

**EPC Methods Guidance:
Avoiding Bias in Selecting Studies**

Key Points

- Inclusion criteria are derived from key questions and must be clear and sufficiently detailed to avoid inconsistent application in study selection.
- The Review team should carefully consider how inclusion or exclusion of specific populations, interventions, comparators, outcomes, timeframes, settings, or study designs and characteristics may affect the review conclusions.
- Dual review helps to reduce bias and can identify inclusion criteria that are not sufficiently clear and where subjective judgment may differ.
- Gray literature (e.g. FDA documents, trial registry reports) is important in assessing publication bias or selective outcome reporting biases.

Introduction

Although systematic reviews are intended to reduce bias compared to narrative reviews, reviewers must carefully consider how each decision in selection and inclusion of studies may introduce bias to the review conclusions. The Methods Guide for AHRQ's Evidence-based Practice Centers (EPCs) incorporates methods to identify and reduce bias in many chapters; here we focus on the methods that EPC authors can use to reduce bias in the selection and assessment of studies to be included in a systematic review (SR).

In the initial stages of a SR, topic development, the EPCs develop potential topics that have been submitted by a broad range of stakeholders. Each topic is then evaluated against predefined criteria by a panel comprising various AHRQ EHC staff, the Scientific Resource Center (SRC), and the John M Eisenberg Center for Clinical Decisions and Communication Science. AHRQ ultimately makes the final decision about topics that will move forward in the process, with subsequent refinement for specific key questions by the EPC with further input from Key informants. (1) The whole topic development and refinement process defines the general scope and specific key questions of the topic for the SR; it is, therefore, directly relevant to determining the ultimate selection of studies for a review. The methods used to develop and further refine these topics work towards reducing potential bias, and are discussed elsewhere. (1) While bias may be introduced at these early stages, this paper instead focuses on the methods that EPC authors use to operationalize key questions after they have been finalized.

Selecting studies directly influences the resulting conclusions. We carried out a search for studies that examined variation in study inclusion across SRs of the same topic and found a very small number of relevant studies. (2-5) The most relevant example was a prospective study designed to examine variation between review groups in determining study inclusion. (5) Two review groups (on different continents) were commissioned to review observational evidence on the same topic. For the papers retrieved in both searches but regarded as relevant by only one center (52 in center A and 20 in center B), 63 of the 72 discrepancies occurred in screening citations (titles and abstracts); 9 of the 72 discrepancies occurred during review of full-text articles. Of 310 relevant articles, 166 (54 percent) were

included by both groups. Of papers included by both groups, 80 percent were described by the same study design. Agreement for inclusion of cohort-type and case-control studies was only about 63 percent, and 50 percent or less for ecological and case series studies. In other studies, published systematic reviews that included trials and appeared to focus on the same research question were examined retrospectively and also differed in their lists of included studies (Table 1). (2-4) These findings led us to conclude that variation in the details or lack of adequate specificity of inclusion criteria and methods used to apply these criteria yielded quite different sets of included studies, contributing to differing conclusions. Additionally, based on the finding that there was significant discrepancy in inclusion decisions of observational studies, there may be separate issues relating to the inclusion of randomized controlled trials (RCTs) versus observational studies.

Although some differences in study inclusion decisions in the studies cited above could have been the result of one review's having used a best evidence approach (for example, using a hierarchy of study designs as reviewed by Treadwell et al [AHRQ Method Guidance – in press, cite when published]) or excluding poor quality studies and another not doing so, such decisions should have been identified in the development of, and clearly stated in, the inclusion criteria. In most cases, we could not completely distinguish the differences among reviews in inclusion decisions related to variations in inclusion criteria and those related to variations in the search strategies.

Other authors have addressed reasons for discrepant results from meta-analyses on the (seemingly) same topics. (6, 7) Ionnaidis 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.

Operationalization of the key questions (decisions about what to include as "evidence") presents an ominous source of bias and in these examples, led to different conclusions. In this chapter we outline the potential for systematic bias and random error in the study selection process of SRs and discuss specific strategies to reduce and avoid potential bias when selecting studies to include in SRs.

Setting Inclusion Criteria

Many of the same principles related to avoiding bias in selecting patients for participation in primary studies apply to avoiding bias in selecting studies for inclusion in a SR. (8) A critical first step toward avoiding bias in selecting studies is to start with valid and explicit Key Questions and inclusion criteria to guide the SR process.

Typically, the Key Questions are framed to specify the scope of the SR in broad terms and the inclusion criteria are designed to provide more detail regarding population(s), intervention(s), comparator(s), outcome(s), timing, and setting (PICOTS) of interest, and study designs or study characteristics of

interest (including specification of non-English language publications, quality criteria, study size, or other study design or features such as inclusion of particular outcomes). Although setting inclusion criteria based on Key Questions may seem straightforward, historically we can find many examples of SRs in which the questions (and thus the resulting study inclusion criteria) seemed clear at the outset, but were revealed to be at least somewhat opaque in application. Therefore, 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. When defining the inclusion criteria, EPCs should consider not only the relevance of the studies to the question, but also the potential impact that exclusion of studies may have on the review conclusions.

The criteria should be set a priori and based on the analytic framework or conceptual model. 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, nor so loose 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. Such changes should be carefully considered with input from AHRQ and the TEP and then described in the report.

PICOTS

Examples of potential biases that can be encountered with the PICOTS criteria are briefly summarized in Table 2.

Population. Inclusion criteria for the population(s) of interest should be defined in terms of relevant demographic variables, disease variables (i.e., disease stage, type, or severity), risk factors for disease, cointerventions, and coexisting conditions. (9) For example, if a 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. Selection or exclusion of a specific population may introduce bias since an intervention may be more or less effective within a specific population. Exclusion of a particular population without careful consideration may overestimate or underestimate the effectiveness or harms of an intervention. 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.

Ionnaidis has written about the potential for bias introduced into meta-analysis by reviewers intentionally varying the population characteristics used in selecting studies.(6) For example his own meta-analysis of invasive versus conservative medical management of stable chronic coronary artery disease indicated no significant benefit of the invasive treatments, a finding that was supported by a subsequent large RCT. Following publication of these results, however, a meta-analysis conducted by interventional cardiologists found a significant benefit with invasive treatment. The primary difference

between the two meta-analyses was that the invasive cardiologists selected a narrower, unstable population in which invasive management is known to be more beneficial.

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. 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. The scope of included outcomes should address both measures of effectiveness and harms. In order to reduce variation in study selection related to outcomes, we recommend that the inclusion criteria clearly identify a SR’s primary outcome when applicable, outline any restrictions on measurement methods or timing, and provide guidance for handling of composite outcomes. The inclusion criteria should document the designation of any outcome as primary, as well as clearly reflect any decision to restrict eligibility to only those studies that report the chosen primary outcome (at a minimum). As many EPC reviewers are aware, for some types of outcomes, such as pain and psychological functioning, there is often incredible variability across studies in types of scales used and timing of measurements. For clinical areas that are notoriously characterized by variability in outcome measurement methods, the risk is greater for inconsistency in study selection. In these cases, it is especially important to specify any restrictions on eligible measurement methods; i.e., only including studies that used measurement scales that have been published or validated. However, on the other hand, 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 non-academic 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, the choice 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. (7) Finally, 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. 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. For example, a decision to exclude composite outcomes may be supported by the research of Ferreira-Gonzalez et al, which demonstrated the problems with interpretation of composite endpoints comprised of component endpoints with large gradients in importance to patients and in magnitude of treatment effect. (10)

Timing and Setting. Setting inclusion criteria for time frame 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. Any decision to limit inclusion of studies based on follow-up duration should be clearly specified and based on sound clinical judgment regarding 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 to general inpatient or outpatient settings, and how inclusion or exclusion of these settings may systematically bias the conclusions.

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 non-allocation of treatment), study conduct (quality or risk of bias of individual study), language of publication, study size, or reporting of relevant data.

While the decision to include RCTs may seem to be fairly straightforward, if the criteria for what counts as an RCT is either unclear such that the review team has to make decisions ad hoc, or so narrowly defined that the review question could not be answered, there may be random error or a bias towards insufficient evidence. For example, if inclusion criteria state that a trial must be ‘properly’ randomized to be eligible, or other vague and overly strict criteria, the review may end up inappropriately concluding that there is no evidence on a particular question. (11)

Decisions about including observational studies, however, are more complicated.(12, 13) Inclusion of observational study designs has been shown to result in variability in the set of included studies, and require special care in developing and testing criteria for determining eligibility.(4) Several different types of observational studies can be considered: e.g., cohort studies, with or without a comparison group; case-control studies, case series, case reports, cross-sectional studies. For that reason, specifying the design(s) being considered for inclusion is essential. Because of the lack of consensus on any single taxonomy for assigning labels to specific types of observational study designs, (14) EPC teams should define study designs with sufficient clarity so that their reviewers can consistently and correctly determine if a given study is eligible. Inclusion decisions about observational studies may also be affected by selective outcome and analysis reporting, where only a subset of the original outcomes

measured and analyzed in a study are fully reported or analyzed. Selective outcome reporting (SOR) may occur more frequently in observational studies than in RCTs, [Norris, et al EPC Guidance Document in peer review; cite when published] such that selection decisions and ultimately the conclusions of a SR can be affected. Approaches for detecting SOR in observational studies are described elsewhere, and EPCs should be aware of these methods when considering individual observational studies for eligibility.

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, 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 comparing 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. A SR about the impact of language restrictions on summary measures in SRs and meta-analyses (MA) concluded that they could not find evidence of a systematic bias from the use of language restrictions in SR/MA of conventional medicine, but that further research was needed particularly in medical specialties and areas where there may be publications in non-English languages that are influential in the topic area. (15)

The way that poor quality studies are handled in SRs also varies and may introduce bias. Once a study has been determined to be poor quality, it may be excluded outright; included 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 included in quantitative analyses using weighting based on quality or sensitivity analysis. Deciding to include studies regardless of their quality refutes the value of the assessment that the studies have a high risk of bias, and thus including them may introduce bias in the SR. However, because assessments of quality or risk of bias are never based entirely on empirical evidence, and are subjective by nature, excluding poor quality studies outright may also introduce bias. EPCs should be explicit about how poor quality studies will be handled, a priori. If poor quality studies 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 as poor quality 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 titles and abstracts (and then full-text studies) 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”. (16) Some form of dual review should be done at each stage to reduce potential 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. Using this method, the sensitivity of the process is increased although the specificity may be somewhat reduced; the trade-off is a potentially larger pool of full-text articles to review but a lower chance of having missed an eligible study. We recommend that review teams start with a small number of citations followed by discussion such that variation in interpretation of how the inclusion criteria should be applied can be resolved early on. 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. (17, 18) 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. (19) 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 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 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. (20)

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 Grey Literature to Assess and Reduce Bias

A comprehensive search of the published literature is essential to reduce bias in a review. However, search and selection of published literature may not be sufficient to reduce potential bias in a systematic review because of the possibilities of both publication bias and outcome reporting bias. Publication bias may occur when only positive studies are published while studies with negative findings are never submitted or rejected for publication. Thus a search of grey literature (i.e. conference abstracts) may indicate when the published literature is unduly rosy. Similarly, outcome reporting bias may occur when study investigators preferentially report only positive results, ignoring outcomes measured, collected, and analyzed that did not have a positive finding. A review of original protocols if registered (i.e. with clinicaltrials.gov) may provide some insight as to whether outcome reporting bias is a problem and thus the systematic review is biased toward favorable findings. Outcome reporting bias is the reason EPC reviewers should be cautious about excluding studies simply because they do not report the outcomes of interest. 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. Grey literature helps to provide some fuzzy information on areas that were previously a blind spot in systematic reviews of only published literature.

While there may be variation in how reviewers define 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 Box for descriptions). (21)

In reviewing gray literature documents, reviewers are seeking to identify unpublished studies and unpublished data supplemental to published studies. 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 evidence of publication bias.(22) In some cases, enough information may be available on unpublished studies for the reviewers to assess study quality and include the study in the SR.

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 grey literature documents.
2. Assess studies in grey literature documents for eligibility in the SR using the key questions and inclusion criteria as discussed above.
3. As some sources of grey literature will have overlap with published literature, e.g. FDA documents and trial registries, reviewers should match studies in grey literature documents to those found in published literature to remove any duplicates.
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 grey 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.

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. (23) 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 non-inclusion of these subgroups was an oversight.

Discussion

Our review of the literature indicates that systematic bias and random error can occur in the selection of studies for systematic reviews. 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 during topic development and topic refinement (discussed in further detail by Whitlock et al(1)). 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 thoughtful 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 systematic review.

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 systematic review 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. 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 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.(13)

Conclusion

Overall, 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. To minimize study selection bias, reviewers are advised to adhere to the processes outlined here.

References

1. Whitlock EP, Lopez SA, Chang S, Helfand M, Eder M, Floyd N. 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.
2. Cook DJ, Reeve BK, Guyatt GH, Heyland DK, Griffith LE, Buckingham L, et al. Stress ulcer prophylaxis in critically ill patients. Resolving discordant meta-analyses. *JAMA.* 1996 Jan 24-31;275(4):308-14.
3. Hopyayan K, Mugford M. Conflicting conclusions from two systematic reviews of epidural steroid injections for sciatica: which evidence should general practitioners heed? *British Journal of General Practice.* 1999 Jan;49(438):57-61.
4. Peinemann F, McGauran N, Sauerland S, Lange S. Disagreement in primary study selection between systematic reviews on negative pressure wound therapy. *BMC Medical Research Methodology.* 2008 Jun 26;8(1):41.
5. Thompson R, Bandera E, Burley V, Cade J, Forman D, Freudenheim J, et al. Reproducibility of systematic literature reviews on food, nutrition, physical activity and endometrial cancer. *Public Health Nutrition.* 2007 Dec 6:1-9.
6. Ioannidis J. Meta-research: The art of getting it wrong. *Research Synthesis Methods.* 2011; 1:169-184.
7. Tendal B, Higgins JP, Juni P, Hrobjartsson A, Trelle S, Nuesch E, et al. Disagreements in meta-analyses using outcomes measured on continuous or rating scales: observer agreement study. *BMJ.* 2009;339:b3128.
8. Oxman AD. Checklists for review articles. *BMJ.* 1994 Sep 10;309(6955):648-51.
9. West S, Gartlehner G, Mansfield AJ, Poole C, Tant E, Lenfestey N, et al. Comparative Effectiveness Review Methods: Clinical Heterogeneity. *Methods Research Paper AHRQ Publication No 10-EHC070-EF Available at <http://effectivehealthcareahrq.gov/>.* 2010;September.
10. Ferreira-Gonzalez I, Permanyer-Miralda G, Busse JW, Bryant DM, Montori VM, Alonso-Coello P, 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.
11. Henderson A, Henderson S. Provision of a surgeon's performance data for people considering elective surgery. *Cochrane Database of Systematic Reviews.* 2011;1:1.
12. Chou R, Aronson N, Atkins D, Ismaila AS, Santaguida P, Smith DH, 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.

13. Norris S, Atkins D, Bruening W, et al. Selecting observational studies for comparing medical interventions. Agency for Healthcare Research and Quality. 2010;Methods Guide for Comparative Effectiveness Reviews . Rockville, MD. Available at:
http://www.effectivehealthcare.ahrq.gov/ehc/products/196/454/MethodsGuideNorris_06042010.pdf.
14. Hartling L, Bond K, Harvey K, Santaguida P, Viswanathan M, Dryden D. Developing and Testing a Tool for the Classification of Study Designs in Systematic Reviews of Interventions and Exposures. Agency for Healthcare Research and Quality. 2010. ;Methods Research Report. AHRQ Publication No. 11-EHC-007. Available at <http://effectivehealthcare.ahrq.gov/>.
15. Moher D, Pham B, Lawson ML, Klassen TP. The inclusion of reports of randomised trials published in languages other than English in systematic reviews. Health Technology Assessment.. 2003;7(41):1-106.
16. Institute of Medicine. Knowing what works in health care: A roadmap for the nation. Washington, DC: National Academies Press; 2008. Available from:
<http://www.iom.edu/Reports/2008/Knowing-What-Works-in-Health-Care-A-Roadmap-for-the-Nation.aspx>.
17. Systematic reviews: CRD's guidance for undertaking reviews in health care. York: Centre for Reviews and Dissemination, University of York; 2009.
18. Cochrane Handbook for Systematic Reviews of Interventions Higgins J, Green S, editors: The Cochrane Collaboration www.cochrane-handbook.org; .2009.
19. 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.
20. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. International Journal of Surgery. 2010;8(5):336-41.
21. Relevo R, Balshem H. Finding Evidence for Comparing Medical Interventions. . Agency for Healthcare Research and Quality 2011;Methods Guide for Comparative Effectiveness Reviews. AHRQ Publication No. 11-EHC021-EF. Available at <http://effectivehealthcare.ahrq.gov/>.
22. Owens DK, Lohr KN, Atkins D, Treadwell JR, Reston JT, Bass EB, 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.
23. Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. New England Journal of Medicine. 2008 Jan 17;358(3):252-60.

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? (3)	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 2 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 and 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. (4)	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 1 of the 5 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 most broad 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	This study aimed to resolve discrepancies in 4 previous	From abstract: "The source of discrepancies between prior meta-analyses included incomplete identification of relevant studies,

<p>Prophylaxis in Critically ill patients: resolving discordant meta-analyses. (2)</p>	<p>systematic reviews and provide estimates of the effect of stress ulcer prophylaxis on gastrointestinal bleeding, pneumonia, and mortality in critically ill patients.</p>	<p>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.”</p> <p>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.</p>
--	--	--

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

PICOTS Criteria	Hypothetical 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.
Time frame	Not defined.	Reviewers may report whatever is available in the literature, which may be short term studies.	Without pre-specifying that longterm outcomes are essential and only reporting short term outcomes, reviewers may overestimate effectiveness of treatment.
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

PICOTS Criteria	Hypothetical 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 non-academic 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 non-academic or urban populations which may have higher severity of disease, and over-estimate 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.

Box .

Sources of Unpublished Information for Comparative Effectiveness Reviews	
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 document 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 Scientific Information Packets (SIPs),(21) 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.