,20] Irrespective of the different classification schemes, the end result identifies associations that are either spurious or related to a variable other than intervention/exposure. We use the taxonomy suggested by Higgins et al. in the Cochrane Handbook as a common, comprehensive, and well-disseminated approach (Table 2).[1] Subsequent sections of this guidance refer to this taxonomy of biases.

Table 2. Taxonomy of core biases in the Cochrane Handbook[1]

Types of bias related to conduct of the study (including analysis and reporting)	Definition	Risk of bias assessment criteria
Selection bias and confounding*	Systematic differences between baseline characteristics of the groups that arise from self-selection of treatments, physician-directed selection of treatments, or association of treatment assignments with demographic, clinical, or social characteristics. Includes Berkson's bias, nonresponse bias, incidence-prevalence bias, volunteer/self-selection bias, healthy worker bias, and confounding by indication/contraindication (when patient prognostic characteristics, such as disease severity or comorbidity, influence both treatment source and outcomes).	Randomization, allocation concealment, sequence generation, control for confounders in cohort studies, and case matching in case-control studies
Performance bias	Systematic differences in the care provided to participants and protocol deviation. Examples include contamination of the control group with the exposure or intervention, unbalanced provision of additional interventions or co-interventions, difference in co-interventions, and inadequate blinding of providers and participants.	Fidelity to protocol, unintended interventions or co-interventions

Table 2. Taxonomy of core biases in the Cochrane Handbook[1] (continued)

Types of bias related to conduct of the study (including analysis and reporting)	Definition	Risk of bias assessment criteria
Attrition bias	Systematic differences in the loss of participants from the study and how they were accounted for in the results (e.g., incomplete follow-up, differential attrition). Those who drop out of the study or who are lost to follow-up may be systematically different from those who remain in the study. Attrition bias can potentially change the collective (group) characteristics of the relevant groups and their observed outcomes in ways that affect study results by confounding and spurious associations.	Completeness of outcome data, intention-to-treat analysis with appropriate imputations for missing data, and completeness of follow-up
Detection bias	Systematic differences in outcomes assessment among groups being compared, including systematic misclassification of the exposure or intervention, covariates, or outcomes because of variable definitions and timings, diagnostic thresholds, recall from memory, inadequate assessor blinding, and faulty measurement techniques. Erroneous statistical analysis might also affect the validity of effect estimates.	Blinding of outcome assessors, especially with subjective outcome assessments, bias in inferential statistics, valid and reliable measures
Reporting bias	Systematic differences between reported and unreported findings (e.g., differential reporting of outcomes or harms, incomplete reporting of study findings, potential for bias in reporting through source of funding).	Selective outcome reporting evaluation by comparing study report and (a) protocol or (b) outcomes prespecified in methods

*One approach defines selection bias as the bias that occurs when selection is conditioned on common effects of exposures and outcomes and confounding as the bias that occurs when exposure and outcome have a common cause.[19] According to another classification scheme, selection bias is differential selection affected by exposure/intervention in the study, while confounding is differential selection that occurs before exposure and disease.[20]

A brief review of *Cochrane Handbook of Systematic Reviews,*[1] *Systems to Rate the Strength of Scientific Evidence,*[21] and *Evaluation of Non-randomized Studies*[22] shows empirical evidence for detection bias, attrition bias, and reporting bias.

Risk of Bias and Precision

One key distinction between risk of bias and quality assessment is in the treatment of precision. As noted earlier, one definition of quality subsumes freedom from nonsystematic bias or random error.[4] Tools relying on this definition of quality have included the evaluation of sample size and power to evaluate the impact of random error on the precision of estimates.[23]

Both GRADE[24] and AHRQ guidance on evaluating the strength of evidence[9] separate the evaluation of precision from that of risk of bias. Systematic reviews now routinely evaluate precision (through consideration of the confidence intervals around a summary effect size from pooled estimates) when grading the strength of the body of evidence.[9] Under such circumstances, the evaluation of degree to which studies were designed to allow a precise enough estimate would constitute double-counting limitations to the evidence from a single source. We recommend that AHRQ reviews exclude evaluation of the ability of the study to obtain a precise estimate when assessing the risk of bias for outcomes that can be pooled in meta-analysis or presented quantitatively for single-study bodies of evidence. When outcomes cannot be pooled (as with highly heterogeneous bodies of evidence) or presented quantitatively, assessing the

extent to which individual studies are designed to obtain precise estimates in addition to (but separately from) risk of bias may be appropriate.

Risk of Bias and Applicability

Many commonly used quality assessment tools evaluate external validity in addition to internal validity (risk of bias). A review of tools to rate observational studies identified 14 "best" tools. Each evaluated core elements of internal validity and included questions on representativeness of the sample (a component of applicability).[22] Guidance for the EPC program on how to address applicability (also known as external validity, generalizability, or relevance) recommends that EPCs provide a summary report of the applicability of the body of evidence separately from their judgment of the applicability of individual studies.[8] This guidance notes that although individual studies may not be representative of the population of interest, consistent findings across studies with individually limited generalizability may suggest broad applicability of the results.

We recommend that AHRQ reviews generally exclude considerations of applicability in risk-of-bias assessments of individual studies. We note, however, that some study features may be relevant to both risk of bias and applicability. Duration of follow-up is one such example: if duration of followup is different across comparison groups within a study, this difference could be a source of bias; the absolute duration of follow-up for the study would be relevant to the clinical context of interest and therefore the applicability of the study. Likewise study population may be considered within both risk of bias and applicability: if the populations are systematically different between comparison groups within a study (e.g., important baseline imbalances) this may be a source of bias; the population selected for the focus of the study (e.g., inclusion and exclusion criteria) would be a consideration of applicability. Reviewers need to clearly separate study features that may be potential sources of bias from those that are concerned with applicability outside of the individual study context.

Risk of Bias and Poor or Inadequate Reporting

In theory, internal validity focuses on design and conduct of a study. In practice, assessing the internal validity of a study requires adequate reporting of the study, unless additional information is obtained by reaching out to investigators. Although new standards on reporting seek to improve reporting of study design and conduct,[25-29] EPC review teams continue to need a practical approach to dealing with poor or inadequate reporting. The Cochrane risk of bias tool judges the risk of bias to be uncertain when information is inadequate. EPC reviews have varied in their treatment of reporting of study design and conduct; for example, some have elected to rate poorly *reported* studies as studies with high risk of bias. In general, we recommend that assessment of risk of bias focus primarily on the design and conduct of studies and not on the quality of reporting. EPCs may choose to select an "unclear risk of bias" category for studies with missing or poorly reported information on which to base risk of bias judgments. When studies include meta-analyses, we recommend that quantitative estimates of effect account, through sensitivity analyses, for the impact of including studies with high or unclear risk of bias.

Risk of Bias and Conflict of Interest From Sponsor Bias

Many studies examining the issue of financial conflict of interest have found that sponsor participation in data collection, analysis, and interpretation of findings can threaten the internal validity and applicability of primary studies and systematic reviews.[30,31] The pathways by which

sponsor participation can influence the validity of the results are manifold. They include the following:

1. selection of designs and hypotheses—for example, choosing noninferiority rather than superiority approaches,[32] picking comparison drugs and doses,[32] choosing outcomes,[31] or using composite endpoints (e.g., mortality and quality of life) without presenting data on individual endpoints;[33]
2. selective outcome reporting—for example, reporting relative risk reduction rather than absolute risk reduction or "cherry-picking" from multiple endpoints;[32]
3. differences in internal validity of studies and adequacy of reporting;[34]
4. biased presentation of results;[33] and
5. publication bias.[35]

EPCs can evaluate these pathways if and only if the relationship between the sponsor(s) and the author(s) is clearly documented; in some instances, such documentation may not be sufficient to judge the likelihood of conflict of interest (for example, authors may receive speaking fees from a third party that did not support the study in question).

Editors have grown increasingly concerned about the practice of ghost authoring (i.e., primary authors or substantial contributors are not identified) or guest authoring (i.e., one or more identified authors are not substantial contributors)[36] sponsored studies, a practice that makes the actual contribution of the sponsor very difficult to discern.[37,38]

All these concerns may lead to the conclusion that sponsorship from industry (i.e., for-profit entities) should be included as an explicit consideration for assessment of risk of bias. We concur that sponsorship of studies should be considered in critically appraising the evidence but caution against equating industry sponsorship with high risk of bias for three reasons. First, sponsor bias is not limited to industry; nonprofit and government-sponsored studies may also be guest- or ghost-authored. Moreover, the researchers may have various financial or intellectual conflicts of interest by virtue of, for example, accepting speaking fees from many sources.[39] Second, financial conflict is not the only source of conflict of interest: other potential conflicts include personal, professional, or religious beliefs, desire for academic recognition, and so on.[30] Third, the multiple pathways by which sponsorship may influence studies are not all solely within the domain of assessment of risk of bias: several of these pathways fall under the purview of other systematic review tasks. For instance, concerns about the choice of designs, hypotheses, and outcomes relate as much or more to applicability than other aspects of reviews. Reviewers can and should consider the likely influence of sponsor bias on selective outcome reporting, but when these judgments may be limited by lack of access to full protocols, the assessment of selective outcome reporting may be more easily judged for the body of evidence than for individual studies.

The biased presentation or "spin" on results, although of concern to the lay reader, if limited to the discussion and conclusion section of studies, should have no bearing on systematic review conclusions because systematic reviews do not rely on interpretation of data by study authors.

Internal validity and completeness of reporting constitute, then, the primary pathway by which sponsors may influence the validity of study results that is entirely within the domain of assessment of risk of bias. We acknowledge that this pathway may not be the most important source of sponsor influence: as standards for conduct and reporting of studies become widespread and journals require that they be met, differences in internal validity and reporting between industry-funded studies and other studies will likely attenuate. In balancing these

considerations with the primary responsibility of the systematic reviewer—objective and transparent synthesis and reporting of the evidence—we make three recommendations: (1) at a minimum, EPCs should routinely report the source of each study's funding; (2) EPCs should consider issues of selective outcome reporting at the individual study level and for the body of evidence; and (3) EPCs should conduct sensitivity analyses for the body of evidence when they have reason to suspect that the source of funding or disclosed conflict of interest is influencing studies' results.[32] One limitation of relying on sensitivity analyses to demonstrate evidence of risk of bias for industry-funded studies when sponsor bias is suspected (rather than assuming higher risk for industry-funded studies) is that newer studies may appear to be biased when compared to older studies, because of changes in journal reporting standards.

Risk of Bias and Selective Outcome Reporting

Selective outcome reporting refers to the selection of a subset of analyses for publication based on results[40] and has major implications for both the risk of bias of individual studies and the strength of the body of evidence. Comparisons of the full protocol to published or unpublished results can help to flag studies that selectively report outcomes. In the absence of access to full protocols,[9,17] Guyatt et al. note as follows:

> Selective reporting is present if authors acknowledge pre-specified outcomes that they fail to report or report outcomes incompletely such that they cannot be included in a meta-analysis. One should suspect reporting bias if the study report fails to include results for a key outcome that one would expect to see in such a study or if composite outcomes are presented without the individual component outcomes.[17,p 409]

Methods continue to be developed for identifying and judging the risk of bias when results deviate from protocols in the timing or measure of the outcome. No guidance currently exists on how to evaluate the risk of selective outcome reporting in older studies with no published protocols or whether to downgrade all evidence from a study where comparisons between protocols and results show clear evidence of selective outcome reporting for some outcomes.

Even when access to protocols is available, the evaluation of selective outcome reporting may be required again at the level of the body of evidence. Selective outcome reporting across several studies for a body of evidence may result in downgrading the body of evidence.[17]

Previous research has established the link between industry funding and publication bias, a form of reporting bias in which the decision to selectively publish the entire study is based on results.[41] Publication bias may be a pervasive problem in some bodies of evidence and should be evaluated when grading the body of evidence. New research is emerging on selective outcome reporting in industry-funded studies.[42] As methods on identifying and weighing the likely effect of selective outcome reporting continue to be developed, this guidance will also require updating. Our current recommendation is to consider the risk of selective outcome reporting for individual studies and the body of evidence, particularly when a suspicion exists that forces such as sponsor bias may influence the reporting of outcomes.

Risk of Bias and Outcome Measures

The use of valid and reliable outcome measures reduces the likelihood of detection bias. For example, studies relying on self-report measures may be rated as having a higher risk of bias than studies with clinically observed outcomes. In addition, differential assessment of outcome

measures by study arm (e.g., electronic medical records for control arm versus questionnaires for intervention arm) constitute a source of measurement bias and should, therefore, be included in assessment of risk of bias. We recommend that assessment risk of bias of individual studies include the evaluation of the validity and reliability of outcome measures, and their variation across study arms.

Recent guidance on the evaluation of applicability by Atkins and colleagues states the importance of considering the relevance of outcome measures for judging applicability (or external validity) of the evidence.[43] For instance, studies that focus on short-term outcomes and fail to report long-term outcomes may be judged as having poor applicability or not being directly relevant to the clinical question for the larger population. The choice of specific outcome measures is a consideration when judging applicability and directness rather than risk of bias; their validity and reliability, on the other hand, is a component of risk of bias, as noted above.

Risk of Bias and Study Design

Some designs possess inherent features (such as randomization and control arms) that reduce the risk of bias and increase the potential for causal inference, particularly when considering benefit of the intervention. Other study designs have specific and inherent risks of biases that cannot be minimized. The clinical question will dictate which study designs are suitable to answer a specific question. EPCs consider these design-specific sources of bias at two points in the systematic review process: (1) when evaluating whether to admit observational studies into the review and (2) when evaluating individual studies for design-specific risks of bias. Norris et al. note that the default strategy in systematic reviews should be to *consider* including observational studies for evidence of benefit and the decision rests on the answer to two questions: (1) are there gaps in the trial evidence for the review questions under consideration? and (2) will observational studies provide valid and useful information to address key questions?[7] In considering whether observational studies provide valid and useful information for benefit, EPCs will need to consider the likelihood that observational studies will generally have more numerous and more serious sources of bias than trials. Once an EPC makes the decision to include observational studies, then the review team needs to evaluate each study based on the risks of bias specific to that design.

Both AHRQ and GRADE approaches to evaluating the strength of evidence include study design and conduct (risk of bias) of individual studies as components needed to evaluate body of evidence. The inherent limitations present in observational designs (e.g., absence of randomization) are factored in when grading the strength of evidence, EPCs generally give evidence derived from observational studies a low starting grade and evidence from randomized controlled trials a high grade. They can then upgrade or downgrade the observational and randomized evidence based on the strength of evidence domains (i.e., risk of bias of individual studies, directness, consistency, precision, and additional domains if applicable).[9]

Because systematic reviews evaluate design-specific sources of bias in selecting studies for inclusion in the review and then use study design as a component of risk of bias in judging the strength of evidence, we recommend that EPCs do not use study design labels as a proxy for assessment of risk of bias of individual studies. In other words, EPCs should not downgrade the risk of bias of *individual* studies on the basis solely of study design because doing so would penalize studies again (i.e., at the level of individual studies and the body of evidence). This approach accounts for the fact that a study can be performed with the highest quality *for that study design* but still have some (if not serious) potential risk of bias.[1] This approach also

acknowledges that quality varies, perhaps widely, within designs and that some study designs do have inherent limitations that can never be fully overcome when considering the validity of their results for benefits. For observational studies, an important consideration is to make a list of possible biases based on the topic and specific design and then evaluate their potential importance for each study.

This approach does not, however, address the fact that no grading system presently accounts for variations in potential risk of bias from different types of observational studies. Under current systems of grading strength of evidence, reviews that consider including observational study designs with highly varying risks of bias (e.g., case reports and data from large registries) for the same clinical question would evaluate all such observational designs together in strength of evidence grades. Under such circumstances, our guidance is to consider the question of value to the review with regard to each study design type: "Will [case reports/case series/case control studies, etc.] provide valid and useful information to address key questions?" Depending on the clinical question, the sources of bias from a particular study design may be so large as to constitute an unacceptably high risk of bias. For instance, EPCs may judge information on benefits from case series of interventions as having a very high risk of bias. In such instances, we recommend that EPCs exclude such designs from the review rather than include the study and then apply a common rating of high risk of bias across all studies with that design without consideration of individual variations in study performance.

In summary, this approach allows EPCs to deal with variations in included studies by study design, for instance by rating outcomes for benefit from individual randomized controlled trials (RCTs), or observational studies, as low, medium, high, or unclear risk of bias. It then defers the issue of study design limitations to assessment of the strength of evidence.

Risk of Bias and Fidelity to the Intervention Protocol

Failure of the study to maintain fidelity to the intervention protocol can influence performance bias; it is, therefore, a component of assessment of risk of bias. We note, however, that the interpretation of fidelity may differ by clinical topic. For instance, some behavioral interventions include "fluid" interventions; these involve interventions for which the protocol explicitly allows for modification based on patient needs; such fluidity does not mean the interventions are implemented incorrectly. When interventions implement protocols that have minimal concordance with practice, the discrepancy may be considered an issue of applicability. This lack of concordance with practice does not, however, constitute risk of bias. We also note that when studies implement an intervention with previously established efficacy in varied settings but are unwilling or unable to maintain fidelity to the original intervention protocol, this deviation may influence the risk of bias of the study and the applicability of the intervention overall. We recommend that EPCs account for the specific clinical considerations in determining and applying criteria about fidelity for assessment of risk of bias. Our recommendation is consistent with the Institute of Medicine guidelines on systematic reviews.[44]

Stages in Assessing the Risk of Bias of Studies

International reporting standards require documentation of various stages in a comparative effectiveness review.[45-47] We lay out recommended approaches to assessment of risk of bias in five steps: protocol development, pilot testing and training, assessment of risk of bias, interpretation, and reporting. Table 3 describes the stages and specific steps in assessing the risk

of bias of individual studies that contribute to transparency through careful documentation of decisions.

Table 3. Stages in assessing the risk of bias of individual studies

Stages in risk-of-bias assessment	Specific steps
1. Develop protocol	• Specify terms (i.e., quality assessment or risk of bias) and included concepts • Explain the inclusion of specific risk-of-bias criteria • Select and justify choice of specific risk-of-bias rating tool(s) • Include tools for assessment of risk of bias that justify research-specific risk-of-bias standards and operational definitions of risk-of-bias criteria • Explain how individual risk-of-bias criteria will be summarized to obtain low, moderate, high, or unclear risk of bias for individual outcomes and justify any use of scales (numerical scores leading to categories of risk of bias) • Explain how inconsistencies between pairs of risk of bias reviewers will be resolved • Explain how the synthesis of the evidence will incorporate assessment of risk of bias (including whether studies with high or unclear risk of bias will be used in synthesis of the evidence)
2. Pilot test and train	• Determine composition of the review team. A minimum of two reviewers must rate the risk of bias of each study, with a third reviewer to serve as arbiter of conflicts • Train reviewers • Pilot test assessment of risk of bias tools using a small subset of studies that represent the range of risk of bias in the evidence base • Identify issues and revise tools or training as needed
3. Perform assessment of risk of bias of individual studies	• Determine study design of each (individual) study • Make judgments about each risk of bias criterion, using the preselected appropriate criteria for that study design and for each predetermined outcome • Make judgments about overall risk of bias for each included outcome of the individual study, considering study conduct, and categorize as low, moderate, high, or unknown risk of bias within study design; document the reasons for judgment and process for finalizing judgment • Resolve differences in judgment and record final rating for each outcome
4. Use assessment of risk of bias in synthesis of evidence	• Conduct preplanned analyses • Consider additional required analyses • Incorporate assessment of risk of bias in quantitative/qualitative synthesis, keeping study design categories separate
5. Report assessment of risk of bias process and limitations	• Cite reports on validation of the selected tool(s), the assessment of risk of bias process (summarizing from the protocol), and limitations to the process • Describe actions to improve assessment of risk-of-bias reliability if applicable

The plan for assessment of risk of bias should be included within the protocol for the entire review. As prerequisites to developing the plan for assessment of risk of bias, EPCs must identify the important intermediate and final outcomes that need assessment of risk of bias and other study descriptors or study data elements that are required for the assessment of risk of bias in the systematic review protocol. Protocols must justify what risk-of-bias criteria will be evaluated and how the reviewers will incorporate risk of bias of individual studies in the synthesis of evidence.

The assessment must include a minimum of two reviewers per study with a third to serve as arbitrator. EPCs should anticipate having to review and revise assessment of risk of bias forms and instructions in response to problems arising in training and pilot testing.

Assessment of risk of bias should be consistent with the analysis plans in registered protocols of the reviews. Published reviews must include risk-of-bias criteria and should describe the

selected tools and their reliability and validity when such information is available. EPC reviews should report all criteria used for each evaluated outcome. The synthesis of the evidence should reflect the *a priori* analytic plan for incorporating risk of bias of individual studies in qualitative or quantitative analyses. EPCs should report the outcomes of all preplanned analyses that included risk-of-bias criteria regardless of statistical significance or the direction of the effect. Published reviews should also include justifications of all *post hoc* decisions to limit synthesis of included studies to a subset with common methodological or reporting attributes.

Design-Specific Criteria To Assess Risk of Bias

We present design-specific criteria to assess risk of bias for five common study designs: RCTs, cohort (prospective, retrospective, and nonconcurrent), case-control (including nested case-control), case series, and cross-sectional (Table 4).[48] Table 4 draws on other instruments,[1,49] was modified based on workgroup consensus and peer review, and is not intended to serve as a one-size-fits-all instrument. Rather, it is intended to remind reviewers of common sources of bias for some common types of study designs. A critical task that reviewers need to incorporate within each review is the careful identification and recording of likely sources of bias for each topic and each included design. Reviewers may select specific criteria or combinations of criteria relevant to the topic. For instance, blinding of outcome assessors may not be possible for surgical interventions but the inability to blind outcome assessors does not obviate the risk of bias from lack of blinding. Reviewers should be alert to the use of self-reported or subjective outcome measures or poor controls for differential treatment in such studies that could elevate the risk of bias further.[1,50]

Table 4. Design-specific criteria to assess for risk of bias for benefits

Risk of bias	Criterion	RCTs	CCTs or cohort	Case-control	Case series	Cross-sectional
Selection bias	Was the allocation sequence generated adequately (e.g., random number table, computer-generated randomization)?	x				
	Was the allocation of treatment adequately concealed (e.g., pharmacy- controlled randomization or use of sequentially numbered sealed envelopes)?	x				
	Were participants analyzed within the groups they were originally assigned to?	x	x			
	Did the study apply inclusion/exclusion criteria uniformly to all comparison groups?		x			x
	Were cases and controls selected appropriately (e.g., appropriate diagnostic criteria or definitions, equal application of exclusion criteria to case and controls, sampling not influenced by exposure status)			x		
	Did the strategy for recruiting participants into the study differ across study groups?		x			
	Does the design or analysis control account for important confounding and modifying variables through matching, stratification, multivariable analysis, or other approaches?	x	x	x*	x	x
Performance bias	Did researchers rule out any impact from a concurrent intervention or an unintended exposure that might bias results?	x	x	x	x	x
	Did the study maintain fidelity to the intervention protocol?	x	x	x	x	
Attrition bias	If attrition (overall or differential nonresponse, dropout, loss to follow-up, or exclusion of participants) was a concern, were missing data handled appropriately (e.g., intention-to-treat analysis and imputation)?	x	x	x	x	x
Detection bias	In prospective studies, was the length of follow-up different between the groups, or in case-control studies, was the time period between the intervention/exposure and outcome the same for cases and controls?	x	x	x		
	Were the outcome assessors blinded to the intervention or exposure status of participants?	x	x	x	x	x
	Were interventions/exposures assessed/defined using valid and reliable measures, implemented consistently across all study participants?	x	x	x	x	x
	Were outcomes assessed/defined using valid and reliable measures, implemented consistently across all study participants?	x	x	x	x	x
	Were confounding variables assessed using valid and reliable measures, implemented consistently across all study participants?		x	x	x	x
Reporting bias	Were the potential outcomes prespecified by the researchers? Are all prespecified outcomes reported?	x	x	x	x	x

*Cases and controls should be similar in all factors known to be associated with the disease of interest, but they should not be so uniform as to be matched for the exposure of interest.

Another example of a criterion that requires topic-specific evaluation is prespecification of outcomes. Depending on the topic, prespecification of outcomes is entirely appropriate and expected, regardless of study design. For other topics, data from observational studies may offer the first opportunity to identify unexpected outcomes that may need confirmation from RCTs. For review topics in search of evidence on rare long-term outcomes, requiring prespecification would be inappropriate. A third example of a criterion requiring topic-specific evaluation is the expected attrition rate. Differential or overall attrition because of nonresponse, dropping out, loss to follow-up, and exclusion of participants can introduce bias when missing outcome data are related to both exposure/treatment and outcome. Reviewers of topics that focus on short-term clinical outcomes may select a low expected attrition rate. We also note that with attrition rate in particular, no empirical standard exists across all topics for demarcating a high risk of bias from a lower risk of bias; these standards are often set within clinical topics. The list of recommended criteria does not represent comprehensive sources of bias for other study designs. For instance, case series studies with repeated time measures may require a question asking whether the study accounted for regression to the mean. Some concepts included in Table 4, particularly intention-to-treat, have been interpreted in a variety of ways. The *Cochrane Handbook of Systematic Reviews* offers a more detailed treatment of intention to treat.[1]

Tools for Assessing Risk of Bias

EPCs can use one of two general approaches to assessing risk of bias in systematic reviews. One method is often referred to as a *components approach*. This involves assessing individual items that are deemed by the systematic reviewers to reflect the methodological risk of bias, or other relevant considerations, in the body of literature under study. For example, one commonly assessed component in RCTs is allocation concealment.[51] Reviewers assess whether the randomization sequence was concealed from key personnel and participants involved in a study before randomization; they then rate the component as adequate, inadequate, or unclear. The rating for each component is reported separately. The second common approach is to use a *composite approach* that combines different components related to risk of bias or reporting into a single overall score.

Many tools have emerged over the past 20 years to assess risk of bias. Some tools are specific to different study designs, whereas others can be used across a range of designs. Some have been developed to reflect nuances specific to a clinical area or field of research. Because many AHRQ systematic reviews typically address multiple research questions, they may require the use of several risk of bias tools or the selection of various different components to address all the study designs included.

- Currently there is no consensus on the best approach or preferred tool for assessing risk of bias, because the components associated with risk of bias are in contention. As such, there are a large number of tools available, and their marked variations and relative merits can be problematic for systematic reviewers. We advocate the following general principles when selecting a tool, or approach, to assessing risk of bias in systematic reviews. EPCs should opt for tools that:were specifically designed for use in systematic reviews;
- have demonstrated acceptable validity and reliability, or show transparency in how assessments are made by providing explicit support for each assessment;
- specifically address items related to risk of bias (internal validity), and preferably are based on empirical evidence of bias;

- where available, are specific to the study designs being evaluated; and
- avoid the presentation of risk-of-bias assessment as a composite score, that is, an overall numeric rating of study risk of bias across items, for example 11 from 15 items.

Although there is much overlap across different tools, there is no single universal tool that addresses all the varied contexts for assessment of risk of bias. Appendix A details a select list of tools that have been shown to be reliable or valid, are widely used, or have been recommended for use in systematic reviews that compared risk-of-bias assessment instruments.[21,22,52-54] We do not discuss tools that have been developed to guide and assess the reporting of studies. These reporting guidelines assess different constructs than what is commonly understood as risk of bias (internal validity). A list of reporting guidelines for different study designs is available through the EQUATOR network at www.equator-network.org.

Assessing the Risk of Bias for Harms

Although the assessment of harms is almost always included as an outcome in intervention studies, the manner of capturing and reporting harms is significantly different than the outcomes of benefit. Harms are defined as the "totality of possible adverse consequences of any intervention, therapy or medical test; they are the direct opposite of benefits, against which they must be compared."[55] For a detailed explanation of terms associated with harms please refer to the AHRQ Methods guide on harms.[56] Systematic reviews of intervention studies need to consider the balance between the harms and benefits of the treatment. Empirical evidence across diverse medical fields indicates that reporting of safety information—including milder harms—receives much less attention than the positive efficacy outcomes.[57,58] Thus, an evaluation of the benefits alone is likely to bias conclusions about the net efficacy or effectiveness of the intervention. Although reviewers recognize the importance of harms outcomes, harms are generally ignored in risk-of-bias assessment checklists. Several recent reviews[21,52-54] of risk-of-bias checklists and instruments do not identify harms as a key criterion within the checklists. We infer that many of the current risk-of-bias scales and checklists have assumed that harms are simply another study "outcome" and that taking this view suggests that the developers assume that no differences exist between harms and benefits in terms of risk-of-bias assessment.

For some aspects of risk-of-bias assessment, this approach may be reasonable. For example, consider an RCT evaluating the outcomes of a new drug therapy relative to those of a placebo control group; improper randomization would increase the risk of bias for measuring both outcomes of benefit and harm. However, unlike outcomes of benefit, harms and other unintended events are unpredictable and methods or instruments used to capture all possible adverse events can be problematic. This implies that that there is a potential for risk of bias for harms outcomes that is distinct from biases applicable to outcomes of benefit.

Because the type, timing, and severity of some harms are not anticipated—especially for rare events—many studies do not specify exact protocols to actively capture events. Standardized instruments used to systematically collect information on harms are often not included in the study methods. Study investigators may assume that patients will know when an adverse event has occurred, accurately recall the details of the event, and then "spontaneously" report this at the next outcome assessment. Thus, harms are often measured using passive methods that are poorly detailed, resulting in potential for selective outcome reporting, misclassification, and failure to capture significant events. Although some types of harms can be anticipated (e.g., pharmacokinetics of a drug intervention may identify body systems likely to be affected) that

include both common (e.g., headache) and rare conditions (e.g., stroke), harms may also occur in body systems that are not necessarily linked to the intervention from a biologic or epidemiologic perspective. In such instances, an important issue is establishing an association between the event and the intervention. The primary study may have established a separate committee to evaluate association between the harm and the putative treatment; as such blinding is not possible in such evaluations. Similarly, evaluating the potential for selective outcome reporting bias is complex when considering harms; some events may be unpredictable or they occur so infrequently relative to other milder effects that they are not typically reported. Given the possible or even probable unevenness in evaluating harms and benefits in most intervention studies, we recommend that EPCs assess the risk of bias of the study separately for benefits and for harms (see Appendix A for suggested tools and approaches).

Summarizing the Risk of Bias of a Study

For any outcomes undergoing assessment of strength of evidence, reviewers must consider all of the items together after completing evaluations of the assessment of risk of bias items for a given study. Then reviewers place risk of bias in a given study for each outcome into a summary category: low, medium or high.[9] Reviewers may conclude unclear risk of bias from poorly reported studies. This section describes methods for achieving that categorization and discusses guidelines for reporting this information. A study's risk of bias category can be different for different outcomes, which means that review teams should record the different outcome-specific categories as necessary. This situation can arise from, for instance, variation in the completeness of data, differential blinding of outcome assessors, or other outcome-specific items. Summarizing risk of bias for each patient-centered outcome within a study is recommended for synthesis of evidence across the studies and evaluating strength of evidence.[1] We do not recommend summarizing risk of bias across several outcomes for a given study because such global assessments across outcomes would involve subjective author judgments about relative importance of patient-centered outcomes and other factors for decision making.

Categories for Outcome-Specific Risk of Bias

An overall rating of low, medium, high, or unclear risk of bias should be made for the most clinically important outcomes as defined in the review protocol. As is true for scoring individual criteria or items, EPCs should do this overall rating within the study design. Observational studies and RCTs should be evaluated separately using recommended domains (Table 4). EPCs should adopt a dual reviewer approach to this step as well. Finally, given that these assessments involve subjective considerations, reviewers must clearly describe their rationale and explicit definitions for all ratings.

A study categorized as low risk of bias implies confidence on the part of the reviewer that results represent the true treatment effects (study results are considered valid). The study reporting is adequate to judge that no major or minor sources of bias are likely to influence results. A study rated as medium risk of bias implies some confidence that the results represent true treatment effect. The study is susceptible to some bias but the problems are not sufficient to invalidate the results (i.e., no flaw is likely to cause major bias).[59] A study categorized as high risk of bias implies low confidence that results represent true treatment effect. The study has significant flaws that imply biases of various types that may invalidate its results; these may arise from serious errors in conduct, analysis, or reporting, large amounts of missing information, or

discrepancies in reporting. A study categorized as "unclear" risk of bias is missing information, making it difficult to assess limitations and potential problems.

Methods and Considerations for Summarizing Risk of Bias

Some outcomes within a systematic review will receive ratings of the strength of evidence. One core component of the strength of a body of evidence for a given outcome is the overall risk of bias of the outcome data in all studies reporting that outcome.[9] This overall risk of bias is dictated by the risk of bias of the individual studies.

Incomplete reporting is an unavoidable challenge in summarizing the risk of bias of individual studies. To categorize the study, the reviewer must simultaneously consider (1) the known strengths, (2) the known weaknesses, and (3) the unknown attributes. A preponderance of unknown attributes may result in the study being categorized as unclear risk of bias; this might occur, for example, when EPC reviewers cannot determine whether the study was prospective or when investigators did not report the proportion of enrollees who provided data. In some cases, however, the unknown attributes are relatively minor; in these cases, EPC reviewers might still deem them of low risk of bias.

One way to assign a category is to make a simple "holistic" judgment; that is, a judgment based on an overall perception of risk of bias rather than an evaluation of all components of bias. Unfortunately, this approach is not transparent and is likely not to be reproducible. The main problem is inconsistent bases for judgment: if the studies were reexamined, the same reviewer might alter the category assignments. Reviewers may also be influenced, consciously or unconsciously, by other unstated aspects of the studies, such as the prestige of the journal or the identity of the authors. EPCs can and should explain how their reviewers made these judgments, but the fact remains that these approaches can suffer from substantial subjectivity. This transparency in terms of providing explicit support for each of the judgments or assessments made is a key feature of the Risk of Bias tool developed by The Cochrane Collaboration. Detailed and explicit support for each assessment not only ensures complete transparency, but allows the reader to (re)evaluate each assessment.

Instead, we recommend that, in aiming for transparency and reproducibility, EPC reviewers use a set of specific rules for assigning a category. These rules can take the form of declarative statements. For instance, in reviews of topics requiring randomization and blinding, one may make a declarative statement such as "adequately randomized and blinded studies are good; adequately randomized but unblinded studies are fair; inadequately randomized and unblinded studies are poor." EPCs could also lay out more complicated rules that reflect the items in the chosen instrument, but the key is transparency. Obviously, many other items could be incorporated into these rules, but, again, the key is transparency. Notice that such declarative statements implicitly assign weights to the different items. In any case, the authors must justify how synthesis of evidence incorporated risk-of-bias criteria or overall rank of risk of bias.

Within rule-based assignment, one option is to use the domains of risk of bias and then the items within those domains as a basis for the rules. For example, studies that met the majority of the items for all domains are good; studies that met the majority of the items for some (previously specified number) of the domains are fair; all other studies are poor. This process relies on an accurate assignment of items into domains. The basic requirement is adequate explanation of the method used.

The use of a quantitative scale is another way to employ a transparent set of rules. For a scale, the weights of different items are explicit rather than implicit. But any weighting system,

whether qualitative or quantitative, must be recognized as subjective and arbitrary, and different reviewers may choose to use different weighting methods. Using transparent rules does not remove the subjectivity inherent in assigning the risk of bias category. Subjectivity remains in the choice of different rules, or rules that assigning items to domains, and if the latter, what proportion of items must be met to earn a given rating. Consequently, reviewers should avoid attributing unwarranted precision (such as a score of 3.42) to ratings or creating subcategories or ambiguous language such as "in the middle of the fair range."

The approaches outlined above reveal two competing concerns: being transparent, and not being too formulaic. Transparency is important so that users can understand how categories were assigned, and also have some assurance that the same process was used for all of the studies. There is a danger, however, in being too formulaic and insensitive to the specific clinical context of the review. For example, if an outcome is unaffected by blinding, then the unconsidered use of a blinding "rule" (e.g., studies must be blinded to be categorized as low risk of bias) would be inappropriate for that outcome. Thus, we recommend careful consideration of the clinical context as reviewers strive for good transparency.

Previous research has demonstrated that empirical evidence of bias differed across individual domains rather than overall risk of bias.[60] Meta-epidemiological studies have demonstrated that treatment effects did not differ across overall categories of high versus low-risk of bias but did differ by criteria such as masking of treatment status or valid statistical methods.[60-62] Reviewers may use meta-analyses to the association between risk of bias domains and treatment effect with subgroup analyses or meta-regression.[61-63]

Conclusion

Assessment of risk of bias is a key step in conducting systematic reviews that informs many other steps and decisions made within the review. It also plays an important role in the final assessment of the strength of the evidence. The centrality of assessment of risk of bias to the entire systematic review task requires that assessment processes be based on sound empirical evidence when possible or on theoretical principles. In assessing the risk of bias of studies, EPCs should specify constructs and risks of bias specific to the content area, use at least two independent reviewers with a defined process for consensus and standards for transparency, and clearly document and justify all processes and decisions.

References

1. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. In: Higgins JPT, Green S, eds. The Cochrane Collaboration; 2011.
2. Cochrane Collaboration Glossary Version 4.2.5. 2005. Available at: http://www.cochrane.org/sites/default/files/uploads/glossary.pdf; http://effectivehealthcare.ahrq.gov/. Accessed January 2011.
3. Juni P, Altman DG, Egger M. Assessing the quality of controlled clinical trials. In: Egger M, Davey SG, Altman DG, eds. Systematic reviews in health care. Meta-analysis in context. 2001/07/07 ed. London: BMJ Books; 2001. p. 87-108.
4. Lohr KN. Rating the strength of scientific evidence: relevance for quality improvement programs. Int J Qual Health Care 2004;16(1):9-18. PMID: 15020556.
5. Balshem H, Helfand M, Schunemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011 Apr;64(4):401-6. PMID: 21208779.
6. U.S. Preventive Services Task Force Procedure Manual. AHRQ Publication No. 08-05118-EF. Available at: http://www.uspreventiveservicestaskforce.org/uspstf08/methods/procmanual.htm. Accessed July 2008.
7. Norris SL, Atkins D, Bruening W, et al. Observational studies in systemic reviews of comparative effectiveness: AHRQ and the Effective Health Care Program. J Clin Epidemiol 2011 Nov;64(11):1178-86. PMID: 21636246.
8. Atkins D, Best D, Briss PA, et al. Grading quality of evidence and strength of recommendations. BMJ 2004 Jun 19;328(7454):1490. PMID: 15205295.
9. Owens DK, Lohr KN, Atkins D, et al. AHRQ series paper 5: grading the strength of a body of evidence when comparing medical interventions—Agency for Healthcare Research and Quality and the effective health-care program. J Clin Epidemiol 2010;63(5):513-23.
10. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol 2011 Apr;64(4):395-400. PMID: 21194891.
11. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines 6. Rating the quality of evidence-imprecision. J Clin Epidemiol 2011 Dec;64(12):1283-93. PMID: 21839614.
12. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 8. Rating the quality of evidence-indirectness. J Clin Epidemiol 2011 Dec;64(12):1303-10. PMID: 21802903.
13. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 7. Rating the quality of evidence-inconsistency. J Clin Epidemiol 2011 Dec;64(12):1294-302. PMID: 21803546.
14. Guyatt GH, Oxman AD, Montori V, et al. GRADE guidelines: 5. Rating the quality of evidence-publication bias. J Clin Epidemiol 2011 Dec;64(12):1277-82. PMID: 21802904.
15. Guyatt GH, Oxman AD, Schunemann HJ, et al. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol 2011 Apr;64(4):380-2. PMID: 21185693.
16. Guyatt GH, Oxman AD, Sultan S, et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol 2011 Dec;64(12):1311-6. PMID: 21802902.
17. Guyatt GH, Oxman AD, Vist G, et al. GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias). J Clin Epidemiol 2011 Apr;64(4):407-15. PMID: 21247734.
18. Delgado-Rodriguez M, Llorca J. Bias. J Epidemiol Community Health 2004 Aug;58(8):635-41. PMID: 15252064.
19. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004 Sep;15(5):615-25. PMID: 15308962.

20. Validity in Epidemiologic Studies. In: Rothman KJ, Greenland S, Lash TL, eds. Modern Epidemiology. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008. p. 418-55, 9129-147.

21. West SL, King V, Carey TS, et al. Systems to rate the strength of scientific evidence. Evidence Report/Technology Assessment No. 47. AHRQ Publication No. 02-E016. Rockville, MD: Agency for Healthcare Research and Quality; 2002.

22. Deeks JJ, Dinnes J, D'Amico R, et al. Evaluating non-randomised intervention studies. Health Technol Assess 2003;7(27):iii-x, 1-173. PMID: 14499048.

23. Cook TD, Campbell DT. Quasi-experimentation: design and analysis issues for field settings. Boston: Houghton Mifflin Company; 1979.

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

25. Little J, Higgins JP, Ioannidis JP, et al. Strengthening the reporting of genetic association studies (STREGA): an extension of the strengthening the reporting of observational studies in epidemiology (STROBE) statement. J Clin Epidemiol 2009 Jun;62(6):597-608 e4. PMID: 19217256.

26. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 2001 Apr 14;357(9263):1191-4. PMID: 11323066.

27. Knottnerus A, Tugwell P. STROBE—a checklist to Strengthen the Reporting of Observational Studies in Epidemiology. J Clin Epidemiol 2008 Apr;61(4):323. PMID: 18313555.

28. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. J Clin Epidemiol 2003 Nov;56(11):1118-28. PMID: 14615003.

29. Davidoff F, Batalden P, Stevens D, et al. Publication guidelines for improvement studies in health care: evolution of the SQUIRE Project. Ann Intern Med 2008 Nov 4;149(9):670-6. PMID: 18981488.

30. Bekelman JE, Li Y, Gross CP. Scope and impact of financial conflicts of interest in biomedical research: a systematic review. JAMA 2003 Jan 22-29;289(4):454-65. PMID: 12533125.

31. Newcastle-Ottawa Quality Assessment Scale: Cohort studies. Available at: http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm. Accessed January 2011.

32. Smith R. Medical journals are an extension of the marketing arm of pharmaceutical companies. PLoS Med 2005 May;2(5):e138. PMID: 15916457.

33. Julian DG. What is right and what is wrong about evidence-based medicine? J Cardiovasc Electrophysiol 2003 Sep;14(9 Suppl):S2-5. PMID: 12950509.

34. Jorgensen AW, Maric KL, Tendal B, et al. Industry-supported meta-analyses compared with meta-analyses with non-profit or no support: differences in methodological quality and conclusions. BMC Med Res Methodol 2008;8:60. PMID: 18782430.

35. Lee K, Bacchetti P, Sim I. Publication of clinical trials supporting successful new drug applications: a literature analysis. PLoS Med 2008 Sep 23;5(9):e191. PMID: 18816163.

36. American Medical Writers Association. AMWA ethics FAQs, publication practices of particular concern to medical communicators. 2009. Available at: http://www.amwa.org/default.asp?Mode=DirectoryDisplay&DirectoryUseAbsoluteOnSearch=True&id=466. Accessed June 2, 2011.

37. Ross JS, Hill KP, Egilman DS, et al. Guest authorship and ghostwriting in publications related to rofecoxib: a case study of industry documents from rofecoxib litigation. JAMA 2008 Apr 16;299(15):1800-12. PMID: 18413874.

38. DeAngelis CD, Fontanarosa PB. Impugning the integrity of medical science: the adverse effects of industry influence. JAMA 2008 Apr 16;299(15):1833-5. PMID: 18413880.

39. Hirsch LJ. Conflicts of interest, authorship, and disclosures in industry-related scientific publications: the tort bar and editorial oversight of medical journals. Mayo Clin Proc 2009 Sep;84(9):811-21. PMID: 19720779.

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

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

42. Vedula SS, Bero L, Scherer RW, et al. Outcome reporting in industry-sponsored trials of gabapentin for off-label use. N Engl J Med 2009 Nov 12;361(20):1963-71. PMID: 19907043.

43. Atkins D, Chang S, Gartlehner G, et al. Assessing the Applicability of Studies When Comparing Medical Interventions. Agency for Healthcare Research and Quality. Methods Guide for Comparative Effectiveness Reviews. AHRQ Publication No. 11-EHC019-EF. Available at: http://effectivehealthcare.ahrq.gov/. Accessed January 2011.

44. Institute of Medicine. Finding what works in health care: standards for systematic reviews. Available at: http://www.nap.edu/openbook.php?record_id=13059&page=R1. Accessed June 2, 2011.

45. Shea BJ, Hamel C, Wells GA, et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. J Clin Epidemiol 2009 Oct;62(10):1013-20. PMID: 19230606.

46. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. J Clin Epidemiol 2009 Oct;62(10):1006-12. PMID: 19631508.

47. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 2009;339:b2700. PMID: 19622552.

48. Hartling L, Bond K, Harvey K, et al. Developing and testing a tool for the classification of study designs in systematic reviews of interventions and exposures. Prepared by the University of Alberta Evidence-based Practice Center under Contract No. 290-02-0023. Rockville, MD: Agency for Healthcare Research and Quality: June 2009. AHRQ Publication No. 11-EHC007-EF.

49. Viswanathan M, Berkman ND. Development of the RTI item bank on risk of bias and precision of observational studies. J Clin Epidemiol 2011 Sep 28;PMID: 21959223.

50. Egger M, Smith DH. Under the meta-scope: potentiatls and limitations of meta-analysis. Evidence based resource in anaesthesia and analgesia. BMJ Publication 2000.

51. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. Obstet Gynecol 2010 May;115(5):1063-70. PMID: 20410783.

52. Olivo SA, Macedo LG, Gadotti IC, et al. Scales to assess the quality of randomized controlled trials: a systematic review. Phys Ther 2008 Feb;88(2):156-75. PMID: 18073267.

53. Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol 2007;36(3):677-8. PMID: 17470488.

54. Whiting P, Rutjes AW, Dinnes J, et al. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. J Clin Epidemiol 2005 Jan;58(1):1-12. PMID: 15649665.

55. Ioannidis JP, Evans SJ, Gotzsche PC, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. Ann Intern Med 2004 Nov 16;141(10):781-8. PMID: 15545678.

56. Chou R, Aronson N, Atkins D, et al. AHRQ series paper 4: assessing harms when comparing medical interventions: AHRQ and the effective health-care program. J Clin Epidemiol 2010 May;63(5):502-12. PMID: 18823754.

57. Ioannidis JP, Lau J. Improving safety reporting from randomised trials. Drug Saf 2002;25(2):77-84. PMID: 11888350.

58. Ioannidis JP, Lau J. Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas. JAMA 2001 Jan 24-31;285(4):437-43. PMID: 11242428.

59. 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. PMID: 11306229.

60. Balk EM, Bonis PA, Moskowitz H, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA 2002 Jun 12;287(22):2973-82. PMID: 12052127.

61. Higgins JP, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 6;327(7414):557-60. PMID: 12958120.

62. Herbison P, Hay-Smith J, Gillespie WJ. Adjustment of meta-analyses on the basis of quality scores should be abandoned. J Clin Epidemiol 2006 Dec;59(12):1249-56. PMID: 17098567.

63. Fu R, Gartlehner G, Grant M, et al. Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care Program. J Clin Epidemiol 2011 Nov;64(11):1187-97. PMID: 21477993.

Appendix A. Tools To Assess Risk of Bias of Individual Outcomes

This appendix provides a brief overview of tools to evaluate randomized controlled trials (RCTs), nonrandomized studies, and harms. This information does not represent a comprehensive systematic synthesis of tools available but provides details for a select list of tools that have been shown to be reliable or valid, are widely used, or have been recommended for use in systematic reviews that compared risk of bias assessment instruments.[1-5] For most tools, the preliminary step in assessing whether a chosen tool is applicable to the specific study is to categorize the study design. We recommend the use of tools such as that developed by Hartling et al. to categorize study designs.[6]

Randomized Controlled Trials

A large number of tools have been developed to assess risk of bias in RCTs. In 2008, Armijo Olivo et al.[1] published a systematic review identifying scales designed to assess the risk of bias of RCTs. They identified 21 scales but found that the majority were not "rigorously developed or tested for validity and reliability."

Armijo Olivo et al. found that the Jadad scale demonstrated the strongest evidence in terms of validity and reliability. The Jadad scale demonstrates face, content, criterion, and construct validity. One limitation regarding the assessment of criterion or concurrent validity for all risk of bias tools is that it depends on a gold standard that does not exist for these tools. Hence, reports of construct validity need to be interpreted in light of the tool used as the reference standard for comparisons. Armijo Olivo et al. found that the Jadad scale was most commonly cited in the medical literature. The Jadad scale was the most commonly used tool in systematic reviews produced by The Cochrane Collaboration until recently, and it is still the most commonly used tool to assess risk of bias of RCTs in AHRQ evidence reports. The Jadad scale addresses three domains (i.e., randomization, blinding, and handling of withdrawals and drop-outs), but does not address adequacy of allocation concealment. The tool includes five questions which take approximately 10 minutes to apply to an individual trial. Although the Jadad scale was developed in the context of pain research it has been tested and used widely in other fields. Although the Jadad scale is the most commonly used tool to assess risk of bias of RCTs, concerns regarding its appropriateness have recently emerged.[7-9] Specifically, there is some evidence that the tool reflects quality of reporting rather than risk of bias.[10]

Armijo Olivo et al. highlighted two other tools that were developed using rigorous methods and tested for validity and reliability. Verhagen et al. developed the Delphi List to assess RCTs in general (i.e., not specific to a clinical area or field of study). It has demonstrated good face, content, and concurrent validity and has been tested for reliability. It includes the following items: inclusion/exclusion criteria of study population defined; randomization; allocation concealment; baseline comparability of study groups; blinding of investigator, subjects, and care providers; reporting of point estimates and variability for primary outcomes; and intention-to-treat analysis.[11]

Yates et al. developed a tool to assess the risk of bias of RCTs of cognitive behavioral therapy for chronic pain. The tool has two parts, one related to the treatment (five items) and the second related to study design and methods (eight items with multiple parts). The latter part of the tool includes questions on the following domains: reporting of inclusion/exclusion criteria;

reporting of attrition; adequate description of the sample; steps to minimize bias (i.e., randomization, allocation, measurement, treatment expectations); outcomes justified, valid, and reliable; length of follow-up (i.e., sustainability of treatment effects); adequacy of statistical analyses; comparability or adequacy of control group. It has shown face, content, and construct validity and good inter-rater reliability.[12] The tool has not been widely used.

In 2005, The Cochrane Collaboration convened a group to address several concerns in the assessment of trial risk of bias. One concern was the growing number of tools being used and inconsistent approaches to risk of bias assessment across different systematic reviews. Participants also recognized that many of the tools being used were not based on empirical evidence showing that the items they included were related to biased results. Moreover, many tools combined elements examining methodological conduct with items related to reporting.

From this work a new tool for randomized trials emerged—the Risk of Bias tool.[6] This tool was released after publication of the review by Armijo Olivo et al. described above. The Risk of Bias tool includes seven domains for which empirical evidence demonstrates associations with biased estimates of effect. The domains are sequence generation; allocation concealment; blinding of participants and personnel; blinding of outcome assessment; missing outcome data; selective outcome reporting; and other sources of bias. The final domain, "other sources of bias," includes design specific risks of bias, baseline imbalance, blocked randomization in unblinded trials, differential diagnostic activity, and other potential biases.[13] The Cochrane Handbook[13] provides guidance on assessing the different domains including "other sources of bias." The Handbook emphasizes that topics within the other domain should focus on issues related to bias and not imprecision, heterogeneity, or other quality measures that are unrelated to bias. Further, these items will vary across different reviews and should be identified and prespecified when developing the review protocol.

Although the Risk of Bias tool is now the recommended method for assessing risk of bias of RCTs in systematic reviews conducted through The Cochrane Collaboration, the tool has not undergone extensive validity or reliability testing. However, one of the unique and critical features of the Risk of Bias tool is its transparency. That is, users are instructed to document explicit support for each assessment alongside the assessment. The developers of the tool argue that this transparency is more important than demonstrations of "reliability" and "validity," because complete transparency is ensured and each assessment can readily be (re)evaluated by the reader.

Nonrandomized Studies

Several systematic reviews have been conducted to identify, assess, and make recommendations regarding risk of bias assessment tools for use in nonrandomized studies (including nonrandomized experimental studies and observational studies). West et al.[5] identified 12 tools for use in observational studies and recommended 6 of these for use in systematic reviews. Deeks et al.[4] identified 14 "best tools" from among 182 and recommended 6 for use in reviews. Of interest is that the two reports identified only three tools in common: Downs and Black,[14] Reisch,[15] and Zaza.[16] These three tools are applicable to a range of study designs; only two were developed for use in systematic reviews.[14,16]

One recent and comprehensive systematic review of risk of bias assessment tools for observational studies identified 86 tools.[2] The tools varied in their development and their purpose: only 15 percent were developed specifically for use in systematic reviews; 36 percent were developed for general critical appraisal and 34 percent were developed for "single use in a

specific context." The authors chose not to make recommendations regarding which specific tools to use; however, they broadly advised that reviewers select tools that

- contain a small number of components or domains;
- are as specific as possible with regard to study design and the topic under study;
- are developed using rigorous methods, evidence-based, and valid and reliable; and
- are simple checklists rather than scales when possible.

The Cochrane Collaboration provides recommendations on use of tools for nonrandomized studies. They acknowledge the abundance of tools available but, like Sanderson et al., make no recommendation regarding a single instrument.[2] They recommend following the domains in the Risk of Bias tool, particularly for prospective studies. A working group within the Cochrane Collaboration is currently modifying the Risk of Bias tool for use in nonrandomized studies.

The Cochrane Handbook highlights two other tools for use in nonrandomized studies: the Downs and Black[14] and Newcastle Ottawa Scale.[17] They implicitly recommend the Newcastle Ottawa Scale over the Downs and Black because the Downs and Black is time-consuming to apply, requires considerable epidemiology expertise, and has been found difficult to apply to case-control studies.[17]

The Newcastle Ottawa Scale is frequently used in systematic reviews for articles about studies with this type of design. It contains separate questions for cohort and case-control studies. It was developed based on threats to validity in nonrandomized studies; these specifically include selection of participants (generalizability or applicability), comparability of study groups, methods for outcome assessment (cohort studies) or ascertainment of exposure (case-control studies), and adequacy of follow-up. The developers have reported face and content validity for this instrument, and they revised it based on experience using the tool in systematic reviews.[17] It has also been tested for inter-rater reliability.[18,19] Examination of its criterion validity and intra-rater reliability is underway and plans are being developed to examine its construct validity.

Other recently developed checklists address the quality of observational, nontherapeutic studies of incidence of diseases or risk factors for chronic diseases[20] or observational studies of interventions or exposures.[21] The checklists have been developed based on a comprehensive literature review,[22] are based on predefined flaws in internal validity, and discriminate reporting from conduct of the studies. These tools are continuing inter-rater reliability tests.

Instruments and Tools To Evaluate Quality of Harms Assessment

No systematic reviews evaluating tools to assess the potential for biases associated with harms were found. However, three tools/checklists were identified and two of these recognize that some biases may arise when capturing and reporting harms that are distinct from the outcomes of benefit and therefore require separate assessment.

One checklist developed by the Cochrane Collaboration offers some guidance, and leaves the final choice up to the reviewer to select items from a list that is stratified by the study design.[13] It assumes that these questions (see Table A-1) can be added to those criteria already detailed in the Cochrane Risk of Bias tool.

Table A-1. Recommendations for elements of assessing quality of the evidence when collecting and reporting harms, by study design

Study design	Quality considerations
RCTs	On study conduct: • Are definitions of reported adverse effects given? • Were the methods used for monitoring adverse effects reported, such as use of prospective or routine monitoring; spontaneous reporting; patient checklist, questionnaire or diary; systematic survey of patients? What was the source to assess harms (self-report vs. medical exam vs. PI opinion)? Who decided seriousness, severity, and causal relation with the treatments? On reporting: • Were any patients excluded from the adverse effects analysis? • Does the report provide numerical data by intervention group? • Which categories of adverse effects were reported by the investigators?
Case series	• Do the reports have good predictive value? • How was causality determined? • Is there a plausible biological mechanism linking the intervention to the adverse event? • Do the reports provide enough information to allow detailed appraisal of the evidence?
Case control	• Consider typical biases for this nonrandomized study design.

From Loke et al., 2011[23]

Chou and Helfand developed a tool for an AHRQ systematic review to assess the risk of bias of studies evaluating carotid endarterectomy; the primary outcome in these studies included adverse events.[24] Four of eight items within this tool were directed specifically to assessing bias associated with adverse events; however, these criteria are applicable to other interventions, although no formal validation has been undertaken.[24] The Chou and Helfand tool has been used in comparative studies (RCTs and observational studies). No formal reliability testing has been undertaken and the tool is interpreted as a summed score across eight items. One advantage of this tool is that it includes elements of study design (for example, randomization, withdrawal) and some items specific to harms. Table A-2 shows the items within this scale.

The McMaster University Harms scale (McHarm) was developed specifically for evaluating harms and is applicable to studies evaluating interventions (both randomized and nonrandomized studies). The criteria within McHarm are detailed in Table A-3. The McHarm tool is used in conjunction with other risk of bias assessment tools that evaluate basic design features (e.g., randomization). The McHarm assumes that some biases to study conduct are unique to harms collection and that these should be evaluated separately from outcomes of benefit; scoring is considered on a per item basis. Reliability was evaluated (in expert and nonexpert raters) in RCTs of drug and surgical interventions. Internal consistency and inter-rater reliability were evaluated and found to be acceptable (greater than 0.75) with the exception of drug studies for nonexperts; in this instance the inter-rater reliability was moderate. An intra-class correlation coefficient greater than 0.75 was set as the acceptable threshold level for reliability. With the exception of nonexpert raters for drug studies, all other groups of raters showed high levels of reliability (Table A-4).

Table A-2. Quality assessment tool for studies reported adverse events[24]

Criterion	Explanation	Score
Quality criterion 1: Nonbiased selection	1: study is a properly randomized controlled trial, or an observational study with a clear predefined inception cohort (that attempted to evaluate all patients in the inception cohort) 0: study does not meet above criteria (e.g., convenience samples)	
Quality criterion 2: Adequate description of population	1: study reports two or more demographic characteristics, presenting symptoms/syndrome and at least one important risk factor for complications 0: study does not meet above criteria	
Quality criterion 3: Low loss to follow-up	1: study reports number lost to follow-up, and the overall number lost to follow-up is low (threshold set at 5% for studies of carotid endarterectomy and 10% for studies of rofecoxib) 0: study does not meet above criteria	
Quality criterion 4: Adverse events prespecified and defined	1: study reports explicit definitions for major complications that allow for reproducible ascertainment (what adverse events were being investigated and what constituted an "event") 0: study does not meet above criteria	
Quality criterion 5: Ascertainment technique adequately described	1: study reports methods used to ascertain complications, including who ascertained, timing, and methods used 0: study does not meet above criteria	
Quality criterion 6: Nonbiased ascertainment of adverse events	1: independent or masked assessment or complications (for studies of carotid endarterectomy, someone other than the surgeon who performed the procedure; for studies of rofecoxib, presence of an external endpoint committee blinded to treatment allocation) 0: study does not meet above criteria	
Quality criterion 7: Adequate statistical analysis of potential confounders	1: study examines one or more relevant confounders/risk factors (in addition to the comparison group in controlled studies), using acceptable statistical techniques such as stratification or adjustment 0: study does not meet above criteria	
Quality criterion 8: Adequate duration of follow-up	1: study reports duration of follow-up and duration of follow-up adequate to identify expected adverse events (threshold set at 30 days for studies of carotid endarterectomy and 6 months for studies of rofecoxib) 0: study does not meet above criteria	
Total quality score = sum of scores (0-8)	>6: Good 4-6: Fair <4: Poor	

Reprinted from Chou R, Fu R, Carson S, et al. Methodological shortcomings predicted lower harm estimates in one of two sets of studies of clinical interventions. J Clin Epidemiol 2007 Jan;60(1):18-28, with permission from Elsevier.

Table A-3. McMaster tool for assessing quality of harms assessment and reporting in study reports (McHarm)

	Question
1.	Were the harms PREDEFINED using standardized or precise definitions?
2.	Were SERIOUS events precisely defined?
3.	Were SEVERE events precisely defined?
4.	Were the number of DEATHS in each study group specified OR were the reason(s) for not specifying them given?
5.	Was the mode of harms collection specified as ACTIVE?
6.	Was the mode of harms collection specified as PASSIVE?
7.	Did the study specify WHO collected the harms?
8.	Did the study specify the TRAINING or BACKGROUND of who ascertained the harms?
9.	Did the study specify the TIMING and FREQUENCY of collection of the harms?
10.	Did the author(s) use STANDARD scale(s) or checklist(s) for harms collection?
11.	Did the authors specify if the harms reported encompass ALL the events collected or a selected SAMPLE?
12.	Was the NUMBER of participants that withdrew or were lost to follow-up specified for each study group?
13.	Was the TOTAL NUMBER of participants affected by harms specified for each study arm?
14.	Did the author(s) specify the NUMBER for each TYPE of harmful event for each study group?
15.	Did the author(s) specify the type of analyses undertaken for harms data?

From: hiru.mcmaster.ca/epc/mcharm.pdf

Note: The answers to each question are yes (implying less risk of bias), no (implying high risk of bias), and unsure.

Table A-4. McMaster tool for assessing quality of harms assessment and reporting in study reports (McHarm): inter rater reliability (intra-class correlation coefficients and confidence intervals) within different groups of raters

	Drug studies	Surgery studies	All studies
Nonexpert Raters	0.69 (0.27, 0.91)	0.92 (0.80, 0.98)	0.88 (0.77, 0.94)
Experts Raters	0.89 (0.73, 0.97)	0.93(0.85,0.98)	0.92 (0.86, 0.97)
All Raters	0.89 (0.75, 0.97)	0.96 (0.92, 0.99)	0.95 (0.91, 0.98)

From: hiru.mcmaster.ca/epc/mcharm.pdf

References

1. Olivo SA, Macedo LG, Gadotti IC, et al. Scales to assess the quality of randomized controlled trials: a systematic review. Phys Ther 2008 Feb;88(2):156-75. PMID: 18073267.
2. Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol 2007;36(3):677-8. PMID: 17470488.
3. Whiting P, Rutjes AW, Dinnes J, et al. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. J Clin Epidemiol 2005 Jan;58(1):1-12. PMID: 15649665.
4. Deeks JJ, Dinnes J, D'Amico R, et al. Evaluating non-randomised intervention studies. Health Technol Assess 2003;7(27):iii-x, 1-173. PMID: 14499048.
5. West SL, King V, Carey TS, et al. Systems to rate the strength of scientific evidence. Evidence Report/Technology Assessment No. 47. AHRQ Publication No. 02-E016. Rockville, MD: Agency for Healthcare Research and Quality; 2002.
6. Hartling L, Bond K, Harvey K, et al. Developing and testing a tool for the classification of study designs in systematic reviews of interventions and exposures. Prepared by the University of Alberta Evidence-based Practice Center under Contract No. 290-02-0023. AHRQ Publication No. 11-EHC007-EF. Rockville, MD: Agency for Healthcare Research and Quality: June 2009.
7. Berger VW. The (lack of) quality in assessing the quality of transplantation trials. Transpl Int 2009 Oct;22(10):1029; author reply 3. PMID: 19497066.

8. Berger VW. Is the Jadad score the proper evaluation of trials? J Rheumatol 2006 Aug;33(8):1710-1; author reply 1-2. PMID: 16881132.

9. Jadad AR. The merits of measuring the quality of clinical trials: is it becoming a Byzantine discussion? Transpl Int 2009 Oct;22(10):1028. PMID: 19740247.

10. Hartling L, Ospina M, Liang Y, et al. Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. BMJ 2009;339:b4012. PMID: 19841007.

11. Verhagen AP, de Vet HC, de Bie RA, et al. The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. J Clin Epidemiol 1998 Dec;51(12):1235-41. PMID: 10086815.

12. Yates SL, Morley S, Eccleston C, et al. A scale for rating the quality of psychological trials for pain. Pain 2005 Oct;117(3):314-25. PMID: 16154704.

13. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. In: Higgins JPT, Green S, eds.: The Cochrane Collaboration; 2011.

14. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Commun Health 1998;52:377–84.

15. Reisch JS, Tyson JE, Mize SG. Aid to the evaluation of therapeutic studies. Pediatrics 1989 Nov;84(5):815-27. PMID: 2797977.

16. Zaza S, Carande-Kulis VG, Sleet DA, et al. Methods for conducting systematic reviews of the evidence of effectiveness and economic efficiency of interventions to reduce injuries to motor vehicle occupants. Am J Prev Med 2001;21(4 Suppl):23-30.

17. Newcastle-Ottawa Quality Assessment Scale: Case control studies. Available at: http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm. Accessed January 2011.

18. An evaluation of the Newcastle Ottawa Scale: an assessment tool for evaluating the quality of non-randomized studies. XI Cochrane Colloquium: Evidence, Health Care and Culture; 2003 Oct 26-31; Barcelona, Spain.

19. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. 3rd Symposium on Systematic Reviews: Beyond the Basics; 2000 Jul 3-5; Oxford, UK.

20. Shamliyan TA, Kane RL, Ansari MT, et al. Development quality criteria to evaluate nontherapeutic studies of incidence, prevalence, or risk factors of chronic diseases: pilot study of new checklists. J Clin Epidemiol 2011 Jun;64(6):637-57. Epub 2010 Nov 11. PMID: 21071174.

21. Viswanathan M, Berkman ND. Development of the RTI item bank on risk of bias and precision of observational studies. J Clin Epidemiol 2012 Feb;65(2):163-78. Epub 2011 Sep 29. PMID: 21959223.

22. Shamliyan T, Kane RL, Dickinson S. A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases. J Clin Epidemiol 2010 Oct;63(10):1061-70. PMID: 20728045.

23. Loke YK, Price D, Herxheimer A. Adverse effects. In: Higgins JPT, Green S, eds. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. updated March 2011: The Cochrane Collaboration; 2011.

24. Chou R, Fu R, Carson S, et al. Methodological shortcomings predicted lower harm estimates in one of two sets of studies of clinical interventions. J Clin Epidemiol 2007 Jan;60(1):18-28. PMID: 17161750.