Methods Guide
for Comparative Effectiveness Reviews

Avoiding Bias in Selecting Studies

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

This report is based on research conducted by the Oregon Health & Science University, McMaster University, and Southern California Evidence-based Practice Centers (EPCs) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-2007-10057-I). The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

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

This report may be used, in whole or in part, as the basis for development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

Persons using assistive technology may not be able to fully access information in this report. For assistance contact info@ahrq.gov.

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

Suggested citation: McDonagh M, Peterson K, Raina P, Chang S, Shekelle P. Avoiding Bias in Selecting Studies. Methods Guide for Comparative Effectiveness Reviews. (Prepared by the Oregon Health & Science University, McMaster University, and Southern California Evidence-based Practice Centers under Contract No. 290-2007-10057-I.) AHRQ Publication No. 13-EHC045-EF. Rockville, MD: Agency for Healthcare Research and Quality; February 2013. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

Authors:

Marian McDonagh, Pharm.D.[a]

Kim Peterson, M.S.[a]

Parminder Raina, Ph.D.[b]

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

Paul Shekelle, M.D., Ph.D., M.P.H.[d]

[a]Oregon Health and Science University Evidence-based Practice Center
[b]McMaster University Evidence-based Practice Center
[c]Agency for Healthcare Research and Quality
[d]Southern California Evidence-based Practice Center

Preface

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

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

The Methods Guide for Comparative Effectiveness Reviews is a living document, and will be updated as further empiric evidence develops and our understanding of better methods improves. We welcome comments on this Methods Guide paper. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by email to epc@ahrq.hhs.gov.

Carolyn M. Clancy, M.D.
Director
Agency for Healthcare Research and Quality

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

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

Contents

Tables

Key Points

- One hypothesis-testing study and numerous case examples indicate that operational criteria guiding the selection of studies into a systematic review (SR) or meta-analysis can influence the conclusions.
- Assessments of how this source of bias can be reduced, or even the magnitude of the bias, are not available.
- In the absence of conclusive evidence about how to reduce this potential for bias, we recommend that inclusion criteria be clearly described in detail sufficient to avoid inconsistent application in study selection and that inclusion criteria be documented in a protocol.
- We propose hypothetical examples that illustrate how selection of inclusion and exclusion criteria may introduce bias.
- Experience suggests that dual review can identify inclusion criteria that are not sufficiently clear and occasions where subjective judgment may differ. Gray literature (e.g., U.S. Food and Drug Administration [FDA] documents, trial registry reports) can help identify and possibly reduce bias from publication bias or selective outcome reporting.

Background

Much has been written about the importance of various aspects of the conduct of a SR: how to best search computerized databases; whether or not reviewers should be masked to the authors and journals and outcomes of studies being reviewed; how to assess studies for the risk of bias; and the strengths and weaknesses of various different methods of statistically combining the results. The Methods Guide for the Agency for Healthcare Research and Quality (AHRQ) Evidence-based Practice Center (EPC) Program has chapters summarizing the literature and best-practices advice on numerous such aspects of a SR.[1]

We are concerned here with the potential for bias at a point upstream in the SR process—namely what is the effect of going from the initial question of interest ("what is the effect of intervention X on condition Y?") to the operational aspects of the review (such as selecting inclusion/exclusion criteria). For example, in a recent Comparative Effectiveness Review on drugs to treat low bone density, the EPC identified nine prior meta-analyses evaluating the antifracture efficacy of alendronate compared with placebo or no treatment.[2] The meta-analyses were published between 1997 and 2009, and included between them 17 randomized controlled trials (RCTs) published between 1994 and 2004. One might expect that all the trials included in earlier meta-analyses would be included in later meta-analyses, but this is not the case. One meta-analysis published in 2002 included 10 trials, while another published in 2004 included only 5: 4 were among the 10 trials in the 2002 meta-analyses, but 1 trial (published in 1998) was not. Some of the differences in trial inclusion could be explained by whether data were included on vertebral and nonvertebral fractures; whether nonvertebral fractures were treated as a general group; whether nonvertebral fractures were split out into fractures of the hip or wrist; or whether patient populations were considered as secondary prevention or as primary prevention. These differences in which trials were included led to differences in conclusions. In one meta-analysis,[3] the conclusion was that the decrease in nonvertebral fractures was not statistically significant. In another meta-analysis[4] published 3 years earlier, the conclusion was that the beneficial effect of alendronate compared with placebo on nonvertebral fractures was statistically significant. All EPCs can tell similar stories.

Conflicting conclusions confuse decisionmakers, especially if all reviews purported to answer the same question and the differences in the applicability of the evidence are not clearly denoted. Bias results from systematic alteration from the truth. Although we do not know the exact truth, different conclusions lead readers to believe that alternate inclusion and exclusion criteria result in biased conclusions. In order to investigate the potential for this source of bias and identify methods studies that investigate how best to reduce it, we searched for studies that examined two or more SRs of the same topic, evaluating the impact of variation in study inclusion.

We found a very small number of relevant studies (Table 1).[5-8] The most relevant example was a prospective study designed to examine reproducibility between two review groups (on different continents) commissioned to review evidence on the same question, using a common methods specification manual.[8] While the manual outlined the important features of inclusion criteria, the specific criteria used by each group are not reported. Search terms were specified a priori, and the groups were instructed to find and include all study designs, including non-English language, case series, ecological, cross-sectional, case-control, cohort, and intervention studies. Both review groups agreed on including 166 articles, but disagreed on 72 articles (Center A included 52 papers not included by Center B, and Center B included 20 papers not included by Center A). Sixty-three of the 72 discrepancies occurred in screening title and abstract; 9 of the 72

discrepancies occurred during review of full-text articles. Other similar retrospective studies also found differences in their lists of included studies and sometimes different conclusions (Table 1). Although the amount of evidence is small to confirm the presence of bias, the potential for bias is possibly quite large.

Table 1. Studies evaluating reasons for discrepancies in included studies among systematic reviews

Study	Study Aims	Evaluation
Hopayian K and Mugford M (1999) Conflicting conclusions from two systematic reviews of epidural steroid injections for sciatica: which evidence should general practitioners heed?[6]	The aim of this study was to find the reasons for the discordance between two reviews focusing on use of epidural steroid injection for treatment of low back pain and sciatica and to draw conclusions for users of these reviews.	Each review excluded two papers that the other included, both of which supported the ultimate conclusions of the review that included them. One of these studies was published in a non-English language journal and was excluded by one review. The other papers, however, were published in well-known journals. One of these papers was excluded from one review due to problems with extracting the data, while the other review was qualitative and did not require these data to come to a conclusion. The outcome measures included, and inclusion of non-English language papers account for at least some of the differences.
Peinemann F, McGauran N, et al (2008). Disagreement in primary study selection between systematic reviews on negative pressure wound therapy.[7]	The objective of this study was to compare systematic reviews on negative pressure wound therapy with regard to their agreement in inclusion of primary studies.	The authors conclude that the reviews differed in inclusion of studies, primarily the inclusion of studies other than nonrandomized controlled trials. They indicate that the differences arise from differences in methodology, classification of study design, and style of reporting excluded studies. Our analysis of this example showed that included study designs varied among reviews. However, only one of the five reviews concluded that evidence supported the use of the treatment, while the others consistently found that the evidence was insufficient, largely due to concerns over quality. The review that found treatment to be effective had the broadest inclusion criteria with respect to study design and ultimately included 25 papers, compared with 14, 6, 6, and 7 included in the other reviews.
Cook DH, Reeve BK, et al. (1996) Stress Ulcer Prophylaxis in Critically ill patients: resolving discordant meta-analyses.[5]	This study aimed to resolve discrepancies in four previous systematic reviews and provide estimates of the effect of stress ulcer prophylaxis on gastrointestinal bleeding, pneumonia, and mortality in critically ill patients.	From abstract: “The source of discrepancies between prior meta-analyses included incomplete identification of relevant studies, differential inclusion of non-English language and nonrandomized trials, different definitions of bleeding, provision of additional information through direct correspondence with authors, and different statistical methods.” Our analysis of these reviews focused on the prevention of stress ulcer bleeding, as this outcome was common across the reviews. The definition of bleeding differed among reviews. Two more recent reviews came to very different conclusions that can be directly related to the inclusion criteria. One review included both randomized and “quasi-randomized controlled trials,” while the other review included randomized controlled trials with at least 10 subjects per arm published in a variety of languages. In this example, the difference in conclusions in appears to be related largely to inclusion of non-English language articles in one but not the other.

Other authors have addressed reasons for discrepant results from meta-analyses on the (seemingly) same topics.[9,10] Ioannidis has examined multiple such scenarios and concluded that the reasons for discrepancy are typically multifactorial, but include differing study questions and inclusion criteria as well as differences in the process of applying the criteria in study selection. He gives examples of situations where inclusion criteria for meta-analyses were apparently specified in way that would obtain results that supported the viewpoints of the authors rather than reflecting questions of clinical uncertainty.[9]

As part of the EPC Methods Guide, we intend that this paper will guide EPCs when selecting studies for inclusion in an SR. Guidance is intended to reduce inconsistencies and risk of bias. Unfortunately, because there are no available studies to guide us how best to reduce this variation, what follows is based on fundamental principles of SRs and the experience of the EPC program.

Inconsistencies and bias can certainly occur during the development of key questions, which define the scope of the review and details the population(s), intervention(s), comparator(s), outcome(s), timing, and setting (PICOTS), and sometimes even the study designs or study characteristics of interest. The methods used by the EPC program at this earlier stage are discussed elsewhere.[11] Likewise, we recognize that bias can also be introduced during the searching stage,[12] or in how reviewers handle assessment of reporting biases,[13] and guidance on these methods are provided elsewhere.[11,12] This paper focuses on what to do with the literature once it is identified. We first describe the types of bias then stratify the guidance on addressing these biases into sections: Setting Inclusion Criteria to Avoid Bias in Selecting Studies, Study Selection Process, and Using Gray Literature to Assess and Reduce Bias.

Types of Potential Biases in Selecting Studies

Spectrum Bias

The inclusion or exclusion of a specific population can have a dramatic impact on the conclusions for the effectiveness of a treatment. For example, while one meta-analysis found no significant benefit of the invasive treatment for coronary artery disease over conservative treatment, a subsequent meta-analysis by invasive cardiologists found significant benefit with invasive treatment when they included patients with unstable angina, a population in which invasive management is known to be more beneficial.[9]

Publication bias and outcome reporting bias can have implications for the conclusions of a review. Bias in selection of studies may overlap with these biases, but methods for avoiding them are addressed in other chapters.[13,14]

Random Error

Even when reviewers have a common understanding of the selection criteria, random error or mistakes may result from individual errors in reading and reviewing studies.

Guidance for Setting Inclusion Criteria To Avoid Bias in Selecting Studies

Although setting inclusion criteria based on key questions may seem straightforward, the experience in the AHRQ EPC program has shown that this is often not the case. The AHRQ EPC program has an explicit process of systematic review development called Topic Refinement. Its goal is the development of inclusion criteria based on the Key Questions via a process that involves the review team and technical expert panel input.

One of the main goals in developing inclusion criteria is to minimize ambiguity. Greater ambiguity in inclusion criteria increases the possibility of poor reproducibility due to many subjective decisions regarding what to include, potentially resulting in at least random error in study selection.

The criteria should be set a priori and based on the analytic framework or conceptual model using a protocol.[15-17] The benefits of using a protocol specific to SRs include improving transparency and rigor of SRs, and important to this chapter, reducing bias in study selection decisions. Requirements for SR protocols for reviews conducted by EPCs are currently undergoing further development in coordination with other organizations (e.g., Institute of Medicine and PROSPERO). The protocol should be based on a standard set of elements, publicly available, ideally through a SR Registry, (e.g. PROSPERO, www.crd.york.ac.uk/prospero/).

However, there is a balance to be struck between making the inclusion criteria so narrow that it is unlikely that eligible evidence will be found and so loosely defined that it increases the possibility of poor reproducibility due to many subjective decisions regarding what to include. EPCs should attempt to strike this balance, but recognize that there will be times when their initial attempt is not working and changes need to be made. All eligibility criteria decisions should be reported transparently in the published SR.

Selecting PICOTS Criteria

In addition to random error from ambiguous definition of criteria, the selection of PICOTS inclusion or exclusion criteria can introduce systematic bias. A systematic review starts with a broad comprehensive search and the choice of which studies to include can directly influence the resulting conclusions. The EPC should carefully consider whether PICOTS criteria are effect modifiers and how inclusion and exclusion criteria may potentially skew the studies and thus results reported in the review.

Table 2 below suggests potential implications or biases that may result from specific hypothetical examples of inclusion and exclusion criteria.

Table 2. Hypothetical examples of potential for bias based on inadequately defined PICOTs

PICOTS Criterion	Inclusion Criterion	Potential for Bias in Selecting Studies for Review	Possible Biased Result
Population	Population is described as patients with heart failure	The reviewer may have to decide which classes of heart failure the question was meant to whether these different severities are meant to be combined or evaluated separately.	Reviewer chooses to include only Class III and IV heart failure and finds that the intervention is effective, whereas conclusions on effectiveness may have been diluted if all severity classes had been included.
Intervention	Intervention described as anticoagulants	Reviewer must make the decision on which interventions are considered anticoagulants; e.g., may combine oral and injectable anticoagulants.	Combining oral and injectable anticoagulants may be inappropriate for short term effectiveness and harms and may overestimate benefits for oral anticoagulants and underestimate harms for short term effects.
Comparator	Not defined	Reviewer makes choice among other interventions include in review, interventions excluded from the review, and how to handle placebo, or no treatment, groups.	Reviewer includes only placebo or no treatment groups and concludes that the intervention is effective, whereas it may be less effective in comparison to existing interventions.
Outcome	Described as effectiveness outcomes	Reviewers determine whether specific outcomes are in fact effectiveness. For example, cognitive testing using laboratory settings.	Reviewers report information on intermediate or surrogate outcomes and fail to report lack of effectiveness outcomes, thus making the intervention seem more effective than if clinical outcomes are considered.
Timeframe	Not defined	Reviewers may report whatever is available in the literature, which may be short-term studies.	Without prespecifying that long term outcomes are essential and only reporting short term outcomes, reviewers may overestimate effectiveness of treatment. Also secular trends may mean that older studies may either over or under estimate the effect of an intervention depending on changes in standard of care, technology, or disease epidemiology.
Setting	Described as outpatient	Reviewers must decide whether various settings are in fact outpatient, such as residential treatment programs.	Patients in residential treatment programs may be patients with more severe symptoms or other comorbidities in which the intervention may be more or less effective.

Table 2. Hypothetical examples of potential for bias based on inadequately defined PICOTs (continued)

PICOTS Criterion	Inclusion Criterion	Potential for Bias in Selecting Studies for Review	Possible Biased Result
Study Designs or Study Characteristics	Randomization or allocation of treatment (RCT vs. observational studies)	Reviewer decides to include RCTs only.	Limitation to RCTs may be more likely to exclude certain types of interventions such as procedures or dietary/nutritional interventions, as well as studies reporting long term outcomes or harms.
	Quality or risk of bias of individual studies	Reviewer decides to exclude low quality studies or those at high risk of bias.	Studies conducted in nonacademic centers or with a null effect may be more likely to rate as "low quality" due to rejection from high impact journals. Exclusion of all low quality studies or those at high risk of bias may exclude large body of consistent studies that may yield valuable information on benefits or harms.
	Study size	Reviewer decides to exclude RCTs less than 50 participants or observational studies less than 1000 patients.	Exclusion of small studies may exclude valuable information. Exclusion of small studies introduce bias such as by excluding studies conducted in nonacademic or urban populations which may have higher severity of disease, and overestimate effectiveness.
	English language	Reviewer decides to exclude non-English studies.	Exclusion of non-English studies may exclude studies that found a null effect and thus overestimate effectiveness.
	Inclusion of necessary information	Reviewer may exclude studies that do not report the primary outcomes listed.	Studies may have measured outcomes, but not reported them in the studies due to null findings. Exclusion of these studies may overestimate effectiveness.

PICOTS = population(s), intervention(s), comparator(s), outcome(s), timing, and setting; RCT = randomized controlled trial

Population

Inclusion criteria for the population(s) of interest should be defined in terms of relevant demographic variables, disease variables (i.e., variations in diagnostic criteria, disease stage, type, or severity), risk factors for disease, cointerventions, and coexisting conditions.[18] For example, if an SR is focusing only on adult populations, then the inclusion criteria should specify the age range of interest. Ambiguity in population inclusion criteria increases the risk that inclusion decisions could be influenced by differing viewpoints about potential relationships between particular demographic or disease factors and outcome. Table 2 illustrates one such example of how inadequate description of inclusion criteria for a heart failure population may bias the results of SR. Inclusion criteria for population subgroups of interest should also be defined with similar specificity.

Intervention and Comparators

Although the Key Questions may frame the interventions in broad terms such as "anticoagulants," it is essential for the inclusion criteria to specify exactly which individual interventions are of interest, including their duration and intensity. Otherwise, reviewers may end

up missing important interventions and thus overestimate or underestimate the effectiveness or harms of an intervention. This is particularly important in reviews of health care delivery programs that are less clearly defined. A review may examine a specific program as a whole, the component parts of a program, or the theoretical mechanism of action of a component part. Defining an intervention too narrowly may increase the confidence in effectiveness, but reduce the relevance of the finding for implementation in other settings.

To enhance readability, key questions may not always define the comparison, which may introduce both random and systematic error. Without specifying the comparator, one reviewer may compare the effectiveness of anticoagulants to compression stockings, another may compare them to early walking, and yet another may compare it to other anticoagulants. Selection of a comparison of known poor effectiveness may systematically bias the effectiveness of the intervention away from the null, whereas poor specification and thus inappropriate combination of comparisons may result in an uninterpretable result.

Outcomes

Regardless of the topic, SRs should focus on assessing a range of patient-centered outcomes, including both benefits and harms. The scope of included outcomes should address both effectiveness and harms on which strength of the evidence will be graded.[19] If intermediate outcomes are included they should be presented in context of how they relate to the clinically important harms and benefits (e.g., via an analytic framework) as outlined in the chapter of grading the strength of the evidence.[19] When there are a large number of outcomes included, EPCs should specify a priori which clinically important outcomes they will grade the strength of evidence. Despite the temptation to exclude studies that only report a specific outcome (e.g., mortality), EPCs should be cautious since this may augment the risk of identifying studies that have selectively published only outcomes with positive results (selective outcome reporting bias).

In order to reduce variation in study selection related to outcomes, we recommend that the inclusion criteria clearly identify and describe outcomes, outline any restrictions on measurement methods or timing of outcome measurement, and provide guidance for handling of composite outcomes. For clinical areas (such has pain and psychological functioning) that are notoriously characterized by variability in outcome measurement methods and a multitude of scales and instruments, the risk is greater for inconsistency in study selection. In these cases, it is especially important to consider how to handle this variation early in the SR process. The EPC may choose to restrict to specific measurement methods (i.e., only including studies that used measurement scales that have been published or validated), but need to consider what studies they will be eliminating and what effect this may have on the review. Study investigators that do not use the most commonly validated instruments may be systematically different from those that do. For example, investigators from different communities may use different instruments and systematic exclusion of these studies may exclude specific populations such as rural or small communities or nonacademic populations.

Lack of specificity on other aspects of outcome measurement may also bias SR conclusions. For example where study reports include multiple time points for outcome measurement, but the SR inclusion criteria are not adequately specific about the relative importance of different time points, the choice of which to include or to emphasize is left to the reviewers. This scenario could lead to important differences in conclusions depending on which outcome-time point pair are selected for inclusion, particularly in a meta-analysis.[10]

Finally, it is ideal to consider individual outcome separately, rather than using composite outcomes. Composite endpoints are often difficult to interpret and may exaggerate the magnitude of treatment effect.[20] EPC reviewers should consider specifying whether composite outcomes are of interest and, if so, whether there is a need to place any restrictions on which combinations of outcomes are acceptable (e.g. those with similar importance to patients and magnitude of treatment effect). Otherwise, there may be variation in selection of studies that, for example, do not separately report mortality and cardiovascular events. EPC review teams should rely on empiric research when available to form the basis of any decisions to limit study selection based on outcomes.

Timeframe and Setting

Setting inclusion criteria for timeframe (duration of study, years of study conduct, etc.) and setting may not apply to all clinical questions. Reviewers should identify the expected time period of study that would be needed to identify effectiveness on patient-important outcomes and harms. Lack of specification for the need for long-term studies may overestimate the effect on short term outcomes, while under-reporting the effect on long term outcomes. EPCs should clearly specify any decision to limit studies based on followup duration and define a priori the most relevant time periods for the interventions, populations, and outcomes of interest. When the focus of a SR is confined to a particular setting, such as a nursing home environment or residential treatment center, the inclusion criteria should include guidance for considering eligibility of studies that include commingled or ill-defined settings. Reviewers should consider how interventions may be different in settings such as nursing homes or other long-term care settings compared with general inpatient or outpatient settings and how inclusion or exclusion of these settings may systematically bias the conclusions. The criterion for study setting may also be considered when setting the selection criteria for population.

Study Designs or Study Characteristics

Due to time, budget, or resource constraints as well as concerns about the validity and relevance of the studies, reviewers often make decisions about excluding studies based on study design features (randomization or nonallocation of treatment), study conduct (quality or risk of bias of individual study), language of publication, study size, or reporting of relevant data.

Observational studies make up the bulk of the published literature. EPCs should refer to the Methods guidance for when to include observational studies.[21,22] However after deciding to include observational studies, EPCs need to take special care in developing and testing criteria for determining eligibility.[4] Because of the lack of consensus on any single taxonomy for assigning labels to specific types of observational study designs,[23] EPC teams should define study designs with sufficient clarity so that their reviewers can consistently and correctly determine if a given study is eligible. Exclusion of observational studies without careful consideration about whether these studies may provide information that would not be available from RCTs (i.e., long-term outcomes or harms and representative populations) may bias the review conclusions.

Reviewers often include other study design or reporting characteristics as eligibility criteria. Reviewers may decide to restrict study inclusion based on sample size (e.g.,> 1,000 patients) or publication language (e.g., English language only). However, smaller studies or non-English studies may be systematically different from larger studies or English-language studies and limiting by these characteristics for convenience may introduce a systematic bias as well. For

example, in a review of surgical and pharmaceutical interventions, studies on surgical interventions may be smaller than studies on pharmaceuticals, thus biasing a review that excludes small studies to find evidence on drugs but insufficient evidence on surgical interventions.

Typically such decisions are taken for reasons of time-efficiency. The assumption is that not employing such limits would yield a very large number of studies that would significantly increase workload without providing additional value in terms of high-quality evidence. Without empirical evidence relative to the topic area under review, it is not possible to rule out systematic bias. For example, the decision to use only English-language publications may be set because the review team does not have the ability to read other languages but the time and cost of translation are not feasible within the report timeline and budget. Studies of language restrictions in SRs have had variable results, from significant impact to very little impact, sometimes depending on the specific topic being studied.[24-34]

The way that high risk of bias studies are handled in SRs also varies and may introduce bias. Once a study has been determined to have high risk of bias, options include outright exclusion; inclusion in evidence tables with or without inclusion in a narrative description of the evidence (possibly depending on whether the study constitutes the only evidence for a given intervention and/or outcome); or inclusion in quantitative analyses using weighting based on quality or sensitivity analysis. Including studies with a high risk of bias without appropriate weighting for their risk of bias may introduce bias in the SR. However, because assessments of risk of bias are never based entirely on empirical evidence, and are subjective by nature, outright exclusion of studies with high risk of bias may also introduce bias. Additionally, weighting in meta-analysis based on risk of bias assessments may introduce bias and has been shown to result in inconsistency.[35] EPCs should be explicit about how such studies will be handled, a priori. If studies with high risk of bias are to be excluded in any way, they should be clearly identified in the text or in an appendix. Such transparency improves the likelihood that erroneous ratings of studies with high risk of bias can be identified.

Study Selection Process

Even with clear, precise inclusion criteria, elements of subjectivity and potential for human error in study selection still exist. For example, inclusion judgments may be influenced by personal knowledge and understanding of the clinical area or study design (or lack thereof).

The study selection process is typically done in two stages; the first stage involves a preliminary assessment of only the titles and abstracts of the search results. The purpose of this step is to eliminate efficiently all obviously ineligible publications. The second stage involves a careful review of the full-text publications.

Dual review—having two reviewers independently assess citations for inclusion—is one method of reducing the risk of biased decisions on study inclusion, as is recommended in the Institute of Medicine's "What works in healthcare: standards for systematic reviews."[36] Some form of dual review should be done at each stage to reduce the potential for random errors and bias. Reviewers compare decisions and resolve differences through discussion, consulting a third party when consensus cannot be reached. The third party should be an experienced senior reviewer. The two stages of assessment are discussed in more detail below. Dual review can help identify misunderstandings of the criteria and resolve them such that the studies included will truly fulfill the intended criteria.

At the title and abstract stage, one alternative to 100 percent dual review is to have one reviewer accept the citations that appear to meet inclusion criteria and send them on to full-text review, with a second reviewer assessing only those citations and abstracts that the first reviewer deemed ineligible. Although there is currently no empiric evidence to support this method, we speculate that the sensitivity of the process is increased although the specificity may be somewhat reduced; the tradeoff is a potentially larger pool of full-text articles to review but a lower chance of having missed an eligible study. Additionally there is a risk of reviewer bias, with the second reviewer's knowledge that the first reviewer had deemed the studies ineligible. A second reasonable alternative is to conduct dual review on a small percentage of the citations, insuring reliability of assessments before going on to have the remainder of citations assessed by a single reviewer. In this situation, we recommend that review teams start with a pilot phase, using screening forms based on the eligibility criteria, to screen a small number of studies (e.g., 10 to 20 percent), followed by discussion such that variation in interpretation of how the inclusion criteria should be applied can be resolved early on. For this calibration process we suggest pairing a methodologist with a clinical expert if possible. For the stage of reviewing of full-text articles we recommend that EPCs undertake a complete independent dual review.

Some experts assert that reviewers' knowledge of the identity of the study authors, institution, or journal, or year of publication may influence their decisions and that masking of these factors might be useful.[37,38] These assertions may be based on the findings of a randomized study conducted by Berlin, et al., where there was considerable disagreement between blinded and unblinded reviewers in selecting studies for meta-analysis in where reviewers were using the same inclusion criteria.[39] However, the conclusions of this study were that masking "during study selection and data extraction had neither a clinically nor a statistically significant effect on the summary odds ratio" and that masking required 1.3 hours per paper. Hence, masking of reviewers to manuscript details is not routinely recommended.

Testing of inter- or intra-rater reliability, using the kappa statistic is sometimes suggested as a necessary component of the dual review strategy. However, because the goal is to include the "right" studies and not necessarily to achieve perfect agreement, and using the usual dual review process should obviate the need for such testing, this approach is not generally recommended.

Documenting and reporting all decisions made in the study selection process at the full-text level provides transparency that is essential in allowing independent assessment of the potential for bias by readers of SRs. SRs should include the numbers of studies screened, assessed for eligibility, and included in the review, ideally in the form of a flow diagram as recommended in the Preferred Reporting Items for Systematic Review and Meta-Analyses (PRISMA) statement.[17]

As a part of this transparency, SRs should include a listing of excluded studies, along with respective reasons for exclusion. The list of excluded studies is meant to document the reason that specific studies reviewed at the full-text level were excluded when a reader may reasonably think they might have been included. An example would be studies in which the population and interventions meet eligibility, but the study design or comparator does not.

Using Gray Literature To Assess and Reduce Bias

In reviewing gray literature documents, reviewers are seeking to identify unpublished studies and unpublished data supplemental to published studies. Just as excluding studies can cause systematic variation, different approaches to finding and including or using grey literature can also affect the studies included and thus the conclusions of a review. While there may be variation in definitions of gray literature in general, EPC guidance outlines the best practices for

identifying gray literature from regulatory data (e.g., the FDA), manufacturers, and other unpublished information such as abstracts or trial registries (see Table 3 for descriptions).[12] At a minimum, knowledge of unpublished studies may lead the EPC to reduce their assessment of the strength of the body of evidence in the review because of the existence of grey literature may suggest evidence of publication bias.[40] There is a risk that the gray literature identified has a high risk of bias; that the reason for lack of publication was due to flaws in the study rather than negative results. In some cases, enough information may be available for the reviewers to assess study quality and include the study in the SR.

A review of original protocols (i.e., registered with clinicaltrials.gov) may identify selective reporting in the published literature for outcomes in which there is a positive result. Comprehensive searches for protocols and identification of selective outcome reporting may lead a reviewer to reduce their confidence in a positive finding. EPC reviewers should be alert to the possibility that the study measured and analyzed the outcome of interest, but did not report the finding due to a negative result. Gray literature helps to provide some fuzzy information on areas that were previously a blind spot in SRs of only published literature.

Reviewing gray literature may be resource intensive, and it is not yet clear if or when the effort required is worth the potential benefit. Despite these limitations, the risk for selective and biased publication of studies makes the inclusion of gray literature a necessary component of high quality SRs until empirical evidence is available to provide further guidance. Given the complexity of gray literature and the likelihood that a given review may not be able to fully search and include all gray literature, we recommend that the review protocol define, a priori, the sources of gray literature (Table 3), and the eligibility criteria applied to them. The following are our recommendations for how to approach selecting studies from gray literature documents in a way that will minimize potential bias in selection of studies:

1. Identify studies for the SR using standard search techniques first and become familiar with these studies before reviewing gray literature documents.
2. Assess studies in gray literature documents for eligibility in the SR using the key questions and inclusion criteria as discussed above.
3. As some sources of gray literature will have overlap with published literature, for example, FDA documents and trial registries, reviewers should match studies in gray literature documents based on characteristics such as unique study identifies, sample size (by group), and study duration, to those found in published literature to remove any duplicates. This information is sometimes readily available, but often matching is difficult.
4. As with assessment of other types of evidence, dual review is a good way to guard against potentially biased inclusion decisions. Reporting on the inclusion of unpublished studies or data is important to ensure transparency and to identify areas about which EPCs have less confidence that the reporting is unbiased because the included information had not been published and, therefore, had not yet been vetted through a peer review process.
5. If gray literature search uncovers studies that were not included in the published literature, EPC must consider whether the studies have sufficient data and are of sufficient quality to be included in the analysis. If not, then consider whether the presence of such studies suggests that the published literature is biased and should be "downgraded" for publication bias in assessing the strength of evidence.

Table 3. Sources of unpublished information for comparative effectiveness reviews

Source	Description
FDA Documents	Documents from the FDA are the reports written by FDA professional staff assigned to review a New Drug Application submitted by a pharmaceutical manufacturer when applying for FDA approval of a drug for a specific indication or set of related indications. Although FDA review documents have multiple parts, the two most relevant sections for the EPC review team are the medical reviewers' and statistical reviewers' reports. By reviewing these sections, the EPC may identify studies that they did not find through their published literature search and that may indicate the presence of publication or outcome reporting bias. Many of the FDA documents currently available are only scanned originals, meaning that EPCs cannot use software search functions on them; moreover, in some sections, the FDA may have redacted some material; finally, in addition to potentially relevant trials, these documents may also include studies that are not relevant to a SR (e.g., studies in healthy subjects). Nonetheless, they can provide data and analyses of Phase 2 and 3 trials that may be more extensive than are available in published manuscripts.
Scientific Information Packets	Through the SIPs,[12] manufacturers may submit published and unpublished data from RCTs and observational studies relevant to clinical outcomes. For unpublished studies, manufacturers are asked to provide a summary that includes study number, study period, design, methodology, indication and diagnosis, drug dose and duration, inclusion and exclusion criteria, primary and secondary outcomes, baseline characteristics, numbers of patients screened/eligible/enrolled/lost to withdrawn/follow-up/analyzed, and effectiveness/efficacy and safety results. For studies registered with ClinicalTrials.gov, the ClinicalTrials.gov identifier, condition, and intervention are also requested.
Trial Registries	Trial registries that contain results from trials registered, such as the ClinicalTrials.gov and Clinicalstudyresults.org, can be useful sources of information for reviewers. Because the study is registered at the beginning of the study, the intended primary outcome measures, sample size, and other trial characteristics are known prior to reading reports of results. While this can be very useful in identifying potential outcome reporting biases, these registries are also useful in identifying completed studies that have not yet been published, and data on outcomes that may not have been reported in the publications of the trial.

EPC = Evidence-based Practice Center; FDA = U.S. Food and Drug Administration; RCT = randomized controlled trial; SIP = scientific information packet; SR = systematic review

Because the studies in the FDA documents and trial registries are referred to by codes and because the publications of these studies may or may not also list these numbers, EPCs must often match up the studies using study characteristics (e.g., numbers of included patients, duration of study). Doing so allows reviewers to identify relevant unpublished studies or additional outcomes or and statistical analyses examined in a known study that had not been not reported in the published literature. This process, although lengthy, can help EPCs identify the full body of evidence that is relevant to the question and better identify or reduce bias in selection of studies. Comparing these documents to published manuscripts of the trials may also uncover changes in the definition the primary outcome or misrepresentation of the primary outcome.[41] Dual review of gray literature documents is recommended when assessing relevance for potential inclusion into the review.

EPCs may determine that unpublished, supplemental data from the documents in the scientific information packets (SIPs) pertaining to studies with publications may be appropriate for inclusion into their review. For example, subgroup analyses may be reported in SIPS that had not appeared in the published manuscript(s); however, EPCs do need to view these data with caution. EPC reviewers should have discussed and established a priori guidance on when to include specific types of unpublished data and how to handle such data when they are included. With respect to subgroup data or analyses, for example, the review team should define the clinically relevant subgroup populations (e.g., characterized by comorbidities and drug co-administration) during topic development and document them a priori in the inclusion criteria document. If SIPs present data on populations other than those identified as clinically relevant, then EPCs would not include them or include them only as hypothesis generating; alternatively,

EPCs may consider formally amending the inclusion criteria if clinical expertise indicates that noninclusion of these subgroups was an oversight.

Discussion

Our review of the literature indicates that systematic bias and random error can potentially occur in the selection of studies for SRs. Methods exist to reduce the likelihood of both problems, as described in this chapter. Some aspects of potential bias in study selection overlap with considerations to reduce bias when defining the key questions (discussed in further detail by Whitlock, et al.[11]). Table 2 highlights some potential sources of bias that reviewers should consider when selecting inclusion and exclusion criteria. However these are only potential sources of bias and need further research to establish which may be more likely to introduce systematic bias into a review. Further, as this is likely topic specific, reviewers need to have a careful and considered approach in selecting inclusion and exclusion criteria. After selection of inclusion and exclusion criteria, reviewers should track the reasons for exclusions of studies and consider at the end whether exclusion of studies due to the reasons identified in Table 2 may have biased the study. The potential effect of excluding or combining studies on the results should be highlighted as a potential limitation in the Discussion section of the SR.

A potential source of bias that was not addressed in this paper is the assessment and management of conflict of interest for authors, funders, and others with input into the SR process, including technical experts, key informants, and peer reviewers. The possible impact of conflicts is unknown at this time, but is the subject of future research, and is addressed in the Institute of Medicine's Standards for Systematic Reviews.[15] EPCs must be aware of not only the possibility of outcome reporting bias of individual studies, but also their own presentation of outcomes and how that may be introduce bias into the interpretation of findings. While some of these issues have been touched on in this paper, they are the subject of future research as well.

EPC reviewers should explicitly consider how they handle the concept of "best evidence" in both inclusion and synthesis of studies. Even when studies technically meet all eligibility criteria, and are correctly identified for inclusion using rigorous assessment procedures, the level of contribution each eligible study will make to the body of evidence can vary importantly. Depending on the availability of the best possible evidence, EPCs may differ in the extent to which they use lower-strength evidence for a given SR.

For example, when the evidence from randomized controlled trials that directly compare interventions has no obvious gaps, then the value of lower-strength evidence from observational studies, indirect comparisons from placebo-controlled trials, and pooled analyses of only a select number of studies is lower than it would be if the EPC reviewers did encounter such gaps. Thus, when gaps exist in the best possible evidence, the value of lower-strength evidence is greater. Reviewers must rely on their expert judgment as to what constitutes a gap in the best possible evidence and to what extent to report the lower-strength evidence. Systematic bias or random error can occur when EPCs do not clearly establish decision rules for utilizing lower-strength evidence.[22]

Conclusion

In summary, EPCs should write the key questions and inclusion criteria in a way that provides their reviewers with detail sufficient to minimize variation in interpretation. Discussion, dual review, and practice will aid in reducing potential bias by establishing consistent interpretation of the criteria. EPCs should disclose the studies evaluated at the full-text level and determined to be ineligible and provide brief reasons for those exclusions.

Reporting the steps taken to avoid bias in selecting studies, such as conducting dual review, tracing the resulting flow of studies through the review (e.g., PRISMA diagram), and reporting potentially relevant studies that were excluded (with reasons for their exclusion) in the SR is essential for transparency. Gray literature can provide evidence on publication bias and outcomes reporting bias; EPCs should use processes similar to those used with published literature in reviewing gray literature to avoid potential bias in selecting unpublished studies or data. Depending on the experience levels of the SR team members, the complexity of the clinical area, the size of the SR, and other factors, the exact approach to operationalizing the study selection process may vary somewhat from SR to SR. Below are some summary points to minimize various types of study selection bias.

- Define inclusion and exclusion criteria by PICOTS clearly and in a protocol. Reduce ambiguity as much as possible.
- Consider the risk of introducing spectrum bias when selecting populations.
- Define interventions with specificity such that they are applicable to the intended user of the review.
- Be cautious about excluding studies based on reporting of outcomes of interest.
- Dual review can help reduce random error in applying inclusion and exclusion criteria
- Examine grey literature for evidence of unpublished data or studies that may indicate the presence of publication bias or selective outcome reporting bias. Consider the risk of bias of this information before using the information in the review or to adjust the strength of evidence of the review.

References

1. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. AHRQ Publication No. 10(11)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2011. http://www.effectivehealthcare.ahrq.gov/ehc/products/60/318/MethodsGuide_Prepublication-Draft_20120523.pdf.

2. MacLean C, Alexander A, Carter J, et al. Comparative Effectiveness of Treatments To Prevent Fractures in Men and Women With Low Bone Density or Osteoporosis. Comparative Effectiveness Review No. 12. (Prepared by Southern California/RAND Evidence-based Practice Center under Contract No. 290-02-0003). Rockville, MD: Agency for Healthcare Research and Quality; 2007. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

3. Sawka A, Papaioannou A, Adachi J. Does alendronate reduce the risk of fracture in men? A meta-analysis incorporating prior knowledge of anti-fracture efficacy in women. BMC Musculoskelet Disord. 2005;6:39. PMID: 16008835.

4. Cranney A, Wells G, Willan A, et al. Meta-analyses of therapies for postmenopausal osteoporosis. II. Meta-analysis of alendronate for the treatment of postmenopausal women. Endocr Rev. 2002;23(4):508-16. PMID: 12202465.

5. Cook DJ, Reeve BK, Guyatt GH, et al. Stress ulcer prophylaxis in critically ill patients. Resolving discordant meta-analyses. JAMA. 1996 Jan 24-31;275(4):308-14. PMID: 8544272.

6. Hopayian K, Mugford M. Conflicting conclusions from two systematic reviews of epidural steroid injections for sciatica: which evidence should general practitioners heed? Br J Gen Pract. 1999 Jan;49(438):57-61. PMID: 10622020.

7. Peinemann F, McGauran N, Sauerland S, et al. Disagreement in primary study selection between systematic reviews on negative pressure wound therapy. BMC Med Res Methodol. 2008 Jun 26;8(1):41. PMID: 18582373.

8. Thompson R, Bandera E, Burley V, et al. Reproducibility of systematic literature reviews on food, nutrition, physical activity and endometrial cancer. Public Health Nutr. 2008 Oct;11(10):1006-4. PMID: 18053295.

9. Ioannidis JPA. Meta-research: The art of getting it wrong. Research Synthesis Methods. 2011;1:169-84.

10. Tendal B, Higgins JP, Juni P, et al. Disagreements in meta-analyses using outcomes measured on continuous or rating scales: observer agreement study. BMJ. 2009;339:b3128. PMID: 19679616.

11. Whitlock EP, Lopez SA, Chang S, et al. AHRQ series paper 3: identifying, selecting, and refining topics for comparative effectiveness systematic reviews: AHRQ and the effective health-care program. J Clin Epidemiol. 2010 May;63(5):491-501. PMID: 19540721.

12. Relevo R, Balshem H. Finding Evidence for Comparing Medical Interventions. Methods Guide for Comparative Effectiveness Reviews. AHRQ Publication No. 11-EHC021-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2011. http://effectivehealthcare.ahrq.gov/.

13. Norris S, Holmer H, Ogden L, et al. Selective Outcome Reporting as a Source of Bias in Reviews of Comparative Effectiveness. (Prepared by the Oregon Evidence-based Practice Center under Contract No. 290-2007-10057-I.) AHRQ Publication No. 12-EHC110-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

14. Viswanathan M, Ansari MT, Berkman ND, et al. Assessing the Risk of Bias of Individual Studies in Systematic Reviews of Health Care Interventions. Methods Guide for Comparative Effectiveness Reviews. AHRQ Publication No. 12-EHC047-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2012. http://www.effectivehealthcare.ahrq.gov/ehc/products/322/998/MethodsGuideforCERs_Viswanathan_IndividualStudies.pdf.

15. Institute of Medicine. Finding What Works In Health Care: Standards for Systematic Reviews. Washington, DC: National Academies Press; 2011.

16. Clarke M, Stewart L. PROSPERO–the new international prospective register of systematic reviews. Cochrane Methods. Cochrane Database of Systematic Reviews 2011;Suppl 1:1-40

17. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Surg. 2010;8(5):336-41. PMID: 20171303.

18. West S, Gartlehner G, Mansfield AJ, et al. Comparative Effectiveness Review Methods: Clinical Heterogeneity. Methods Research Paper. AHRQ Publication No. 10-EHC070-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2010. http://effectivehealthcare.ahrq.gov/.

19. Owens DK, Lohr KN, Atkins D, et al. Chapter 10. Grading the Strength of a Body of Evidence When Comparing Medical Intervention. In: Methods Guide for Effectiveness and Comparative Effectiveness Reviews. AHRQ Publication No. 10(11)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality; August 2011. Chapters available at: www.effectivehealthcare.ahrq.gov

20. Ferreira-Gonzalez I, Permanyer-Miralda G, Busse JW, et al. Methodologic discussions for using and interpreting composite endpoints are limited, but still identify major concerns. J Clin Epidemiol. 2007 Jul;60(7):651-7; discussion 8-62. PMID: 17573977.

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

22. Norris S, Atkins D, Bruening W, et al. Selecting Observational Studies for Comparing Medical Interventions. In: Methods Guide for Comparative Effectiveness Reviews. AHRQ Publication No. 10(11)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2010. http://www.effectivehealthcare.ahrq.gov/ehc/products/196/454/MethodsGuideNorris_06042010.pdf.

23. 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. Methods Research Report. Rockville, MD: Agency for Healthcare Research and Quality; 2010. AHRQ Publication No. 11-EHC-007. http://effectivehealthcare.ahrq.gov/.

24. Nylenna M, Riis P, Karlsson Y. Multiple blinded reviews of the same two manuscripts. Effects of referee characteristics and publication language. JAMA. 1994 Jul 13;272(2):149-51. PMID: 8015129.

25. Gregoire G, Derderian F, Le Lorier J. Selecting the language of the publications included in a meta-analysis: is there a Tower of Babel bias? J Clin Epidemiol. 1995 Jan;48(1):159-63. PMID: 7853041.

26. Moher D, Fortin P, Jadad AR, et al. Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews. Lancet. 1996 Feb 10;347(8998):363-6. PMID: 8598702.

27. Egger M, Zellweger-Zahner T, Schneider M, et al. Language bias in randomised controlled trials published in English and German. Lancet. 1997 Aug 2;350(9074):326-9. PMID: 9251637.

28. Moher D, Pham B, Klassen TP, et al. What contributions do languages other than English make on the results of meta-analyses? J Clin Epidemiol. 2000 Sep;53(9):964-72. PMID: 11004423.

29. Juni P, Holenstein F, Sterne J, et al. Direction and impact of language bias in meta-analyses of controlled trials: empirical study. Int J Epidemiol. 2002 Feb;31(1):115-23. PMID: 11914306.

30. Moher D, Pham B, Lawson ML, et al. The inclusion of reports of randomised trials published in languages other than English in systematic reviews. Health Technol Assess. 2003;7(41):1-106. PMID: 14670218.

31. Pham B, Klassen TP, Lawson ML, et al. Language of publication restrictions in systematic reviews gave different results depending on whether the intervention was conventional or complementary. J Clin Epidemiol. 2005 Aug;58(8):769-76. PMID: 16086467.

32. Pilkington K, Boshnakova A, Clarke M, et al. "No language restrictions" in database searches: what does this really mean? J Altern Complement Med. 2005 Feb;11(1):205-7. PMID: 15750383.

33. Baussano I, Brzoska P, Fedeli U, et al. Does language matter? A case study of epidemiological and public health journals, databases and professional education in French, German and Italian. Emerg Themes Epidemiol. 2008;5:16. PMID: 18826570.

34. Morrison A, Moulton K, Clark M, et al. English-language restriction when conducting systematic review-based meta-analyses: systematic review of published studies. Ottawa: Canadian Agency for Drugs and Technologies in Health; 2009. p. 1-17.

35. Greenland S, O'Rourk K. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. Biostatistics. 2001;2:463-71. PMID: 12933636.

36. Institute of Medicine. Knowing what works in health care: A roadmap for the nation. In: Eden J, Wheatley B, McNeil B, et al., eds. Washington, DC: National Academies Press; 2008.

37. Systematic reviews: CRD's guidance for undertaking reviews in health care. York: Centre for Reviews and Dissemination, University of York, UK; 2009.

38. Cochrane Handbook for Systematic Reviews of Interventions. The Cochrane Collaboration. 2009.www.cochrane-handbook.org.

39. Berlin JA. Does blinding of readers affect the results of meta-analyses? University of Pennsylvania Meta-analysis Blinding Study Group. Lancet. 1997 Jul 19;350(9072):185-6. PMID: 9250191.

40. 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. May;63(5):513-23. PMID: 19595577.

41. Turner EH, Matthews AM, Linardatos E, et al. Selective publication of antidepressant trials and its influence on apparent efficacy. N Engl J Med. 2008 Jan 17;358(3):252-60. PMID: 18199864.