Methods Research

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
www.ahrq.gov

Contract Number:

Prepared by:

Investigators:

AHRQ Publication No.
<Month Year>

Table of Contents

Abstract iii

Introduction 1

Objectives 2

Methods/Approaches 2

Task 1: Lists of inclusion criteria and factors that might affect a reviewer's decision to use each criterion 3

Task 2: List of evidence prioritization strategies 13

Task 3: Methods for evaluating evidence prioritization strategies 16

Question 1: Do different strategies lead to similar or different conclusions? 17

Question 2: Which strategy (or strategies) leads to the most appropriate conclusions? 18

References 19

Tables

Table 1. Inclusion criteria pertaining to publication format 3

Table 2. Inclusion criteria/analysis criteria pertaining to study design, study conduct and reporting, and study relevance..... 4

Table 3. Tabulation of inclusion variables and modifying factors 8

Table 4. Strategies for defining the "best evidence" set for a given key question..... 16

Figures

Figure 1. Flow chart of application of inclusion criteria..... 12

Abstract

One of the major challenges facing systematic reviewers is determining the criteria for study inclusion and analysis. Reviewers often employ a “best evidence” approach to address the key questions in the reviews. However, what is meant by “best” is often unclear, and several factors may influence a reviewer’s decision to broaden or narrow the criteria. 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 directive, but rather to provide a conceptual framework to enhance the transparency of inclusion decisions made during the course of a systematic review. Accordingly, 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, and (3) list the ways in which evidence prioritization strategies might be formally evaluated. Consensus-based methods were used to achieve the stated goals of the project. The end products of tasks 1 and 2 were the primary constructs of the decision framework. Task 3 identified a number of potential approaches that might be used to formally evaluate these strategies in the future.

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* 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.

Existing guidance from the 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 observational (i.e., non-randomized) 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 simply 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 that do not meet the inclusion criteria have effectively been given zero priority. In this way, 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 non-randomized 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 (despite randomization), 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 one examines the methods and results of randomized studies, and if they are insufficient to permit a conclusion, to then relax the study inclusion criteria so that non-randomized evidence can be considered. This staged approach can be planned *a priori*, to avoid the possible bias of study 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 sub-prioritized randomized trials in various ways. One approach is to include only *double-blind* randomized trials, thereby employing a best-evidence approach at the level of the inclusion criteria (for examples, see the following 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 non-randomized studies, and a sub-prioritization of these can be easily justified (e.g., based on whether the authors matched groups at baseline). For examples of reviewers sub-prioritizing nonrandomized studies, see the following references: [8-12].

Thus, 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 lack of an organizing

framework in the methodology of evidence prioritization was the impetus behind the project summarized in this document.

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 of them 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 non-randomized studies), whereas others do not. None of them, however, can guarantee an answer to the Key Question.

Objectives

This project seeks to *outline a framework for evidence prioritization*, i.e., for defining the "best evidence." This framework can improve transparency and also suggest alternatives to reviewers as they make difficult decisions about the "best" evidence. As noted above, this document is not intended to be directive, but rather to provide a conceptual construct with options to aid the decision-making process. It is only designed for evaluation of evidence on interventions; procedures for evaluating other types of evidence bases (e.g., diagnostic studies) are beyond the scope of this report. This is Phase I of a larger project (Phase II would involve a formal evaluation of the impact of variations in inclusion criteria on a review's conclusions). Led by the ECRI Institute 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 website.

Methods/Approaches

Consensus-based methods were used to achieve the stated goals of the project. 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 website.

Task 1: Lists of inclusion criteria and factors that might affect a reviewer's decision to use each criterion

There are two basic types of inclusion criteria that are typically used in systematic reviews. The first set includes criteria pertaining to publication format (such as those in Table 1). These criteria are usually unaffected by subsequent decisions regarding "best evidence" and analysis.

Table 1. Inclusion criteria pertaining to publication format

Criterion
English language (may depend on the topic)
Peer-reviewed
Year of publication sufficiently recent
Full article, not just an abstract
Data on these study participants were not reported in other included articles (to avoid double-counting of study participants)

The second set includes criteria pertaining to study design, study conduct and reporting, and study relevance (or applicability) (Table 2). Unlike the criteria in Table 1, these criteria may not be used in all systematic reviews. Indeed, certain criteria (such as outcome assessor blinding) may be rarely used by reviewers, but they are listed because of the possibility that they might be used in certain circumstances (e.g., when this design characteristic is particularly important, or when it effectively delineates the evidence). The right-hand column of the table includes factors that may affect a reviewer's decision to use each criterion. Note that this is not intended to be a hierarchy based on study design. Instead, the use of these criteria may vary by Key Question, depending on what evidence is available or necessary to address the question.

We note that reviewers may select a subset of these criteria for study presentation (encompassing studies that will be presented 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 control 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 control 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 be analyzed. Either way, reviewers may choose from the list of criteria presented in Table 2.

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

Criterion	Factors Influencing the Decision to Employ this Criterion (Modifying Factors)
Criteria Pertaining to Study Design	
1. Random assignment to intervention groups	• Number of studies that did this (if none, then don't require it; if many, then require it) • Existence of important unmeasured confounders (if so, require randomization) • Degree of applicability in studies that did this (if low, don't require it) • Whether the outcome is a harm (if so, don't require it) • Is randomization ethical (e.g., is there equipoise between interventions)? (if not, don't require it)
2. Blinding of study participants to which intervention they received	• Number of studies that did this (if none, then don't require it; if many, then require it) • How difficult it is to blind study participants or maintain blinding (if too difficult, then don't require it) • Whether study participant knowledge of intervention group can influence outcomes (if so, then require it)(this subsumes outcome subjectivity and confounding) • Degree of applicability in studies that did this (if low, don't require it) • Whether the outcome is a harm (if so, don't require it) • Is blinding ethical? (e.g., would it require an unethical control)(if not, don't require it)
3. Blinding of providers to which intervention they provided	• Number of studies that did this (if none, then don't require it; if many, then require it) • How difficult it is to blind providers or maintain blinding (if too difficult, then don't require it) • Whether provider knowledge of intervention group can influence outcomes (if so, then require it) • Degree of applicability in studies that did this (if low, don't require it) • Whether the outcome is a harm (if so, don't require)
4. Blinding of outcome assessors was provided	• Number of studies that did this (if none, then don't require it; if many, then require it) • How difficult it is to blind outcome raters or maintain blinding (if too difficult, then don't require it) • Whether outcome raters' knowledge of intervention group can influence outcomes (if so, then require it) • Degree of applicability in studies that did this (if low, don't require it) • Whether the outcome is a harm (if so, don't require)
5. Presence of independent control group	• Number of studies that did this (if none, then don't require it; if many, then require it) • Whether disease course allows intervention effects to be predicted accurately without an independent control group • If there would be substantial carryover effect if study participants received both interventions sequentially, then require an independent group • Whether the outcome is a harm (if so, don't require it)

Criterion	Factors Influencing the Decision to Employ this Criterion (Modifying Factors)
6. Presence of concurrent control group	• Number of studies that did this (if none, then don't require it; if many, then require it) • If concomitant interventions differed between periods, then require concurrence • If other aspects of intervention and/or follow-up changed over time (i.e., institutional changes), then require concurrence • Whether the outcome is a harm (if so, don't require it)
7. Presence of a before-after comparison	• Number of studies that did this (if none, then don't require it; if many, then require it) • Whether disease course allows intervention effects to be predicted accurately without a comparison • If there would be substantial carryover effect if study participants received both interventions sequentially, then require an independent group • Whether the outcome is a harm (if so, don't require it)
8. Adequate washout period	(Note: this criterion applies only to studies where the same study participants receive different interventions in different time periods, such as crossover studies) • Number of studies that did this (if none, then don't require it; if many, then require it)) • Possibility of substantial carryover effect (if likely, require it; if unlikely, don't require it)
9. The intervention comparison must be direct, not indirect	• Number of studies that did this (if none, then don't require it; if many, then require it) • If there exists evidence in which study participants and interventions were sufficiently similar in different types of studies, then don't require direct evidence
10. Good baseline comparability	• Number of studies that did this (if none, then don't require it; if many, then require it)
11. Prospective planning (the study question was determined before any data were collected)	• Number of studies that did this (if none, then don't require it; if many, then require it)
12. Consecutive enrollment	• Number of studies that did this (if none, then don't require it; if many, then require it)
Criteria Pertaining to Study Conduct and Reporting	

Criterion	Factors Influencing the Decision to Employ this Criterion (Modifying Factors)
13. The minimum length of follow-up was x	• Number of studies that did this (if none, then don't require it; if many, then require it) • If outcomes before this timepoint are very unlike from outcomes after this timepoint, then require it (e.g., how long do you need to follow the study participant to determine whether they benefited or were harmed from intervention?)
14. The minimum number of study participants per group who provided data was x	• Number of studies that did this (if none, then don't require it; if many, then require it) • Frequency of outcome occurrence (if frequent, don't require it; if low, then require it)
15. The minimum % of enrollees who provided data was x%	• Number of studies that did this (if none, then don't require it; if many, then require it) • If the enrollees who did not provide data could be very unlike those who did provide data, then require it
16. Percent difference between groups in proportion of study participants who had usable data (controlled studies only)	• Number of studies that did this (if none, then don't require it; if many, then require it)
17. Study must have reported at least one outcome of interest	• No modifying factors
18. Calculable or imputable effect size	• Number of studies that did this (if none, then don't require it; if many, then require it) • If meta-analysis of this outcome will not be performed, then do not require this
19. Study must have used validated method of outcome measurement	• Number of studies that did this (if none, then don't require it; if many, then require it)
Criteria Pertaining to Relevance	

Criterion	Factors Influencing the Decision to Employ this Criterion (Modifying Factors)
20. Study participant characteristics are representative of the relevant study participants specified in the Key Question (e.g., adults vs. children)	• Number of studies that did this (if none, then don't require it; if many, then require it)
21. Intervention characteristics are representative of the relevant intervention specified in the Key Question (e.g., percent study participants with adjunct intervention)	• Number of studies that did this (if none, then don't require it; if many, then require it)
22. Setting characteristics are representative of the relevant intervention specified in the Key Question (e.g., inpatient vs. outpatient)	• Number of studies that did this (if none, then don't require it; if many, then require it)

A tabulation between each inclusion criterion and each modifying factor is illustrated in Table 3. A checkmark indicates that a specific criterion (row) is influenced by a specific modifying factor (column).

Table 3. Tabulation of inclusion variables and modifying factors

Inclusion Variables	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 applicability	Frequency of outcome occurrence	Feasibility	Ethical considerations	Minimum time needed to evaluate intervention effect	Is outcome an adverse effect 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?
Criteria Related to Study Design													
1. Randomization	✓			✓	✓			✓		✓			
2. Blinding of study participants	✓		✓		✓		✓	✓		✓			
3. Blinding of providers	✓		✓		✓		✓			✓			
4. Blinding of outcome assessors	✓		✓		✓		✓			✓			
5. Presence of independent control group	✓	✓								✓		✓	
6. Presence of concurrent control group	✓	✓		✓							✓		
7. Presence of any comparison (e.g., before-after, time series)	✓	✓								✓		✓	

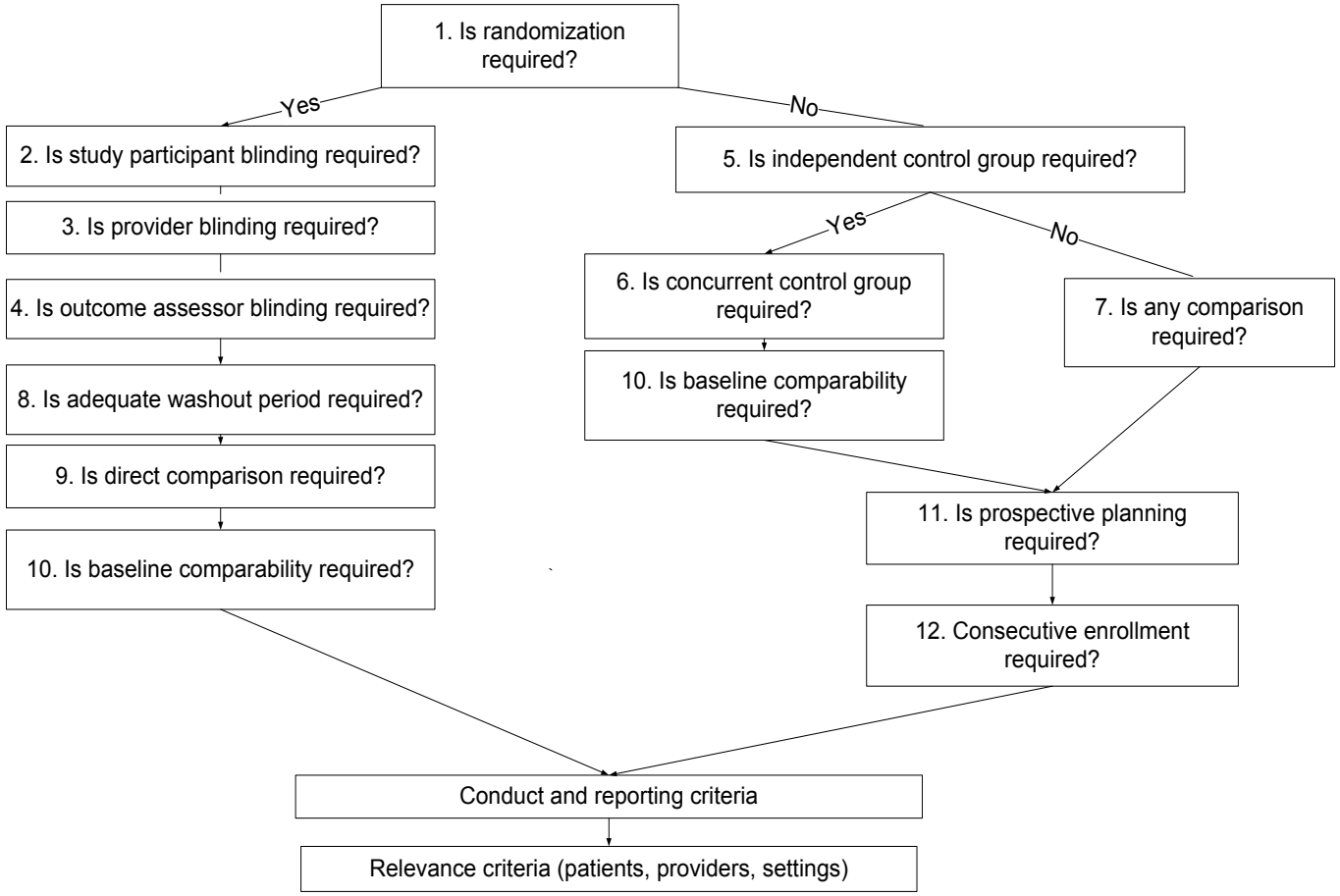
Inclusion Variables	Modifying Factors												
	Number of studies (for each relevant outcome)	Disease course	Outcome subjectivity (can knowledge of intervention group affect outcome)	Potential for unmeasured confounders	Degree of applicability	Frequency of outcome occurrence	Feasibility	Ethical considerations	Minimum time needed to evaluate intervention effect	Is outcome an adverse effect 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?
8. Adequate washout period	✓											✓	
9. Directness of comparison	✓												
10. Baseline comparability of comparison groups	✓												
11. Prospective planning	✓			✓									
12. Consecutive enrollment	✓												
Criteria for Study Conduct and Reporting													
13. Minimum length of follow-up	✓	✓							✓				
14. Percent study participants with usable data	✓												
15. Percent differential attrition (concurrent controlled studies)	✓												
16. Presence of any outcome of interest	✓												

Inclusion Variables	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 applicability	Frequency of outcome occurrence	Feasibility	Ethical considerations	Minimum time needed to evaluate intervention effect	Is outcome an adverse effect 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?
17. Minimum number of study participants/group	✓					✓							
18. Calculable or imputable effect size	✓												✓
19. Validated measuring instruments	✓												
Criteria Pertaining to Relevance													
20. Study participant characteristics (e.g., adults vs. children)	✓				✓								
21. Intervention characteristics (e.g., percent study participants with adjunct intervention)	✓				✓								
22. Setting characteristics (e.g., inpatient vs. outpatient)	✓				✓								

Figure 1 illustrates the logical flow for application of inclusion criteria from a best evidence perspective. The process begins with consideration of criteria related to study design, beginning with randomization. If enough RCTs are available, one may move down the left branch of the pathway and consider whether blinding (of study participants, providers, and/or outcome assessors) is required. If few or no RCTs are available, the reviewer moves down the right branch of the pathway and considers other types of controlled studies. The next decisions are about requirements for other aspects of study design such as baseline comparability, prospective planning, and consecutive enrollment. Regardless of which pathway is initially chosen, the next stages of decision-making are identical: consideration of criteria related to study conduct and reporting followed by consideration of criteria related to study relevance (applicability).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 control group), #6 (presence of independent control 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), #3 (provider blinding), and #4 (outcome rater blinding). While in theory non-randomized trials could employ any of these types of blinding, in practice this almost never occurs.

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 EHC methods manual,[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 or 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 non-randomized 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.

Figure 1. Flow chart of application of inclusion criteria



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 reach a valid answer to 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 (referred to as the "tabled" set). 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 delineation 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 lesser quality. This 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 precluding an answer to the Key Question (resulting in an evidence grade of Insufficient).

A second strategy is to add studies that, relative to the single best study, are more susceptible to bias and/or less applicable. This increases power, allows for possible replication, permits measurement of cross-study consistency. However, the statistical power of this larger set may still be too low to answer the Key Question.

This insufficiency may motivate a third strategy, which involves a further lowering of the bar: admit still lower-quality studies into the formal analysis, up to the point where the combined evidence permits an answer to the 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. Being able to answer the question implies that the evidence strength rating was not "Insufficient." An example of this third strategy can be found in the EHC 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).

One core component of the strength of the evidence is precision (i.e., a wide confidence interval around an effect size would not be considered High strength). Thus, adding still more studies to the evidence base will increase the chances of obtaining a narrow confidence interval (unless the results show substantial heterogeneity), without affecting the other components of the strength rating, thereby increasing the strength of the evidence. This represents a fourth strategy: defined as "best evidence" the set of studies that collectively yield the highest strength-of-evidence rating.

For strategies 3 and 4, the potential disadvantage of adding lower-level studies (increased chance of bias) is somewhat minimized in that lower-level studies can only increase the strength of the evidence if the findings are consistent with the findings of higher-level studies. For

example, if two higher-level studies together lead to a low strength of evidence, additional lower-level studies could boost the strength to moderate if all of the studies have consistent findings. A potential disadvantage is that observational studies with biases in opposite directions might have similar (consistent) effect sizes, which could lead to an overestimate in the strength of evidence. Furthermore, if the only available studies are lower-level studies then the risk of bias is not reduced by consistency in the findings.

Table 4 (see following page) outlines the four strategies. Checkmarks indicate which aspects of the evidence are required components of each strategy. 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% of study participants, define the "Best evidence" set as those that reported data on at least 80% 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; #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. The rightmost three columns display how the four strategies differ with respect to feasibility, the risk of an inappropriate conclusion, and the risk of an inappropriate lack of conclusion.

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 (a type I error), it also has a high risk of an inappropriate lack of conclusion (a type II error due to inadequate statistical power). 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 type I error due to inclusion of lower-quality studies; however, due to its greater statistical power it has the lowest risk of a type II error. 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, 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, meta-analysis of multiple bodies of evidence (e.g. RCTs and non-randomized studies) should be performed using an appropriate statistical model

for generalized synthesis.[13-19] Alternatively, a reviewer may choose to synthesize different bodies of evidence separately and then decide whether the lower-level 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	Factors considered by the strategy					Feasibility	Risk of inappropriate conclusion (Type I error)	Risk of inappropriate lack of conclusion (Type II error)
	Risk of bias	Applicability	Replication and cross-study consistency	Permit a conclusion[1]	Maximize evidence strength			
1. Single best study	✓	✓				High	Low	High
2. Best set	✓	✓	✓					
3. Best set that yields an answer to the Key Question	✓	✓	✓	✓				
4. Best set that yields an answer to the Key Question and maximizes the evidence strength	✓	✓	✓	✓	✓	Low	High	Low

[1] 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.

Task 3: Methods for evaluating evidence prioritization strategies

The tradeoffs inherent in the choice of prioritization strategy raise at least two important questions. One 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)[20] 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 follow-up); changing criteria about study design (e.g., whether to include randomized studies with unblinded outcome assessment).[20] Similar recommendations have been made by several other authors.[21-25]

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 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. One 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," one needs to define the correct answer to a given clinical question. This could be based on meta-analysis of a mature evidence base.

Meta-analysis of a mature evidence base. Perform a meta-analysis of all well-done trials of a given clinical topic (using individual patient-level data if available). Define criteria for which of the published trials are actually entered into this meta-analysis (only randomized blinded trials, or any direct comparison trials, etc.). This represents the reference-standard evidence prioritization strategy. Then compare the listed strategies with this one, and see which strategy best agrees with the reference standard.

An alternative approach is to systematically review evidence for a clinical question at different points in time, including a time when the evidence base was not "mature" (i.e., relatively few published studies) and a later time when substantially more evidence was available. One could then determine which evidence prioritization strategy used at the earlier time would have best predicted the findings of the later, more "mature" evidence base. The summary effect size of the mature 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 not changed during the chosen time interval.

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.

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. Guide for conducting comparative effectiveness reviews [draft for public comment]. Rockville (MD): Agency for Healthcare Research and Quality (AHRQ); 2007 Oct 10. Selecting evidence: controlled trials. p. 16-25.

3. Norris S, Atkins D, Bruening W, Fox S, Johnson E, Kane R, Morton SC, Oremus M, Ospina M, Randhawa G, Schoelles K, Shekelle P, Viswanathan M. Observational studies in systematic reviews of comparative effectiveness. J Clin Epidemiol (in press):1-25.

4. Dundar Y, Hill R, Dickson R, Walley T. 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, Henry DH, Richardson GE, Rodriquez GI, Bourgeois HP, Ke C, Jun S, Dansey RD. 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;Epub ahead of print. 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 ahead of print. PMID: 18955332

11. Morshed S, Bozic KJ, Ries MD, Malchau H, Colford JM Jr. 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, Solomon MJ. 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. Spiegelhalter DJ, Best NG. Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. Stat Med 2003 Dec 15;22(23):3687-709. PMID: 14652869

14. Prevost TC, Abrams KR, Jones DR. Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening. Stat Med 2000 Dec 30;19(24):3359-76. PMID: 11122501

15. Dominici F, Parmigiani G, Wolpert RL, Hasselblad V. Meta-analysis of migraine headache treatments: combining information from heterogeneous designs. J Am Stat Assoc 1999 Mar;94(445):16-28.

16. Begg CB, Pilote L. A model for incorporating historical controls into a meta-analysis. Biometrics 1991 Sep;47(3):899-906. PMID: 1742445

17. Li Z, Begg CB. Random effect models for combining results from controlled and uncontrolled studies in a meta-analysis. J Am Stat Assoc 1994;89:1523-7.

18. Droitcour J, Silberman G, Chelimsky E. A new form of meta-analysis for combining results from randomized clinical trials and medical-practice databases. Int J Technol Assess Health Care 1993 Summer;9(3):440-9. PMID: 8340208

19. Turner RM, Spiegelhalter DJ, Smith GC, Thompson SG. Bias modelling in evidence synthesis. J R Stat Soc Ser A Stat Soc 2009 Jan;172(1):21-47. PMID: 19381328

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

21. 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

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

23. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. Methods for meta-analysis in medical

research. London: John Wiley; 2000. Sensitivity analysis. (Wiley series in probability and statistics: applied probability and statistics). p. 147-52.

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

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