Draft Methods Report
Number XX

A Framework for Assessing the Strength of Methodological Recommendations for Systematic Review and Meta-analysis

Version [#]

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD20850
www.ahrq.gov

This information does not represent and should not be construed to represent an Agency for

Healthcare Research and Quality or Department of Health and Human Services determination or

Contract No. [Redacted]

Prepared by:
[Redacted]

Investigators
[Redacted]

This document is in the public domain and may be used and reprinted without permission except those copyrighted materials noted, for which further reproduction is prohibited without the specific permission of copyright holders.

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

Authors [redacted] are involved in developing open source software for meta-analysis, but this software is available at no charge and the authors receive no financial benefit.

This report is based on research conducted under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD. The findings and conclusions in this document are those of the author(s), who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help health care researchers and funders of research make well-informed decisions in designing and funding research and thereby improve the quality of health care services. This report is not intended to be a substitute for the application of scientific judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical research and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances.

This report may be used, in whole or in part, as the basis for research design or funding opportunity announcements. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

AHRQ Contacts

Stephanie Chang, MD, MPH
Director, EPC Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality
540 Gaither Road
Rockville, MD 20850

Contents

Abstract 8

1. Background 9

2. Developing the framework 10

3. Description, explanation and elaboration 11

3.1. Define the background context (Step 1) 14

3.2. Decompose the recommendation into testable and nontestable parts (Step 2) 19

3.3. Opine on the acceptability of nontestable parts (Step 3) 25

3.4. Assess testable parts in four dimensions (Step 4) 26

3.5. Provide an overall strength for the recommendation (Step 5) 35

4. Discussion 38

References 42

Tables

Table 1. Example methodological recommendation statements 12

Table 2. Background context for R1: “Use two or more members of the review team, working independently, to screen and select studies” 14

Table 3. Background context for R2: “Use the Peto odds ratio or the Manten-Haenszel method for odds ratios rather than inverse variance weighting for meta-analysis of rare events” 15

Table 4. Background context for R3: “Use diagnostic test meta-analysis models that treat sensitivity and specificity as random parameters rather than as fixed unknowns” 16

Table 5. Background context for a composite recommendation statement constructed for exposition. R4 “Use a generalized mixed effects linear model (random effects logistic regression) to obtain summary odds ratios in a meta-analysis of binary data” 22

Table 6. Background context for a recommendation constructed for exposition: R5 “Use a regression of effect size versus its precision to test for publication bias in a meta-analysis” 25

Table 7. Description of the evidentiary basis of R2 (a testable statement of statistical nature) 31

Table 8. Description of the evidentiary basis of R1 (a testable statement for which empirical data are necessary) 32

Table 9. Summary of the overall strength of three example recommendations 38

Figures

Figure 1. Overview of the proposed framework for rating the strength of methodological recommendations 13

Abstract

We propose a framework for rating the overall strength of methodological recommendations, and for organizing and communicating the logic behind them. We start by defining the background context in which the recommendations are made. We distinguish recommendations that are testable (in that their likelihood to hold can be informed by theory or data) from nontestable ones, which represent beliefs or assumptions that are unknowable. Nontestable statements can be justified, but their validity cannot be demonstrated. Testable statement can be assessed in terms of the adequacy of their evidentiary basis, the feasibility of following them, the expected impact of following them versus not, and their congruence with the desiderate of the background context. Considering these four dimensions, one can opine on the overall strength of the recommendation: high (a minimum standard, or a mandatory item) and low, (a best practice, or a highly desirable item).

1. Background

Evidence-based processes such as systematic reviews and meta-analyses are increasingly used to inform healthcare decisions. For their findings to be trustworthy, evidence-based methodologies should be based on sound principles and theories, and supported by empirical data. Following this tenet, many entities have generated methodological guidance for performing systematic reviews, meta-analyses, and decision or economic analyses.[1-4] Among them, the United States Institute of Medicine's (IOM) has published 21 minimum standards (corresponding to 82 elements of performance) for publicly funded systematic reviews,[1] and the Cochrane Collaboration is developing the 80-recommendation-strong Methodological Expectations of Cochrane Intervention Reviews (MECIR) project.[2]

The majority of such recommendations are derived from experts' experience and knowledge of the literature, and their appreciation of the norms in the field. But wide acceptance of methodologies does not necessarily imply soundness. Indeed, in several cases evidence-based medicine methodologies have been recommended for widespread use without ever having been held to rigorous standards of evaluation. While such *de facto* standard methods and processes may appear to be sound in principle, they are often applied dogmatically due to popularity rather than methodological appropriateness. Authoritative bodies such as the Cochrane Collaboration or the Evidence-based Practice Centers (EPC) program may then formally endorse popular methods, further encouraging widespread adoption. Consider the case of the funnel plot, first proposed in the mid-1980s in the social science literature as a means to detect publication bias.[5] The MOOSE group recommended the use of funnel plots "to aid in the detection of publication bias",[6] and numerous meta-analysts have done so, despite legitimate concerns regarding its

interpretation[7,8] and statistical properties.[9-11] The same is true for the use of summary quality scores in assessing methodological soundness,[12] and also for statistical methods for meta-analysis such as the use of continuity corrections for meta-analyses in studies with rare events.[13-15]

Similar to clinical practice guidelines, methodological recommendations should be transparent in their rationale, evidence base, and strength. Preferred methodologies must be identified, and the rationale for their selection and how important it is to abide by them should also be communicated. There is therefore a need for a framework for assessing the evidence supporting the methodological recommendations for research synthesis applications. Specifically, the details of the context in which a methodological recommendation is made should be made explicit. For a specific methodological recommendation, the approach to assess its scientific rigor, evidence base, applicability, and feasibility should be elucidated. This work is a foray into the problem of providing consistent and transparent methodological guidance.

2. Developing the framework

In brief, we started by generating several dozens of hypothetical recommendation statements on methodological decisions often encountered when performing systematic reviews and meta-analyses. The covered topics spanned the entire systematic review process, including: developing and refining key questions; identifying and appraising relevant evidence; extracting data; conducting meta-analyses or qualitative syntheses and presenting findings. The recommendation statements we considered varied in the degree of specificity and technical sophistication. We went through an iterative process of analyzing and discussing our exercise recommendation statements to identify an initial set of components for an evaluative framework.

Early versions of the framework were modified based on feedback from (a) two discussion sessions with researchers and faculty at the Center for Clinical Evidence Synthesis, Tufts Medical Center and Tufts University; (b) individual teleconference sessions with five internationally recognized technical experts in systematic review and meta-analysis; (c) and in a 40 minute discussion at the 2011 Society for Research Synthesis Methodology meeting. While the version of the framework presented here incorporates feedback obtained from these sessions, it is the opinion and construct of the authors, and does not necessarily agree with the opinions of the technical expert panel or other experts who were engaged during its development.

3. Description, explanation and elaboration

Methodological recommendations provide guidance for action when facing a methods problem. Table 1 shows three example recommendations for systematic review and meta-analysis obtained from IOM, MECIR or AHRQ guidance. The first (R1) prescribes an optimal approach (optimal in some yet unstated sense) to identifying eligible studies for a systematic review among the citations returned by the searches, as compared with other (yet unstated) approaches. The other two recommendations (R2 and R3) pertain to questions of statistical nature. R2 provides guidance for choosing a meta-analytic method for studies with rare events. R3 advises on the choice between random effects and fixed effects modeling for studies of diagnostic test accuracy.

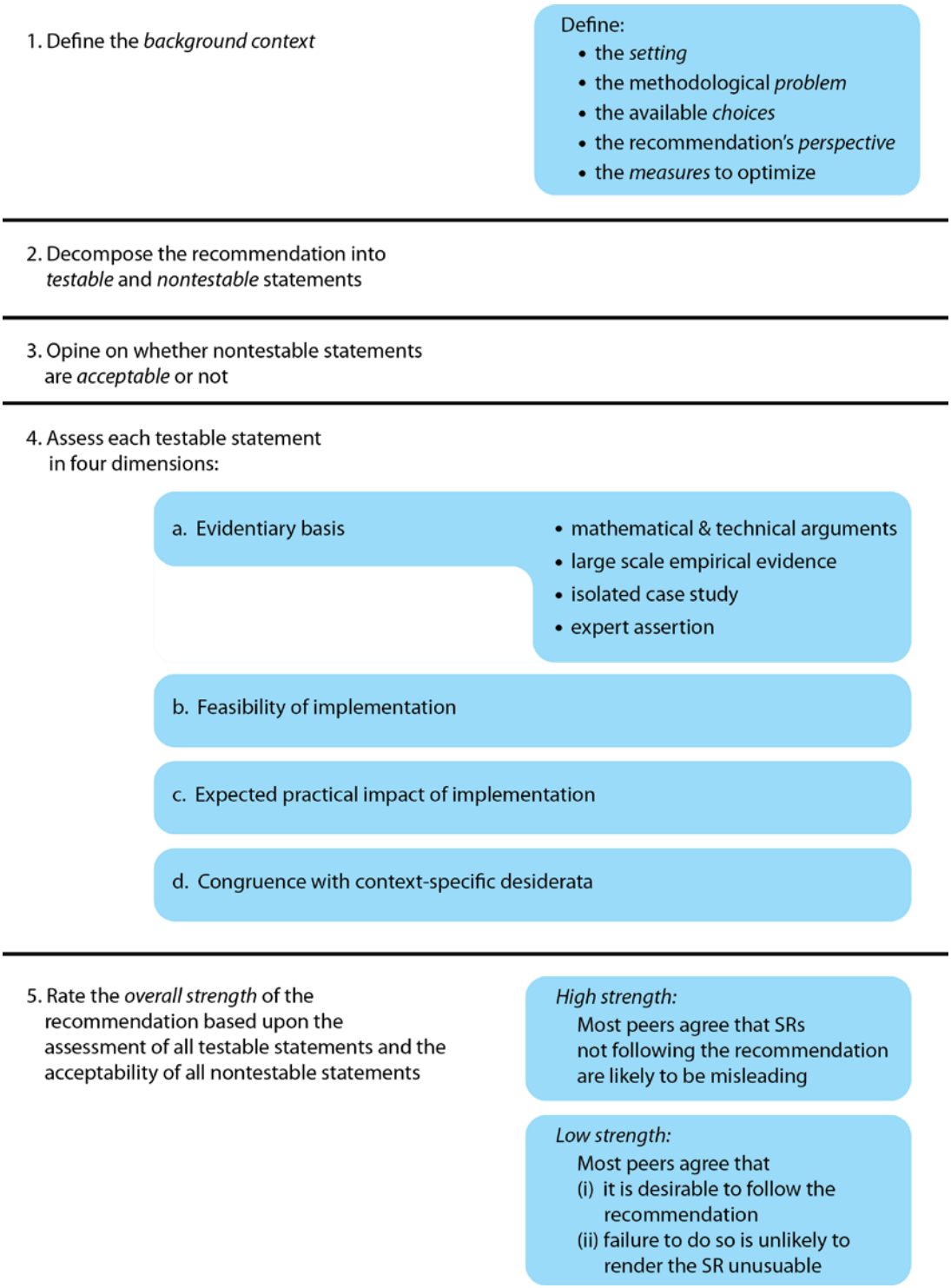
Table 1. Example methodological recommendation statements

#	*Recommendation*	*Source**
R1	Use two or more members of the review team, working independently, to screen and select studies	IOM, element 3.3.3. [1] and MECIR, #39 [2]
R2	Use the Peto odds ratio or the Manten-Haenszel method for odds ratios rather than inverse variance weighting for meta-analysis of rare events	AHRQ guide (reworded) [3]
R3	Use diagnostic test meta-analysis methods that treat the overall sensitivity and specificity as random parameters rather than as fixed unknowns	AHRQ guide (reworded) [4]

AHRQ: Agency for Healthcare Research and Quality; IOM: Institute of Medicine.
*Similar recommendations have been made by other entities (not cited).

The framework outlined in Figure 1 is intended to help those developing methodological guidance to explain the rationale behind their choices, and to rate the strength for their recommendation statements. It comprises a series of five steps described in Sections 3.1 through 3.5. Specifically, Section 3.1 explains what is needed to define the methodological challenge addressed by the recommendation. Section 3.2 explains how to break down (decompose) complex recommendations into more manageable parts. Sections 3.3 and 3.4 deal with assessing the logic behind the recommendation statement, its support, applicability, and anticipated impact; and Section 3.5 describes how to provide a rating for the overall strength of the recommendation.

Figure 1. Overview of the proposed framework for rating the strength of methodological recommendations



SR: systematic review. Bold horizontal lines define sequential steps. See text for explanation and elaboration.

3.1. Define the background context (Step 1)

All recommendation statements require a background context that specifies the problem being addressed and prescribes the desired characteristics of an optimal solution to the problem. The background context describes (1) the *setting* and (2) the *problem* at hand; (3) a finite set of *alternative choices* for addressing the problem; (4) the *perspective* of the recommendation; (5) and, *a (set of) measure(s)* with respect to which one can rank alternative choices. As an example, Table 2 reconstructs the presumed background context of R1 using information from the IOM report and from other sources.[1,16] Table 3 and Table 4 provide background contexts for R2 and R3, respectively.

Table 2. Background context for R1: "Use two or more members of the review team, working independently, to screen and select studies"

#	*Element of the background context*	*Description*
1	Setting	Publicly funded SRs in healthcare
2	Problem	Identifying eligible studies for a SR among citations returned by searches
3	Alternative choices	a) Single human reviewer b) Computer-assisted screening c) At least two human reviewers (independently) d) At least two nonindependent human reviewers (nonindependently) e) Human plus computer-assisted screening (independently) f) Human plus computer-assisted screening (not independently)
4	Perspective	That of the funder of the SR (public agency such as AHRQ) or of the user of the SR (including decisionmakers and other consumers). Desired attributes of SRs: • Credibility • Generalizability • Efficiency • Patient centeredness • Scientific rigor • Timeliness • Transparency
5	Measures to optimize	Minimize the likelihood of missing eligible research; maximize the efficiency and timeliness of SR; maximize SR credibility.

SR: systematic review. This background context was constructed using information from the IOM report.

3.1.1. Setting

Recommendation R1 refers to publicly funded systematic reviews in healthcare. It is not necessary that R1, or any recommendation for that part, is transferable to settings other than the intended one. For example, researchers in fields such as evolutionary biology and ecology (in which funding for systematic reviews from any source is scarce compared to healthcare) might have stated this recommendation differently [personal communication, Prof. Jessica Gurevitch, State University of NY]. The observation that methodological recommendations are embedded in settings is directly analogous to clinical practice guidelines that are setting-specific. For example, the National Comprehensive Cancer Network guidelines for breast cancer screening,[17] or prostate cancer screening and management[18,19] referred primarily to healthcare systems in advanced economies and were modified for developing countries.[20,21]

Table 3. Background context for R2: "Use the Peto odds ratio or the Manten-Haenszel method for odds ratios rather than inverse variance weighting for meta-analysis of rare events"

#	*Element of the background context*	*Description*
1	Setting	[Same as R1]
2	Problem	Combining odds ratios in meta-analyses of at least 5 studies with rare events (<5% per arm).
3	Alternative choices	a) No assumptions on homogeneity or heterogeneity i) Peto method ii) Manten-Haenszel method b) Assuming homogeneity i) Inverse-variance method (fixed effect) ii) Logistic regression (fixed effect) c) Assuming heterogeneity, normal distribution for within-study variability i) Noniterative estimation of heterogeneity (DerSimonian Laird) ii) Iterative estimation of heterogeneity, restricted maximum likelihood d) Assuming heterogeneity, binomial distribution for within-study variability e) Iterative estimation of heterogeneity in a generalized mixed effects linear model
4	Perspective	[Same as R1]
5	Measures to optimize	Statistical bias; coverage probability; mean squared error; maximize feasibility of implementation; maximize SR/meta-analysis credibility.

SR: systematic review. The Table is for exposition, and was constructed using information from EPC Methods guidance documents.[3]

Table 4. Background context for R3: "Use diagnostic test meta-analysis models that treat sensitivity and specificity as random parameters rather than as fixed unknowns"

#	*Element of the background context*	*Description*
1	Setting	[Same as R1]
2	Problem	Deciding between fixed and random effects modeling for meta-analysis of diagnostic tests.
3	Alternative choices	a) Models treating sensitivity and specificity as fixed unknowns b) Models treating sensitivity and specificity as random parameters
4	Perspective	[Same as R1]
5	Measures to optimize	Maximize the generalizability of SR/meta-analysis findings.

SR: systematic review. The Table is for exposition, and was constructed using information from EPC methods guidance documents.[4]

3.1.2. Problem

A concise and unambiguous description of the challenge addressed by the recommendation.

3.1.3. Alternative choices

For R1, we have at least six reasonable options, which are combinations of (1) whether the screening is done in a single pass or with redundant efforts (e.g., in duplicate); (2) whether the redundant screening efforts are independent or not; (3) and whether screening is conducted by unaided humans, or by humans using computer-assisted processes.[22,23] Currently, all six options are feasible alternatives; they have pros and cons, which can be captured by appropriate measures (see below). Considering all realistic alternative choices is important for an analysis of any decisional problem,[24,25] and this includes making methodological recommendations. For example, choices that have not been included in the list of alternatives are not being considered, and if the optimal choice is among them, it will not be selected. The IOM report alluded to only two of the six choices, namely, independent redundant screening versus single screening by unaided humans (choices c versus a in Table 2; we will return to this in Section 3.5).

Table 3 and Table 4 list alternative choices for R2 (methods for meta-analyzing odds ratios when events are rare) and R3 (choosing between alternative modeling approaches), respectively.

3.1.4. Perspective

Funders of systematic reviews and meta-analyses, those who conduct them, and their consumers have different perspectives on which attributes of a systematic review or a meta-analysis are most important. Presumably all would agree that systematic reviews should be credible, scientifically rigorous, timely and generalizable. However, the cost associated with the production of systematic reviews is probably more important to funders and those who conduct the research than to consumers.

Many methodological recommendations are developed from the perspective of the entity that oversees the production of the systematic reviews. This can be a funding agency such as AHRQ, an entity such as the Cochrane Collaboration, or a professional society. The perspective prescribes the desired attributes of systematic reviews that will be produced using the methodological guidance. In turn, the desired attributes of the systematic reviews prescribe which measures are important for choosing among the alternative choices. Helfand and Balshem summarize that systematic reviews produced by AHRQ should fully explore the clinical logic behind the reviewed questions, make use of best evidence for each type of question, provide a balanced assessment of benefits and harms, and manage conflict of interest responsibly.[16] For our examples, we specify seven desired attributes of systematic reviews (Table 2, row 4) based on the rationales articulated in the introduction of the IOM report. (A subtle difference: the IOM panel describes these seven attributes as “criteria” for developing their methodology standards,

in the sense that they selected standards that promote reviews with these attributes. We will return to this difference in the Discussion)

3.1.5. Measures to optimize

Finally, the background context should indicate one or more measures for assessing the optimality of each of the alternative choices. According to our interpretation of the background context for R1, plausible measures include the likelihood of missing eligible research (which should be minimized); the efficiency of the process, and the timeliness and credibility of the systematic reviews that will be generated (which should be maximized) (Table 2). It is reasonable to choose other measures in place of those proposed here. It is also reasonable to assign different importance or weight to each measure. The four measures proposed for R1 are not readily quantifiable in a practical manner. However stating them promotes transparency, and using them in a qualitative fashion helps communicate the rationale for the final recommendation.

By contrast, for technical problems it may be feasible to define quantitative measures. For example, for R2 it is reasonable to choose measures that quantify the statistical performance of alternative meta-analysis methodologies, including the method's statistical bias (desired to be zero), actual coverage probability of "95% confidence intervals" (desired to be the nominal 0.95), and mean square error (desired to be as small as possible).

Finally, we have specified a single measure for R3, i.e., that the findings of the synthesis should be generalizable to all studies that are similar to those included in the meta-analysis. Observe that when there is only one measure to optimize the alternative choices can be ranked.

For R1 and R2, where we have more than one measures to optimize, it is possible that no single choice is optimal on all measures. Ideally we have to define how much we care for each one of the measures explicitly, or implicitly, by explaining how we chose between methods that are preferable according to some metrics but not others.

3.2. Decompose the recommendation into testable and nontestable parts (Step 2)

We distinguish two types of statements that can be found in methodological recommendations, namely statements that are testable in a practical manner, and statements that are not. It is important to distinguish between these two types, as explained below.

3.2.1. Testable statements

For testable statements a procedure ("test") exists that can inform on the likelihood that the statement holds. R1 and R2 contain such testable statements. For R1 it is practical to conduct an empirical comparison of the six alternative choices in Table 2 in a sizable number of systematic reviews. In fact, human screening in one pass has been compared with double independent screening,[26] and computer-assisted screening has been compared with human-only screening in limited settings.[22,23] Similarly, for R2 it is practical to compare the alternative choices in Table 3 according to their statistical performance in simulation studies.[13-15]

In addition, it is expected that most peers will agree that properly selected procedures or tests can inform choices between the alternatives in each background context, although they may still disagree on exactly how to interpret the information.

3.2.2. Nontestable statements

By contrast, R3 is different in a subtle but important way. R3 uses statistical terminology to describe a belief about the unobserved and unknowable process by which diagnostic accuracy studies are generated. Treating sensitivity and specificity as "random parameters" implies that the true sensitivity and specificity in each study are believed to be generated from an unobserved joint distribution. Learning (inferring) the parameters of this distribution enables one to generalize beyond the studies that are included in a meta-analysis, and to state conclusions about all studies generated by the same process. By contrast, under a fixed effect assumption, one is not willing to assume a particular generating process for the studies at hand. Instead, one is interested in drawing conclusions only on the collection of studies already available, and does not care to state conclusions that apply more generally.

We clarify that *both approaches in R3 are valid* – in that they give answers to different questions. Hence it is not meaningful to "test" which belief system is "optimal". It is the judgment of those making the recommendation that it is an acceptable recommendation in the given background context. In the example, those making the recommendation R3 operate in a background setting where it is desirable to infer conclusions about all diagnostic studies that can be generated in the same way as the studies already at hand (Table 4).

Therefore, peers can legitimately disagree about the acceptability of nontestable statements. In R3, peers can disagree about whether the findings of a meta-analysis should be generalizable beyond the studies at hand, and thus about the content of the recommendation.

Ideally, methodological recommendations should contain testable statements, and keep nontestable statements to the necessary minimum. Further, as described next, it is preferable to avoid recommendations that contain both testable and nontestable statements.

3.2.3. Decomposition of recommendations that include testable and nontestable statements

A simple example

Recommendations R1, R2 and R3 are simple, in that they contain only testable or only nontestable statements. Composite recommendations include both testable and nontestable statements. For exposition, we construct a composite hypothetical recommendation, within the background context in Table 5:

R4: *Use a generalized mixed effects linear model (random effects logistic regression) to obtain summary odds ratios in a meta-analysis of binary data*.

We can decompose R4 into a nontestable statement:

R4.1: *Use meta-analysis models that treat the effect size in each study as a random parameter rather than a fixed unknown,*

and a testable statement:

R4.2: *Given R4.1, use a generalized mixed effects linear model (random effects logistic regression) to obtain summary odds ratios.*

The first statement (R4.1) is similar to R3, in that it prescribes a belief about the process by which studies are generated. In itself, R4.1 is not testable in a practical manner. Like R3, from the perspective of those making the recommendation R4.1 is either acceptable or not. By contrast R4.2 pertains to a statistical methodology, and its likelihood to hold (for example, in terms of statistical performance) can be evaluated with well-established methods. It is easier to think about the decompositions R4.1 and R4.2, rather than the composite statement R4 on its own. A real example is provided in the next paragraph.

Table 5. Background context for a composite recommendation statement constructed for exposition. R4 "Use a generalized mixed effects linear model (random effects logistic regression) to obtain summary odds ratios in a meta-analysis of binary data"

#	*Element of the background context*	*Description*
1	Setting	[Same as R1]
2	Problem	Combining odds ratios in meta-analyses of at least 10 studies (>5% event rate per arm)
3	Alternative choices	c) No assumptions on homogeneity or heterogeneity i) Peto method ii) Manten-Haenszel method d) Assuming homogeneity i) Inverse-variance method (fixed effect) ii) Logistic regression (fixed effect) e) Assuming heterogeneity, normal distribution for within-study variability i) Noniterative estimation of heterogeneity (DerSimonian Laird) ii) Iterative estimation of heterogeneity, restricted maximum likelihood f) Assuming heterogeneity, binomial distribution for within-study variability i) Iterative estimation of heterogeneity in a generalized mixed effects linear model
4	Perspective	[Same as R1]
5	Measures to optimize	a) Maximize the generalizability of SR findings b) Statistical bias; coverage probability; mean squared error; maximize feasibility of implementation; maximize SR/meta-analysis credibility.

SR: systematic review. This recommendation is a composite recommendation (see text).

Asymmetry in funnel plots

Composite recommendations can be difficult to appreciate, and can lead to protracted and nonproductive discourse. The case of the funnel plot as an aid to detect publication bias is such an example. Funnel plots were introduced in the mid 80's as a means to assess the likelihood of publication bias.[5] In brief, if publication bias is present, studies with statistically significant results are more likely to be published compared to studies with statistically nonsignificant results, biasing meta-analyses of published data. The motivation for funnel plots is simple. If all studies are generated from a single underlying distribution, a scatterplot of observed study effects on the horizontal axis and their precision on the vertical axis would resemble a symmetric inverted funnel in the absence of publication bias, or an asymmetric inverted funnel in the presence of publication bias (and assuming a specific selection process).[5,27] A seminal paper in the British Medical Journal popularized a method for testing for funnel plot asymmetry.[27] Its authors were very careful in explaining that publication bias is only one of many explanations of funnel plot asymmetry,[27,28] but many others have chosen to ignore this advice.

The misuse of funnel plots in the literature has spurred an academic exchange with very little cross-talking between those who do not accept that funnel plot asymmetry is particularly informative on the likelihood of publication bias,[7-9,29] and others who opted to focus on developing improved statistical tests and other methods for assessing funnel plot asymmetry.[10,11,28,30-33] The parallel monologues continued in a methodological discussion in the Cochrane Collaboration, where the discussants (Profs Christopher Schmid and Alex Sutton) agreed on technical aspects of testing for asymmetry [personal communication with Prof Schmid], but disagreed about their informativeness. Academics from both sides of the discussion

published consensus guidance on the interpretation of funnel plot asymmetry and the use of pertinent tests.[34]

To put these considerations in the proposed framework, we paraphrase the MOOSE group's advice that "methods should be used to aid in the detection of publication bias, e.g. fail-safe methods or funnel plots",[6] and construct a hypothetical recommendation within the background context in Table 6:

R5: *Use a regression of effect size versus its precision to test for publication bias in a meta-analysis*.

As above, we can decompose R5 into a nontestable statement:

R5.1 *Interpret funnel plot asymmetry as a (strong) indication of publication bias,*

and a testable statement:

R5.2 *Given R5.1, use the regression of effect size versus its precision to test for funnel plot asymmetry.*

The first statement (R5.1) is nontestable. It expresses the belief that the most likely interpretation for asymmetric funnel plots (or at least, of asymmetric funnel plots with specific characteristics[33]) is publication bias rather than other biases, heterogeneity, statistical artifacts, or chance; and that a specific selection process operates. However, R5.2 is testable in a practical manner, e.g., by means of simulation studies. The aforementioned consensus publication succeeds in focusing on testable statements, and avoids specifics on nontestable ones.[34]

Table 6. Background context for a recommendation constructed for exposition: R5 "Use a regression of effect size versus its precision to test for publication bias in a meta-analysis"

#	*Element of the background context*	*Description*
1	Setting	[Same as R1]
2	Problem	Assessing the likelihood of publication bias in a meta-analysis of at least 10 studies
3	Alternative choices	a) Search clinical trial registries and other sources and contact manufacturers to identify unpublished data b) Infer on the likelihood of publication bias based on funnel plot asymmetry i) Use a regression test of effect size versus its precision ii) Use the trim-and-fill method iii) Use visual inspection of contour enhanced funnel plots c) Infer on the likelihood of publication bias using selection models […] [… another 4 options, combination of (a) with each of (b.i), (b.ii), (b.iii) or (c)]
4	Perspective	[Same as R1]
5	Measures to optimize	Minimize the likelihood of missing eligible research; maximize the efficiency and timeliness of SR; maximize SR credibility.

SR: systematic review. This recommendation is a composite recommendation.

3.3. Opine on the acceptability of nontestable statements (Step 3)

If applicable, those making the methodological recommendation should provide a justification for the nontestable recommendations (or the nontestable parts of composite recommendations). The justification can help readers understand the viewpoint of the nontestable recommendation, and inform their decision about whether or not they accept this point of view.

For example, our own justification for R3 is that it is not particularly interesting to learn only about the mean sensitivity and specificity in a specific of diagnostic test studies. It is much more interesting to generalize findings to all studies similar to the meta-analyzed ones. R4.1 (the nontestable part of R4) is the corresponding of R3 for treatment studies, and has an analogous justification. We constructed R5 for exposition, and it so happens that we do not agree with it.

We believe that, in the absence of additional information, it is unknowable whether asymmetry in a funnel plot is more suggestive of publication bias than for competing explanations. Very respected peers disagree with our position,[35] and for the sake of the discussion, we "coined" R5 to represent beliefs other than our own.

3.4. Assess testable statements in four dimensions (Step 4)

The next step is to assess whether recommendations containing testable statements achieve a high with respect to four dimensions: (a) *scientific rigor* or (adequacy of) *evidentiary basis*, (b) *feasibility* of implementation, (c) expected *practical impact* of implementation, and (d) congruence with *context-specific* desiderata. It is unclear whether qualitative descriptions are sufficient, or whether they should be further summarized by an explicit score for each dimension. For clarity, we opt to also score each dimension as "high", "low" or "unclear".

3.4.1. Scientific rigor or evidentiary basis

This dimension informs on the likelihood that the (testable) statement holds, by evaluating the arguments or data that support it. As is the case for clinical practice guidance, where the current thinking favors different study designs for measuring the effectiveness versus the safety of interventions or the accuracy of medical testing, no single hierarchy of arguments or data applies to all testable methodological recommendations. We first describe broad categories of supporting information, and then suggest which of these should be valued more than others for commonly encountered types of testable recommendations.

Categories of supporting information

For description, we distinguish four categories of supporting information: (i) mathematical and technical arguments, (ii) empirical data of large scale, (iii) isolated case studies, (iv) and expert opinion.

The first category of **mathematical and technical arguments** includes knowledge and information that can be generated largely without collecting empirical data. This includes *mathematical derivations* and *proofs*, and theoretical explorations in the form of *simulation analyses*. Mathematical and technical arguments are most pertinent to recommendations of technical nature. For example, it is easy to show that in a pairwise meta-analysis of intervention studies, naïve pooling (where one first combines within treatments, and then calculates differences between treatments) yields a summary treatment effect estimate that is confounded by study. In another example, a theorem proves that normalized weights that are inversely proportional to the sampling variance in each study yield the most efficient estimator of the summary effect in a meta-analysis of normally distributed data compared to other weighting schemes (e.g., by sample size, or by inverse-standard error).[36]

In the same vein, simulation studies can be very informative, especially when mathematical proofs are too difficult or impossible to derive. Simulations are very useful for comparing the statistical performance of methods, but they can only examine a relatively small number of scenarios. To derive more general conclusions one must be interpolate or extrapolate the simulation findings to address scenarios that have not been examined. Simulation studies have to follow good design practices to be robust and applicable to real-life situations.[37]

Empirical data of large scale are obtained by means of randomized, case-control, cohort, survey or other designs. Thus there is substantial variation in terms of evidentiary strength. Randomized trials comparing research methodologies are relatively uncommon. They have been used to compare variants of processes in which the human factor renders a paired design impractical. For example, a randomized comparison of blinding versus not blinding of data extractors to the names and affiliations of primary study authors showed no difference in the accuracy of data extraction, but a substantial difference in the overhead and cost of performing a meta-analysis (higher in the blinding arm).[38] Not directly relevant to meta-analysis, randomized trials have been used to examine the effects of blinding reviewer to the authors identities versus not in journal peer-review,[39-41] and having eponymous versus anonymous peer-reviewers.[39,41] An ingenious study randomized participants to receive random versus nonrandom assignments, to empirically assess whether a nonrandomized experiment yields accurate answers.[42] Much more common are observational studies, such as those informing on the associations between study treatment effect and study-level characteristics – much like evaluating risk factors in the medical domain. For example, a reanalysis of three meta-epidemiological studies of 146 meta-analyses (1346 trials) found an association between stronger treatment effect and absence of allocation concealment for subjective outcomes such as pain or quality of life, but not for objective outcomes such as mortality.[43] Finally, empirical data in the form of surveys of the literature can document current practices and inform on their prevalence.[44-46]

Isolated case studies are analogous to case reports and case series in the medical domain. These are often feasibility studies demonstrating application of a new method (see for example reference[47]). Information from isolated case studies is generally not sufficient to describe practice, document associations, or support the use of a methodology. However, isolated

case studies are valuable in demonstrating the catastrophic implications or the infeasibility and inconsistency of a practice that should be avoided. An instance of the former is the documentation of extreme publication bias in a meta-analysis of trials of antipsychotics, where including unpublished studies together with published data yielded highly contradictory conclusions.[48] In an example of the latter, rating the methodological quality of trials in a meta-analysis using 25 published quality scores resulted in very fragile study rankings and very inconsistent subgroup analyses.[12] Documentation of a very undesirable event is valuable information, but it is seldom sufficient on its own to support methodological guidance. It is generally important to have an estimate of the probability of the undesirable event, to decide whether a comprehensive but expensive approach should be as minimum standard to minimize the likelihood of the undesirable event.

Finally, **expert opinion** as a summary of the experts' experience and intuition can be invoked in supporting methodological recommendations when no other information is available or when it is necessary to interpolate or extrapolate from existing data. Much of the existing guidance is based on expert opinion.

Scoring the adequacy of the supporting information

Based on the above, we propose the following for scoring the adequacy of supporting information for methodological problems.

Arguably, **technical or mathematical problems, including those relevant to the statistical performance of methodologies**, are most appropriately addressed by theoretical and technical arguments. Consider the problem addressed by R2, as defined in Table 3, where one

has to choose between four methods for the meta-analysis of studies with rare events. Table 7 summarizes the evidentiary basis of R2. An analytical answer would be the strongest possible supporting information to the problem, but one does not exist or cannot be obtained in a practical manner. Simulation studies are the only practical way to assess the statistical performance of the four alternative choices for R2. Three simulation studies have explored the statistical properties of several methods for estimating the summary effect for meta-analyses of rare events, providing sufficient information for the comparisons in Table 3.[13-15]

We are not aware of empirical studies comparing the four methods in a large number of meta-analyses including mostly studies with rare events. An empirical comparison cannot distinguish e.g., which method has the least statistical bias, because in an empirical sample the "true" effects are unknowable. Empirical data from a survey of meta-analyses though can inform on how often the problem is encountered in practice, and how often the alternative methods would lead to contradictory conclusions. Finally, isolated case reports and expert opinion are probably much less useful in the presence of other information.[49-51] They can demonstrate disagreement between methods, but cannot identify the better performing one.

We score the adequacy of the evidentiary base of R2 as "high", because it (1) includes supporting information of the appropriate type (i.e., mathematical or technical arguments in the form of simulation studies); (2) the simulation studies are well performed; and (3) they correspond to realistic scenarios (Table 7). We would assign a "low" score for the adequacy of the evidence base if there were no well-performed simulation studies, even in the presence of empirical studies of large scale. This is because empirical evidence cannot inform on the statistical performance of the compared meta-analytic methods (the chosen measures to

optimize). As discussed in Section 3.5, a "high" score for the adequacy of the evidence base does not automatically imply that the overall strength of the recommendation is also "high".

Table 7. Description of the evidentiary basis of R2 (a testable statement of statistical nature)

Supporting information	*Description*	*Comments*
Mathematical and technical arguments	No analytical answer available or practical Three simulation studies have explored questions relevant to the meta-analysis of studies with rare events. [13-15] In sum, the Peto or the Manten-Haenszel method for the odds ratio were least biased and had coverage probability closest to the nominal across simulation scenarios.	The simulation studies do not cover all possible scenarios or even all commonly used metrics. For example, they do not explore the risk ratio metric. The simulation studies are well-done, and are generalizable to the background context at hand. The simulated scenarios are applicable to real life settings.
Empirical data of large scale	We do not know of large scale empirical evidence.	Empirical evidence informs on the concordance of alternative choices in real life situations, but cannot inform on which are closer to the truth.
Isolated proof of concept studies	Isolated examples (nonexhaustive list references[49-51]).	These are examples that alternative methods can yield different results.
Expert opinion	An example is reference.[50]	Generic comments; no consideration of simulation data or empirical evidence.

Most other problems, however, cannot be addressed satisfactorily only by relying on theory or simulations, because deciding which of the alternative choices to recommend is **directly informed by empirical data** (i.e., at least one of the measures defined in the background context must be obtained empirically). For example, consider the problem addressed by R1 (choosing between processes for screening citations for inclusion in a systematic review). The probability of missing at least one eligible study, and the resources needed for each choice must be obtained empirically. The ideal empirical study would use a paired design to compare the six alternative options in Table 2, i.e., apply all six options to a sizable number of systematic reviews. Such a study does not exist (Table 8). Instead, all-human screening has been compared

with computer assisted screening in data from a case series of eight systematic reviews in clinical and molecular medicine;[22,23,52] and in another limited assessment, single screening by a human was estimated to miss between 0 and 24% of finally eligible studies compared to more comprehensive screening (estimation based on a capture-recapture approach).[26]

We score the adequacy of the evidentiary base of R1 as "low", because there are no large-scale empirical studies (i.e., no supporting information of the appropriate type). We cannot use a mathematical or technical argument for R1, because in the background context in Table 2 the measures to be optimized necessitate specific empirical data. As before, a "low" score for the adequacy of the evidence base does not imply that the overall strength of the recommendation is also "low".

Table 8. Description of the evidentiary basis of R1 (a testable statement for which empirical data are necessary)

Supporting information	*Description*	*Comments*
Mathematical and technical arguments	Not applicable.	Empirical data are necessary for ranking choices in Table 2
Empirical data of large scale	We do not know of empirical data of large scale.	An ideal study would compare all six options in a sizable number of systematic reviews.
Isolated proof of concept studies	In an analysis of four extractors in a large systematic review, the estimated proportion of eligible citations missed by a single reviewer was between 0 and 24%.[26] Computer-assisted screening has been compared with single or duplicate human screening in a small number of examples.[22,23,52]	No study compared all six alternatives in the same samples.
Expert opinion	IOM panel and MECIR consider redundant independent screening a minimum standard and a mandatory item, respectively.[1,2]	No references are provided.

3.4.2. Feasibility of implementation

This refers to the feasibility of implementing (following) the methodological recommendation in the setting of interest. Recommendations that require substantial expertise or considerable resources will be of course more challenging to implement. For example, Cochrane and EPC guidance[4] recommend the bivariate model with binomial likelihood,[53,54] or the hierarchical summary ROC (HSROC) model[55] to synthesize diagnostic accuracy, rather than independent meta-analyses of sensitivity and specificity. The recommended methods can be implemented as hierarchical regression models or generalized linear mixed effects models, but this presupposes access to advanced statistical software, facility with statistical programming, and at least basic understanding of the advanced meta-analysis models.[56] The same is true for meta-analyses of mixed treatment comparisons (network meta-analyses). Similarly, unless publicly available resources make computer-assisted screening easily accessible, three of the six options in Table 2 that include computed-assisted processes will be widely feasible.[57] By contrast, the recommended method in R2 is implemented in many widely available software tools, and requires only a very basic understanding of statistical concepts.[58] Thus, we assign a "high" feasibility score to R2.

The feasibility of implementing the recommendation depends on the background context in which the recommendation is made. The recommendation for at least double independent screening in R1 is probably feasible for systematic reviews with adequate funding, or even for unfunded reviews if there is substantial organizational and other support. In our case, we assign a "high" score for the feasibility of implementing R1 for (most if not all) publicly funded reviews

in healthcare. As mentioned already, it is debatable whether this recommendation is feasible as a minimum standard for systematic reviews in all settings.

3.4.3. Expected practical impact of implementation

The practical impact of a methodological recommendation can be large or marginal, and can be conjectured using empirical data. For example, according to the Cochrane diagnostic test accuracy workgroup and to EPC guidance, meta-analyses of diagnostic accuracy tests should use advanced methods.[4,59] The rationale is that the advanced methods (bivariate syntheses of sensitivity and specificity and HSROC analysis) respect the multivariate nature of test performance metrics, allow for the nonindependence between sensitivity and specificity across studies ("threshold effect") and also for between-study heterogeneity. However, in a large empirical evaluation of over 308 meta-analyses of diagnostic test accuracy, the difference in the estimates and the confidence intervals between advanced methods and simpler approaches (e.g., separate univariate meta-analyses of sensitivity and specificity) was almost always modest or negligible [redacted reference and reference[60]]. An analogous observation was made when comparing univariate versus multivariate meta-analysis for estimating marginal treatment effects [redacted reference and references[61,62]]. In some sense, in both examples, the choice between more advanced and simpler methods is probably of academic rather than of practical interest.

The expected impact of following versus not following recommendations R1 or R2 is not clear. Specifically for R1, the differential expected impact of double independent screening by reviewers over, e.g., other options for redundant screening in Table 2 is not obvious. For R2, it is unclear how often conclusions would change using alternative methods. We therefore assign an "unclear" scores in this dimension.

3.4.4. Congruence with context-specific desiderata

This dimension captures desiderata of the background context that are not adequately captured by the other three dimensions. For example, from the perspective of the funder of systematic reviews, one goal is to produce reviews that are credible (appear to be free of conflicts of interest), generalizable, patient-centered, rigorous, timely and transparent. Additional goals may also be important to a funder, even if their importance is not immediately apparent from different perspectives. For example, it is desirable that all systematic reviews produced by a program follow standardized approaches. Therefore, a recommendation can favor one of two otherwise similar choices, on the basis of standardization across reviews. This would be less of a concern from the perspective of the consumer of the review, or of those conducting it.

In our case, R1 is not relevant to generalizability, patient centeredness or transparency of a systematic review, but its implementation would promote the generation of reviews with other desirable attributes (most importantly, credibility and scientific rigor); and would serve program consistency (standardization) as well. One can make similar comments for R2.

3.5. Provide an overall strength for the recommendation (Step 5)

Those producing methodological guidance expect that *ceteris paribus* and on average, one is better off following the guidance than not. However, some recommendations represent **truly minimum standards** or **truly mandatory items**, in that they are essential for producing high quality systematic reviews and meta-analyses. If such minimal standards or mandatory items are not met, most peers would have concerns on the comprehensiveness or the validity or the systematic review, its findings and conclusions.

Another category is recommendations that represent **best practices** and are **highly desirable** –but not mandatory. For this category, most peers would agree that (1) following the recommendation would improve the systematic review or meta-analysis, and (2) that not following the recommendation is unlikely to render the review unusable.

Assuming that one would never recommend practices that are generally harmful, there still remains the possibility for a third category, namely recommendations for which it is unclear whether they are useful or superfluous. We believe that this third category should be an empty set. Otherwise put, we believe that **one should only make recommendations that one deems clearly beneficial** given the background context at hand.

In accordance with others,[1,2] we propose that recommendations in the first category should be assigned a "high" overall strength, and those in the second category should be assigned a "low" overall strength. Our position is that a recommendation should not be made if it belongs to the third category, and therefore we do not allow for other ratings (e.g., "insufficient"). This means that no methodological recommendations should be given for problems that are not well understood. As a consequence, methodological recommendations will be sparse in difficult problems where guidance in most needed, and generic rather than specific.

Characterizing the overall strength of a recommendation

Characterizing the overall strength of a recommendation is subjective by nature. We believe that a careful and balanced discussion of the aforementioned dimensions within the background context of the recommendation will (1) clarify the rationale for the recommendation;

and (2) help deduce whether the recommendation is a minimum standard/mandatory item (high strength), or a best practice/highly desirable item (low strength).

R1 ("use two or more members of the review team, working independently, to screen and select studies") was considered of "high" overall strength by the IOM report (it is a minimum standard) and MECIR (it is a mandatory item). Table 9 summarizes our analysis of R1 through the proposed framework. We would consider R1 a recommendation of "low" overall strength, meaning that it is probably a highly desirable process, but it is unclear that it "outperforms" the other five alternatives in Table 2 on the basis of the selected measures. We would probably assign a "high" overall strength to the following variant of R1:

R1' *Use redundant screening processes rather than single screening to screen and select studies,*

because processes that have built in a serious quality control (redundancy), more likely to outperform single-screening-based processes. It is unclear whether IOM and MECIR considered all six alternatives in Table 2, or only two (redundant independent screening versus single screening). If the background context in Table 2 included only these two options, we would probably give a "high" strength to R1 as well.

We assigned a "high" strength to R2, because it has an adequate evidentiary support, it is feasible, and congruent with the desiderata of scientific rigor and credibility, and despite the fact that we are not clear on its practical impact. In conjunction with the above, R2 might have to be modified if summary metrics other than the odds ratio are of interest, or if additional meta-analysis methods were to be included among the alternatives.

Finally, R3 was a nontestable recommendation. As explained in Section 3.3, we opined that it is an acceptable recommendation, and do not assign it an overall strength rating.

Table 9. Summary of the overall strength of three example recommendations

Question/Step	*R1*	*R2*	*R3*
Phrasing	Use two or more members of the review team, working independently, to screen and select studies	Use the Peto odds ratio or the Manten-Haenszel method for odds ratios rather than inverse variance weighting for meta-analysis of rare events	Use diagnostic test meta-analysis methods that treat the overall sensitivity and specificity as random parameters rather than as fixed unknowns
Is the recommendation testable? (Section 3.2 & 3.3)	Yes	Yes	No*
Scores			
Adequacy of evidentiary basis or scientific rigor (Section 3.4.1)	Low (Table 8)	High (Table 7)	NA
Feasibility of implementation (Section 3.4.2)	High	High	NA
Expected impact of implementation (Section 3.4.3)	Unclear	Unclear	NA
Congruence with context-specific desiderata (Section 3.4.4)	High	High	NA
Overall strength of the recommendation			
Authors of this report	Low**	High	NA
IOM	High	NA	NA
MECIR	High	NA	NA

NA: not applicable
* We deem that this nontestable statement is acceptable within the background setting in Table 4. See Section 3.3.
**We would assign a “High” strength to a different recommendation (see text).

4. Discussion

We propose a framework for rating the overall strength of methodological recommendations, and for organizing and communicating the logic behind them. We start by defining the background context in which the recommendations are made. We distinguish recommendations that are testable (in that their likelihood to hold can be informed by theory or

data) from nontestable ones, which represent beliefs or assumptions that are unknowable. Nontestable statements can be justified, but their validity cannot be demonstrated. Testable statement can be assessed in terms of the adequacy of their evidentiary basis, the feasibility of following them, the expected impact of following them versus not, and their congruence with the desiderate of the background context. Considering these four dimensions, one can opine on the overall strength of the recommendation: high (a minimum standard, or a mandatory item) and low, (a best practice, or a highly desirable item).

We do not allow for additional ratings of the overall strength of methodological recommendations because we believe that a recommendation should be made only for problems that are well understood, and when the recommended choice is on average and under usual conditions a clearly "beneficial" one. This is somewhat unfulfilling, in that recommendations will be sparse for problems for which guidance is most necessary. Conversely, recommendations would be more specific for problems where the optimal solutions are known. However, our position is that recommendations should generally be followed, and therefore a stringent approach is reasonable. We believe that this is observed in the IOM and MECIR recommendations: they are more numerous and more specific on processes (e.g., forming the review team, refining the topic, managing conflict of interest), but provide only generic guidance on exactly how to synthesize evidence. Guidance can (and arguably should) be provided when dealing with difficult problems, with the understanding that that the suggested approaches may not be optimal.

We are ambivalent on whether there should be an explicit scoring of testable statements in the four dimensions or not. Explicit scoring of individual dimensions is as subjective as rating

the overall strength of the recommendation. It is conceivable that the same recommendation can receive different scores in some dimensions (e.g., feasibility of implementation, congruence with context-specific desiderata), for different background contexts. On the other hand, explicit scoring can be used as a means to abstract the logic for rating the overall strength as "high" or "low", and thus as a tool of communication. Further, quantification often forces one to think carefully about tradeoffs. We lean in favor of explicitly scoring the individual dimensions, as the GRADE system does for clinical practice guidelines.

The framework calls for a structured set of subjective judgments. Extensive as it may be, the subjectivity in rating the dimensions of testable statements, and the fuzziness of the dimensions themselves are no worse, and probably much less extensive than the subjectivity and fuzziness of recommendations that do not have a defined background context; that are composites of testable and nontestable statements; lack an organized categorization of their supporting information; and fail to distinguish considerations on feasibility from the expected impact of following the recommendation. We therefore maintain that more structured approaches such as the one proposed here will serve the methodological community at least as well, if not better than less structured ones.

We believe that our framework represents a more general and more structured alternative to the approaches used to date by those who make methodological recommendations (for example the IOM panel, the MECIR authors, or the PCORI methodology committee[1,2,63]), but it does not represent a dissent. Using the framework can help communicate subtle but important aspects of the recommendation at hand. For example, we selected element 3.3.3 in the IOM

report (example R1,[a] which is similar to standard #39 in MECIR) to illustrate our framework. We would probably agree with IOM and MECIR in rating the strength of this recommendation as "high" if the only alternative is single screening. However, we do not think that such a specific recommendation constitutes a minimum standard when a broader set of alternatives is considered.

Finally, the feasibility and ease of using the framework itself is not clear. Guidance on deriving scores for individual dimensions and for the overall strength of the recommendation will have to be developed, perhaps following the paradigm of GRADE.[64-75] Further, the framework is quite general, in that it could be applied to methodological recommendations in other domains, beyond systematic review and meta-analysis. Broader application may provide insights or make obvious the need for modifications or larger changes. It is the nature of methods to change, to evolve. And this is the fate of the current proposal as well.

---

[a] "Use two or more members of the review team, working independently, to screen and select studies"

References

1. *Finding what works in healthcare. Standards for systematic reviews.* . Washington, DC: Institute of Medicine of the National Academies;2011.

2. Chandler J, Churchill R, Higgins J, Lasserson T, Tovey D. Methodological standards for the conduct of new Cochrane intervention reviews. 2011; http://www.editorial-unit.cochrane.org/mecir.

3. Fu R, Gartlehner G, Grant M, et al. Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care Program. *Journal of clinical epidemiology.* Nov 2011;64(11):1187-1197.

4. Trikalinos TA, Balion CM, Coleman CI, et al. Chapter 8: meta-analysis of test performance when there is a "gold standard". *Journal of general internal medicine.* Jun 2012;27 Suppl 1:S56-66.

5. Light RJ, Pillemer DB. *Summing up: The science of reviewing research.* Cambridge, MA: Harvard University Press; 1984.

6. Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. *JAMA : the journal of the American Medical Association.* Apr 19 2000;283(15):2008-2012.

7. Lau J, Ioannidis JP, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. *BMJ.* Sep 16 2006;333(7568):597-600.

8. Terrin N, Schmid CH, Lau J. In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *Journal of clinical epidemiology.* Sep 2005;58(9):894-901.

9. Terrin N, Schmid CH, Lau J, Olkin I. Adjusting for publication bias in the presence of heterogeneity. *Statistics in medicine.* Jul 15 2003;22(13):2113-2126.

10. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of clinical epidemiology.* Nov 2000;53(11):1119-1129.

11. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *Journal of clinical epidemiology.* Sep 2005;58(9):882-893.

12. Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA : the journal of the American Medical Association.* Sep 15 1999;282(11):1054-1060.

13. Bradburn MJ, Deeks JJ, Berlin JA, Russell Localio A. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in medicine.* Jan 15 2007;26(1):53-77.

14. Rucker G, Schwarzer G, Carpenter J, Olkin I. Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. *Statistics in medicine.* Feb 28 2009;28(5):721-738.

15. Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in medicine.* May 15 2004;23(9):1351-1375.

16. Helfand M, Balshem H. AHRQ series paper 2: principles for developing guidance: AHRQ and the effective health-care program. *Journal of clinical epidemiology.* May 2010;63(5):484-490.

17. Bevers TB, Anderson BO, Bonaccio E, et al. NCCN clinical practice guidelines in oncology: breast cancer screening and diagnosis. *Journal of the National Comprehensive Cancer Network : JNCCN.* Nov 2009;7(10):1060-1096.

18. Kawachi MH, Bahnson RR, Barry M, et al. NCCN clinical practice guidelines in oncology: prostate cancer early detection. *Journal of the National Comprehensive Cancer Network : JNCCN.* Feb 2010;8(2):240-262.

19. Mohler J, Bahnson RR, Boston B, et al. NCCN clinical practice guidelines in oncology: prostate cancer. *Journal of the National Comprehensive Cancer Network : JNCCN.* Feb 2010;8(2):162-200.

20. Abulkhair O, Saghir N, Sedky L, et al. Modification and implementation of NCCN guidelines on breast cancer in the Middle East and North Africa region. *Journal of the National Comprehensive Cancer Network : JNCCN.* Jul 2010;8 Suppl 3:S8-S15.

21. Hassen WA, Karsan FA, Abbas F, et al. Modification and implementation of NCCN guidelines on prostate cancer in the Middle East and North Africa region. *Journal of the National Comprehensive Cancer Network : JNCCN.* Jul 2010;8 Suppl 3:S26-28.

22. Wallace BC, Small K, Brodley CE, et al. Toward modernizing the systematic review pipeline in genetics: efficient updating via data mining. *Genetics in medicine : official journal of the American College of Medical Genetics.* Apr 5 2012.

23. Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH. Semi-automated screening of biomedical citations for systematic reviews. *BMC bioinformatics.* 2010;11:55.

24. Philips Z, Ginnelly L, Sculpher M, et al. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. *Health Technol Assess.* Sep 2004;8(36):iii-iv, ix-xi, 1-158.

25. Ramsey SD, Sullivan SD. Weighing the economic evidence: guidelines for critical assessment of cost-effectiveness analyses. *The Journal of the American Board of Family Practice / American Board of Family Practice.* Nov-Dec 1999;12(6):477-485.

26. Edwards P, Clarke M, DiGuiseppi C, Pratap S, Roberts I, Wentz R. Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records. *Statistics in medicine.* Jun 15 2002;21(11):1635-1640.

27. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ.* Sep 13 1997;315(7109):629-634.

28. Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *Journal of clinical epidemiology.* Oct 2001;54(10):1046-1055.

29. Ioannidis JP, Trikalinos TA. The appropriateness of asymmetry tests for publication bias in meta-analyses: a large survey. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne.* Apr 10 2007;176(8):1091-1096.

30. Rucker G, Schwarzer G, Carpenter J. Arcsine test for publication bias in meta-analyses with binary outcomes. *Statistics in medicine.* Feb 28 2008;27(5):746-763.

31. Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Statistics in medicine.* Oct 30 2006;25(20):3443-3457.

32. Moreno SG, Sutton AJ, Thompson JR, Ades AE, Abrams KR, Cooper NJ. A generalized weighting regression-derived meta-analysis estimator robust to small-study effects and heterogeneity. *Statistics in medicine.* Jun 30 2012;31(14):1407-1417.

33. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of clinical epidemiology.* Oct 2008;61(10):991-996.

34. Sterne JA, Sutton AJ, Ioannidis JP, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ.* 2011;343:d4002.

35. Sutton AJ. Dealing with publication bias: Funnel plots and beyond. *14th Cochrane Colloquium.* Freiburg, Germany2008.

36. Hedges LV, Olkin I. *Statistical methods for meta-analysis.* San Diego, CA: Academic Press; 1985.

37. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Statistics in medicine.* Dec 30 2006;25(24):4279-4292.

38. Berlin JA. Does blinding of readers affect the results of meta-analyses? University of Pennsylvania Meta-analysis Blinding Study Group. *Lancet.* Jul 19 1997;350(9072):185-186.

39. Godlee F, Gale CR, Martyn CN. Effect on the quality of peer review of blinding reviewers and asking them to sign their reports: a randomized controlled trial. *JAMA : the journal of the American Medical Association.* Jul 15 1998;280(3):237-240.

40. McNutt RA, Evans AT, Fletcher RH, Fletcher SW. The effects of blinding on the quality of peer review. A randomized trial. *JAMA : the journal of the American Medical Association.* Mar 9 1990;263(10):1371-1376.

41. van Rooyen S, Godlee F, Evans S, Smith R, Black N. Effect of blinding and unmasking on the quality of peer review: a randomized trial. *JAMA : the journal of the American Medical Association.* Jul 15 1998;280(3):234-237.

42. Shadish WR, Clark MH, Steiner PM. Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments. *J Am Stat Assoc.* Dec 2008;103(484):1334-1343.

43. Wood L, Egger M, Gluud LL, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ.* Mar 15 2008;336(7644):601-605.

44. Dahabreh IJ, Chung M, Kitsios GD, et al. Comprehensive Overview of Methods and Reporting of Meta-Analyses of Test Accuracy. *Agency for Healthcare Research and Quality (US).* 2012.

45. Willis BH, Quigley M. Uptake of newer methodological developments and the deployment of meta-analysis in diagnostic test research: a systematic review. *BMC medical research methodology.* 2011;11:27.

46. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in medicine.* May 30 2008;27(12):2037-2049.

47. Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F, Colditz GA. Meta-analysis of multiple outcomes by regression with random effects. *Statistics in medicine.* Nov 30 1998;17(22):2537-2550.

48. Turner EH, Knoepflmacher D, Shapley L. Publication bias in antipsychotic trials: an analysis of efficacy comparing the published literature to the US Food and Drug Administration database. *PLoS medicine.* 2012;9(3):e1001189.

49. Dahabreh IJ, Economopoulos K. Meta-analysis of rare events: an update and sensitivity analysis of cardiovascular events in randomized trials of rosiglitazone. *Clin Trials.* 2008;5(2):116-120.

50. Hernandez AV, Walker E, Ioannidis JP, Kattan MW. Challenges in meta-analysis of randomized clinical trials for rare harmful cardiovascular events: the case of rosiglitazone. *American heart journal.* Jul 2008;156(1):23-30.

51. Friedrich JO, Adhikari NK, Beyene J. Inclusion of zero total event trials in meta-analyses maintains analytic consistency and incorporates all available data. *BMC medical research methodology.* 2007;7:5.

52. Wallace BC, Small K, Brodley CE, Trikalinos TA. Who should label what? Instance allocation in multipple expert active learning. *Siam International Conference on Data Mining (SDM).* 2011:176-187.

53. Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of clinical epidemiology.* Dec 2006;59(12):1331-1332; author reply 1332-1333.

54. Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *Journal of clinical epidemiology.* Sep 2004;57(9):925-932.

55. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in medicine.* Oct 15 2001;20(19):2865-2884.

56. Wallace BC, Dahabreh IJ, Trikalinos TA, Lau J, Trow P, Schmid CH. Closing the Gap between Methodologists and End-Users: R as a Computational Back-End. *J Stat Softw.* Jun 2012;49(5):1-15.

57. Wallace BC, Small K, Brodley CE, Lau J, Trikalinos TA. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium.* 2012:819-824.

58. Wallace BC, Schmid CH, Lau J, Trikalinos TA. Meta-Analyst: software for meta-analysis of binary, continuous and diagnostic data. *BMC medical research methodology.* 2009;9:80.

59. Diagnostic test accuracy working group: Handbook for diagnostic test accuracy reviews. 2011; http://srdta.cochrane.org/handbook-dta-reviews.

60. Harbord RM, Whiting P, Sterne JA, et al. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. *Journal of clinical epidemiology.* Nov 2008;61(11):1095-1103.

61. Trikalinos TA, Olkin I. Meta-analysis of effect sizes reported at multiple time points: A multivariate approach. *Clin Trials.* Aug 7 2012.

62. Trikalinos TA, Olkin I. A method for the meta-analysis of mutually exclusive binary outcomes. *Statistics in medicine.* Sep 20 2008;27(21):4279-4300.

63. Methodological standards and patient-centeredness in comparative effectiveness research: the PCORI perspective. *JAMA : the journal of the American Medical Association.* Apr 18 2012;307(15):1636-1640.

64. Balshem H, Helfand M, Schunemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. *Journal of clinical epidemiology.* Apr 2011;64(4):401-406.

65. Brunetti M, Shemilt I, Pregno S, et al. Grade guidelines: 10. Considering resource use and rating the quality of economic evidence. *Journal of clinical epidemiology.* Aug 3 2012.

66. Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *Journal of clinical epidemiology.* Apr 2011;64(4):383-394.

67. Guyatt G, Oxman AD, Sultan S, et al. GRADE guidelines 11-making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. *Journal of clinical epidemiology.* Apr 27 2012.

68. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. *Journal of clinical epidemiology.* Apr 2011;64(4):395-400.

69. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines 6. Rating the quality of evidence--imprecision. *Journal of clinical epidemiology.* Dec 2011;64(12):1283-1293.

70. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 8. Rating the quality of evidence--indirectness. *Journal of clinical epidemiology.* Dec 2011;64(12):1303-1310.

71. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. *Journal of clinical epidemiology.* Dec 2011;64(12):1294-1302.

72. Guyatt GH, Oxman AD, Montori V, et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. *Journal of clinical epidemiology.* Dec 2011;64(12):1277-1282.

73. Guyatt GH, Oxman AD, Santesso N, et al. GRADE guidelines 12. Preparing Summary of Findings tables-binary outcomes. *Journal of clinical epidemiology.* May 18 2012.

74. Guyatt GH, Oxman AD, Sultan S, et al. GRADE guidelines: 9. Rating up the quality of evidence. *Journal of clinical epidemiology.* Dec 2011;64(12):1311-1316.

75. Guyatt GH, Oxman AD, Vist G, et al. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). *Journal of clinical epidemiology.* Apr 2011;64(4):407-415.