Methods Guide
for Comparative Effectiveness Reviews

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

This report is based on research conducted by the Agency for Healthcare Research and Quality (AHRQ) Evidence-based Practice Centers' Methods Workgroup. The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

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

This research was funded through contracts from the Agency for Healthcare Research and Quality to the following Evidence-based Practice Centers: RTI (290-2015-00011-I), University of Alberta (290-2015-00001-I), ECRI-Penn (290-2015-00005-I), The Johns Hopkins University (290-2015-00006-I), Brown University (290-2015-00002-I), Mayo Clinic (290-2015-00013-I), Minnesota University (290 -2015-00008-I), and Kaiser Permanente Center for Health Research (290-2015-00007-I).

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

AHRQ or U.S. Department of Health and Human Services endorsement of any derivative products that may be developed from this report, such as clinical practice guidelines, other quality enhancement tools, or reimbursement or coverage policies may not be stated or implied. Persons using assistive technology may not be able to fully access information in this report. For assistance, contact epc@ahrq.hhs.gov.

Suggested citation: Viswanathan M, Patnode C, Berkman ND, Bass EB, Chang S, Hartling L, Murad HM, Treadwell JR, Kane RL. Assessing the Risk of Bias in Systematic Reviews of Health Care Interventions. Methods Guide for Comparative Effectiveness Reviews. (Prepared by the Scientific Resource Center under Contract No. 290-2012-0004-C). AHRQ Publication No. 17(18)-EHC036-EF. Rockville, MD: Agency for Healthcare Research and Quality; December 2017. Posted final reports are located on the Effective Health Care Program search page.
DOI: https://doi.org/10.23970/AHRQEPCMETHGUIDE2.

Prepared by:
Scientific Resource Center
Portland, OR

Investigators:
Meera Viswanathan, Ph.D.
Carrie D. Patnode, Ph.D., M.P.H.
Nancy D. Berkman, Ph.D.
Eric B. Bass, M.D., M.P.H.
Stephanie Chang, M.D., M.P.H.
Lisa Hartling, Ph.D.
M. Hassan Murad, M.D., M.P.H.
Jonathan R. Treadwell, Ph.D.
Robert L. Kane, M.D.

Preface

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

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

The Methods Guide for Comparative Effectiveness Reviews is a living document, and will be updated as further empiric evidence develops and our understanding of better methods improves.

If you have comments on this Methods Guide paper, they may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 5600 Fishers Lane, Rockville, MD 20857, or by email to epc@ahrq.hhs.gov.

Gopal Khanna, M.B.A.
Director
Agency for Healthcare Research and Quality

Arlene S. Bierman, M.D., M.S.
Director
Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Director
Evidence-based Practice Center Program
Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality

Acknowledgments

The authors gratefully acknowledge the following individuals for their contributions to this project: Issa J. Dahabreh, M.D., M.S., Celia Fiordalisi, M.S., Makalapua Motu'apuaka, B.S., Robin Paynter, M.L.I.S., Edwin Reid, M.S., and Lyndzie Sardenga, B.S.

Peer Reviewers

Prior to publication of the final evidence report, EPCs sought input from independent Peer Reviewers without financial conflicts of interest. However, the conclusions and synthesis of the scientific literature presented in this report does not necessarily represent the views of individual reviewers.

Peer Reviewers must disclose any financial conflicts of interest greater than $10,000 and any other relevant business or professional conflicts of interest. Because of their unique clinical or content expertise, individuals with potential non-financial conflicts may be retained. The Task Order Officer and the EPC work to balance, manage, or mitigate any potential non-financial conflicts of interest identified.

The list of Peer Reviewers follows:

Roger Chou, M.D., FACP
Director, Pacific Northwest EPC
Portland, OR

Susanne Hempel, Ph.D.
Co-Director, Southern California EPC
RAND Corporation
Santa Monica, CA

Jennifer Lin, M.D., M.C.R.
Director, Kaiser Permanente EPC
Portland, OR

Terri Pigott, Ph.D.
Associate Provost for Research
Loyola University Chicago
Chicago, IL

Gillian Sanders Schmidler, Ph.D.
Director, Duke University EPC
Durham, NC

P. Lina Santaguida, Ph.D., M.Sc.
Assistant Professor, McMaster University
South Hamilton, ON

Karen Schoelles, M.D., S.M., FACP
Director, ECRI Institute-Penn Medicine EPC
Plymouth Meeting, PA

Jeffrey C. Valentine, Ph.D.
Professor
College of Education and Human Development
University of Louisville
Louisville, KY

C. Michael White, Pharm.D., FCP, FCCP
Director, University of Connecticut EPC
Storrs, CT

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

Structured Abstract

Objective. Risk-of-bias assessment is a central component of systematic reviews but little conclusive empirical evidence exists on the validity of such assessments. In the context of such uncertainty, we present pragmatic recommendations that can be applied consistently across review topics, promote transparency and reproducibility in processes, and address methodological advances in the risk-of-bias assessment.

Study Design. Epidemiological study design principles; available empirical evidence, risk-of-bias tools, and guidance; and workgroup consensus

Results. We developed recommendations for assessing the risk of bias of studies of health care interventions specific to framing the focus and scope of risk-of-bias assessment; selecting risk of bias categories; choosing assessment instruments; and conducting, analyzing, and presenting results of risk-of-bias assessments. Key recommendations include transparency and reproducibility of judgments, separating risk of bias from other constructs such as applicability and precision, and evaluating risk of bias per outcome. We recommend against certain past practices, such as focusing on reporting quality, relying solely on study design, or numerical quality scores, and automatically downgrading for industry sponsorship.

Conclusion. Risk-of-bias assessment remains a challenging but essential step in systematic reviews. We presented standards to promote transparency of judgments.

Contents

Tables

Key Recommendations

- Recommendations regarding focus and scope of risk-of-bias assessment
  - Clearly separate assessing the risk of bias from other important and related activities such as assessing the degree of congruence between the research questions of a systematic review and designs of included studies, the precision of an effect estimate, and the applicability of the evidence.
  - The methodology for assessing risk of bias should be transparent and reproducible. This requires the review's protocol to include clear definitions of the types of biases that will be assessed and *a priori* decision rules for assigning the risk of bias for each individual study. New or changed processes developed over the course of the review should be documented clearly.
  - Assess risk of bias based on study design-specific criteria and conduct rather than quality of reporting of methods and results. Poorly reported studies may be judged as unclear risk of bias.
  - Allow for separate risk-of-bias ratings for each outcome to account for outcome-specific variations in potential types or extent of bias. For some studies, all outcomes may have the same sources of bias; for other studies, the sources of bias may vary by outcome.
  - Use risk of bias assessments to explore heterogeneity of results, to interpret the estimate of effect through sensitivity analysis (quantitatively if studies can be pooled, qualitatively otherwise), and to grade the strength of evidence.
  - Do not rely solely on study design label (e.g., randomized controlled trial [RCT] or cohort, case-control) as a proxy for assessment of risk of bias of individual studies.
  - Reviewers who incorporate existing systematic reviews in new reviews or subgroup analyses from individual studies should evaluate the credibility of these sources of information.
- Recommendations for selecting risk of bias categories
  - Select risk of bias categories as appropriate for the topic and study design because not all categories of bias matter equally for all topics and designs.
  - When selecting risk of bias categories, consider bias arising in the randomization process or due to confounding; departures from intended interventions; missing data; measurement of outcomes; and selective outcome reporting in all studies. Additionally, biased participant selection and misclassification of interventions may influence results in nonrandomized or poorly randomized studies.
  - Do not use poor or incomplete reporting, industry funding, or disclosed conflict of interest to rate an outcome or study as high risk of bias; do, however, report these issues transparently and consider their impact on bias.
- Recommendations for choosing instruments for assessing risk of bias
  - Choose risk-of-bias instruments that are based on epidemiological study design principles, established measurement properties (e.g., reliability, internal consistency) or empirical evidence (when available).
  - Choose instruments that include items assessing specific concerns related to each of the risk of bias categories that pose threats to the accuracy of the effect estimate.

- Recommendations for conducting, analyzing, and presenting results of risk of bias assessments
  - Use processes to reduce uncertainty in individual judgments such as dual independent assessment of risk of bias with an unbiased reconciliation method. First-order assessments of risk of bias by machine-learning methods require secondary human review.
  - Balance the competing considerations of simplicity of presentation and burden on the reader when presenting results of risk of bias assessments. An overall study or outcome-specific risk of bias rating alone, without supporting details, offers simplicity but lacks transparency. Provide enough detail to make the rationale for the assessment clear
  - Consider both the direction and magnitude of possible bias on the effect estimate when possible, rather than leaving the burden to the reader.
  - Avoid the presentation of risk of bias assessment solely as a numerical score; at minimum, consider sensitivity analyses of these scores.
  - When summarizing the evidence, consider conducting sensitivity analyses to evaluate whether including studies with high or unclear risk of bias (overall or in specific categories) influences the estimate of effect or heterogeneity.
  - Systematic reviewers who choose to exclude high risk-of-bias studies from their analysis should explain and justify the criteria used to identify excluded studies.

Introduction

Assessing the risk of bias of studies included in the body of evidence is a foundational part of all systematic reviews.[1, 2] It is distinct from other important and related activities of assessing the degree of the congruence of the research question with the study design and the applicability of the evidence. The specific use of risk-of-bias assessments can vary. Assessment of risk of bias (labeled as unclear, high, moderate, or low) are intended to help interpret findings and explain heterogeneity; in addition, EPC reviews use risk-of-bias assessments of individual studies in grading the strength of the body of evidence. Some EPC reviews may exclude studies assessed as high risk of bias.

Despite the importance of risk-of-bias assessments in systematic reviews, evidence on the validity of such assessments is available only for a few risk-of-bias categories.[3-5] Specifically, evidence suggests that effect sizes may be inaccurate when allocation is inappropriately concealed; random sequences are inadequately generated; and patients, clinicians, or outcome assessors (particularly for subjective outcomes) are not blinded.[4, 6] The influence on estimates of effect can be inconsistent and difficult to predict for other bias categories such as confounding, fidelity to the protocol, and attrition bias, possibly because meta-epidemiological studies are inadequately powered.[5] In addition to concerns regarding the validity of such assessments, methodological studies have raised concerns about the limited reliability of risk-of-bias judgments.[7, 8]

We do not attempt, in this document, to address the underlying and important sources of uncertainty related to the validity or reliability of risk-of-bias assessment. 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. This update adds areas of guidance (e.g., evaluating subgroup analyses and including systematic reviews as evidence), modifies guidance to reflect new thinking (e.g., risk-of-bias categories), and offers guidance to promote clarity and consistency. As with other AHRQ methodological guidance, our intent is to present standards that can be applied consistently across EPCs and review topics, promote transparency and reproducibility in processes, and account for methodological changes in the systematic review process. These standards are based on epidemiological study design principles, available empirical evidence, or workgroup consensus. As greater evidence accumulates in this methodological area, our standards will continue to evolve. When possible, our guidance offers flexibility to account for the wide range of AHRQ EPC review topics and included study designs, but also offers parameters within which this flexibility can be applied.

In this guidance document, we define terms as appropriate for the EPC program, explore the potential overlap in different steps of the systematic review, and offer recommendations on the inclusion and exclusion of constructs that may apply to multiple steps of the systematic review process. This guidance applies to systematic reviews exploring the link between an intervention or exposure and outcome. (Reviewers focusing on diagnostic tests,[9] prognosis,[10-12] prevalence, or qualitative[13] analysis should also consult guidance specific to these topics.) Later sections of this guidance document provide advice on minimum design-specific criteria to evaluate risk of bias and the stages involved in assessing risk of bias. We conclude with guidance on summarizing risk of bias.

Terminology

We interpret the "risk of bias" of an intervention study as the **likelihood** of **inaccuracy** in the **estimate** of **causal effect in that study**. This interpretation has five components:

1. **"Likelihood."** The actual bias of a study is unknowable, because the true effect size is unknowable. Further, poor study reporting can make important aspects of study design and conduct unclear. A risk-of-bias assessment offers a qualitative judgment of likelihood of bias.
2. **"Inaccuracy."** A study can either overestimate or underestimate the true effect, and EPC reviewers should consider both possibilities.
3. **"Estimate."** This word places the focus of risk of bias on the study's point estimate of the effect, not the precision of that point estimate. We discuss this in more detail in the next section ("Constructs Included and Excluded in Risk of Bias Assessment").
4. **"Causal effect."** In assessing the efficacy or effectiveness of one intervention versus another or versus a control, a key goal is to assess the extent to which an observed outcome difference can be directly attributed to the treatment difference.
5. **"In that study."** This phrase is meant to exclude the concept of applicability from risk of bias. Whether the study results apply to other contexts is outside the scope of risk-of-bias assessment.[14-16]

We use the phrase "risk of bias" rather than "quality assessment," because the meaning of the term *quality* varies, depending on the source of the guidance. Quality has been defined 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."[17] The Grading of Recommendations Assessment, Development and Evaluation Working Group (GRADE) uses the term quality to

refer to judgments based about the strength of the *body of evidence.*[18] 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.[19] Cochrane 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 [this emphasis] does not overcome the problem of having to rely on reports to assess the underlying research)."[14]

Because of inconsistency and potential misunderstanding in the use of the term "quality," this guidance uses risk of bias as the preferred terminology. Assessing the risk of bias of a study can be thought of as assessing the risk that the results are skewed by bias in study design or execution. This assessment process should be tailored to the specific research and clinical context of the review. **We recommend that EPCs define the terms selected in their systematic review protocols and describe the risk-of-bias categories included in the assessment.**

In the remainder of this document, we refer to components of risk of bias as categories and elements within each category as criteria (or items, if we are referring specifically to a tool). Because ideas on risk-of-bias categories have evolved, the next section describes debated larger constructs that either continue or are no longer considered to be risk-of-bias categories.

Constructs To Include and Exclude From Risk-of-Bias Assessment

Past guidance has not been consistent on which constructs to include in tools to assess bias or quality.[20, 21] The types of constructs included in tools in the past have included one or more of the following:

1. conduct of the study or internal validity,
2. precision,
3. applicability or external validity,
4. poor reporting of study design and conduct,
5. selective reporting of outcomes,
6. choice of outcome measures,
7. design of included studies,
8. fidelity to the intervention protocol, 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 issues. 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 clear guidance on related components of systematic reviews (such as selection of evidence,[22] assessment of applicability,[23] or grading the strength of evidence[18, 24-32]), some review groups continue to use practices that have served well in the past.

In the absence of strong empirical evidence, methodological decisions in this guidance document rely on epidemiological study design principles.[1] 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.

The constructs selected in the assessment of risk of bias may differ because of the clinical topic, academic orientation of the reviewers, and guidelines by sponsoring organizations. In AHRQ-sponsored reviews, 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 assessment in the past now overlap with or fall within other systematic review tasks. **Table 1** illustrates which constructs to include for each systematic review task when 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. Specific *categories* to consider when assessing risk of bias are noted separately below. Constructs wholly or partially excluded from risk-of-bias assessment continue to play an important role in the overall assessment of the evidence. The remainder of this section describes these constructs in greater detail and the rationale for including or excluding them in risk-of-bias assessments.

Table 1. Addressing precision, applicability, and bias within a systematic review

Construct	Included in Appraisal of Individual Study Risk of Bias?	Included in Assessing Applicability of Studies and the Body of Evidence?	Included Separately in Grading Strength of the Body of Evidence?
Precision	No	No	Yes (required domain)
Applicability	No	Yes	Depends on approach. GRADE includes applicability as part of strength of evidence assessment (within directness) whereas AHRQ-EPC reports applicability separately, (with the exception of rating surrogate outcomes as indirect evidence)[24]
Poor or inadequate reporting of study design and conduct	Yes (specific risk-of-bias categories and entire studies may be rated as having unclear risk of bias)	No (but could influence ability to judge applicability)	Yes
Selective outcome reporting	Yes	Not directly (however, selective reporting of results might limit the applicability of available results)	Yes (reporting bias)
Choice of outcome measures	Yes (potential for outcome measurement bias; specifically validity, reliability, and variation across study arms)	Yes (applicability of outcomes measures)	Yes (directness of outcome measures)
Study design	Yes (stronger study designs generally have lower risk of bias. However, study design should not be a proxy for risk of bias)	Not directly (however, applicability may be limited in studies with very narrow inclusion criteria)	Yes (overall risk of bias is rated separately for randomized and nonrandomized studies)

Construct	Included in Appraisal of Individual Study Risk of Bias?	Included in Assessing Applicability of Studies and the Body of Evidence?	Included Separately in Grading Strength of the Body of Evidence?
Fidelity to the intervention protocol	Yes	Yes (to the extent that fidelity or lack of fidelity influences applicability of intervention to other settings)	No
Conflict of interest	Not directly (however, conflict of interest may increase the likelihood of one or more sources of bias)	Not directly (however, conflict of interest may limit applicability if study authors or sponsors restrict study participation based on other interests)	Not directly (however, conflict of interest may influence domains of risk of bias, directness, and publication bias)

Abbreviations: AHRQ-EPC, Agency for Healthcare Research and Quality-Evidence-Based Practice Centers; GRADE, Grading of Recommendations Assessment, Development and Evaluation.

Precision

Precision refers to the degree of uncertainty surrounding an effect estimate with respect to a given outcome, based on the sufficiency of sample size and number of events.[24] Both GRADE[33] and AHRQ guidance on evaluating the strength of evidence[24] separate the evaluation of precision from that of the summary of risk of bias for a body of evidence (study limitations). Systematic reviews now routinely evaluate precision (through consideration of the optimal information size or required information size and confidence intervals around a summary effect size from pooled estimates) when grading the strength of the body of evidence.[24] Thus, the inclusion of precision as a construct under risk of bias would constitute double-counting limitations to the evidence from a single source. **We recommend that AHRQ reviews exclude considerations of power and precision of the effect estimate when assessing the risk of bias**.

Applicability

Applicability refers to the extent to which the effects observed in published studies are likely to reflect the expected results when a specific intervention is applied to the population of interest under "real-world" conditions.[34] Both GRADE[33] and AHRQ guidance on evaluating the strength of evidence[24] 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 follow-up 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, the 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. **We recommend that reviewers clearly separate**

study features that may be potential sources of bias from those that are concerned with the applicability of the individual study to the intervention, population, and context of interest.

Poor or Inadequate Reporting

In theory, risk of bias focuses on the design and conduct of a study. In practice, assessing the risk of bias of a study depends on the availability of a clear and complete description of how the study was designed and conducted, and may require additional information by reviewing clinical trials registries or study protocols or reaching out to investigators. Although new standards seek to improve reporting of study design and conduct,[35-39] EPC review teams continue to need a practical approach to dealing with poor or inadequate reporting. Empirical studies suggest that unclear or poor reporting may not always reflect poor study conduct.[40]

EPC reviews have varied in their treatment of reporting of study design and conduct. Some have elected to rate outcomes from poorly *reported* studies as having high risk of bias. Other EPCs have chosen to select an "unclear risk-of-bias" category for studies with missing or poorly reported information on which to base risk-of-bias judgments. In other cases, EPCs have judged that specific bias components, although poorly reported, have no material effect on overall risk of bias. **We recommend that assessment of risk of bias focus primarily on the design and conduct of studies and not on the quality of reporting. However, we recognize that poor reporting can impede judgments of risk of bias. Therefore, we also recommend that EPCs clearly document inadequate reporting for all risk of bias domains. When reviews include meta-analyses, we recommend that systematic reviewers consider sensitivity analyses to assess the impact of including studies with poorly reported risk-of-bias components; when studies cannot be pooled, consider qualitative analyses.**

Selective Outcome Reporting

Reporting bias occurs when the nature and direction of the results influences their dissemination.[1] Reporting bias includes bias in whether to publish or not (publication bias), when to publish (time lag bias), where to publish (location bias [selecting venues with greater or lesser ease of access depending on the direction of results]) and what to publish (selective outcome reporting). Many of these sources of bias are best addressed at the level of the body of evidence; patterns of bias may not be discernable at the level of the individual study. Selective outcome reporting, specifically, has[41] major implications for both the risk of bias of individual studies and the strength of the body of evidence[41] and can be discerned in some instances at the level of the individual study. Comparisons of the full protocol to published and unpublished results can help to flag studies that selectively report outcomes. In the absence of access to full protocols,[24, 32] Guyatt et al. note that "[o]ne 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."[32] Note that selective outcome reporting includes selective reporting of planned analyses and selective reporting of results.

Methods continue to be developed for identifying and judging the risk of bias when results deviate from protocols in the timing or measurement 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 within a body of evidence may result in downgrading the body of evidence.[32]

Previous research has established the link between funding source and an array of consequential study decisions on design, conduct, and dissemination of results (sponsor bias).[42-44] Publication bias may be a pervasive problem in some bodies of evidence and should be evaluated when grading the body of evidence, as should time lag and location bias.[43]As methods on identifying and weighing the likely effect of selective outcome reporting and other reporting biases continue to be developed, this guidance will also require updating. **We recommend considering the risk of selective outcome reporting for both individual studies and the body of evidence, particularly when a suspicion exists that forces such as sponsor bias may influence the reporting of analyses and results.**

Choice of Outcome Measures

The use of valid and reliable outcome measures reduces the likelihood of bias in measuring outcomes. For example, some self-report measures may be rated as having a higher risk of bias than clinically observed outcomes in unblinded designs; at the same time, patient-reported outcomes may also be more applicable to the general population. In addition, use of different outcome measures for each 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 of risk of bias of individual studies include the evaluation of the validity and reliability of outcome measures overall, and differences in validity and reliability between study arms.**

The validity and reliability measures across treatment arms are criteria for judging the risk-of-bias, but the choice of specific outcome measures should also be considered when judging the directness of the outcome and applicability of the study. Directness of outcomes (or comparisons) refers to whether the evidence directly links interventions to important health outcomes and is a key domain in assessing the strength of the body of evidence[24] or applicability.[34]

Study Design

In general, stronger study designs will have lower risk of bias. 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, often included in EPC reviews, have specific and inherent risks of biases that cannot be minimized. However, instead of equating risk of bias solely with study design, the bias represented by study design features may be considered at the overall strength of evidence level. For example, 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 the body of evidence. The inherent limitations present in nonrandomized designs are factored in when grading the strength of evidence. EPCs generally give evidence derived from nonrandomized studies a lower starting grade and evidence from randomized controlled trials a high grade. They can then upgrade or downgrade the nonrandomized 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).[24]

We recommend that EPCs do not use study design labels (e.g. observational studies) 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 based solely on the study design label but should use risk-of-bias categories or criteria that consider the role of the design element and the subsequent risk of bias. A study can be conducted well but still have some (if not serious) potential risk of bias because of underlying design flaws.[1]

EPCs may consider whether to exclude evidence from study designs with limited ability to address causal inference, such as case studies and case series. 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, etc.] provide valid and useful information to address key questions?" Depending on the clinical question and the context, EPCs may judge that the information provides value or that the risk of bias from a particular study design may be unacceptably high. **If such nonrandomized studies are included, we recommend that EPCs consider the risk of bias of individual studies, rather than applying a single common rating based on design without considering study-specific variations in design and conduct**.

Fidelity to the Intervention Protocol

Failure of the study to maintain fidelity to the intervention protocol can bias performance; it is, therefore, a component of risk of bias assessment. We note, however, that the interpretation of fidelity may differ by clinical topic and the nature of the outcome evaluated. For instance, some behavioral interventions include "fluid" interventions; these involve interventions for which the protocol explicitly allows for modification based on patient needs or concomitant treatments. Such fluidity does not mean the interventions are implemented incorrectly, and an intention-to-treat analysis will capture the effect of the intervention as assigned. 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. When systematic reviewers are interested in the effect of starting and adhering to interventions (the per-protocol effect), deviations from the intervention protocol (including lower-than-expected adherence) can bias results. **We recommend that EPCs account for the specific clinical and outcome considerations in determining and applying criteria about fidelity for assessment of risk of bias.**

Conflict of Interest

Studies have found that conflicts of interest (financial and nonfinancial) can threaten the internal validity and applicability of primary studies and systematic reviews.[45, 46] Conflicts of interest can arise from when investigators or funders of studies deploy strategies that influence the results such as (1) selecting specific designs and hypotheses—for example, choosing noninferiority rather than superiority approaches,[47] picking comparison drugs and doses,[47] choosing outcomes,[46] or using composite endpoints (e.g., mortality and quality of life) without presenting data on individual endpoints;[48] (2) selectively reporting outcomes —for example, reporting relative risk reduction rather than absolute risk reduction; selecting from multiple

endpoints[47] or reporting on subscales of larger scales; reporting inappropriately developed categorical variables, based on selected cut-points in continuous measures;[49] (3) presenting results in a biased[48] or inadequate manner[49] and (4) failing to publish results, thereby contributing to publication bias.[50]

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). In other instances, 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)[51] makes the actual contribution of the sponsor very difficult to discern.[52, 53]

Given these concerns, conflicts of interest should be considered when critically appraising the evidence because they may serve as an indirect marker of risk of bias. For several reasons, we caution against simple-to-follow rules such as equating industry sponsorship with high risk of bias. First, financial conflicts of interest are not limited to industry; nonprofit and government-sponsored studies may also have conflicts of interest. Researchers may have various financial or intellectual conflicts of interest by virtue of, for example, accepting speaking fees from many sources.[54] 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.[45] Third, the multiple pathways by which conflicts of interest 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 conflicts of interest on selective outcome reporting for individual studies, 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.

Conflicts of interest may be particularly apparent in conclusions of studies.[55] Although of concern to the general reader, biased presentation or "spin" on results, if limited to the discussion and conclusion section of studies, should have no bearing on systematic review conclusions because systematic reviews should not rely solely on interpretation of data by study authors. Nonetheless, biased presentation of results may serve as a flag to evaluate the potential for risk of bias closely.

Internal validity and completeness of reporting constitute, then, the primary pathway by which conflicts of interest may influence the validity of study results that is entirely within the purview of assessment of risk of bias. We acknowledge that this pathway may not be the most important source of conflict of interest: as standards for conduct and reporting of studies become widespread and journals require that they be met, differences in internal validity and reporting between studies with and without inherent conflicts of interest 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 recommend: (1) at a minimum, EPCs should routinely report the source of each study's funding (or the failure of the study to report such information); (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 (quantitative or qualitative) 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.[47]

Stages in Assessing the Risk of Bias of Studies

International reporting standards require documentation of various stages in a systematic review.[56-58] We lay out recommended approaches to assessing risk of bias in five steps: protocol development, pilot testing and training, assessment of risk of bias, interpretation, and reporting. **Table 2** describes the stages and specific steps in assessing the risk of bias of individual studies that contribute to transparency through careful documentation of decisions.

The plan for assessing 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 outcomes that need risk-of-bias assessment and other study descriptors or study data elements that are required to assess risk of bias in the systematic review protocol. Protocols must describe and justify what risk-of-bias categories and tools will be used and how the reviewers will incorporate risk of bias of individual studies in the synthesis of evidence.

The assessment should include a minimum of two independent reviewers per study with an unbiased reconciliation method such as a third person serving 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. Although we recommend that risk-of-bias assessment be performed in duplicate, reviewers should be aware of recent software developments that may improve the efficiency of the process. A study by Marshall et al. (2014)[59, 60] applied text-mining software to 2,200 full-text publications and their parent Cochrane reviews. The software analyzed textual patterns between full-text articles and the eventual risk-of-bias assessments of Cochrane authors (e.g., the occurrence of the phrase "sealed envelopes" in a full article is likely an accurate predictor of "low" risk of bias with respect to concealment of allocation). Although the software should not be used to completely replace reviewers (as it did make some erroneous predictions), other possible uses include the production of first-pass judgments (with subsequent human review), or the automation of text flagging to support reviewers' risk-of-bias judgments. **First order assessments of risk of bias by machine-learning require secondary human review.**

Assessment of risk of bias should be consistent with the registered protocols of the reviews. The synthesis of the evidence should reflect the *a priori* plan in the protocol 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. When reviewers exclude high risk-of-bias studies from their analysis entirely without any sensitivity analyses, we recommend that reviewers explain their decision.

Table 2. Stages in assessing the risk of bias of individual studies

Stages in Risk-of-Bias Assessment	Specific Steps
1. Develop protocol	• Specify risk-of-bias categories (including sources of potential confounding for nonrandomized studies) and criteria and explain their inclusion • Select and justify choice of specific risk-of-bias rating tool(s), including validity of selected tools (use risk-of-bias assessment tools that can identify potential risk-of-bias categories specific to the content area and study design) • Explain how individual risk-of-bias categories (or items from a tool) will be presented or summarized (e.g., individually in tables, incorporated in sensitivity analysis, combined in an algorithm to obtain low, moderate, high, or unclear risk of bias for individual outcomes) • 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 excluded from synthesis of the evidence and implications of such exclusions)
2. Pilot test and train	• Determine composition of the review team. Teams should include methods and content experts. A minimum of two reviewers must rate the risk of bias of each study, and an approach developed for the arbitration of conflicts. • Train reviewers • Pilot test assessment of risk-of-bias tools using a small subset of studies that are likely to represent the range of risk-of-bias concerns 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 • For nonrandomized study designs, consider specifying a "target" trial[a] to assist in considering how results from a nonrandomized study may differ from those expected in an RCT; such specification may help identify specific sources of bias. Clarify whether the effect of interest is in relation to assignment to the intervention (intention-to-treat) OR starting and adhering to the intervention (e.g., per-protocol effect) • For nonrandomized studies, specify likely sources of potential confounding • Make judgments about each risk-of-bias category, using the preselected appropriate criteria for that study design and for each predetermined outcome • Present judgment criteria on individual categories or items or as a summary for each outcome • If presenting a summary, make judgments about overall risk of bias for each included outcome of the individual study, considering study conduct, and rate as low, moderate, high, or unknown risk of bias within study design; document the reasons for judgment and process for finalizing judgment • If separately presenting risk-of-bias for individual items, assess the implications for direction and magnitude of bias. Resolve differences in judgment and record final rating for each outcome
4. Use risk-of-bias assessments in synthesizing evidence	• Conduct preplanned analyses based on a priori criteria for including or excluding studies based on risk-of-bias assessments • Consider and conduct, as appropriate, additional analyses (e.g., quantitative or qualitative sensitivity analyses or exploration of heterogeneity) to assess impact of risk of bias on findings. • Summarize individual study risk of bias into overall strength of evidence study limitations domain.
5. Report risk-of-bias findings, process and limitations	• Describe the risk-of-bias process (summarizing from the protocol), post-protocol deviations, and limitations to the process. • Present findings and conclusions transparently, balancing the competing considerations of simplicity of presentation with burden on the reader

[a]A target trial is a hypothetical randomized controlled trial of the intervention; feasibility or ethics do not play a role in constructing such a hypothetical trial.[61]

Identifying, Selecting, and Assessing Categories of Risk of Bias

Identifying Categories of Risk of Bias

Different categories of bias are often described by a host of different terms and the same terms are sometimes used to refer to different categories of bias depending on the study design of interest. Here, we rely and expand on the newly developed ROBINS-I tool[61] to outline specific categories of risk of bias (termed "domains" in the ROBINS-I tool) for assessment in systematic reviews (**Table 3**). We chose this tool because it offers a comprehensive array of bias categories that captures recent advances in epidemiological thinking. Despite the focus on assessing the risk of bias in nonrandomized studies (e.g., controlled nonrandomized clinical trials, prospective or retrospective cohort studies, and case-control studies) in the ROBINS-I tool, the core categories of risk of bias apply to randomized trials. The key additions relate to biases occurring before or at the start of the intervention. The categories outlined here specifically relate to designs that allow a causal interpretation of the effect of the intervention on outcomes and suggest a preliminary set of criteria for RCTs, nonrandomized cohort designs (nonrandomized controlled designs, prospective and retrospective cohorts with comparisons), and case-control studies. It excludes case studies, case series and cross-sectional studies, although some systematic reviews may choose to include information from such studies. If a study that claims to be an RCT is determined to be better classified as a nonrandomized study (e.g., due to major problems with "randomization"), reviewers may elect to classify the study as nonrandomized, and thus assess risk of bias based on criteria for nonrandomized studies.

In the ROBINS-I taxonomy of bias, pre-intervention sources of bias arise from confounding and selection of participants into the study. Biases arising at the start of the intervention can occur when intervention status is misclassified (i.e., intervention groups are not clearly defined or recorded at the start of the intervention, classification of the intervention status is affected by knowledge of the outcome). Biases occurring after the initiation of the intervention may arise from departures in intended interventions, missing data, measurement of outcomes, and selective reporting. The authors propose evaluating potential sources of bias in a nonrandomized study against a "target" trial that avoids biases arising lack of randomization in assignment. A target trial is a hypothetical randomized controlled trial of the intervention; feasibility or ethics do not play a role in constructing such a hypothetical trial.[61]

Selecting and Assessing Relevant Categories of Bias For a Review

Determining the risks of bias that are most salient or that require special consideration is often dependent on the focus of the clinical topic being reviewed. For example, in the table below, biases arising from departures from intended interventions are particularly relevant for outcomes for which the exposure of interest is starting and adhering to interventions.[61] Reviewers should determine *a priori* whether the intervention of interest is assignment to the intervention at baseline, or assignment and adherence to the assigned intervention. Prespecification of outcomes (as it relates to bias in reporting results) is another example that requires topic- or outcome-specific evaluation. For example, prespecification of *benefits* within a study is entirely appropriate and expected, regardless of study design. The prespecification of

particular *harms*, however, may not be possible for all topics; in these cases, data from observational studies may offer the first opportunity to identify unexpected outcomes. Likewise, for review topics in search of evidence on rare long-term outcomes, requiring prespecification would be inappropriate. Another 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 and outcome. Reviewers of topics that focus on short-term clinical outcomes may expect a low rate of attrition. We 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. Some criteria included in **Table 3**, particularly intention-to-treat, have been interpreted in a variety of ways. The *Cochrane Handbook of Systematic Reviews* offers a more detailed description of intention to treat.[1]

Reviewing the risk of bias within individual studies often begins by looking at a study as a whole for potential biases (e.g., valid randomization and allocation procedures, confounding) and then focusing on risks that might occur at an outcome-specific level as not all sources of bias will influence all outcomes measured in a study in the same degree or direction. For instance, biases in the measurement of outcomes (e.g., blinding of outcome assessors) and biases due to missing data may be different for each outcome of interest. That is, blinding of outcome assessors may be particularly important for self-reported measures that are interviewer-administered but may not be a central risk for objectively-measured clinical outcomes. Likewise, in cases of high attrition within a study or for particular outcomes, the appropriateness and effect of procedures to account for missing data (e.g., baseline or last observation carried forward methods) should be considered at an outcome-specific level.

Table 3 is not intended to be used as an instrument. We recommend selecting the most important categories of bias for the outcome(s) and topic at hand. No checklist can replace a thoughtful consideration of all relevant issues. A hypothetical consideration of a target trial can help identify the most important risk-of-bias considerations.[61] In particular, in relation to assessing non-randomized studies, a combination of methods and topical expertise will be necessary to anticipate the most important sources of bias, assess risk of bias, and interpret the effect of potential sources of bias on estimates of effect.

Table 3. Description of risk-of-bias categories and study design-specific assessment criteria for randomized and nonrandomized studies of interventions (adapted from ROBINS-I)[a]

Categories of Bias Related to Design and Conduct of the Study[b]	Description of Bias	Study Design or Conduct Factors to Avoid Bias	RCTs[c]	Nonrandomized Studies[d]	Case-Controls
Bias arising in the randomization process or due to confounding	When one or more prognostic variables (factor that predict the outcome of interest) influences whether study participants receive one or the other intervention	• Random sequence generation	X		
		• Allocation concealment: approach that precludes researchers enrolling participants from knowing their assignment	X	X[e]	
		• Balance in baseline characteristics, or appropriate adjustment for differences in baseline characteristics	X	X	X[f]
		• No baseline confounding (i.e., participant characteristics such as disease severity or comorbidity are unlikely to influence the intervention and outcome) or appropriate analysis methods are used to adjust for important baseline confounding	X	X	X
		• No time-varying confounding (i.e., participant prognostic variables are unlikely to influence discontinuations or switches between interventions) or appropriate analysis methods are used to adjusted for important time-varying confounding		X	X
Bias in selecting participants into the study[g]	When participants (or initial followup time for some participants only) are selected into the study based on characteristics observed after the start of the intervention/exposure	• Selection of participants is independent of characteristics observed after the start of the intervention that are likely to be associated with the intervention[h]		X	X
		• Start of follow-up and start of intervention coincide		X	X
		• If potential for selection bias, appropriate analysis methods are used to account for participants who were inappropriately excluded		X	X
Bias in classifying interventions	When participant intervention status is misclassified because the intervention status was not recorded in a valid and reliable manner at the start of the intervention	• Participant intervention status is clearly and explicitly defined and measured		X	X
		• Information used to define intervention group status is recorded at the start of the intervention		X	X
		• Classification of intervention status is unaffected by knowledge of the outcome or risk of the outcome		X	X
Bias due to departures from intended interventions[i, j]	Differences between the intended and actual intervention	• Implementation of the intervention as intended and adherence to assigned intervention regimen	X	X	X
		• Co-interventions are balanced between intervention groups	X	X	X
		• No or minimal contamination between groups	X	X	X

Categories of Bias Related to Design and Conduct of the Study[b]	Description of Bias	Study Design or Conduct Factors to Avoid Bias	RCTs[c]	Nonrandomized Studies[d]	Case-Controls
		• Participants are blinded to intervention group assignment	X	X[e]	
		• Providers are blinded to participant intervention group assignment	X		
		• Analysis appropriately accounts for the intended intervention assignment for all participants	X	X	
		• If deviation from intended intervention, analysis adjusts for imbalance between groups in co-interventions that could affect outcomes	X	X	X
Bias from missing data	Overall or systematic differences between study groups in loss of participants from the study that are not accounted for in the analyses	• Outcome data are reasonably complete[h] and proportion of participants and reasons for missing data are similar across groups	X	X	X
		• Confounding variables that are controlled for in the analysis are reasonably complete across participants		X	X
		• Appropriate statistical methods are used to account for missing data (i.e., intention-to-treat analyses using appropriate imputation techniques)	X	X	X
		• Intervention status is reasonably complete and does not differ systematically between groups	X	X	
Bias in measurement of outcomes	Overall or systematic differences between study groups in assessment of outcomes	• Outcome assessors are blinded to intervention status of participants[k]	X	X	
		• Outcomes are measured using valid and consistent procedures and instruments across all study participants	X	X	X
		• Errors in measurement of the outcome are unrelated to the intervention received (i.e., no differential misclassification of outcomes)	X	X	X
		• Appropriate use of inferential statistics[l]	X	X	X
Bias in reporting outcomes selectively	Selectively reporting outcomes based on the findings	• Outcomes are prespecified and all prespecified outcomes are reported	X	X	X
		• No evidence that the intended measures, analyses, or subgroup analyses are selectively concealed	X	X	X

RCT = randomized clinical trial

[a]Details on categories, definitions, and items can be found in Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. The BMJ. 2016;355:i4919. doi:10.1136/bmj.i4919. Note that the first 3 categories of biases presented in the Table occur before or at the time of the intervention or exposure. The remaining categories of biases occur after the intervention. Adapted table used with permission.

[b]Bias arising from design occur before or at the intervention and include bias due to confounding, selection, and classification of interventions. Bias arising from conduct occur after the intervention and can arise from departures from intended interventions, missing data, measurement of outcomes, and reporting.

[c]Bias arising from design for RCTs arise principally from randomization flaws, as contrasted with other designs that have multiple potential sources of bias before or at intervention. Bias arising from conduct for RCTs are similar to bias for other designs.

[d]Includes nonrandomized controlled studies with investigator-allocated treatment and observational studies of prospective or retrospective cohorts with comparison arms

[e]Relevant only for nonrandomized experimental studies where the investigator allocates treatment

[f]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.

[g]Refers to biases that are internal to the study only, and does not refer to issues of applicability (e.g., restricting the sample to a specific clinical population). Selection bias results when the study design results in a biased estimate of the effect because the design of the study resulted in the exclusion of some participants or their data. For example, studies that evaluate the effect of folic acid supplementation on neural tube on live births only selectively exclude outcomes from pregnancies resulting in fetal deaths. Selection bias can also arise in retrospective studies that do not have complete data for all potential participants at inception or do not restrict their design to "naïve" drug users – by design, these designs potentially exclude eligible participants.

[h]Although we do not expect selection bias to occur routinely in trials, informative censoring in trials with different baseline times could potentially result in selection bias.

[i]This category is relevant only when the review is evaluating the effect of starting and adhering to interventions.

[j]There are no established rules for determining a threshold for appropriate completeness of outcome data. Reviewers should establish what is meant by "Reasonably complete" based on the specific topic and outcome.

[k]Blinding of outcome assessors is especially important with subjective outcome assessments.

[l]Reviewers do not need to evaluate inferential statistics used in studies that report results in a manner that permits meta-analyses or other independent analyses. When reviewers need to rely solely on the results as presented by authors, they may elect to review the use of inferential statistics in the study.

Tools for Assessing Risk of Bias

Many tools have emerged over the past 25 years to assess risk of bias; several reviews that describe and compare the most commonly used risk-of-bias instruments.[20, 62-66] 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 categories to address all the study designs included. Although there is much overlap across different tools, no single universal tool addresses all the varied contexts for assessment of risk of bias. If reviewers choose to use or adapt an existing risk of bias tool/instrument, thoughtful consideration of all relevant issues specific to the topic at hand is crucial. We advocate the following general principles when selecting a tool, or approach, to assessing risk of bias in systematic reviews. **EPCs should use tools that:**

- were specifically designed for use in systematic review,
- are specific to the study designs being evaluated,
- show transparency in how assessments are made by providing explicit support for each assessment,
- specifically address items related to risk-of-bias categories,
- are, at minimum, based on theory and are preferably based on empirical evidence that risk-of-bias categories are associated with biased effect estimates or have reasonable face validity, and
- avoid the presentation of risk-of-bias assessment solely as a numerical score (or, if numerically scored, conduct sensitivity analyses of these scores at minimum).

Direction and Magnitude of Bias

Reviewers should consider both the direction and magnitude of possible bias on the effect estimate in arriving at a risk of bias rating. Regarding direction, reviewers should be careful not to assume that all study biases result in overestimation of effect sizes. As defined earlier, bias is any mis-estimation of an effect size, and both underestimation and overestimation are problematic for decision makers. Although the task of considering the direction and magnitude of bias can be challenging, it helps reviewers judge whether the potential bias is consequential or should be ignored because it is unlikely to materially alter results. It also helps reviewers judge whether deficiencies noted for different areas of bias are related. For example, baseline imbalances in observational studies that have no relationship with the outcome may not be consequential.

The likely direction of bias depends on the risk-of-bias category being considered as well as specific considerations within that category. In the case of confounding—as described by ROBINS-I ("pre-intervention prognostic factor that predicts whether an individual receives one or the other intervention of interest")—effect size is often overestimated, and a classic case is "confounding-by-indication," since patients with different medical indications would have had different outcomes regardless of treatment. In the category of missing data, on the other hand, the direction of bias depends on whose data are missing and why they are missing. If one treatment group had a larger rate of missing quality-of-life data and the reason for missing data was that

those patients were cured and felt no reason to attend follow-up appointments, then the available data are biased against the group with the larger rate of missing data. But if the reason for missing data was deteriorating health (e.g., did not feel well enough to attend follow-up appointments), the available data are biased in favor of the group with more missing data.

Further complicating matters is the possibility of different biases cancelling each other out. If a study has two clear biases but they appear to work in opposite directions, reviewers may infer that the effect size estimate may be fairly accurate. This inference depends on numerous assumptions, including (1) that the reviewer has correctly judged the direction of bias in both cases; (2) that the two biases have similar magnitude; and (3) that the reviewer has correctly judged that no other biases play an important role. All three of these are subjective judgments. Thus, the claim of "cancelling out," while theoretically possible, would require strong consensus within a review team.

Regarding the magnitude of bias, an ideal scenario is when one can use existing research to quantify the risk of bias of each effect size estimate, and then adjust the estimates accordingly ("bias adjustment"). Rarely will a review team have the necessary evidence and resources to support this endeavor. Despite the likely lack of empirical evidence for quantitatively adjusting estimates to account for bias, we believe that considerations of magnitude of bias matter. We also note, however, that current review processes entail several implicit judgments about the magnitude of bias. For example, when reviewers decide which risk-of-bias items to use, they are attempting to capture the biases that have the largest influence on effect size. Also, some risk-of-bias items use numerical thresholds (e.g., did at least 85% of enrolled patients provide data to the time point of interest?), and studies meeting that threshold are considered to have no bias for that item. Our recommendation, then, is to consider the implications of the risk of bias carefully rather than in a formulaic fashion. Such an effort will help focus reviewers on consequential sources of bias. It will also help understand how different sources of bias might be related.

Assessing the Credibility of Subgroup Analyses

Systematic reviewers routinely consider benefits and harms in specified subpopulations or other subgroups (e.g., by specific route of administration of a drug). Subgroup analyses can help to improve understanding of factors that contribute to heterogeneity of study results. A misleading subgroup analysis may not fit the classic description of bias or confounding, but can have the same effect by providing evidence users with incorrect conclusions. Therefore, **when systematic reviewers report or synthesize subgroup analyses, they should inform readers of their assessment of the credibility (trustworthiness) of inferences derived from such analyses.**

Studies rated as having a high risk of bias for the main analysis of benefits or harms will also likely have a high risk of bias for subgroup analysis. However, studies with low risk of bias for their overall analysis of benefits or harms may not necessarily have credible subgroup analysis. In fact, empiric evaluation shows that the credibility of subgroup effects, even when overall claims are strong, is usually low.[67]

Assessing the credibility of subgroup analyses in primary studies requires paying attention to issues such as whether: (1) chance can explain the apparent subgroup effect (i.e., an interaction test can be conducted to demonstrate whether the difference in effect size between subgroups is less likely to be caused by chance); (2) the subgroup effect is consistently observed in several studies; (3) the subgroup hypothesis is one of a small number of hypotheses developed *a priori* with a specified direction; (4) there is strong preexisting biological rationale for the effect; and

(5) the evidence supporting the subgroup effect is observed within studies (as opposed to only being observed in comparisons across studies; which is less credible).[68] There is no specific tool or checklist that has been validated for assessing the credibility of subgroup analysis although criteria have been proposed for preventive clinical services[69]and for randomized controlled trials.[70]

In addition to challenges that relate to spurious subgroup effects that are demonstrated to be statistically significant (but may not be credible), there are other challenges that relate to the fact that subgroup analyses are usually underpowered.[69] Therefore, a statistically nonsignificant subgroup interaction cannot rule out a true interaction.

Assessing the Risk of Bias for Harms

Although harms are almost always included as an outcome in intervention studies that requires a risk-of-bias assessment, the manner of capturing and reporting harms is significantly different from 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."[71] For a detailed explanation of terms associated with harms please refer to the AHRQ Methods Guide on harms.[72] Decisionmakers 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 receives much less attention than the positive efficacy outcomes.[73, 74] Bias for harms from observational studies continue to be a major concern. Design and analytic choices can substantially alter results.[75]

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; 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, reviewers may elect to evaluate the risk of bias for benefits and harms in different ways. **We recommend that EPCs be explicit about whether they plan to apply the same methods for risk of bias to both benefits and harms and justify the choice of methods.**

Assessing the Credibility of Existing Systematic Reviews

This guide focuses on assessing risk of bias of primary studies; however, it is becoming more common to use existing systematic reviews in evidence synthesis products. There are two main

approaches to using systematic reviews. First, if there are systematic reviews on the interventions (or topics) of interest, reviewers may choose to conduct an overview of reviews. Overviews are defined by Cochrane as knowledge synthesis products that bring together "multiple systematic reviews addressing a set of related interventions, conditions, population, or outcomes."[76] In overviews, "the unit of searching, inclusion and data analysis is the systematic review."[76] Second, systematic reviews may be integrated into *de novo* reviews, i.e., parts of the systematic review(s) may be used as a basis for information in a new systematic review.[77, 78] For example, the list of included studies may be used as a starting point for a new systematic review, with additional searching that builds upon the search in the existing review. Other parts of an existing systematic review may also be used, such as risk-of-bias assessments, data extraction, and/or data analyses conducted by those who produced the original systematic review. More details on integrating systematic reviews can be found in another EPC Methods Guide.[77, 78]

When conducting an overview of reviews, it is important to assess the credibility of the included systematic reviews, as well as evaluate the procedures for and document the results of risk-of-bias assessments of the included studies. Likewise, when considering whether or not to integrate systematic review results into *de novo* reviews, it may also be important to assess their credibility to guide decisions about whether to use elements of the review (i.e., what confidence do we have in the methodological rigor with which the review was conducted) and to report on the risk of bias if elements are used and reported in a *de novo* review (i.e., informing the reader about the methodological rigor of the information that has been incorporated).

Several tools have been developed to determine how trustworthy systematic reviews are; these tools have used variable terms including "risk of bias" and "methodological quality." The term "credibility" was suggested to replace "risk of bias" when dealing with determining how trustworthy the review process was.[79, 80] The rationale for this differentiation is that a very well conducted systematic review of poorly conducted trials can produce biased estimates but the review itself may have been well done. Conversely, a review with a poor search strategy may lead to estimates that do not represent the totality of evidence, yet, the estimates are not necessarily biased towards one particular direction (overestimation or underestimation of the treatment effect). Therefore, the credibility of the process of a systematic review can be defined as the extent to which its design and conduct are likely to have protected against misleading results.[79] Credibility may be undermined by inappropriate eligibility criteria, inadequate literature search, or failure to optimally synthesize results. On the other hand, the term "risk of bias" remains as a descriptor of possible bias in individual studies or a body of studies.

Several tools are available to assess the credibility of systematic reviews; although some without much uptake.[81-84] The more commonly used tool, developed in 2007, is the Assessing the Methodological Quality of Systematic Reviews Evaluations (AMSTAR) tool. The developers of the original AMSTAR tool are currently working on modifying the tool.[85] A second more recent tool is ROBIS, Risk of Bias in Systematic Reviews, which was released in 2015.[86] ROBIS focuses on risk of bias as opposed to the rigor of the process of a systematic review, which is the focus of AMSTAR.[87]

In addition to the above tools, there are at least two reporting guidelines for systematic reviews: Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) and Meta-analysis of Observational Studies in Epidemiology (MOOSE).[57, 88] Both are available at www.equator-network.org, along with variations/extensions and guidelines for other types of reviews (e.g., meta-narrative reviews and realist syntheses). These may provide a proxy for

methodological quality/risk of bias/credibility and an indication of the extent or comprehensiveness of reporting.[a]

Reporting the Risk of Bias

During the protocol phase, reviewers should decide on the best approach for reporting the results of the risk-of-bias assessments. The approach used to summarize risk-of-bias assessments should balance considerations of simplicity of presentation and burden on the reader. Risk-of-bias results of individual studies can be reported using a *composite* or a *components* approach. In a *composite* approach, systematic reviewers combine the results of category-specific risk-of-bias assessments to produce a single overall assessment. This assessment often results in a judgement of low, moderate, high, or unclear risk-of-bias. Because a study's risk-of-bias category or "rating" can be different for different outcomes, review teams may opt to record the overall assessments by outcome. Alternatively, if the risk-of-bias assessments were generally uniform across outcomes, an overall study-level risk-of-bias rating could be generated for the study as a whole that can be applied to all outcomes.

Although creating a summary risk-of-bias judgment for each study or outcome may be a necessary step for strength of evidence judgment, such a summary runs the risk of ignoring or overweighting important sources of bias. In a *components* approach, reviewers report the risk-of-bias assessment for each study for each bias category or even each item. Previous research has demonstrated that empirical evidence of bias differed across individual categories rather than overall risk of bias.[89] Reviewers may use meta-analyses to examine the association between risk-of-bias categories or items and treatment effect with subgroup analyses or meta-regression.[90-92]

An approach that relies solely on presentation of judgment on the components (categories or items) alone, however, devolves the burden of effort of interpretation of a study's risk of bias from the systematic reviewer to the readers. Therefore, we suggest that reviewers carefully consider composite (outcome- or study-specific) summary risk-of-bias judgements as well as component (category)-specific assessments. When presenting the results, reviewers should focus

---

[a] A number of critical appraisal tools and checklists also exist for systematic reviews, for example:

Critical Appraisal Skills Program (CASP) systematic review checklist
(http://www.casp-uk.net/checklists)

Health Evidence Quality Assessment Tool (HE-QAT)
(http://www.healthevidence.org/documents/our-appraisal-tools/QA_tool&dictionary_18.Mar.2013.pdf)

JBI (Joanna Briggs Institute) critical appraisal instrument for Systematic reviews and Research Syntheses
(http://joannabriggs.org/assets/docs/jbc/operations/criticalAppraisalForms/JBC_Form_CritAp_SRsRs.pdf)

National Institute for Health and Care Excellence (NICE) systematic reviews and meta-analyses methodology checklist
(https://www.nice.org.uk/process/pmg10/chapter/appendix-b-methodology-checklist-systematic-reviews-and-meta-analyses)

Scottish Intercollegiate Guidelines Network (SIGN) Systematic Reviews and Meta-Analysis Checklist
(http://www.sign.ac.uk/checklists-and-notes.html).

A detailed discussion of these tools is beyond the scope of this guide.

on the elements of risk of bias of greatest relevance to understanding and interpreting the evidence.

Transparency is important so that users can understand how final assessments were assigned. Transparency also helps to ensure that risk-of-bias results can be reproduced and assures that the same process was used for all included studies. In applying the same rules across all outcomes to ensure consistency, there is a danger, however, in being too formulaic and insensitive to the specific clinical context of the outcome. 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. **The presentation of risk-of-bias assessments should be done in a way that allows readers not only to determine whether each type of bias is present, absent, or unknown for each study, but also the most likely direction and magnitude of bias when bias is likely to be present (when possible).**

Again, we recommend that, in aiming for transparency and reproducibility, EPC reviewers use a set of specific rules for assigning risk-of-bias "ratings". These rules should take the form of declarative statements that indicate any judgments or weighting that was applied to specific risk-of-bias items or domains. Though the use of quantitative scales is a way to employ a transparent set of results, any weighting system, whether qualitative or quantitative, must be recognized as subjective and arbitrary, and different reviewers may choose to use different weighting methods. Consequently, we believe that 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."

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 theoretical principles at minimum, and sound empirical evidence when possible. In assessing the risk of bias of studies, EPCs should prioritize transparency of judgment through careful documentation of processes and decisions.

References

1. Higgins J, Green S. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. In: Higgins JPT, Green S, eds.: The Cochrane Collaboration; 2011. http://handbook.cochrane.org. Accessed on November 1, 2017.
2. Institute of Medicine. Finding what works in health care: standards for systematic reviews. http://www.nap.edu/openbook.php?record_id=13059&page=R1. Accessed on November 1, 2017.
3. Armijo Olivo S, Ospina M, da Costa BR, et al. Poor Reliability between Cochrane Reviewers and Blinded External Reviewers When Applying the Cochrane Risk of Bias Tool in Physical Therapy Trials. PloS one. 2014;9(5):e96920. PMID: 24824199. http://dx.doi.org/10.1371/journal.pone.0096920
4. Savovic J, Jones HE, Altman DG, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Ann Intern Med. 2012 Sep 18;157(6):429-38. PMID: 22945832. http://dx.doi.org/10.7326/0003-4819-157-6-201209180-00537
5. Hempel S, Suttorp MJ, Miles JNV, et al. Empirical Evidence of Associations Between Trial Quality and Effect Sizes. Methods Research Report (Prepared by the Southern California Evidence-based Practice Center under Contract No. 290-2007-10062-I). AHRQ Publication No. 11-EHC045-EF. Rockville, MD: Agency for Healthcare Research and Quality. Available at: http://effectivehealthcare.ahrq.gov. 2011.
6. Berkman ND, Santaguida PL, Viswanathan M, et al. The Empirical Evidence of Bias in Trials Measuring Treatment Differences. Prepared for: Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services, Contract No. 290-2007-10056-I. AHRQ Publication No. 14-EHC050-EF. Prepared by: RTI-UNC Evidence-based Practice Center, Research Triangle Park, NC. Rockville, MD: 2014.
7. Hartling L, Hamm MP, Milne A, et al. Testing the Risk of Bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. J Clin Epidemiol. 2013;66(9):973-81. PMID: 22981249. http://dx.doi.org/10.1016/j.jclinepi.2012.07.005
8. Applying the risk of bias tool in a systematic review of combination long-acting beta-agonists and inhaled corticosteroids for maintenance therapy of persistent asthma. 17th Cochrane Colloquium; 2009 2009; Singapore; London. Cochrane Collaboration.
9. Santaguida PL, Riley CR, Matchar DB. Chapter 5: Assessing risk of bias as a domain of quality in medical test studies. AHRQ Publication No. 12-EHC077-EF. Chapter 5 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare .ahrq.gov/reports/final.cfm. Also published in a special supplement to the Journal of General Internal Medicine, July 2012. Rockville, MD: 2012.
10. Hayden JA, van der Windt DA, Cartwright JL, et al. Assessing bias in studies of prognostic factors. Ann Intern Med. 2013;158(4):280-6. PMID: 23420236. http://dx.doi.org/10.7326/0003-4819-158-4-201302190-00009
11. Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). Ann Intern Med. 2015;162(10):735-6. PMID: 25984857. http://dx.doi.org/10.7326/L15-5093-2
12. Moons KG, de Groot JA, Bouwmeester W, et al. Critical Appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 2014;11(10):e1001744. PMID: 25314315. http://dx.doi.org/10.1371/journal.pmed.1001744

13. Lockwood C, Munn Z, Porritt K. Qualitative research synthesis: methodological guidance for systematic reviewers utilizing meta-aggregation. Int J Evid Based Healthc. 2015;13(3):179-87. PMID: 26262565. http://dx.doi.org/10.1097/XEB.0000000000000062

14. Crowe M, Sheppard L. A review of critical appraisal tools show they lack rigor: Alternative tool structure is proposed. J Clin Epidemiol. 2011;64(1):79-89. PMID: 21130354. http://dx.doi.org/10.1016/j.jclinepi.2010.02.008

15. Cochrane Collaboration. Glossary of Terms in the Cochrane Collaboration. Version 4.2. 5. Updated May. 2005.

16. 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; 2001:87-108.

17. Lohr KN, Carey TS. Assessing "best evidence": issues in grading the quality of studies for systematic reviews. Joint Commission Journal on Quality Improvement. 1999;25(9):470-9. PMID: 10481816.

18. Balshem H, Helfand M, Schunemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011 Apr;64(4):401-6. PMID: 21208779. http://dx.doi.org/10.1016/j.jclinepi.2010.07.015

19. U.S. Preventive Services Task Force. U.S. Preventive Services Task Force Procedure Manual. AHRQ Publication No. 08-05118-EF. Available at http://www.uspreventiveservicestaskforce.org/uspstf08/methods/procmanual.htm; 2008. Accessed on November 1, 2017.

20. Deeks JJ, Dinnes J, D'Amico R, et al. Evaluating non-randomised intervention studies. Health Technology Assessment. 2003;7(27):1-173. PMID: 14499048. http://dx.doi.org/10.3310/hta7270

21. Turner L, Boutron I, Hróbjartsson A, et al. The evolution of assessing bias in Cochrane systematic reviews of interventions: celebrating methodological contributions of the Cochrane Collaboration. Syst Rev. 2013;2(1):79. PMID: 24059942. http://dx.doi.org/10.1186/2046-4053-2-79

22. 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. http://dx.doi.org/10.1016/j.jclinepi.2010.04.027

23. 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. http://dx.doi.org/10.1136/bmj.328.7454.1490

24. Berkman ND, Lohr KN, Ansari MT, et al. Grading the strength of a body of evidence when assessing health care interventions: an EPC update. J Clin Epidemiol. 2015 Nov;68(11):1312-24. PMID: 25721570. http://dx.doi.org/10.1016/j.jclinepi.2014.11.023

25. 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. http://dx.doi.org/10.1016/j.jclinepi.2010.09.012

26. 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. http://dx.doi.org/10.1016/j.jclinepi.2011.01.012

27. 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. http://dx.doi.org/10.1016/j.jclinepi.2011.04.014

28. 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. http://dx.doi.org/10.1016/j.jclinepi.2011.03.017

29. 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. http://dx.doi.org/10.1016/j.jclinepi.2011.01.011

30. 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. http://dx.doi.org/10.1016/j.jclinepi.2010.09.011

31. 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. http://dx.doi.org/10.1016/j.jclinepi.2011.06.004

32. 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. http://dx.doi.org/10.1016/j.jclinepi.2010.07.017

33. 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. http://dx.doi.org/10.1136/bmj.39489.470347.AD

34. 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/; 2011.

35. 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. http://dx.doi.org/10.1016/j.jclinepi.2008.12.004

36. 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.

37. 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. http://dx.doi.org/10.1016/j.jclinepi.2007.11.006

38. 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.

39. 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.

40. Mhaskar R, Djulbegovic B, Magazin A, et al. Published methodological quality of randomized controlled trials does not reflect the actual quality assessed in protocols. J Clin Epidemiol. 2012 Jun;65(6):602-9. PMID: 22424985. http://dx.doi.org/10.1016/j.jclinepi.2011.10.016

41. 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. http://dx.doi.org/0.1136/bmj.c365

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

43. 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. http://dx.doi.org/10.1056/NEJMsa0906126

44. Dickersin K, Chalmers I. Recognizing, investigating and dealing with incomplete and biased reporting of clinical research: from Francis Bacon to the WHO. J R Soc Med 2011;104(12):532-8. PMID: 22179297. http://dx.doi.org/10.1258/jrsm.2011.11k042

45. 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.

46. Wells GA, Shea B, O'Connell D, et al. Newcastle-Ottawa Quality Assessment Scale: Cohort studies. http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm. Accessed on Novemeber 1st 2017.

47. Smith R. Medical journals are an extension of the marketing arm of pharmaceutical companies. PLoS Med. 2005 May;2(5):e138. PMID: 15916457. http://dx.doi.org/10.1371/journal.pmed.0020138

48. 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.

49. 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. http://dx.doi.org/10.1186/1471-2288-8-60

50. 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. http://dx.doi.org/10.1371/journal.pmed.0050191

51. American Medical Writers Association. AMWA ethics FAQs, publication practices of particular concern to medical communicators. 2009. http://www.amwa.org/default.asp?Mode=DirectoryDisplay&DirectoryUseAbsoluteOnSearch=True&id=466. Accessed on November 1, 2017.

52. 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. http://dx.doi.org/10.1001/jama.299.15.1800

53. 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. http://dx.doi.org/10.1001/jama.299.15.1833

54. 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. http://dx.doi.org/10.4065/84.9.811

55. Yank V, Rennie D, Bero LA. Financial ties and concordance between results and conclusions in meta-analyses: retrospective cohort study. BMJ 2007;335(7631):1202-5. PMID: 18024482. http://dx.doi.org/10.1136/bmj.39376.447211.BE

56. 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;62(10):1013-20. PMID: 19230606. http://dx.doi.org/10.1016/j.jclinepi.2008.10.009

57. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med. 2009;151(4):264-9. PMID: 19622511. http://dx.doi.org/10.7326/0003-4819-151-4-200908180-00135

58. 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. http://dx.doi.org/10.1136/bmj.b2700

59. Marshall et al. Automating risk of bias assessment for clinical trials. Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics; 2014 Newport Beach, California. ACM; pp. 88-95.

60. Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. J Am Med Inform Assoc. 2015(Journal Article). http://dx.doi.org/10.1093/jamia/ocv044

61. Sterne JAC, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016 10/12;355:i4919. PMID: 27733354. http://dx.doi.org/10.1136/bmj.i4919

62. Olivo SA, Macedo LG, Gadotti IC, et al. Scales to assess the quality of randomized controlled trials: a systematic review. Phys Ther. 2008;88(2):156-75. PMID: 18073267. http://dx.doi.org/10.2522/ptj.20070147

63. 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. International Journal of Epidemiology. 2007;36(3):666-76. PMID: 17470488. http://dx.doi.org/10.1093/ije/dym018

64. 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. http://dx.doi.org/10.1016/j.jclinepi.2004.04.008

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

66. Zeng X, Zhang Y, Kwong JS, et al. The methodological quality assessment tools for preclinical and clinical studies, systematic review and meta-analysis, and clinical practice guideline: a systematic review. J Evid Based Med 2015;8(1):2-10. PMID: 25594108. http://dx.doi.org/10.1111/jebm.12141

67. Sun X, Briel M, Busse JW, et al. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. BMJ. 2012;344:e1553. PMID: 22422832. http://dx.doi.org/10.1136/bmj.e1553

68. Sun X, Ioannidis JP, Agoritsas T, et al. How to use a subgroup analysis: users' guide to the medical literature. JAMA. 2014;311(4):405-11. PMID: 24449319. http://dx.doi.org/10.1001/jama.2013.285063

69. Whitlock EP, Eder M, Thompson JH, et al. An Approach to Addressing Subpopulation Considerations in Systematic Reviews. . Submitted to Systematic Reviews, 10/2016 2016.

70. Sun X, Briel M, Walter SD, et al. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. BMJ. 2010;340(Journal Article):c117. PMID: 20354011. http://dx.doi.org/10.1136/bmj.c117

71. 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.

72. 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. http://dx.doi.org/10.1016/j.jclinepi.2008.06.007

73. Ioannidis JP, Lau J. Improving safety reporting from randomised trials. Drug Saf. 2002;25(2):77-84. PMID: 11888350. http://dx.doi.org/250202 [pii]

74. 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.

75. Madigan D, Ryan PB, Schuemie M. Does design matter? Systematic evaluation of the impact of analytical choices on effect estimates in observational studies. Ther Adv Drug Saf. 2013;4(2):53-62. PMID: 25083251. http://dx.doi.org/10.1177/2042098613477445

76. Foisy M, Fernandes RM, Tianjing L, et al. Chapter 22 Overviews of Reviews. In: Higgins JPT, Green S, eds. The Cochrane Handbook for Systematic Reviews of Healthcare Interventions. Update 2016, under review.: Cochrane; 2016.

77. Robinson KA, Chou R, Berkman ND, et al. Twelve recommendations for integrating existing systematic reviews into new reviews: EPC guidance. J Clin Epidemiol. 2016;70:38-44. PMID: 26261004. http://dx.doi.org/10.1016/j.jclinepi.2015.05.035

78. Robinson K, Chou R, Berkman N, et al. Integrating Bodies of Evidence: Existing Systematic Reviews and Primary Studies. Methods Guide for Comparative Effectiveness Reviews (Prepared by the Scientific Resource Center under Contract No. 290-2012-00004-C). AHRQ Publication No. 15-EHC007-EF. Rockville, MD: Agency for Healthcare Research and Quality. February 2015. www.effectivehealthcare.ahrq.gov/reports/final.cfm. 2008.

79. Murad MH, Montori VM, Ioannidis JP, et al. How to read a systematic review and meta-analysis and apply the results to patient care: users' guides to the medical literature. JAMA. 2014 Jul;312(2):171-9. PMID: 25005654. http://dx.doi.org/10.1001/jama.2014.5559

80. Alkin MC. Evaluation roots: Tracing theorists' views and influences: Sage; 2004.

81. Kung J, Chiappelli F, Cajulis OO, et al. From systematic reviews to clinical recommendations for evidence-based health care: validation of revised assessment of multiple systematic reviews (R-AMSTAR) for grading of clinical relevance. Open Dent J. 2010;4(1); PMID: 21088686. http://dx.doi.org/10.2174/1874210601004020084

82. Pieper D, Buechter RB, Li L, et al. Systematic review found AMSTAR, but not R (evised)-AMSTAR, to have good measurement properties. J Clin Epidemiol. 2015;68(5):574-83. PMID: 25638457. http://dx.doi.org/10.1016/j.jclinepi.2014.12.009

83. Higgins J, Lane PW, Anagnostelis B, et al. A tool to assess the quality of a meta-analysis. Res Synth Methods. 2013;4(4):351-66. PMID: 26053948. http://dx.doi.org/10.1002/jrsm.1092

84. Donegan S, Williamson P, Gamble C, et al. Indirect comparisons: a review of reporting and methodological quality. PloS one. 2010;5(11):e11054. PMID: 21085712. http://dx.doi.org/10.1371/journal.pone.0011054

85. Shea B. AMSTAR Tool, personal communication.

86. Whiting P, Savović J, Higgins JP, et al. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol. 2016;69(1):225-34. PMID: 26092286. http://dx.doi.org/10.1016/j.jclinepi.2015.06.005

87. Faggion CM, Jr. Critical appraisal of AMSTAR: challenges, limitations, and potential solutions from the perspective of an assessor. BMC Med Res Methodol. 2015;15(Journal Article):63. PMID: 26268372. http://dx.doi.org/10.1186/s12874-015-0062-6

88. Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. JAMA. 2000;283(15):2008-12. PMID: 10789670.

89. 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.

90. 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. http://dx.doi.org/10.1016/j.jclinepi.2010.08.010

91. Higgins JP, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. BMJ. 2003 Sep 6;327(7414):557-60. PMID: 12958120. http://dx.doi.org/10.1136/bmj.327.7414.557

92. Herbison P, Hay-Smith J, Gillespie WJ. Adjustment of meta-analyses on the basis of quality scores should be abandoned. J Clin Epidemiol. 2006;59(12):1249-56. PMID: 17098567. http://dx.doi.org/10.1016/j.jclinepi.2006.03.008