Assessing the Risk of Bias of Individual Studies when Comparing Medical Interventions

Introduction

In this document, we update existing AHRQ guidance for systematic reviews on assessment of risk of bias of individual studies. As with other AHRQ methodological guidance, our intent is to present standards that can be applied consistently across Evidence-based Practice Centers (EPCs) and topics, promote transparency in processes, and account for other steps in the systematic review process.

EPCs, in synthesizing a body of evidence during a systematic review (SR) or comparative effectiveness review (CER), rely heavily on assessment of risk of bias for several steps in the process including interpreting their results and grading the strength of the body of evidence (SOE). Assessment of risk of bias may also guide other decisions in the review process, such as study inclusion (selection criteria for the review overall, and for qualitative and quantitative synthesis) and interpretation of heterogeneous findings.

This guidance document begins by defining terms as appropriate for the EPC program, explores the potential overlap in various constructs used in different steps of the systematic review, and offers recommendations on the inclusion and exclusion of constructs that may apply to multiple steps of the systematic review process. We note that this guidance applies to reviews (such as AHRQ-funded reviews) that separately assess the risk of bias of individual studies, the strength of the body of evidence, and applicability of the findings. This guidance may not hold relevance for reviews that combine evaluations of risk of bias or quality of individual studies with applicability.

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

Key Messages

- The task of assessing the risk of bias of individual studies is part of assessing the strength and applicability of a body of evidence. Reviewers should separate criteria for assessing risk of bias of individual studies from those that assess precision, directness, and applicability.
- EPCs may chose to use the terms "assessment of risk of bias" or "quality assessment". EPCs should define clearly the term used in their SR and CER protocols and describe the constructs included as part of the assessment of the risk of bias.
- We recommend that AHRQ reviews:
    - Do not use study design labels as a proxy for assessment of risk of bias of individual studies.
    - Opt for tools that were: specifically designed for use in systematic reviews; have demonstrated acceptable validity and reliability; specifically address items related to methodological quality (internal validity), and preferably are based on empirical evidence of bias; where available, are specific to the study designs being evaluated; and avoid the presentation of results as a composite score.
    - Explicitly evaluate risk of bias from selection, performance, attrition, detection, and selective outcome reporting.
    - Select items from recommended criteria for each included study design, as appropriate for the topics.
    - Consider validity and reliability of outcome measures and fidelity to the protocol as components of detection bias and performance bias, respectively.
    - Beware of double jeopardy. Generally speaking, exclude precision and applicability when assessing the risk of bias since these are assessed in other domains when evaluating the strength of a body of evidence.
    - Assess risk of bias based on study design and conduct rather than reporting. The EPC should not base risk of bias ratings for individual studies on poor reporting, source of funding, or disclosed conflict of interest, although they should report these issues transparently.
    - Conduct sensitivity analyses, when appropriate, for the body of evidence to evaluate whether source of funding or disclosed conflict of interest is influencing studies' results.
    - Define decision rules for assessing the overall risk of bias score for an individual study.

Terminology and Constructs

Variations in Terminology

In conducting systematic reviews, despite the central role of risk of bias assessment of individual studies, use of the term has varied considerably across review groups. A common alternative to "risk of bias" is "quality assessment," but the meaning of the term *quality* varies, depending on the source of the guidance. GRADE uses the term quality to refer to *an individual study* as well as judgments based about the strength of the *body of evidence* (quality of evidence);[1] USPSTF equates quality with internal validity of individual studies.[2] In contrast, the Cochrane collaboration argues for wider use of the phrase "risk of bias" instead of "quality", reasoning that "an emphasis on risk of bias overcomes ambiguity between the quality of reporting and the quality of the underlying research (although does not overcome the problem of having to rely on reports to assess the underlying research)."[3]

Because of inconsistency and potential misunderstanding in the use of the tem quality, we refer to the extent to which a single study's design and conduct protect against all bias in the estimate of effect using the more precise terminology: "assessment of risk of bias." Thus, assessing the risk of bias of a study can be thought of as assessing the risk that the study results reflect bias in study design or execution rather than the true effect of the intervention or exposure under study. Risk of bias (defined as the risk of "a systematic error or deviation from the truth, in results or inferences")[3] is interchangeable with internal validity (defined as "the extent to which the design and conduct of a study are likely to have prevented bias"[4] or "the extent to which the results of a study are correct for the circumstances being studied.")[5] and may overlap to a great extent with quality, "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."[6]

Guidance on Terminology

EPCs may choose to use any of these terms—risk of bias, quality, or internal validity—in describing critical appraisal of individual studies. We recognize the competing demands for flexibility across reviews to account for specific clinical contexts and consistency within review teams and across EPCs. We advocate transparency of planned methodological approach and documentation of decisions and therefore recommend that EPCs define the term selected in their SR and CER protocols and describe the constructs included as part of the assessment.

Variations in Constructs

An additional source of variation arises from the fact that assessment of quality or risk of bias been used to refer to evaluations of one or more of the following issues: (1) conduct of the study/internal validity, (2) random error, (3) external validity or applicability, (4) completeness of reporting, (5) selective outcome reporting, (6) choice of outcome measures, (7) study design, (8) fidelity of the intervention, and (9) conflict of interest in the conduct of the study.

The variation in underlying constructs stems from two sources. First, no strong empirical evidence supports one approach over another; this gap leads to a proliferation of approaches based on the practices of different academic disciplines and the needs of different clinical topics. Second, in the absence of updated guidance on risk of bias assessment that accounts for how new guidance on related components of systematic reviews (such as selection of evidence,[7]

assessment of applicability,[8] or grading the strength of evidence[9]) relate to, overlap with, or are distinct from risk of bias assessment of individual studies, some review groups continue to use quality practices that have served well in the past.

In the absence of strong empirical evidence, methodological decisions in this guidance document rely on epidemiological principles.[3] Thus, this guidance document presents a conservative path forward. Because absence of evidence is not evidence of absence and systematic reviewers have the responsibility to evaluate potential sources of bias and error if these concerns could plausibly influence study results, we include these concerns even if no empirical evidence exists that they influence study results.

Guidance on Constructs to Include or Exclude from Risk of Bias Assessment

The constructs included in the assessment of risk of bias may differ because of the academic orientation of the reviewers, guidelines by sponsoring organizations, and clinical topic. New guidance and requirements for systematic reviews from AHRQ have reduced the variability in other related steps of the systematic review process and, therefore, allow for greater consistency in risk of bias assessment as well. Some constructs that EPCs may have considered part of risk of bias (or quality) assessment in the past now overlap with or fall within the domains of other systematic review tasks. Table 1 illustrates which constructs to include for each systematic review task when systematic reviews separately assess the risk of bias of individual studies, the strength of the body of evidence (using AHRQ guidance), and applicability of the findings for individual studies.

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

Construct	Included in Appraisal for Individual Studies?	Included in Assessing Applicability for Individual Studies?	Included in Grading Strength of the Body of Evidence?
Risk of bias (study conduct)/internal validity	Yes	No	Yes (required domain of risk of bias)
Precision	Only when no quantitative pooling or presentation is possible	No	Yes (required domain of precision)
Applicability/external validity	Only when components of applicability influence risk of bias (e.g., duration of follow-up varies across intervention arms)	Yes	Yes (component of applicability [surrogacy of outcomes] fall within required domain of directness)
Completeness of Reporting	Yes, as prerequisite to judgment rather than component of risk of bias	No	No
Selective outcome reporting (SOR)	Yes, only when judgments can be made about the impact of differences between outcomes listed full protocol and published materials	Yes	Yes (optional domain of publication bias)
Outcome measures	Yes (validity, reliability, variation across study arms)	Yes (applicability of choice of outcomes)	Yes (directness of measures under required domain of directness)
Study design	Assessment should account for varied sources of bias by design rather than rate individual studies for study design per se	No	Yes (required domain of risk of bias)
Fidelity to protocol	Yes	No	No
Conflict of interest	No	No	Yes (optional domain of publication bias)

Types of Risks of Bias included in Assessment of Risk of Bias

Although numerous classification schemes exist for classifying and defining biases,[10] we elect to use the taxonomy suggested by Higgins et al. in the Cochrane Handbook as a common, comprehensive, and well-disseminated approach (Table 2).[3] Subsequent sections of this guidance refer to this taxonomy of biases.

A brief review of three sources (*Cochrane Handbook of Systematic Reviews,*[3] *Systems to Rate the Strength of Scientific Evidence,*[11] *Evaluation of Non-randomized Studies*[12] show empirical evidence for detection bias, attrition bias, and reporting bias.

Table 2. Taxonomy of core biases in the Cochrane Handbook[3]

Types of Bias Related to Conduct of the Study (Including Analysis and Reporting)	Definition	Risk of Bias Assessment Criteria
Selection Bias	Systematic differences that arise from self-selection of treatments, physician-directed selection of treatments, or association of treatment assignments with demographic, clinical, or social characteristics. Includes confounding by indication (when patient prognostic characteristics, such as disease severity or co-morbidity, influence both treatment source and outcomes.)	Randomization, allocation concealment, sequence generation, control for confounders in cohort studies, and case matching in case-control studies
Performance Bias	Systematic differences in the care provided to participants and protocol deviation. Examples include: contamination of the control group with the exposure or intervention, unbalanced provision of additional interventions or co-interventions, difference in co-interventions, and inadequate blinding of providers and participants	Fidelity to protocol, unintended interventions or co-interventions
Attrition Bias	Systematic differences in the loss of participants from the study and how they were accounted for in the results, e.g., incomplete follow-up, differential attrition. Those who drop out of the study or who are lost to follow-up may be systematically different from those who remain in the study. Attrition bias can potentially change the collective (group) characteristics of the relevant groups and their observed outcomes in ways that affect study results by confounding and spurious associations.	Incomplete outcome data, intention-to-treat analysis, and completeness of follow-up
Detection Bias	Systematic differences in outcomes assessment among groups being compared, including systematic misclassification of the exposure or intervention, covariates, or outcomes because of variable definitions and timings, diagnostic thresholds, recall from memory, inadequate assessor blinding, and faulty measurement techniques. Erroneous statistical analysis might also affect the validity of effect estimates.	Blinding of outcome assessors, especially with subjective outcome assessments, bias in inferential statistics, valid and reliable measures
Reporting Bias	Systematic differences between reported and unreported findings, e.g., differential reporting of outcomes or harms, incomplete reporting of study findings, potential for bias in reporting through source of funding	Selective outcome reporting evaluation by comparing study report and (a) protocol or (b) outcomes prespecified in methods

Risk of Bias and Precision

One key distinction between risk of bias and quality assessment is in the treatment of precision. Quality assessment—the evaluation of systematic bias, nonsystematic bias, and inferential error[6]—subsumes nonsystematic bias or random error. The impact of random error on the precision of estimates can be reduced by increasing sample size.[13] In keeping with the inclusion of random error in this definition of quality, quality assessment tools have included sample size evaluation as an explicit component in the past.

Both GRADE[14] and recent AHRQ guidance on evaluating the strength of evidence[9] separate the evaluation of precision from that of risk of bias. Systematic reviews now routinely evaluate precision (through consideration of the confidence intervals around a summary effect size from pooled estimates) when grading the strength of the body of evidence.[9] Under such circumstances,

the evaluation of precision in assessing the quality of individual studies as well as the body of evidence would constitute "double jeopardy." We recommend that AHRQ reviews exclude precision when assessing the risk of bias for outcomes that can be pooled in meta-analysis or presented quantitatively (for single studies). When outcomes cannot be pooled (as with highly heterogeneous bodies of evidence) or presented quantitatively, assessing precision in addition to (but separately from) risk of bias in appraising individual studies may be appropriate.

Risk of Bias and Applicability

Many commonly used quality assessment tools measure external validity. A review of tools to rate observational studies identified 14 "best" tools. Each evaluated both core elements of internal validity and also included questions on representativeness of the sample.[12] New guidance for the EPC program on how to address applicability (sometimes known as external validity, generalizability, or relevance) recommends that EPCs provide a summary report of the applicability of the body of evidence separately from their judgment of the applicability of individual studies.[8] This guidance also notes that although individual studies may not be representative of the population of interest, consistent findings across studies with individually limited generalizability may suggest broad applicability of the results.

We recommend that AHRQ reviews exclude overall applicability in risk of bias assessments of individual studies. We note, however, that some components of applicability, such as duration of follow-up or population source, may also be relevant for evaluating risk of bias; EPCs may, therefore, elect to include them in assessment of risk of bias individual studies. For instance, when duration of follow-up differs between intervention arms, this difference results in a heightened risk of performance bias and may also affect the applicability of findings. However, when duration of follow-up is inadequate to establish the clinical relevance of the outcome, systematic reviewers may infer poor applicability (rather than high risk of bias).

Risk of Bias and Completeness of Reporting

In theory, internal validity focuses on design and conduct of a study. In practice, assessing the internal validity of a study requires adequate reporting of the study, unless additional information is obtained via some "gray literature" effort. Although new standards on reporting seek to improve reporting of study design and conduct,[15-19] EPC review teams continue to need a practical approach to dealing with poor or inadequate reporting. The Cochrane risk of bias tool judges the risk of bias to be uncertain when information is inadequate. EPC reviews have varied in their treatment of reporting of study design and conduct; for example, some have elected to rate poorly *reported* studies as studies with high risk of bias In general, we recommend that assessment of risk of bias focus primarily on the design and conduct of studies and not on the quality of reporting. Nevertheless, we also recognise the importance of evaluating reporting in the context of the clinical topic and the study. For that reason, we recommend that EPCs set up clearly stated and consistent standards within their own reviews to deal with the issue of poor reporting. We provide further guidance on how to address incomplete reporting in a later section.

Risk of Bias and Selective Outcome Reporting

Selective outcome reporting is a special subset of inadequate reporting; it has major implications for both the quality of individual studies and the strength of the body of evidence. Guyatt et al. note that selective outcome reporting, that is, the "incomplete or absent reporting of some outcomes and not others based on results,"[20] may be intuitively regarded by some as

belonging with publication bias (or bias resulting from selective reporting of positive results). Publication bias is a component of the evaluation of the strength of the body of evidence rather than of individual study quality.[9] Guyatt et al. (p. 409) note that "selective reporting is present if authors acknowledge pre-specified outcomes that they fail to report or report outcomes incompletely such that they cannot be included in a meta-analysis. One should suspect reporting bias if the study report fails to include results for a key outcome that one would expect to see in such a study or if composite outcomes are presented without the individual component outcomes."[20] Without access to the full protocol, judgments of selective outcome reporting at the individual study level may be difficult to justify; a consideration of these issues at the level of the body of evidence, when evaluating publication bias, may be more appropriate.

An additional consideration is how to evaluate selective outcome reporting in the context of studies with multiple outcomes, some of which may be reported selectively, and others reported completely. This scenario may be addressed either in evaluating risk of bias or in evaluating the strength of evidence. For the former, EPCs may assume a higher risk of bias for all reported outcomes in the presence of clear evidence of selective outcome reporting for another outcome. Alternatively, with no evidence of selective outcome reporting for an individual outcome, EPCs may still judge that selective outcome reporting exists for the body of evidence for that outcome alone.

Risk of Bias and Outcome Measures

The use of valid and reliable outcome measures reduces the likelihood of detection bias. In addition, variation in outcome measures by study arm constitutes a source of measurement bias and should, therefore, be included in assessment of risk of bias. We recommend that assessment risk of bias of individual studies include the evaluation of the validity and reliability of outcome measures, and their variation across study arms. Recent guidance on the evaluation of applicability by Atkins and colleagues states the importance of considering the relevance of outcome measures for judging applicability (or external validity) of the evidence.[21] The choice of specific outcome measures is a consideration for applicability and for strength of evidence. For example, studies relying on self-report measures may be rated as having a higher risk of bias than studies with clinically observed outcomes. Studies that focus on short-term outcomes and fail to report long-term outcomes may be judged as having poor applicability or not being directly relevant to the clinical question.

Risk of Bias and Study Design

Some designs possess inherent features (such as randomization and control arms) that reduce the risk of bias and increase the validity of causal inference. Each study design has specific risks of bias that may differ depending on the clinical question.

EPCs consider these design-specific sources of bias at two points in the systematic review process: (1) when evaluating whether to admit classes of evidence into the review and (2) when evaluating individual studies for design-specific risks of bias. Norris et al. note that the default strategy in systematic reviews should be to *consider* including observational studies and the decision rests on the answer to two questions: (1) are there gaps in the trial evidence for the review questions under consideration? and (2) will observational studies provide valid and useful information to address key questions?[7] In considering whether or not observational studies provide valid and useful information, EPCs will need to consider the likelihood that observational studies will generally have more numerous and more serious sources of bias than

trials. Once an EPC makes the decision to include observational studies, then the review team needs to evaluate each study based on the risks of bias specific to that design.

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

Because systematic reviews evaluate design-specific sources of bias in selecting studies for inclusion in the review and then use study design as a component of risk of bias in judging the strength of evidence, we recommend that EPCs do not use study design labels as a proxy for assessment of risk of bias of individual studies. In other words, EPCs should not downgrade the risk of bias of *individual* studies on the basis solely of study design because doing so would penalize studies again (i.e., at the level of individual studies and the body of evidence). This approach accounts for the fact that a study can be performed with the highest quality *for that study design* but still have some (if not serious) potential risk of bias.[3] This approach also acknowledges that quality varies, perhaps widely, within designs and that study designs do have inherent limitations.

Depending upon the clinical question, the sources of bias from a particular study design may be so large as to constitute a high risk of bias. For instance, EPCs may judge information on benefits from case series of interventions as having a high risk of bias. In such instances, we recommend that EPCs exclude such designs from the review.

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

Risk of Bias and Fidelity to the Protocol

Failure of the intervention to maintain fidelity to the protocol can influence performance bias; it is, therefore, a component of assessment of risk of bias. We note, however, that the interpretation of fidelity may differ by clinical topic. For instance, some behavioral interventions include "fluid" interventions; these involve interventions for which the protocol explicitly allows for modification based on patient needs; such fluidity does not mean the interventions are implemented incorrectly. When interventions implement protocols that have minimal concordance what can be adopted in practice, the discrepancy may be considered an issue of applicability, but would not be evaluated under fidelity of the implemented intervention to the protocol. We recommend that EPCs account for the needs of the topic in determining and applying criteria about fidelity for assessment of risk of bias. Our recommendation is consistent with the Institute of Medicine guidelines on systematic reviews.[22]

Risk of Bias and Conflict of Interest

Many studies examining the issue of financial conflict of interest have found that sponsor participation in data collection, analysis, and interpretation of findings can threaten the internal

validity of the primary studies and systematic reviews.[23,24] The pathways by which sponsor participation can influence the validity of the results are manifold. They include:

1. selection of designs and hypotheses – for example, choosing noninferiority rather than superiority approaches,[25] picking comparison drugs and doses,[25]choosing outcomes[24], or using composite endpoints(e.g., , mortality and quality of life) without presenting data on individual endpoints; [26]
2. selective outcome reporting—for example, reporting relative risk reduction rather than absolute risk reduction or "cherry-picking" from multiple endpoints;[25]
3. differences in quality (meaning, internal validity) of studies and adequacy of reporting;[27]
4. biased presentation of results;[26] and
5. publication bias.[28]

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

Editors have grown increasingly concerned about the practice of ghost authoring (i.e., primary authors or substantial contributors are not identified) or guest authoring (i.e., one or more identified authors are not substantial contributors)[29] sponsored studies, a practice that makes the actual contribution of the sponsor very difficult to discern.[30,31]

All these concerns may lead one to conclude that sponsorship from industry (i.e., for-profit entities) should be included as an explicit consideration for assessment of risk of bias. We concur that sponsorship of studies should be considered in critically appraising the evidence but caution against equating industry sponsorship with high risk of bias or poor quality for three reasons.

- First and foremost, sponsor bias is not limited to industry; nonprofit and government-sponsored studies may also have instances of guest or ghost authoring; moreover, the researchers may have various financial or intellectual conflicts of interest by virtue of, for example, accepting speaking fees from many different sources.[32]
- Second, financial conflict is not the only source of conflict of interest: other types of conflict of interest may include personal, professional, or religious beliefs, desire for academic recognition, and so on.[23]
- Third, the multiple pathways by which sponsorship may influence studies are not all solely within the domain of assessment of risk of bias.
- Several of these pathways fall under the purview of other systematic review tasks. For instance, concerns about the choice of designs, hypotheses, and outcomes relate as much or more to applicability than other aspects of reviews. Selective outcome reporting may not always be possible to judge at the individual study level, as noted earlier, and it may be more easily judged for the body of evidence.

The biased presentation or "spin" on results, if limited to the discussion and conclusion section of studies, should have no bearing on judgments of internal validity because systematic reviews do not rely on interpretation of data by study authors. Publication bias lies within the purview of grading the strength of the body of evidence.

Internal validity and completeness of reporting constitute, then, the primary pathway by which sponsors may influence the validity of study results that is entirely within the domain of

assessment of risk of bias. We acknowledge that this pathway may not be the most important source of sponsor influence: as standards for conduct and reporting of studies become widespread and journals require that they be met, differences in internal validity and reporting between industry-funded studies and other studies will likely attenuate. Appraisal of studies for other pathways of sponsor influence may constitute a "double" or "triple jeopardy" if the same considerations are being taken into account during appraisal of strength of evidence and applicability. In balancing these considerations with the primary responsibility of the systematic reviewer, that of objective and transparent synthesis and reporting of the evidence, we make three recommendations: (1) at a minimum, EPCs should routinely report the source of each study's funding; (2) EPCs should consider issues of selective outcome reporting at the individual study level; and (3) EPCs should conduct sensitivity analyses for the body of evidence when they have reason to suspect that the source of funding, or disclosed conflict of interest is influencing studies' results.[25]

Stages in Assessing the Risk of Bias of Studies

International reporting standards require documentation of various stages in a comparative effectiveness review.[33-37] We lay out recommended approaches to assessment of risk of bias in five steps: protocol development, pilot testing and training, assessment of risk of bias, interpretation, and reporting. Table 3 describes the stages and specific steps in assessing the quality of individual studies that contribute to transparency through careful documentation of decisions.

Protocols for assessment of risk of bias build on the protocol for the entire review. As prerequisites to developing the protocol for assessment of risk of bias, EPCs must identify in the overall protocol the important intermediate and final outcomes that need assessment of risk of bias and other study descriptors or study data elements that are required for the assessment of risk of bias. Protocols must justify what quality criteria will be evaluated and how the reviewers will incorporate quality of individual studies in the synthesis of evidence.[38-40]

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

Assessment of risk of bias should be consistent with the analysis plans in registered protocols of the reviews.[41,42] Published reports must include quality criteria and should describe the selected tools and their reliability and validity when such information available EPC reviews should report all criteria used for each outcome and study evaluated. The synthesis of the evidence should reflect the *a priori* analytic plan for incorporating quality of individual studies in qualitative or quantitative analyses. EPCs should report the results of all preplanned analyses that included quality criteria regardless of statistical significance or the direction of the effect. Published reviews should also include justifications of all *post hoc* decisions to synthesize evidence by methodological or reporting quality of studies.

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

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

Design-Specific Recommended Criteria to Assess Risk of Bias

We present design-specific recommended criteria to assess risk of bias for four common study designs: RCTs, cohort (prospective, retrospective, and non-concurrent), case-control (including nested case-control), and case series (Table 4).[43] Reviewers may select specific criteria relevant to the topic. For instance, blinding of outcome assessors may not be possible for surgical interventions. Other criteria may need to be modified for the specific review. For instance, reviewers of topics that focus on short-term clinical outcomes may select a low expected attrition rate. We also note that with attrition rate in particular, no empirical standard exists across all topics for demarcating a high risk of bias from a lower risk of bias; these standards are often set within clinical topics. The list of recommended criteria do not represent comprehensive sources of bias for other study designs, For instance, time series studies may require a question asking whether the study accounted for regression to the mean.

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

Risk of Bias	Criterion	RCTs	Cohort	Case-control	Case series	Cross-sectional
Selection bias	Was treatment adequately randomized (e.g., random number table, computer-generated randomization)?	x				
	Was the allocation of treatment adequately concealed (e.g.,. pharmacy-controlled randomization or use of sequentially numbered sealed envelopes)?	x				
	Any attempt to balance the allocation between the groups?		x			
	Did the study apply inclusion/exclusion criteria uniformly to all comparison groups?		x	x		
	Is the selection of the comparison group appropriate?		x	x		
	Did the strategy for recruiting participants into the study differ across study groups?	x	x			
	Are baseline characteristics similar between groups? If not, did the analysis control for differences?	x	x			
	Does the design or analysis control account for important confounding and modifying variables?		x	x	x	x
Performance bias	Did researchers rule out any impact from a concurrent intervention or an unintended exposure that might bias results?	x	x	x	x	x
	Did variation from the study protocol compromise the conclusions of the study?	x	x	x	x	
Attrition bias	In cohort studies, is the length of follow-up different between the groups, or in case-control studies, is the time period between the intervention/exposure and outcome the same for cases and controls?		x	x		
	Was there a high rate of differential or overall attrition?	x	x	x		
	Did attrition result in a difference in group characteristics between baseline (or randomization) and follow-up?	x	x	x	x	x
	Is the analysis conducted on an intention-to-treat (ITT) basis?	x	x			
Detection bias	Were the outcome assessors blinded to the intervention or exposure status of participants?	x	x	x	x	x
	Are the inclusion/exclusion criteria measured using valid and reliable measures, implemented consistently across all study participants?	x	x	x	x	x
	Are interventions/exposures assessed using valid and reliable measures, implemented consistently across all study participants?	x	x	x	x	x
	Are primary outcomes assessed using valid and reliable measures, implemented consistently across all study participants?	x	x	x	x	x
	Are confounding variables assessed using valid and reliable measures, implemented consistently across all study participants?		x	x	x	x
Reporting bias	Are the potential outcomes pre-specified by the researchers? Are all pre-specified outcomes reported?	x	x	x	x	x

Tools for Assessing Quality

EPCs can use one of two general approaches to assessing study quality in systematic reviews. One method is often referred to as a *components approach*. This involves assessing individual items that are deemed by the systematic reviewers to reflect the methodological quality, or other relevant considerations, in the body of literature under study. For example, one commonly assessed component in RCTs is allocation concealment.[35] Reviewers assess whether the randomization sequence was concealed from key personnel and participants involved in a study before randomization; they then rate the component as adequate, inadequate, or unclear.

The second common approach is to use a tool or *composite approach* that combines different components related to methodological quality, risk of bias, or reporting. A plethora of tools has emerged over the past 20 years to assess quality. 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. Since many AHRQ systematic reviews typically address multiple research questions, they may require the use of several quality assessment or risk of bias tools or the selection of various different components to address all the study designs included.

Currently there is no consensus on the best approach or preferred tool for assessing quality, as the components associated with methodological quality or risk of bias are in contention. As such, there are a large number of tools available, and their marked variations and relative merits can be problematic for systematic reviewers. We advocate the following general principles when selecting a tool, or approach, to assessing quality in systematic reviews. EPCs should opt for tools that:

- were specifically designed for use in systematic reviews;
- have demonstrated acceptable validity and reliability;
- specifically address items related to methodological quality (internal validity), and preferably are based on empirical evidence of bias;
- where available, are specific to the study designs being evaluated; and
- avoid the presentation of results as a composite score (an overall numeric rating of study quality across items, for example 11 from 15 items).

Although, there is much overlap across different tools, there is no single universal tool that addresses all the varied contexts for assessment of risk of bias. Appendix A details a select list of tools that have been shown to be reliable or valid, are widely used, or have been recommended for use in systematic reviews that compared quality assessment instruments.[11,12,44-46] We do not discuss tools that have been developed to guide and assess the reporting of studies. These reporting guidelines assess different constructs than what is commonly understood as methodological quality or risk of bias (internal validity). These reporting guidelines/ checklists assist in adequately assessing study methods and are widely endorsed by journal editors. A list of reporting guidelines for different study designs is available through the EQUATOR network at www.equator-network.org.

Summarizing the Risk of Bias or Quality of a Study

For outcomes that undergoing assessment of strength of evidence, EPC reviewers must consider all of the items together after completing evaluations of the assessment of risk of bias items for a given study (article or articles) and then place the study into a summary category. This will be

one of three ordinal categories: or low, medium or high for risk of bias rating or good, fair, or poor for quality assessment.[9] This section describes methods for achieving that categorization and discusses guidelines for reporting this information. A study's risk of bias or quality category can be different for different outcomes, which means that EPCs should record the different outcome-specific categories as necessary. This situation can arise from, for instance, variation in the completeness of data, differential blinding of outcome assessors, or other outcome-specific quality items.

Categories for Outcome-Specific Risk of Bias or Quality

An overall rating of good, fair or poor quality (or the equivalent low, medium and high risk of study bias) should be made for the most clinically important outcomes as defined in the review protocol. As is true for scoring individual criteria or items, EPCs should do this overall rating within study design; for instance, a well-conducted observational study could be assigned a rating of good quality (or low risk of bias (good quality) as could a well-conducted RCT. As with the earlier steps, EPCs should adopt a dual reviewer approach to this step as well. Finally, given that these assessments involve subjective considerations, reviewers must clearly describe their rationale for all ratings.

A study categorized as "low" risk of bias or good quality implies confidence on the part of the reviewer that results represent the true treatment effects (study results are considered valid). The study reporting is adequate to judge that no major or minor sources of bias are likely to influence results. A study rated as "medium" risk of bias implies some confidence that the results represent true treatment effect. The study is susceptible to some bias the problems are not sufficient to invalidate the results (i.e., no flaw is likely to cause major bias.[47] The study may be missing information, making it difficult to assess limitations and potential problems. A study categorized as "high" risk of bias implies low confidence that results represent true treatment effect. The study has significant flaws that imply biases of various types that may invalidate its results; these may arise from serious errors in conduct, analysis, or reporting, large amounts of missing information, or discrepancies in reporting.

Methods and Considerations for Summarizing Risk of Bias or Quality

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

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

One way to assign a category is to make a simple "holistic" judgment, that is, a judgment based on an overall perception of risk of bias rather than an evaluation of all components of bias. Unfortunately, this approach is not transparent, and it is likely not to be reproducible. The main

problem is inconsistent bases for judgment: if the studies were re-examined, the same reviewer might alter the category assignments. Reviewers may also be influenced (consciously or unconsciously) by other unstated aspects of the studies, such as the prestige of the journal or the identity of the authors. EPCs can and should explain how their reviewers made these judgments, but such the fact remains that these approaches can suffer from substantial subjectivity.

Instead, we recommend that, in aiming for transparency and reproducibility, EPC reviewers use a set of specific rules for assigning a quality category. These rules can take the form of declarative statements, such as "randomized and blinded studies are good; randomized but unblinded studies are fair; inadequately randomized and unblinded studies are poor." EPCs could also lay out more complicated rules, ones that reflect the items in the chosen instrument, but the key is transparency. Obviously, many other quality items could be incorporated into these rules, but the key is transparency. Notice that such declarative statements implicitly assign weights to the different items. Previous research has demonstrated that quantitative synthesis of evidence with quality criteria is the optimal approach CER.[48,49] In any case, the authors must justify how synthesis of evidence incorporated risk of bias criteria or overall rank of risk of bias.

Within rule-based assignment, one option is to use the domains of risk of bias and then the items within those domains as a basis for the rules. For example: studies that met the majority of the items for all domains are good; studies that met the majority of the items for more than half (but not all) of the domains are fair; all other studies are poor. This process relies on an accurate assignment of items into domains. The basic requirement is adequate explanation of the method used.

The use of a quantitative scale is another way to employ a transparent set of rules. For a scale, the weights of different items are explicit rather than implicit. But any weighting system, whether qualitative or quantitative, must be recognized as subjective and arbitrary, and different reviewers may choose to use different weighting methods. Using transparent rules does not remove the subjectivity inherent in assigning the risk of bias category. Subjectivity remains in the choice of different rules, or rules that assigning items to quality domains, and if the latter, what proportion of items must be met to earn a given rating. Consequently, reviewers should avoid attributing unwarranted precision (such as a quality score of 3.42) to quality ratings or creating subcategories or ambiguous language such as "in the middle of the fair range".

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

Conclusion

Assessment of risk of bias is a key step in conducting systematic reviews that informs many other steps and decisions made within the review. It also plays an important role in the final assessment of the strength of the evidence. The centrality of assessment of risk of bias to the entire systematic review task requires that assessment processes be based on sound empirical evidence or theoretical principles. In carrying out assessment of risk of bias, EPCs should specify consistent parameters across different content areas, use at least two independent reviewers with

a defined process for consensus and standards for transparency, and clearly document and justify all processes, decisions, and results.

References

1. Balshem H, Helfand M, Schunemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. Journal of Clinical Epidemiology. In Press, Corrected Proof.
2. 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.
3. Higgins JPT, Green S, eds. Cochrane handbook for systematic reviews of interventions. Version 5.0.2. [updated September 2009]. The Cochrane Collaboration. Available from www.cochrane-handbook.org 2009.
4. Cochrane Collaboration Glossary Version 4.2.5. 2005 [cited; Available from: http://www.cochrane.org/sites/default/files/uploads/glossary.pdf.
5. 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.
6. Lohr KN. Rating the strength of scientific evidence: relevance for quality improvement programs. Int J Qual Health Care. 2004;16(1):9-18.
7. Norris S AD, Bruening W, et al. Selecting observational studies for comparing medical interventions. In: Agency for Healthcare Research and Quality. Methods Guide for Comparative Effectiveness Reviews [posted June 2010]. Rockville, MD. Available at: http://www.effectivehealthcare.ahrq.gov/ehc/products/196/454/MethodsGuideNorris_06042010.pdf. 2010.
8. Atkins D, Best D, Briss PA, et al. Grading quality of evidence and strength of recommendations. BMJ. 2004 Jun 19;328(7454):1490.
9. Owens DK, Lohr KN, Atkins D, et al. AHRQ series paper 5: grading the strength of a body of evidence when comparing medical interventions--Agency for Healthcare Research and Quality and the effective health-care program. J Clin Epidemiol. 2010;63(5):513-23.
10. Delgado-Rodriguez M, Llorca J. Bias. J Epidemiol Community Health. 2004 Aug;58(8):635-41.
11. 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.
12. Deeks JJ, Dinnes J, D'Amico R, et al. Evaluating non-randomised intervention studies. Health Technol Assess. 2003;7(27):iii-x, 1-173.
13. Cook TD, Campbell DT. Quaisi-experimentation: design and analysis issues for field settings. Boston: Houghton Mifflin Company; 1979.
14. 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.
15. 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.

16. 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.

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

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

19. 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.

20. Guyatt GH, Oxman AD, Vist G, et al. GRADE guidelines: 4. Rating the quality of evidencedstudy limitations (risk of bias) and publication bias. J Clin Epidemiol. In Press.

21. 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.

22. Institute of Medicine. Finding what works in health care: standards for systematic reviews. [cited June 2, 2011]; Available from: http://www.nap.edu/openbook.php?record_id=13059&page=R1.

23. 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.

24. Newcastle-Ottawa Quality Assessment Scale: Cohort studies. Available from: http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm.

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

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

27. 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.

28. 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.

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

30. 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.

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

32. 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.

33. Simera I, Moher D, Hoey J, et al. A catalogue of reporting guidelines for health research. Eur J Clin Invest. 2010 Jan;40(1):35-53.

34. Simera I, Moher D, Hirst A, et al. Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network. BMC Med. 2010;8:24.

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

36. Vandenbroucke JP. STREGA, STROBE, STARD, SQUIRE, MOOSE, PRISMA, GNOSIS, TREND, ORION, COREQ, QUOROM, REMARK... and CONSORT: for whom does the guideline toll? J Clin Epidemiol. 2009 Jun;62(6):594-6.

37. Simera I, Moher D, Hoey J, et al. The EQUATOR Network and reporting guidelines: Helping to achieve high standards in reporting health research studies. Maturitas. 2009 May 20;63(1):4-6.

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

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

40. 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.

41. Straus S, Moher D. Registering systematic reviews. CMAJ. 2010 Jan 12;182(1):13-4.

42. Higgins J, Green S. The Cochrane Collaboration. The Cochrane Handbook for Systematic Reviews of Interventions. 2005 [cited 2006]; Available from: http://www.cochrane.org/resources/handbook/handbook.pdf.

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

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

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

46. 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.

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

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

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

Appendix A. Tools to Assess Risk of Bias or Quality of Individual Outcomes

This appendix provides a brief overview of tools to evaluate randomized controlled trials (RCTs), nonrandomized studies, medical tests, and harms. For most tools, the preliminary step in assessing whether or not a chosen tool is applicable to the specific study is to categorize the study design. We recommend the use of tools such as that developed by Hartling et al. to categorize study designs.[1]

Randomized Controlled Trials

A large number of tools have been developed to assess the methodological quality or risk of bias in RCTs. In 2008, Armijo Olivo et al.[2] published a systematic review identifying scales designed to assess the quality of RCTs. They identified 21 scales but found that the majority were not "rigorously developed or tested for validity and reliability."

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

Armijo Olive et al. highlighted two other tools that were developed using rigorous methods and tested for validity and reliability. Verhagen et al. developed the Delphi List to assess RCTs in general (i.e., not specific to a specific clinical area or field of study). It has demonstrated good face, content, and concurrent validity and has been tested for reliability. It includes the following nine items: inclusion/exclusion criteria of study population defined; randomization; allocation concealment; baseline comparability of study groups; blinding of investigator, subjects, and care providers; reporting of point estimates and variability for primary outcomes; and, intention-to-treat analysis.[3]

Yates et al. developed a tool to assess the quality of RCTs of cognitive behavioral therapy for chronic pain. The tool has two parts, one on treatment quality (five items) and the second on quality of study design and methods (eight items with multiple parts). The latter part of the tool includes questions on the following domains: reporting of inclusion/exclusion criteria; reporting of attrition; adequate description of the sample; steps to minimize bias (i.e., randomization, allocation, measurement, treatment expectations); outcomes justified, valid, and reliable; length of follow-up (i.e., sustainability of treatment effects); adequacy of statistical analyses; comparability or adequacy of control group. It has shown face, content, and construct validity and good inter-rater reliability.[4] The tool has not been widely used.

In 2005, The Cochrane Collaboration convened a group to address several concerns in the assessment of trial quality. One concern was the growing number of tools being used and

inconsistent approaches to quality assessment across different systematic reviews. Participants also recognized that many of the tools being used were not based on empirical evidence showing that the items they included were related to biased results. Moreover, many tools combined elements examining methodological conduct with items related to reporting.

From this work a new tool for randomized trials emerged—the Risk of Bias tool.[6] This tool was released after publication of the review by Armijo Olivo et al. described above. The Risk of Bias tool includes six domains for which empirical evidence demonstrates associations with biased estimates of effect. The domains are sequence generation; allocation concealment; blinding; missing outcome data; selective outcome reporting; and other sources of bias. The Risk of Bias tool is now the recommended method for assessing methodological quality, or risk of bias, of RCTs in systematic reviews conducted through The Cochrane Collaboration. The tool has not undergone extensive validity or reliability testing. A working group in The Cochrane Collaboration is currently re-examining the tool based on initial user feedback and making modifications to it.

Nonrandomized Studies

Several systematic reviews have been conducted to identify, assess, and make recommendations regarding quality assessment tools for use in nonrandomized studies (including nonrandomized experimental studies and observational studies). West et al. identified 12 tools for use in observational studies and recommended 6 of these for use in systematic reviews. Deeks et al. identified 14 "best tools" from among 182 and recommended 6 for use in reviews. Of interest is that the two reports identified only three tools in common: Downs and Black,[5] Reisch,[6] and Zaza.[7] These three tools are applicable to a range of study designs; only two were developed for use in systematic reviews.[5]

One recent and comprehensive systematic reviews of quality assessment tools for observational studies identified 86 tools.[8] The tools varied in their development and their purpose: only 15 percent were developed specifically for use in systematic reviews; 36 percent were developed for general critical appraisal and 34 percent were developed for "single use in a specific context." The authors chose not to make recommendations regarding which specific tools to use; however, they broadly advised that reviewers select tools that

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

The Cochrane Collaboration provides recommendations on use of tools for nonrandomized studies. They acknowledge the abundance of tools available but, like Sanderson et al., make no recommendation regarding a single instrument.[8] They recommend following the domains in the Risk of Bias tool, particularly for prospective studies. A working group within the Cochrane Collaboration is currently modifying the Risk of Bias tool for use in nonrandomized studies.

The Cochrane Handbook highlights two other tools for use in nonrandomized studies: the Downs and Black [5] and Newcastle Ottawa Scale.[9] They implicitly recommend the Newcastle Ottawa Scale over the Downs and Black because the Downs and Black is time-consuming to apply, requires considerable epidemiology expertise, and has been found difficult to apply to case-control studies.[9]

The Newcastle Ottawa Scale is the most frequently used in systematic reviews for articles about studies with this type of design. It contains separate questions for cohort and case-control studies. It was developed based on threats to validity in nonrandomized studies; these specifically include selection of participants (generalizability or applicability), comparability of study groups, methods for outcome assessment (cohort studies) or ascertainment of exposure (case-control studies), and adequacy of follow-up. The developers have reported face and content validity or this instrument, and they revined it based on experience using the tool in systematic reviews.[9] It has also been tested for inter-rater reliability. Examination of its criterion validity and intra-rater reliability is under way and plans are being developed to examine its construct validity.

Other recently developed checklists address the quality of observational, nontherapeutic studies of incidence of diseases or risk factors for chronic diseases[10] or observational studies of interventions or exposures.[11] The checklists have been developed based on a comprehensive literature review,[12] are based on predefined flaws in internal validity, and discriminate reporting from conduct of the studies. These tools are continuing inter-rater reliability tests.

Medical Tests

The majority of medical test studies aim to evaluate the accuracy of diagnosing or screening a health condition and typically compare results with those from another medical test. However, medical test studies can also evaluate the value of using tests to monitor disease status, assist in choice of therapy, and for planning subsequent interventions. A detailed explanation of quality assessment for studies evaluating diagnostic tests is shown in Chapter 4 of the AHRQ Medical Test guide. We provide a synopsis of key aspects here. Genetic and prognostic marker tests, has some unique attributes. Genetic tests can be used for diagnostic purposes as well as: (1) assess risk or susceptibility in asymptomatic individuals (identify individuals at risk for future health conditions; (2) identify prognostic factors to potentially guide future treatment, and (3) predict the level of response to treatments or lifestyle and environmental exposures (including diet, infectious agents, chemicals, physical agents, and behavioral factors). A prognostic test or indicator (predictor) is one in which the factor of interest (or a combination of factors) is measured and evaluated to make probabilistic predictions about the likelihood of an event's occurrence. The value of the results of a prognostic test or indicators comes from its accuracy in estimating probabilities of specific outcomes; clinically, this implies that patients classified with higher probabilities may receive a different course of treatment relative to those with lower probabilities. Quality assessment specific to studies evaluating genetic tests used for diagnostic purposes or for prognostic tests are not addressed in this document. We refer the reader to the AHRQ Medical Test Guidance).

As noted previously, variety of study designs types can be used to evaluate the accuracy, predictive ability, or other properties of medical tests. The design types can include RCTs, case series, assay reliability studies, and laboratory studies (assessing analytic validity). In contrast to other quality assessment instruments that are generally based on the study design type, those used to assess diagnostic accuracy of medical tests tend to focus predominantly on the potential for risk of bias related to the test itself rather than aspects of study design type.[13] Two systematic reviews have evaluated quality assessment instruments specifically in the context of diagnostic accuracy. West et al.[13] evaluated 18 tools (6 scales, 9 guidances, and 3 EPC rating systems). They note that all of the tools are to be used in conjunction with other tools relevant for judging the design specific attributes of the study (for example quality of RCTs or observational studies).

Three scales met all 6 criteria considered to be important and these included the Cochrane Working group checklist,[14] the tool by Lijmers et al,[15] and the NHMRC checklist.[16] Whiting et al (2005) undertook a systematic review and identified 91 different instruments, checklists, and guidance documents.[17] Of these 91 quality-related tools, 67 were tools designed specifically for diagnostic accuracy studies and 21 provided guidance for interpretation, conduct, or reporting, or lists of criteria to consider when assessing diagnostic accuracy studies. The majority of these 91 tools do not explicitly state a rationale for inclusion or exclusion of items; neither have the majority of these scales and checklists been subjected to formal test-retest reliability evaluation. Similarly, the majority do not provide a definition of the components of quality considered in the tool. These variations are a reflection of inconsistency of understanding quality assessment within the field of evidence-based medicine. The authors did not recommend any particular checklist or other tool, but rather they used this information to develop their own checklist the "Quality Assessment of Diagnostic Accuracy Studies" (QUADAS). The QUADAS developers employed rigorous development methods and have established validity and reliability.

The QUADAS tool is comprised of 14 criteria that cover 12 biases and 2 reporting items when assessing studies of diagnostic accuracy. The development of QUADAS included a formal Delphi consensus exercise with experts to select items for inclusion, and they also conducted reliability testing. QUADAS has demonstrated good inter-rater reliability for most items (kappa varied from 55% to 100% for 7 or 8 items from 14);[18,19] areas of greatest disagreement included withdrawals, selection criteria, and indeterminate results. The inter-rater reliability findings suggest the need for explicit contextualization of some criteria; that is, the intent of the item is not modified, rather the reviewers need to provide specific examples in the context of the specific medical test showing when the bias is likely to be present or absent.

Two organizations (Centre for Reviews and Dissemination and the Cochrane Collaboration)[20,21] that undertake large number of systematic reviews of diagnostic tests have endorsed the QUADAS. As noted previously by West et al (2002), quality assessment tools for medical tests tend to focus predominantly on the potential for risk of bias related to the "intervention" or medical tests rather than attributes of bias associated with the study design type. A variety of study designs types can be used to evaluate the accuracy, predictive ability, or other properties of medical tests. The design types can include RCTs, case series, assay reliability studies, and laboratory studies (assessing analytic validity). As such, the QUADAS currently does not address some design specific attributes (for example randomization). The QUADAS is currently under redevelopment and is considering addressing this limitation. In the interim, other design specific components may have to be used in conjunction with the QUADAS. Please refer to the medical test guide for further discussion on the components or tools recommended for genetic test studies.

Generally, the majority of the 14 items within the QUADAS can be applied to most types of diagnostic tests; the developers acknowledge some variability for some items that may allow the item to be excluded from the checklist (detailed in the instructions by the developers). Since its original development, the QUADAS tool has been modified for evaluating diagnostic tests in before-after studies[22] and in studies that use technologies or tests that provide comprehensive analysis of the complete or near- complete cellular constituents (such as DNA, proteins, and intermediary metabolites).[23] In the latter case, the "QUADOMICS" added two items (for a total of 16). Note that the original developers of QUADAS did not undertake these modifications and these adapted versions do not have expanded validity and reliability testing. Currently the

original QUADAS tool is in a phase of redevelopment by the original developers; changes are expected to be available in June 2011.

The QUADAS may represent a "minimum" set of criteria that should be consistently evaluated irrespective of the diagnostic test,[13,20] but it may not always be comprehensive. For example, consider the case in which an EPC needs to evaluate a diagnostic accuracy study within the context of an RCT. The QUADAS does not evaluate the process of randomization or allocation concealment; these two biases related to the conduct of randomized trials may be an important source of methodological heterogeneity that the EPC reviewers also need to appraise. EPCs may need to find other ways to evaluate additional criteria related to the particular study design, as this limitation seems to be unique to tools used to assess diagnostic accuracy studies.[13]

Instruments and Tools to Evaluate Quality of Harms Assessment

Although the assessment of harms is almost always included as an outcome in intervention and medical test studies, the manner of capturing, and reporting harms is significantly different than the outcomes of benefit. Harms are defined as the "totality of possible adverse consequences of any intervention, therapy or medical test; they are the direct opposite of benefits, against which they must be compared". (CONSORT Harms extension, 2004).[24] For a detailed explanation of terms associated with harms please refer to the AHRQ Methods guide on harms.[25] Systematic reviews of intervention studies need to consider the balance between the harms and benefits of the treatment. Empirical evidence across diverse medical fields indicates that reporting of safety information (including milder harms) receives much less attention than the positive efficacy outcomes.[26,27] Thus, an evaluation of the benefits alone is likely to bias conclusions about the net efficacy or effectiveness of the intervention. Although reviewers recognize the importance of harms outcomes, harms are generally ignored in quality assessment checklists. Several recent reviews[2,8,13,17] of quality checklists and instruments do not identify harms as a key criterion within the checklists. We infer that many of the current quality scales and checklists have assumed that harms are simply another study "outcome" and that taking this view suggests that the developers assume that no differences exist between harms and benefits in terms of quality assessment.

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

Since many harms are not anticipated (the severity, the type of event -especially rare events, the timing of the event, etc.), many studies do not specify exact protocols to actively capture events. Often standardized instruments used to systematically collect information on harms are often not included in the study methods, and there is the expectation 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 (passive reporting). Thus, harms are often measured using passive methods that are poorly detailed and there is potential for selective outcome reporting, misclassification, and failure to capture significant events. Although, some types of harms can be anticipated (for example, pharmokinetics of a drug intervention may identify body systems likely to be affected) and these typically reflect several possible outcomes

(both common and rare symptoms, such as headache and stroke); there is the potential for harms in body systems not necessarily linked to the intervention from a biologic or epidemiologic perspective. There is also the issue of establishing an association between the event and the intervention. For example, some harms may need to be adjudicated by a separate committee to establish association with the putative treatment, and as such blinding is not possible. 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. As such, there is a trend towards including elements of quality assessment directed specifically at the collection and reporting of harms. Given the possible (indeed probable) unevenness in evaluating harms and benefits in most intervention or medical test studies, we recommend that EPCs assess the quality of the study separately for benefits and for harms.

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

One checklist developed by the Cochrane Collaboration offers some guidance, and leaves the final choice up to the reviewer to select items from a list of that is stratified by the study design.[28] It assumes that these questions (see Table A-1) can be added to those criteria already detailed in the Cochrane Risk of Bias tool.

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

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

Chou and Helfand developed a tool for an AHRQ systematic review to assess the quality of studies evaluating carotid endoarterectomy; the primary outcome in these studies included adverse events.[29] Four from eight items within this tool were directed specifically to assessing bias associated with adverse events; however, these criteria are applicable to other interventions or medical tests, although no formal validation has been undertaken.[29] The Chou and Helfand tool has been used in comparative studies (RCTs and observational studies). No formal reliability testing has been undertaken and the tool is interpretated as a summed score across 8 items. One advantage of this tool is that it includes elements of study design (for example, randomization,

withdrawal, etc.) as well as some items specific to harms. Table A-2 shows the items within this scale.

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

Table A-2. Chou and Helfand quality assessment tool

Criterion	Explanation	Score
Quality criterion 1: Non-biased selection	1: study is a properly randomized controlled trial, or an observational study with a clear pre-defined inception cohort	
	0: study does not meet above criteria	
Quality criterion 2: Adequate description of population	1: study reports 2 or more demographic characteristics, presenting symptoms/syndrome and at least 1 important risk factor for complications	
	0: study does not meet above criteria	
Quality criterion 3: Low loss to follow-up	1: study reports number lost to follow-up, and the overall number lost to follow-up is low (threshold set at 5% for studies of carotid endarterectomy)	
	0: study does not meet above criteria	
Quality criterion 4: Adverse events pre-specified and defined	1: study reports explicit definitions for major complications that allow for reproducible ascertainment	
	0: study does not meet above criteria	
Quality criterion 5: Ascertainment technique adequately described	1: study reports methods used to ascertain complications, including who ascertained, timing, and methods used	
	0: study does not meet above criteria	
Quality criterion 6: Non-biased ascertainment of adverse events	1: independent or masked assessment or complications	
	0: study does not meet above criteria	
Quality criterion 7: Adequate statistical analysis of potential confounders	1: study examines 1 or more relevant confounders/risk factors using acceptable statistical techniques such as stratification or adjustment	
	0: study does not meet above criteria	
Quality criterion 8: Adequate duration of follow-up	1: study reports duration of follow-up and duration of follow-up adequate to identify expected adverse events (threshold set at 30 days for studies of carotid endarterectomy)	
	0: study does not meet above criteria	
	Total quality score = sum of scores (0-8)	

Table A-3. Inter rater reliability (ICC and confidence interval) within different groups of raters.

	Drug Studies	Surgery Studies	All Studies
Non-expert Raters	0.690 (0.267, 0.910)	0.916(0.803, 0.976)	0.876 (0.770, 0.944)
Experts Raters	0.887 (0.733, 0.967)	0.934(0.845,0.981)	0.922(0.855, 0.965)
All Raters	0.888 (0.750, 0.967)	0.962 (0.915, 0.989)	0.947 (0.905, 0.976)

Table A-4. McMaster tool for assessing quality of harms assessment and reporting in study reports (McHarm). The answers to each question are yes (implying less risk of bias), no (implying high risk of bias), and unsure.

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

References

1. Hartling L, Bond K, Harvey K, et al. Developing and testing a tool for the classification of study designs in systematic reviews of interventions and exposures. (Prepared by the University of Alberta Evidence-based Practice Center under Contract No. 290-02-0023.) Rockville, MD: Agency for Healthcare Research and Quality 2009.
2. Olivo SA, Macedo LG, Gadotti IC, et al. Scales to assess the quality of randomized controlled trials: a systematic review. Phys Ther. 2008 Feb;88(2):156-75.
3. Verhagen AP, de Vet HC, de Bie RA, et al. The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. J Clin Epidemiol. 1998 Dec;51(12):1235-41.
4. Yates SL, Morley S, Eccleston C, et al. A scale for rating the quality of psychological trials for pain. Pain. 2005 Oct;117(3):314-25.
5. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Commun Health. 1998;52:377–84.
6. Reisch JS, Tyson JE, Mize SG. Aid to the evaluation of therapeutic studies. Pediatrics. 1989 Nov;84(5):815-27.
7. Zaza S, Carande-Kulis VG, Sleet DA, et al. Methods for conducting systematic reviews of the evidence of effectiveness and economic efficiency of interventions to reduce injuries to motor vehicle occupants. Am J Prev Med. 2001;21(4 Suppl):23-30.
8. Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol. 2007;36(3):677-8.
9. Newcastle-Ottawa Quality Assessment Scale: Case control studies. Available from: http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm.
10. Shamliyan TA, Kane RL, Ansari MT, et al. Development quality criteria to evaluate nontherapeutic studies of incidence, prevalence, or risk factors of chronic diseases: pilot study of new checklists. J Clin Epidemiol. 2010 Nov 9.
11. Viswanathan M, Berkman, N.D. Assessing the Risk of Bias and Precision of Observational Studies of Intervention or Exposures: Development, Validation, and Reliability Testing of the RTI Item Bank. RTI International–University of North Carolina Evidence-based Practice Center, Contract No. 290200710056I. (AHRQ to provide the remainder); 2011.
12. Shamliyan T, Kane RL, Dickinson S. A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases. J Clin Epidemiol. 2010 Oct;63(10):1061-70.
13. 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.
14. Cochrane Methods Working Group on systematic review of screening and diagnostic tests; 1996.
15. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999 Sep 15;282(11):1061-6.

16. National Health and Medical Research Council (NHMRC). How to review the evidence: systematic identification and review of the scientific citerature. Canberra, Australia: NHMRC; 2000.
17. 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.
18. Mann R, Hewitt CE, Gilbody SM. Assessing the quality of diagnostic studies using psychometric instruments: applying QUADAS. Soc Psychiatry Psychiatr Epidemiol. 2009 Apr;44(4):300-7.
19. Whiting PF, Weswood ME, Rutjes AW, et al. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC Med Res Methodol. 2006;6:9.
20. Leeflang MM, Deeks JJ, Gatsonis C, et al. Systematic reviews of diagnostic test accuracy. Ann Intern Med. 2008 Dec 16;149(12):889-97.
21. Centre for Reviews and Dissemination UoY. Systematic Reviews: CRD guidance for undertaking reviews in healthcare. York: CRD, University of York; 2008.
22. Meads CA, Davenport CF. Quality assessment of diagnostic before-after studies: development of methodology in the context of a systematic review. BMC Med Res Methodol. 2009;9:3.
23. Lumbreras B, Porta M, Marquez S, et al. QUADOMICS: an adaptation of the Quality Assessment of Diagnostic Accuracy Assessment (QUADAS) for the evaluation of the methodological quality of studies on the diagnostic accuracy of '-omics'-based technologies. Clin Biochem. 2008 Nov;41(16-17):1316-25.
24. 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.
25. 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.
26. Ioannidis JP, Lau J. Improving safety reporting from randomised trials. Drug Saf. 2002;25(2):77-84.
27. 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.
28. Higgins JPT, Green S, eds. Cochrane handbook for systematic reviews of interventions. Version 5.0.2. [updated September 2009]. The Cochrane Collaboration. Available from www.cochrane-handbook.org 2009.
29. Chou R, Fu R, Carson S, et al. Methodological shortcomings predicted lower harm estimates in one of two sets of studies of clinical interventions. J Clin Epidemiol. 2007 Jan;60(1):18-28.