Methods Research Report

A Framework for "Best Evidence" Approaches in Systematic Reviews

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

Methods Research Report

A Framework for "Best Evidence" Approaches in Systematic Reviews

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
http://www.ahrq.gov

Contract No. HHSA 290-2007-10063-I

Prepared by:
ECRI Institute Evidence-based Practice Center
Plymouth Meeting, PA

Investigators:
Jonathan R. Treadwell, Ph.D.
Sonal Singh, M.D., M.P.H.
Ripple Talati, Pharm.D.
Melissa L. McPheeters, Ph.D., M.P.H.
James T. Reston, Ph.D., M.P.H.

AHRQ Publication No. 11-EHC046-EF
June 2011

This report is based on research conducted by the ECRI Institute Evidence-based Practice Center under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. HHSA 290-2007-10063-I). The findings and conclusions in this document are those of the author(s), who are responsible for its content, and do not necessarily represent the views of AHRQ. 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 clinicians, employers, policymakers, and others make informed decisions about the provision of health care services. This report is intended as a reference and not as a substitute for clinical judgment.

This report may be used, in whole or in part, as the basis for the 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 or actions may not be stated or implied.

No investigators have any affiliations or financial involvement (e.g., employment, consultancies, honoraria, stock options, expert testimony, grants or patents received or pending, or royalties) that conflict with material presented in this report.

Suggested citation: Treadwell J, Reston J, Singh S, Talati R, McPheeters M. A Framework for "Best Evidence" Approaches in Systematic Reviews. Methods Research Report. (Prepared by the ECRI Institute Evidence-based Practice Center under Contract No. HHSA 290-2007-10063-I.) AHRQ Publication No. 11-EHC046-EF. Rockville, MD: Agency for Healthcare Research and Quality. June 2011. Available at: www.effectivehealthcare.ahrq.gov/reports/final.cfm.

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.

To improve the scientific rigor of these evidence reports, AHRQ supports empiric research by the EPCs to help understand or improve complex methodologic issues in systematic reviews. These methods research projects are intended to contribute to the research base in and be used to improve the science of systematic reviews. They are not intended to be guidance to the EPC program, although may be considered by EPCs along with other scientific research when determining EPC program methods guidance.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers as well as the health care system as a whole by providing important information to help improve health care quality. The reports undergo peer review prior to their release as a final report.

We welcome comments on this Methods Research Project. 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 e-mail 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, Evidence-based Practice Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Christine Chang, M.D., M.P.H.
Task Order Officer
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Acknowledgments

The research team would like to acknowledge the efforts of Eileen Erinoff, M.S.L.I.S., and Helen Dunn for providing literature retrieval and documentation management support; and Janice Kaczmarek, M.S., Gina Giradi, M.S., and Kitty Donahue for program management, editorial support, and administrative support.

Peer Reviewers

Laura J. Fochtmann, M.D.
Professor of Psychiatry and Behavioral Sciences
Stony Brook University School of Medicine
Stony Brook, NY

Susan Norris, M.D., M.Sc., M.P.H.
Assistant Professor, Department of Medical Informatics and Clinical Epidemiology
Oregon Evidence-based Practice Center
Oregon Health & Science University
Portland, OR

Diana Pettiti, M.D., M.P.H.
Professor, Biomedical Informatics
Arizona State University
Phoenix, AZ

A Framework for "Best Evidence" Approaches in Systematic Reviews

Structured Abstract

Objectives. Reviewers often employ a "best evidence" approach to address the key questions, but what is meant by "best" is often unclear. The goal of this project was to create a decision framework for "best evidence" approaches in systematic reviews. This document is not intended to be prescriptive, but rather to provide a conceptual construct to enhance the transparency of inclusion decisions made during the course of a systematic review.

Review Methods. We set out to accomplish the following tasks: (1) create a list of possible inclusion criteria, and for each criterion, create a list of factors that might affect a reviewer's decision to use it, (2) create a list of evidence prioritization strategies, and (3) list the ways in which evidence prioritization strategies might be formally evaluated. In a series of conference calls, collaborators from five Evidence-based Practice Centers discussed methods for accomplishing the tasks noted above. After the initial conference call, the project leaders prepared a series of discussion documents specific to the first three tasks. Subsequent conference calls were scheduled to discuss comments and suggestions from the collaborators, whose feedback was incorporated in revisions in the task documents. The document was then externally reviewed by experts from other institutions, and revisions were made based on reviewer comments.

Results. For Task 1, we identified 21 potential inclusion criteria and 15 modifying factors a reviewer should consider when deciding which criteria to employ. The inclusion criteria were divided into three categories: criteria pertaining to study design, criteria pertaining to study conduct and reporting, and criteria pertaining to relevance. A flow chart of the decision process provides a guide to reviewers, and tables illustrate the factors influencing decisions about each inclusion criterion. For Task 2, we identified four strategies for prioritizing evidence. For Task 3, we identified a number of potential approaches that might be used to formally evaluate these strategies in the future.

Conclusions. Systematic reviewers routinely prioritize evidence in numerous ways. This paper provides a framework for understanding the possibilities, considering influential factors, and choosing among the myriad options. This will help enhance the transparency of review processes, which in turn may help users determine how different reviews of the same topic can reach different conclusions.

Contents

Introduction

Systematic reviewers often employ a "best evidence" approach to address the key questions in the reviews. What is meant by "best," however, is often unclear. Clearly, some manner of evidence prioritization (i.e., prioritizing some studies over others) is employed by all systematic reviews. This prioritization can help ensure (but cannot guarantee) that the review's conclusions will stand the test of time.

The phrase "best evidence" was used by Slavin in a 1995 article as an "intelligent alternative" to meta-analysis.[1] Instead, this paper uses "best evidence" to refer to any strategy for prioritizing evidence, regardless of whether that evidence is combined quantitatively in a meta-analysis.

This paper encompasses different interpretations of the phrase "best evidence." Some reviewers may interpret it to mean the "best available evidence," and would therefore always include at least one study in the evidence base for a Key Question. Other reviewers may believe this approach to be too lenient, because the best available evidence may be too biased and potentially misleading, thus sometimes no studies should be included. These latter reviewers can be said to use a threshold interpretation of "best evidence." Both interpretations fit within the larger framework of this paper.

Existing guidance from the Agency for Healthcare Research and Quality (AHRQ) Effective Health Care (EHC) Program addresses the notion of "best evidence" in at least two areas: the study inclusion criteria[2] and the inclusion of nonrandomized studies of beneficial effects. Granted, the study inclusion criteria are not normally considered to define the "best evidence," but rather they typically define the relevant evidence, and the "best evidence" is a subset of what is relevant. Nevertheless, inclusion criteria implicitly prioritize evidence, for example, the inclusion of studies only of a certain design or a certain minimum number of study participants. Studies failing the inclusion criteria receive zero priority. Thus, we place inclusion criteria within the relatively large network of decisions encompassing "best evidence."

The second relevant area of existing EHC guidance is the chapter on when to include nonrandomized studies of beneficial effects.[3] This chapter acknowledges that for many topics in comparative effectiveness, the randomized evidence is insufficient to answer the Key Question. This may be due to poor applicability, low precision, risk of bias (based on other problems with the study's design or conduct), or other factors. The insufficiency of randomized evidence necessitates the consideration of nonrandomized evidence, which may or may not lead to a conclusion, but at least it should be considered in an effort to reach a conclusion. This represents a specific example of a "best evidence" approach in which a reviewer may potentially include nonrandomized evidence as long as their risk of bias is not too high. This approach involves a consideration of the results of randomized trials (i.e., their conclusiveness) when considering whether to include nonrandomized evidence. However, this staged approach should be planned a priori to avoid the possible bias of trial results directly influencing study inclusion decisions.

Using randomization alone as a basis for prioritization is one example, and many other prioritization schemes are possible. For example, within a set of identified randomized trials, the variation in risk-of-bias can be considerable, and many systematic reviewers have subprioritized randomized trials in various ways. One approach is to include only blinded randomized trials, thereby employing a best-evidence approach at the level of the inclusion criteria (for examples, see references 4–7). Clearly, this is only possible if an evidence base contains many randomized studies and the reviewer has the luxury of excluding unblinded randomized trials. Conversely, in

the absence of randomized trials, there can be considerable variability in designs of nonrandomized studies, and a subprioritization of these can be easily justified (e.g., based on whether the authors matched groups at baseline). For examples of reviewers subprioritizing nonrandomized studies, see references 8–12.

In addition to using the inclusion criteria as a vehicle for prioritizing evidence, many other approaches are possible. Some reviews may contain a set of studies that are included and tabled, but not actually analyzed. Some may distinguish between qualitative analysis and quantitative analysis, and perform a meta-analysis on only the highest priority subset of studies. Some may formulate the review conclusions based only on a higher priority subset, or rate the strength of evidence based only on the higher priority subset. These activities, while very different in implementation, all serve to draw the reader's attention towards some studies and away from other studies, and they are discussed in a later section of this report.

Overall, evidence prioritization is a common and necessary practice in systematic reviews. However, the variety of dilemmas facing reviewers, some of which are unanticipated, has spawned innumerable approaches, with no organizing framework. This absence of guidance was the impetus behind this project.

Before we describe the objectives and methods of the project, we list three caveats:

1. Different topics demand different approaches, and it is not the purpose of this document to recommend any single approach. Thus, we do not recommend some prioritization strategies over others.
2. None of the strategies require meta-analysis, and also none preclude meta-analysis. Thus, the framework is independent of how the results of different studies are considered together.
3. Any of the strategies can potentially result in a judgment that the evidence is insufficient to answer the Key Question. Some strategies do consider the conclusiveness of the evidence when prioritizing evidence (such as the aforementioned EHC chapter on when to include nonrandomized studies), whereas others do not. None, however, can guarantee an answer to the Key Question.

Essentially, this paper addresses a reviewer's decisions about lowering the evidence threshold. Why might reviewers do this? How can it be done? When does one stop lowering the bar? The following sections flesh out answers to these questions and are intended to map out numerous options for systematic reviewers.

Objectives

This project seeks to outline a framework for evidence prioritization, that is, for defining the "best evidence." This framework can improve transparency and also suggest alternatives to reviewers as they make difficult decisions. The intended audience for this document is systematic reviewers with an interest in methodology. As noted above, this document is not intended to be prescriptive, but rather to provide a conceptual construct with options to aid the decisionmaking process. It is only designed for evaluation of evidence on the benefits and harms of interventions; procedures for evaluating other types of evidence bases (e.g., diagnostic studies) are beyond the scope of this report. This is phase 1 of a larger project (phase 2 would involve a formal evaluation of the impact of variations in inclusion criteria on a review's conclusions). Led by the ECRI Institute Evidence-based Practice Center (EPC), this project set out to accomplish the following tasks:

1. Create a list of possible inclusion criteria, and for each criterion, create a list of factors that might affect a reviewer's decision to use it.
2. Create a list of evidence prioritization strategies.
3. List the ways in which evidence prioritization strategies might be formally evaluated.
4. Prepare a summary report for posting on the AHRQ Web site.

Methods/Approaches

In a series of conference calls, the two project leaders from ECRI Institute and three collaborators from three other EPCs (Vanderbilt University, University of Connecticut, and Johns Hopkins University) discussed methods for accomplishing the tasks noted above. After the initial conference call, the project leaders prepared a series of discussion documents specific to the first three tasks. Subsequent conference calls were scheduled to discuss comments and suggestions from the collaborators, whose feedback was incorporated in revisions in the task documents. The latter were combined into a single draft summary report approved by all members of the group prior to submission to AHRQ. The document was then externally reviewed by experts from other institutions, revisions were made based on reviewer comments, and the final report was re-submitted for posting on the AHRQ Web site.

Task 1: Lists of Inclusion Criteria and Factors That Might Affect a Reviewer's Decision to use Each Criterion

Two basic types of inclusion criteria are typically used in systematic reviews. The first set includes criteria pertaining to publication characteristics, such as full-article publication (not just an abstract), peer-reviewed publication, year of publication sufficiently recent (to ensure exclusion of outdated technologies), English-language publication (depending on the topic), and exclusion of duplicate publications (to avoid double-counting study participants) unless duplicate studies contain unique outcome data. These criteria are usually unaffected by subsequent decisions regarding "best evidence" and analysis.

The second set includes criteria pertaining to study design, study conduct and reporting, and study relevance to the Key Question(s). These criteria are context sensitive and require clinical and methodological judgments from the review team; in addition, the decision to use certain criteria may be influenced by the limitations of the available evidence (discussed in more detail later in this section). Given their importance, our focus in task 1 was this latter set of inclusion criteria.

Figure 1 illustrates the logical flow for application of inclusion criteria from a best evidence perspective. Note that this figure depicts a sequence of decisions rather than a hierarchy based on study design. The layout of the figure (with randomization at the top) may give an unintended impression: that the most important consideration is whether to limit the evidence to randomized controlled trials (RCTs). In reality, that may not be the most important consideration, particularly for Key Questions that do not address causation. The figure is structured in this manner because the decision to require or not require RCTs is often the first one made, and therefore leads naturally to other types of decisions about the inclusion criteria.

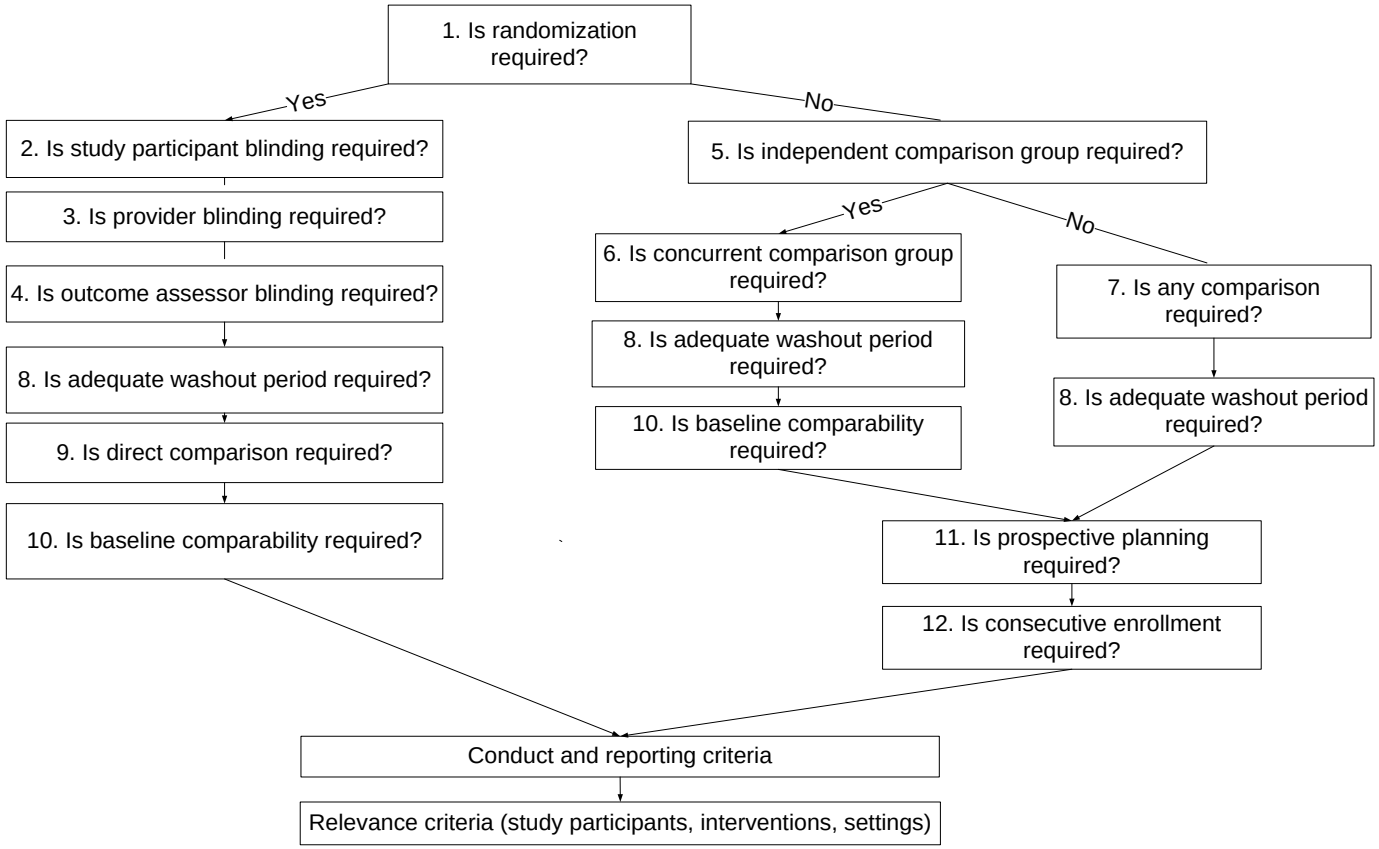
The relevance criteria involve whether the study participants, interventions, and settings are relevant to the Key Question. For example, suppose a Key Question specifies that the population of interest is adults with type 2 diabetes, and some studies enrolled not only these participants but also some adults with type 1 diabetes (and only presented combined results for the two populations). The reviewer must decide whether the combined results are sufficiently relevant to the Key Question. By "relevance," we do not mean relevance to typical clinical practice, which is a concept we refer to as applicability, and is generally addressed at a later stage in the review (see task 2 in this paper).

Table 1, Table 2, and Table 3 list inclusion criteria and the factors that may affect a reviewer's decision to use each criterion. For example, if no RCTs are identified, the reviewer may consider inclusion of nonrandomized studies (if the risk of bias is not too high). Likewise, if the outcome is a harm related to treatment, the reviewer may believe that nonrandomized studies still provide useful information. This is not to imply that well-designed studies measuring harms are suboptimal for determining the true risks of harms. Rather, in some instances, less reliable evidence (even case reports) of rare harms associated with an intervention may be useful in decisionmaking.

For some reviewers, all criteria may be influenced by the number of studies that met that criterion. Rigid adherence to criteria that none of the available studies meet may result in exclusion of a considerable amount of lower quality evidence that might have provided some (albeit weak) evidence to address a Key Question. Ultimately, the reviewer must decide whether modifying initial criteria to allow for inclusion of lower quality evidence would result in inclusion of evidence with an unacceptably high risk of bias. In the latter instance, the reviewer may decide to keep the initial criteria, even if they result in no included studies for the Key Question.

Some reviewers may select a subset of these criteria for study presentation (encompassing studies whose data will be tabled but not necessarily analyzed) and a different subset of criteria for study analysis (studies that met the criteria for presentation and also criteria for analysis). We refer to the latter subset of criteria as analysis criteria. For example, a reviewer might choose "concurrent comparison group" as a criterion for study presentation and "random assignment to intervention groups" as an analysis criterion. In this case the reviewer would tabulate information from all studies with concurrent comparison groups, but only analyze data from RCTs. Alternatively, some reviewers may choose to have only one set of criteria such that any studies that are included will also be analyzed (quantitatively if appropriate, qualitatively if not). In either case, reviewers may choose from the list of criteria presented in Table 1, Table 2, and Table 3.

Figure 1. Process chart of application of inclusion criteria*



* This figure depicts a sequence of decisions about study inclusion criteria. Randomization is listed first because often the decision to require or not require RCTs is the first one made, and therefore leads naturally to other types of decisions about the inclusion criteria. The figure also illustrates the dependencies among criteria. For example, if a reviewer requires 1 (randomization to groups), then logically the reviewer is also requiring 5 (presence of concurrent comparison group), 6 (presence of independent comparison group), 11 (prospective), and 12 (consecutive). Similarly, if a reviewer does not require 1 (randomization to groups), then logically the reviewer is not requiring 2 (study participant blinding), and 3 (provider blinding). While in theory nonrandomized studies could employ any of these types of blinding, in practice this rarely occurs.

Table 1. Inclusion criteria/analysis criteria pertaining to study design

Criterion	Factors Influencing the Decision to Employ This Criterion (Modifying Factors)
Random assignment to intervention groups	• Number of studies that did this • Existence of important unmeasured confounders • Degree of relevance to Key Questions in studies that did this • Whether the outcome is a harm • Is randomization ethical (e.g., is there equipoise between interventions)?
Blinding of study participants to which intervention they received	• Number of studies that did this • How difficult it is to blind study participants or maintain blinding • Whether study participant knowledge of intervention group can influence outcomes • Degree of relevance in studies that did this • Whether the outcome is a harm • Is blinding ethical? (e.g., would it require an unethical comparison group)
Blinding of providers to the intervention that they provided	• Number of studies that did this • How difficult it is to blind providers or maintain blinding • Whether provider knowledge of intervention group can influence outcomes • Degree of relevance in studies that did this • Whether the outcome is a harm
Blinding of outcome assessors to the assigned intervention	• Number of studies that did this • How difficult it is to blind outcome raters or maintain blinding • Whether outcome raters' knowledge of intervention group can influence outcomes • Degree of relevance in studies that did this • Whether the outcome is a harm
Presence of independent comparison group	• Number of studies that did this • Whether disease course allows intervention effects to be predicted accurately without an independent comparison group • If there would be substantial carryover effect if study participants received both interventions sequentially • Whether the outcome is a harm
Presence of concurrent comparison group	• Number of studies that did this • Whether disease course allows intervention effects to be predicted accurately without a concurrent comparison • If concomitant interventions differed between periods • If other aspects of intervention and/or followup changed over time (i.e., institutional changes) • Whether the outcome is a harm

Table 1. Inclusion criteria/analysis criteria pertaining to study design (continued)

Criterion	Factors Influencing the Decision to Employ This Criterion (Modifying Factors)
Presence of a before-after comparison	• Number of studies that did this • Whether disease course allows intervention effects to be predicted accurately without a comparison • If concomitant interventions differed between periods • If there would be substantial carryover effect if study participants received both interventions sequentially • Whether the outcome is a harm
Adequate washout period	• Number of studies that did this • Possibility of substantial carryover effect
The intervention comparison must be direct*, not indirect	• Number of studies that did this • If there exists evidence in which study participants and interventions were sufficiently similar in different types of studies
Good baseline comparability	• Number of studies that did this
Prospective planning (the study question was determined before any data were collected)	• Number of studies that did this
Consecutive enrollment	• Number of studies that did this

*By direct comparison, we mean that a study or studies contains the comparison groups specified in the Key Question (e.g., treatment A vs. treatment B). Indirect comparison means that the studies only contain one of the necessary comparison groups (e.g., treatment A vs. placebo, or treatment B vs. placebo), and the comparison groups of interest can only be compared across studies. Indirect comparisons require careful scrutiny to determine whether patients and interventions are sufficiently similar across studies.

Table 2. Inclusion criteria/analysis criteria pertaining to study conduct and reporting

Criterion	Factors Influencing the Decision to Employ This Criterion (Modifying Factors)
The minimum length of followup was x	• Number of studies that did this • Minimum time needed to evaluate intervention effect
The minimum number of study participants per group who provided data was x	• Number of studies that did this • Frequency of outcome occurrence
The minimum % of enrollees who provided data was x%	• Number of studies that did this • If the enrollees who did not provide data could be very unlike those who did provide data
Percent difference between groups in proportion of study participants who had usable data (only studies with comparison groups)	• Number of studies that did this
Calculable or imputable effect size	• Number of studies that did this • Whether meta-analysis of this outcome will be performed
Study must have used validated method of outcome measurement	• Number of studies that did this

Table 3. Inclusion criteria/analysis criteria pertaining to relevance

Criterion	Factors Influencing the Decision to Employ This Criterion (Modifying Factors)
Study participant characteristics are sufficiently similar to the participants specified in the Key Question (e.g., adults vs. children)	• Number of studies that did this
Interventions (both treatment and comparator) sufficiently similar to the interventions specified in the Key Question	• Number of studies that did this
Setting characteristics are sufficiently similar to the setting specified in the Key Question (e.g., inpatient vs. outpatient)	• Number of studies that did this

A tabulation between each inclusion criterion and each modifying factor is illustrated in Appendix A. In that representation, a check mark indicates that a specific criterion (row) is influenced by a specific modifying factor (column).

Inclusion criteria developed when a reviewer has insufficient knowledge of an evidence base sometimes require modification based upon findings of the initial literature searches or even review of retrieved study data. As noted in the AHRQ Methods Guide for Comparative Effectiveness Reviews,[3] conditional modification of inclusion criteria can still be a priori as long as it is specified in the review protocol (e.g., if Type A studies are not available, Type B studies will be included). For many topics, the best possible evidence (e.g., RCTs with the lowest risk of bias) does not exist, and this absence may not be discovered until the reviewer scans the literature search results. If the initial inclusion criteria specified studies directly comparing specific interventions, the criteria might be modified to allow for indirect comparisons. Conversely, for other topics, overly broad inclusion criteria (e.g., allowing nonrandomized studies or indirect comparisons) may be impractical within the restrictions of time and budget; these criteria may be narrowed to include only the "best" evidence.

Task 2: List of Evidence Prioritization Strategies

After the set of included studies for the Key Question is determined (based on task 1), a reviewer must decide which studies comprise the "best evidence" set. We define this as the set of studies that will be assessed and/or analyzed in an attempt to answer the Key Question. Reaching this answer may or may not involve meta-analysis.

Studies not considered as part of the "best evidence" set, but still included, would be tabled but not otherwise used. Some reviewers may choose to use all included studies in the attempt to draw evidence-based conclusions. If so, then the full list of included studies already defines the "best evidence" set.

Sometimes, however, the included studies are so variable in their risk of bias and/or applicability that some further prioritization is necessary. In this effort, several strategies can be employed. The simplest strategy would be to take the single "best" study, and using it alone, determine what conclusions can be drawn. The definition of "best" would be based on a careful balance of both risk of bias and applicability. For example, this strategy might be employed when evaluating an evidence base that contains a single, high-quality mega-trial and a few smaller trials of clearly lesser quality. Alternatively, a single smaller high-quality trial might represent the best evidence in other circumstances.

The single-best-study approach has the advantage of maximizing quality (i.e., minimizing risk of bias and maximizing applicability). However, it has three disadvantages: (1) the lack of scientific replication of findings, (2) the inability to determine consistency across studies (e.g., heterogeneity of effect sizes), and (3) the likelihood of low statistical power (if the study is not a mega-trial) precluding an answer to the Key Question (resulting in an evidence grade of Insufficient). However, this latter consideration should not influence a reviewer's choice if the remaining evidence (outside of the "best" study) is inapplicable or has an unacceptably high risk of bias.

A second strategy is to add studies that, relative to the single best study, are more susceptible to bias and/or less applicable. This permits measurement of cross-study consistency, and increases power. However, this strategy does not explicitly consider whether the "best set" actually permits a conclusion.

This suggests a third strategy, which involves a further lowering of the bar: admit still lower quality studies into the formal analysis, to increase the chance of obtaining an answer to the Key Question. This approach underscores a tradeoff: increasing quantity in this way will also increase the risk of an inappropriate conclusion, because the just-added studies are of lower quality. An example of this third strategy can be found in the AHRQ Methods Guide chapter and recent paper by Norris et al.,[3] which recommended that study inclusion decisions be influenced by whether the results of RCTs permit a conclusion (see Introduction for more details on this chapter). Note that this third strategy does not guarantee an answer to the Key Question, but does consider conclusiveness for the purpose of evidence prioritization.

One core component of the strength of the evidence is precision. For example, strength cannot be considered high if there is a wide confidence interval with respect to a decision threshold (e.g., precision is too wide to determine whether a difference can be considered clinically significant). Adding still more studies to the evidence base will increase the chance of obtaining a narrow confidence interval (unless the results show substantial heterogeneity), which may in turn increase the overall strength of the evidence. The resulting increase in overall risk-of-bias, however, may negate this possibility. This represents a fourth strategy: consider not only conclusiveness, but also the strength of the evidence, when making prioritizations.

With any of the above strategies, a decision to include lower quality evidence means that all studies on that level should be included. The selective inclusion based on study results or observed consistency with higher quality evidence would introduce bias. Thus, neither strategy 3 nor 4 involve the exclusion of outlier studies in an attempt to reach a conclusion or increase evidence strength. With strategy 4, if the newly included studies inclusion do not increase precision and thereby increase the overall strength of evidence, then the reviewer should exclude all studies on this level and only evaluate evidence from higher quality studies.

For strategies 3 and 4, the potential disadvantage of adding lower quality studies (increased risk of bias) is somewhat minimized in that lower quality studies can only increase the strength of the evidence if the findings are consistent with the findings of higher-quality studies. For example, if two higher quality studies together lead to a low strength of evidence, additional lower level studies would only boost the strength to moderate if they generally agreed with higher level studies.

Additional issues occur when the only available evidence is low in quality. Studies of low quality with biases in opposite directions might have similar (consistent) effect sizes, which could lead to an overestimate in the strength of evidence. Furthermore, consistency or precision in the findings of a low-quality evidence base does not change the fact that the evidence is low quality.

Table 4 outlines the four strategies. Checkmarks indicate which facets of the evidence are explicitly considered during evidence prioritization. All four strategies consider both risk of bias and applicability in prioritizing evidence. The specific implementations could involve:

- The use of a criterion that was not employed for study inclusion (e.g., in a group of included RCTs, define the “best evidence” set as those studies that blinded study participants)
- The use of a more stringent threshold (e.g., in a group of studies that all reported data on at least 50 percent of study participants, define the “best evidence” set as those that reported data on at least 80 percent of study participants),
- The combination of several criteria involving risk-of-bias and applicability

Note that these implementation approaches are derived from the earlier list of potential inclusion criteria for selection of individual studies (Task 1).

Strategies 2–4 further consider both replication and cross-study consistency; strategies 3 and 4 consider whether the evidence is sufficient to permit an answer to the Key Question; and only 4 attempts to maximize the strength of the evidence underpinning the conclusion for that Key Question. Note that a conclusion is still possible using strategies 1 and 2, but strategies 3 and 4 are the only strategies that explicitly use the conclusiveness of the evidence as a factor.

Selecting an evidence prioritization strategy involves a number of tradeoffs. Although strategy 1 (best single study) is the most feasible and has a low risk of leading to an inappropriate conclusion, it also has a high risk of an inappropriate lack of conclusion. At the other extreme, strategy 4 is the least feasible as it may require analysis of a large number of studies, and it has a high risk of an inappropriate conclusion due to inclusion of lower quality studies; however, due to its greater statistical power it has the lowest risk of an inappropriate lack of conclusion. Strategies 2 and 3 allow for an intermediate level of tradeoffs between these two extremes.

A reviewer may specify in the protocol that they will initially use a more stringent strategy regarding study inclusion, but if the resulting evidence is insufficient to permit a conclusion, they may choose a less stringent strategy to increase the chances of reaching a conclusion. However, there is no guarantee that inclusion of lower quality studies will permit a conclusion. Even if a large amount of evidence is available, problems in quality or consistency or precision may preclude a conclusion.

As noted in the Introduction, the reviewer is free to decide whether meta-analysis is appropriate for a given evidence base. If so, a reviewer may choose to synthesize different bodies of evidence (e.g., RCTs and nonrandomized studies) separately and then decide whether the lower quality body of evidence may be used to enhance the overall strength-of-evidence rating.

Table 4. Strategies for defining the "best evidence" set for a given Key Question

Strategy	Risk of bias*	Applicability*	Replication*	Conclusiveness*	Overall Evidence Strength*
1. Single best study	Yes	Yes	No	No	No
2. Best set of studies	Yes	Yes	Yes	No	No
3. Best set of studies, and also consider conclusiveness	Yes	Yes	Yes	Yes	No
4. Best set of studies, and also consider conclusiveness and evidence strength	Yes	Yes	Yes	Yes	Yes

* During the prioritization of evidence, is this factor explicitly considered by a given strategy?

Task 3: Methods for Evaluating Evidence Prioritization Strategies

The tradeoffs inherent in the choice of prioritization strategy raise at least two important questions. The first is whether different strategies on average would lead to similar or different conclusions. The second question is which strategy leads to the "most appropriate" conclusions. For a meta-analysis, this would include the best estimate of the true effect size with the highest strength of evidence.

The use of an alternate prioritization strategy can be viewed as a sensitivity analysis of inclusion criteria. For example, the Cochrane Handbook for Systematic Reviews of Interventions (version 5.0.2, September 2009)[13] recommends numerous sensitivity analyses, especially for review decisions that were arbitrary or unclear. These include the addition or removal of studies wherein poor reporting made it difficult to determine whether they met inclusion criteria; changing criteria about participants (e.g., age range); changing criteria about interventions (e.g., doses); changing criteria about outcomes (e.g., length of followup); changing criteria about study design (e.g., whether to include randomized studies with unblinded outcome assessment).[13] Similar recommendations have been made by several other authors.[14-18]

Note that one can consider two types of conclusions: one about the size of the effect, or one about the direction of the effect. One possible output of a strategy is that there is insufficient evidence, which reflects a non-conclusion about the size and direction of the effect. This may be the most appropriate reviewer decision. Also note that one could compare strategies not only on the conclusions drawn, but also on the strength of the conclusions drawn.

Question 1: Do Different Strategies Lead to Similar or Different Conclusions?

Addressing this question requires using methods that compare these strategies in systematic reviews. Three alternatives might be used.

Method 1: Compare Published Systematic Reviews

A literature search could identify and compare the conclusions of different systematic reviews that used different prioritization strategies to address the same clinical question. The advantage of this method is its relative ease of implementation. Provided a reviewer can find published reviews that addressed the same clinical question using different strategies, the comparison of the reviews' conclusions can be done relatively quickly. Although this would be the least labor-intensive method, it has some drawbacks. First, it may be difficult to identify clinical questions where different systematic reviews used different prioritization strategies. Second, the systematic reviews may have differed in other methodological areas, such as risk-of-bias assessment and strength of evidence assessment, which could then lead to differences in conclusions among reviews. This would make it difficult to determine whether different evidence prioritization strategies truly led to different conclusions, or whether they would have led to the same conclusion if the reviews had been similar in other methodological areas. In addition, methodology is not always well reported in published systematic reviews, often simply due to article length limitations in journals.

Method 2: Test the Robustness of an Existing Systematic Review

A reviewer could identify a single existing systematic review, determine its evidence prioritization strategy (by examining the report inclusion criteria), and test other prioritization strategies on the same evidence base, while keeping all other methodology the same. The advantage of this method over method 1 is that other methodological aspects of review (e.g., risk-of-bias assessment) would no longer confound the comparison. However, this method is more labor-intensive than method 1, as it requires performing independent research synthesis using the other prioritization strategies.

Method 3: Initiate a Systematic Review and Compare Prioritization Strategies

A reviewer could initiate a systematic review of a given clinical question and compare the conclusions generated by two or more different evidence prioritization strategies. Similar to method 2, the reviewer would use the same methods for risk-of-bias assessment and strength-of-evidence rating, so that any differences in conclusions could be attributable only to differences in the evidence prioritization strategies. The advantage over method 2 is that the reviewer would not be dependent on the quality of reporting of a published review, which may often lack important information. However, this would be more labor-intensive than methods 1 and 2 since there would be no reliance on already-published reviews.

Although methods 2 and 3 address the inherent drawbacks of comparing already published systematic reviews, they do not address the more important question of what is the "most appropriate" conclusion (or non-conclusion) to reach.

Question 2: Which Strategy (or Strategies) Leads to the Most Appropriate Conclusions?

In order to measure "appropriateness," a reviewer needs to define the correct answer to a given clinical question. This could be based on meta-analysis of a complete evidence base on a well-understood clinical topic. However, we note that a meta-analysis is not a prerequisite for reaching the most appropriate conclusion.

Meta-analysis of a Complete Evidence Base

Perform a meta-analysis of all well-done studies of a given clinical topic (using participant-level data if available). Define criteria for which of the published studies are actually entered into this meta-analysis (e.g., only randomized blinded trials, or any direct comparison studies, etc.). This represents the reference standard.

Next, define a set of partial evidence that excludes the most recent *x* years of studies. The question is: which prioritization strategy best estimates the reference standard using only this partial evidence? The summary effect size of the complete evidence base would be the benchmark for comparison. However, reviewers should check to determine whether the standard of care that would be used in intervention comparisons has changed during the chosen time interval. Reviewers should also check to ensure that factors other than inclusion criteria, such as selective outcome reporting, publication bias, or changes in implementation strategies, are not potential explanations for observed changes in evidence base findings over time.

Limitations of this approach include the lack of agreement on reliable validity standards for meta-analysis and the possibility of incorporation bias due to testing the validity of a subset of evidence using the whole evidence as gold standard. In some instances, a small evidence base (consisting of one or a few well-designed, appropriately powered studies) may be sufficient to reach the most appropriate conclusion.

Summary

Systematic reviewers routinely prioritize evidence in numerous ways. Our goal in this paper has been to provide a framework for understanding the possibilities, considering influential factors, and choosing among the myriad of options. This will help enhance the transparency of review processes, which in turn may help users determine how different reviews of the same topic can reach different conclusions.

References

1. Slavin RE. Best evidence synthesis: an intelligent alternative to meta-analysis. J Clin Epidemiol 1995 Jan;48(1):9–18. PMID: 7853053

2. Agency for Healthcare Research and Quality. Guide for conducting comparative effectiveness reviews [draft for public comment]. Chapter 3, Selecting evidence: controlled trials. Rockville (MD): Agency for Healthcare Research and Quality; 2007.

3. Norris S, Atkins D, Bruening W, et al. Observational studies in systematic reviews of comparative effectiveness. J Clin Epidemiol (in press):1–25.

4. Dundar Y, Hill R, Dickson R, et al. Comparative efficacy of thrombolytics in acute myocardial infarction: a systematic review. QJM 2003 Feb;96(2):103–13. PMID: 12589008

5. Schmieder RE, Martus P, Klingbeil A. Reversal of left ventricular hypertrophy in essential hypertension. A meta-analysis of randomized double-blind studies. JAMA 1996 May 15;275(19):1507–13. PMID: 8622227

6. Porzio F. Meta-analysis of three double-blind comparative trials with sustained-release etodolac in the treatment of osteoarthritis of the knee. Rheumatol Int 1993;13(2 Suppl):S19–24. PMID: 8210920

7. Lipton A, Stopeck A, van Moos R, et al. A meta-analysis of results from two randomized, double-blind studies of denosumab versus zoledronic acid (ZA) for treatment of bone metastases. J Clin Oncol 2010;28(15 Suppl):9015.

8. Ostermann T, Raak C, Bussing A. Survival of cancer patients treated with mistletoe extract (Iscador): a systematic literature review. BMC Cancer 2009;9:451. PMID: 20021637

9. Abbas S, Seitz M. Systematic review and meta-analysis of the used surgical techniques to reduce leg lymphedema following radical inguinal nodes dissection. Surg Oncol 2009 Dec 9. PMID: 20005090

10. Ziegler R, Grossarth-Maticek R. Individual patient data meta-analysis of survival and psychosomatic self-regulation from published prospective controlled cohort studies for long-term therapy of breast cancer patients with a mistletoe preparation (Iscador). Evid Based Complement Alternat Med 2008 Apr 11 (Epub). PMID: 18955332

11. Morshed S, Bozic KJ, Ries MD, et al. Comparison of cemented and uncemented fixation in total hip replacement: a meta-analysis. Acta Orthop 2007 Jun;78(3):315–26. PMID: 17611843

12. Abraham NS, Byrne CM, Young JM, et al. Meta-analysis of non-randomized comparative studies of the short-term outcomes of laparoscopic resection for colorectal cancer. ANZ J Surg 2007 Jul;77(7):508–16. PMID: 17610681

13. Higgins JP, Green S, eds. Cochrane handbook for systematic reviews of interventions 5.0.2. The Cochrane Collaboration; 2008 [updated 2009 Sep 1]. Available at: http://www.cochrane-handbook.org/.

14. Cook DJ, Sackett DL, Spitzer WO. Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam Consultation on Meta-Analysis. J Clin Epidemiol 1995 Jan;48(1):167–71. PMID: 7853043

15. Egger M, Smith GD, Phillips AN. Meta-analysis. Principles and procedures. BMJ 1997 Dec 6;315(7121):1533–7.

16. Sutton AJ, Abrams KR, Jones DR, et al. Methods for meta-analysis in medical research. Sensitivity analysis. London: John Wiley; 2000. (Wiley series in probability

and statistics: applied probability and statistics). p. 147–52.

17. Egger M, Smith GD. Principles of and procedures for systematic reviews. In: Egger M, Smith GD, Altman DG, eds. Systematic reviews in health care: meta-analysis in context. 2nd ed. London: BMJ Books; 2001. p. 23–42.

18. Higgins JP, Green S. Cochrane handbook for systematic reviews of interventions 4.2.6. The Cochrane Collaboration; 2006 [updated 2006 Sep 1]. Available at: http://www.cochrane.org/sites/default/files/uploads/Handbook4.2.6Sep2006.pdf.

Appendix A. Tabulation of Inclusion Criteria and Modifying Factors

Table A1. Criteria related to study design and modifying factors

	Number of Studies (for Each Relevant Outcome)	Disease Course	Outcome Subjectivity (can Knowledge of Intervention Group Affect Outcome)	Potential for Unmeasured Confounders	Degree of Relevance	Frequency of Outcome Occurrence	Feasibility	Ethical Considerations	Minimum Time Needed to Evaluate Intervention Effect	Is outcome a Harm of intervention?	Concomitant Interventions or Processes of Care Changed Over Time	Possibility of Substantial Carryover Effect if Interventions Given Sequentially	Will Meta-analysis be Performed?	Degree of Similarity Among Different Studies (Participants, Interventions)	Degree of Similarity Among Patients who did and did not Provide Data
Random assignment to intervention groups	✓			✓	✓			✓		✓					
Blinding of study participants to which intervention they received	✓		✓		✓		✓	✓		✓					
Blinding of providers to the intervention that they provided	✓		✓		✓		✓			✓					
Blinding of outcome assessors to the assigned intervention	✓		✓		✓		✓			✓					
Presence of independent comparison group	✓	✓								✓		✓			
Presence of concurrent comparison group	✓	✓									✓				
Presence of a before-after comparison	✓	✓								✓	✓	✓			
Adequate washout period	✓											✓			

Table A1. Criteria related to study design and modifying factors (continued)

	Number of Studies (for Each Relevant Outcome)	Disease Course	Outcome Subjectivity (can Knowledge of Intervention Group Affect Outcome)	Potential for Unmeasured Confounders	Degree of Relevance	Frequency of Outcome Occurrence	Feasibility	Ethical Considerations	Minimum Time Needed to Evaluate Intervention Effect	Is outcome a Harm of intervention?	Concomitant Interventions or Processes of Care Changed Over Time	Possibility of Substantial Carryover Effect if Interventions Given Sequentially	Will Meta-analysis be Performed?	Degree of Similarity Among Different Studies (Participants, Interventions)	Degree of Similarity Among Patients who did and did not Provide Data
The intervention comparison must be direct, not indirect	✓													✓	
Good baseline comparability	✓														
Prospective planning (the study question was determined before any data were collected)	✓														
Consecutive enrollment	✓														

Table A2. Criteria pertaining to study conduct and reporting, and modifying factors

	Number of Studies (for Each Relevant Outcome)	Disease Course	Outcome Subjectivity (can Knowledge of Intervention Group Affect Outcome)	Potential for Unmeasured Confounders	Degree of Relevance	Frequency of Outcome Occurrence	Feasibility	Ethical Considerations	Minimum Time Needed to Evaluate Intervention Effect	Is outcome a Harm of intervention?	Concomitant Interventions or Processes of Care Changed Over Time	Possibility of Substantial Carryover Effect if Interventions Given Sequentially	Will Meta-analysis be Performed?	Degree of Similarity Among Different Studies (Participants, Interventions)	Degree of Similarity Among Patients who did and did not Provide Data
The minimum length of followup was x	✓	✓							✓						
The minimum number of study participants per group who provided data was x	✓					✓									
The minimum % of enrollees who provided data was x%	✓														✓
Percent difference between groups in proportion of study participants who had usable data (only studies with comparison groups)	✓														
Calculable or imputable effect size	✓												✓		
Study must have used validated method of outcome measurement measuring instruments	✓														

Table A3. Criteria pertaining to relevance and modifying factors

	Number of Studies (for Each Relevant Outcome)	Disease Course	Outcome Subjectivity (can Knowledge of Intervention Group Affect Outcome)	Potential for Unmeasured Confounders	Degree of Relevance	Frequency of Outcome Occurrence	Feasibility	Ethical Considerations	Minimum Time Needed to Evaluate Intervention Effect	Is outcome a Harm of intervention?	Concomitant Interventions or Processes of Care Changed Over Time	Possibility of Substantial Carryover Effect if Interventions Given Sequentially	Will Meta-analysis be Performed?	Degree of Similarity Among Different Studies (Participants, Interventions)	Degree of Similarity Among Patients who did and did not Provide Data
Study participant characteristics are sufficiently similar to the participants specified in the Key Question (e.g., adults vs. children)	✓														
Interventions (both treatment and comparator) sufficiently similar to the interventions specified in the Key Question	✓														
Setting characteristics are sufficiently similar to the setting specified in the Key Question (e.g., inpatient vs. outpatient)	✓														