Methods Research Report

Towards a Framework for Communicating Confidence in Methodological Recommendations for Systematic Reviews and Meta-Analyses

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

Methods Research Report

Towards a Framework for Communicating Confidence in Methodological Recommendations for Systematic Reviews and Meta-Analyses

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 No. 290-2007-10055-I

Prepared by:
Tufts Evidence-based Practice Center
Boston, MA

Investigators:
Thomas A. Trikalinos, M.D.
Issa J. Dahabreh, M.D., M.S.
Byron C. Wallace, Ph.D.
Christopher H. Schmid, Ph.D.
Joseph Lau, M.D.

AHRQ Publication No. 13-EHC119-EF
September 2013

This report is based on research conducted by the Tufts Evidence-based Practice Center, (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-2007-10055-I). The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

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

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

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

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

Suggested citation: Trikalinos TA, Dahabreh IJ, Wallace BC, Schmid CH, Lau J. Towards a Framework for Communicating Confidence in Methodological Recommendations for Systematic Reviews and Meta-Analyses. Methods Research Report. (Prepared by Tufts Evidence-based Practice Center under Contract No. 290-2007-10055-I). AHRQ Publication No. 13-EHC119-EF. Rockville, MD: Agency for Healthcare Research and Quality; September 2013. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

Preface

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

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

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

We welcome comments on this Methods Research Project. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by email to epc@ahrq.gov.

Richard G. Kronick, Ph.D.
Director
Agency for Healthcare Research and Quality

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

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

Acknowledgments

The authors gratefully acknowledge the Technical Expert Panel members, the peer and public reviewers for their constructive comments.

Technical Expert Panel

Benjamin Djulbegovic, M.D., Ph.D.
Distinguished Professor of Medicine and Oncology at the University of South Florida & H. Lee Moffitt Cancer Center & Research Institute
Tampa, FL

David Moher, Ph.D.
University Research Chair in Systematic Reviews, University of Ottawa
Ottawa, Canada

Sally Morton, Ph.D.
Professor of Biostatistics and Department Chair, University of Pittsburgh
Pittsburgh, PA

Ingram Olkin, Ph.D.
Professor Emeritus of Statistics and of Education, Stanford University
Stanford, CA

Peer and Public Reviewers

Benjamin Djulbegovic, M.D., Ph.D.
Distinguished Professor of Medicine and Oncology at the University of South Florida & H. Lee Moffitt Cancer Center & Research Institute
Tampa, FL

Julian Higgins, Ph.D.
Professor of Evidence Synthesis, University of York*

Sally Morton, Ph.D.
Professor of Biostatistics and Department Chair, University of Pittsburgh
Pittsburgh, PA

Members of the Institute for Quality and Efficiency in Healthcare (in German: Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesenis)

Members of the International Society for Clinical Biostatistics Subcommittee on Statistics in Regulatory Affairs

*Affiliation during the peer review period

Towards a Framework for Communicating Confidence in Methodological Recommendations for Systematic Reviews and Meta-Analyses

Abstract

We propose a framework for organizing and describing the rationale behind methodological recommendations, and for communicating one's confidence in 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 theoretical arguments or empirical data) from nontestable ones, which represent beliefs or assumptions that are not verifiable. Nontestable statements can be justified, but their validity cannot be demonstrated. Testable statements can be assessed in terms of the adequacy of their evidentiary basis.

Both testable and nontestable statements can be evaluated regarding their feasibility of implementation, the expected impact of following them versus not, and their congruence with the desired characteristics of the background context. Considering these four dimensions, one can indicate one's confidence (along a continuum) about how closely a methods recommendation should be followed: some recommendations may be perceived and communicated as mandatory items (minimum standards), while others as highly desirable but not mandatory items. Finally, giving specific methods guidance for addressing difficult or ill-defined problems can be premature pending more research or clearer problem specification. In such cases, describing the problem and laying out attributes of a satisfactory resolution can serve until actionable guidance can be offered.

We view the proposed framework strictly as a communication tool to describe rationale for the recommendations to the intended audience and not as a device to deduce the "correctness" of a recommendation. Nonetheless, application of the framework can facilitate the latter, because methodologists can most effectively and honestly critique recommendations whose rationale is transparent. We anticipate that this initial instantiation of the framework for making methods recommendations will evolve.

Contents

Tables

Figures

Background

Evidence-based processes such as systematic reviews and meta-analyses are increasingly used to inform health care decisions. For their findings to be trustworthy, methodologies used in evidence synthesis 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. Among them, the United States Institute of Medicine (IOM) has published 21 minimum standards (corresponding to 82 elements of performance) for publicly funded systematic reviews, and the Cochrane Collaboration has developed the 80-recommendation-strong Methodological Expectations of Cochrane Intervention Reviews (MECIR) project.

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, many methodologies have found widespread use in evidence-based medicine without ever having been subjected to rigorous evaluation. While such standard methods and processes may appear to be sound in principle, they are often applied uncritically, without due consideration of their methodological appropriateness. Authoritative bodies, such as the Cochrane Collaboration or the Evidence-based Practice Centers (EPC) program, may then 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. The Meta-analysis of Observational Studies in Epidemiology (MOOSE) group recommended the use of funnel plots "to aid in the detection of publication bias", and numerous meta-analysts have done so, despite legitimate concerns regarding its interpretation and statistical properties. Other practices are common despite strong indications that they are not optimal, without having been recommended explicitly (at least to our knowledge). Examples include using only English-language studies in systematic reviews (e.g., documented by Dahabreh et al.), using summary quality scores for assessing methodological soundness, and using continuity corrections for meta-analyses in studies with rare events.

Similar to clinical practice guidelines, methodological recommendations for those conducting systematic reviews and meta-analyses 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 to practitioners. To achieve this goal, a framework for assessing the evidence supporting the methodological recommendations for research synthesis applications would be useful. Through this framework, the details of the context in which a methodological recommendation is made could be made explicit, and for a specific methodological recommendation, the approach to assess its scientific rigor, evidence base, applicability, and feasibility could be elucidated.

This work is a foray into the problem of providing consistent and transparent methodological guidance. The aim of the proposed framework is to facilitate the presentation of the rationale for methods recommendations, and the communication of the confidence of those making the recommendation that their suggestions should be followed. The framework is not a tool for deciding which recommendation is optimal, nor a checklist for verifying that due diligence was used in coming up with a specific recommendation. It is meant strictly as a transparency and communication device. We present the framework by means of examples, most of which are inspired by actual recommendation statements. In the descriptions that follow, we assume the role of those making the methodological recommendations, irrespective of whether we agree with the content or not.

Developing the Framework

We started by generating several dozens of hypothetical recommendation statements on methodological decisions often encountered when performing systematic reviews and meta-analyses. The methodological decisions 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 these candidate recommendation statements to identify an initial set of potential components for an evaluative framework.

Early versions of the framework were modified based on feedback from (1) two discussion sessions with researchers and faculty at the Center for Clinical Evidence Synthesis, Tufts Medical Center and Tufts University; (2) individual teleconference or in-person sessions with a Technical Expert Panel comprising of five internationally recognized experts in systematic review and meta-analysis; (3) a 40 minute discussion at the 2011 Society for Research Synthesis Methodology[a] meeting; and (4) an external peer and public review process. 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 reflect the opinions of the TEP or other experts who were engaged during its development.

[a] The authors of the report are members of said society.

Description, Explanation, and Elaboration

Methodological recommendations provide guidance for action when facing a choice between alternative methods. Table 1 shows three example recommendations for systematic review and meta-analysis obtained from IOM, MECIR or Agency for Healthcare Research and Quality (AHRQ) guidance. The first (R1) prescribes an optimal approach (optimal in some yet unstated sense) to identify 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 a 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 equal ("fixed") effects modeling for studies of diagnostic test accuracy. Throughout this work we use example recommendation statements for exposition, such as the ones in Table 1. We describe them from the perspective of someone who would make such a recommendation, even in cases where we are not in complete agreement with the contents of the recommendation being discussed.

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.† and MECIR, #39‡
R2	Use the Peto or the Mantel-Haenszel method for meta-analysis of rare events	AHRQ guide (reworded)§
R3	Use a random effects model for diagnostic test meta-analysis	AHRQ guide (reworded)‖

AHRQ = Agency for Healthcare Research and Quality; IOM = Institute of Medicine; MECIR = Methodological Expectations of Cochrane Intervention Reviews

*Similar recommendations have been made by other entities (not cited).

†Finding what works in healthcare. Standards for systematic reviews.[1]

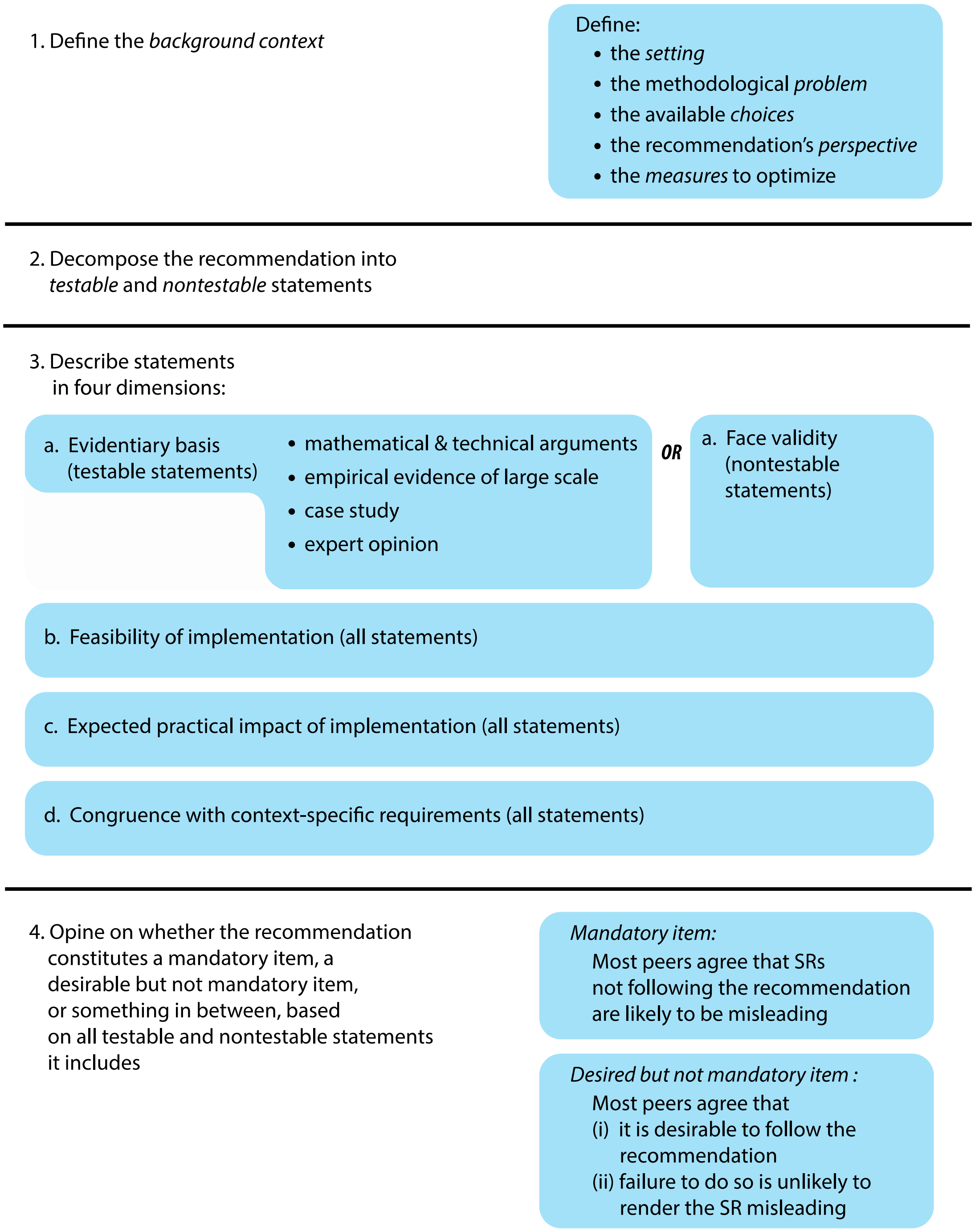
‡ Methodological standards for the conduct of new Cochrane intervention reviews.[2]

§ Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care Program.[3]

‖ Meta-analysis of test performance when there is a "gold standard."[4]

The framework outlined in Figure 1 is intended to help those developing methodological guidance to explain the rationale behind their choices. It comprises a series of four steps described in the following sections. Specifically, Define the Background Context (Step 1) explains what is needed to define the methodological challenge addressed by the recommendation. Decompose the Recommendation into Testable and Nontestable Parts (Step 2) explains how to break down (decompose) complex recommendations into more manageable parts. Describe Statements in Four Dimensions (Step 3) deals with assessing the logic behind the recommendation statement, its support, applicability, and anticipated impact; and Opine on Whether a Recommendation Constitutes a Mandatory Item or a Desirable (but not Mandatory) Item (Step 4) describes how to communicate one's confidence that the recommendation should be followed.

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


1. Define the *background context*
Define:
• the *setting*
• the methodological *problem*
• the available *choices*
• the recommendation's *perspective*
• the *measures* to optimize
2. Decompose the recommendation into *testable* and *nontestable* statements
3. Describe statements in four dimensions:
a. Evidentiary basis (testable statements)
• mathematical & technical arguments
• empirical evidence of large scale
• case study
• expert opinion
OR
a. Face validity (nontestable statements)
b. Feasibility of implementation (all statements)
c. Expected practical impact of implementation (all statements)
d. Congruence with context-specific requirements (all statements)
4. Opine on whether the recommendation constitutes a mandatory item, a desirable but not mandatory item, or something in between, based on all testable and nontestable statements it includes
Mandatory item:
Most peers agree that SRs not following the recommendation are likely to be misleading
Desired but not mandatory item :
Most peers agree that
(i) it is desirable to follow the recommendation
(ii) failure to do so is unlikely to render the SR misleading

SR = systematic review

Note: Bold horizontal lines define sequential steps. See text for explanation and elaboration.

Define the Background Context (Step 1)

All recommendations 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*; (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 measures* 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. Table 3 and Table 4 provide background contexts for R2 and R3, respectively.

Table 2. Background context for recommendation 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 health care
2	Problem	Identify all 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 decision makers 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.

AHRQ = Agency for Healthcare Research and Practice; SR = systematic review
Note: This background context was constructed using information from the IOM report "Finding What Works in Healthcare."
*Finding What Works in Healthcare. Standards for Systematic Reviews. Washington, DC: Institute of Medicine of the National Academies; 2011.

Setting

Recommendation R1 refers to publicly funded systematic reviews in health care. It is not necessary that R1, or any recommendation for that part, be transferable to settings other than the intended one. Minimum standards in the setting of publicly funded systematic reviews may be held to be desirable but not necessary in settings of very limited resources. For example, researchers in fields such as evolutionary biology and ecology (in which funding for systematic reviews from any source is scarce compared with health care) might find this recommendation exceedingly challenging to follow [personal communication, Prof. Jessica Gurevitch, State

University of NY]. The observation that methodological recommendations are embedded in specific settings is directly analogous to clinical practice guidelines that are also setting-specific. For example, the National Comprehensive Cancer Network guidelines for breast cancer screening, or prostate cancer screening and management referred primarily to health care systems in developed economies and had to be modified for use in developing countries.

Table 3. Background context for recommendation R2 ("Use the Peto or the Mantel-Haenszel method 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, most of which have rare events (<5% event rate per arm).
3	Alternative choices	Equal effect ("fixed effects") model
		Peto method
		Mantel-Haenszel method
		Inverse-variance method (fixed effect)
		Logistic regression (fixed effect)
		Exact logistic regression
		Random effects model
		Normal within-/normal between-study variance, noniterative estimation (e.g., DerSimonian-Laird)
		Normal within-/normal between-study variance, iterative estimation (e.g., REML)
		Binomial within-/normal between-study variance, iterative estimation (e.g., ML)
4	Perspective	[Same as R1]
5	Measures to optimize	Statistical bias; coverage probability; mean squared error; feasibility of implementation; SR/meta-analysis credibility.

ML = maximum likelihood; REML = restricted maximum likelihood; SR = systematic review
Note: The Table is for exposition, and was constructed using information from EPC Methods guidance documents and from our own experience as authors of EPC Methods guidance (see reference: Fu R, Gartlehner G, Grant M, et al. Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care Program. J Clin Epidemiol. 2011 Nov;64(11):1187-97.) A preference for using the odds ratio as the metric of interest is implicit here.

Table 4. Background context for recommendation R3 (“Use a random effects model for diagnostic test meta-analysis”)

#	Element of the Background Context	Description
1	Setting	[Same as R1]
2	Problem	Estimating average sensitivity and specificity in a meta-analysis of at least 10 studies
3	Alternative choices	Always use random effects modeling
		Always use equal effect modeling
		Choose between random and equal effect modeling on the basis of model best fit to the data at hand
4	Perspective	[Same as R1]
5	Measures to optimize	Maximize the generalizability of SR/meta-analysis findings.

SR = systematic review
Note: The table is for exposition, and was constructed using information from EPC methods guidance documents (see reference 4: Trikalinos TA, Balion CM, Coleman CI, et al. Chapter 8: meta-analysis of test performance when there is a "gold standard". J Gen Intern Med. Jun 2012;27 Suppl 1:S56-66.)

Problem

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

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[b] 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. Currently, all six options have been implemented in applied and methodological research projects. The advantages and disadvantages of the six options can to some extent be captured by specific measures (see below). Considering all realistic alternatives is important for an analysis of any decisional problem, and this includes making methodological recommendations. Only choices included in the list of alternatives will be considered. Consequently, if the optimal choice is not among them, it will not be selected. In this example, 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 the section on step 4).

Table 3 and Table 4 list alternative choices for R2 (methods for meta-analyzing odds ratios when events are rare) and R3 (choosing between random and equal [“fixed”] effects models for diagnostic test meta-analysis), respectively. These two examples provide the full choice sets considered by those making the recommendations.

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-

[b] We use “redundant efforts” in the Quality Control sense, i.e., processes that occur in parallel to increase the reliability of a system; we do not use the colloquial meaning of redundancy as a synonym of inefficiency.

analysis are most important. Presumably all would agree that systematic reviews should be credible, scientifically rigorous, transparent, 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. On the other hand, consumers have a key interest in patient-centered conclusions.

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, a nonprofit entity such as the Cochrane Collaboration, a professional society, a for-profit contract research organization, or an industrial firm performing an in-house assessment of one of their products. 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 deemed important for choosing among the alternative choices. Helfand and Balshem state 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. For our examples, we specify seven desired attributes of systematic reviews (Table 2, row 4) based on the rationale articulated in the introduction of the IOM report.[c]

Measures To Optimize

The background context should specify one or more measures with respect to which the alternative choices may be compared. A plausible interpretation of the background context for R1 is that the goal is to minimize the likelihood of missing research that would change the decisions informed by the systematic review. Plausible measures that can be used to operationalize whether these goals are achieved thus can include the number of eligible studies that are missed (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 (row 5 in Table 2) are not readily quantifiable in a practical manner (though other, readily quantifiable measures can be used instead). However stating them promotes transparency, and using them in a qualitative fashion helps communicate the rationale for the final recommendation.

By contrast, for technical (e.g., statistical) 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 nominally 0.95), and mean square error (desired to be as small as possible).

Finally, the background context for R3 specified a single measure to optimize, i.e., that the findings of the synthesis should be generalizable to all studies that are similar to those included in the meta-analysis. When there is only one measure to optimize the alternative choices can be ranked readily. R1 and R2 specified more than one measure to optimize, and it is possible that no single choice is optimal on all measures. Those making recommendations R1 and R2 (or any other recommendations with multiple measures to optimize) have to specify how much they care

[c] A subtle difference: the IOM panel describes these seven attributes as “criteria” for developing their methodology standards, in the sense that the panel selected standards that promote reviews with these attributes.

for each measure, explicitly or implicitly. Specifically, for R2, bias and mean square error can operate in opposite directions, because mean squared error incorporates both bias and variance. Thus, those making the recommendation must decide which measure they care most about, or more generally, how to incorporate the multiple measures in their decisionmaking. This challenge is more formally studied in the context of decision making with multiple objectives, or multi-attribute or multi-criteria decision analysis.

Decompose the Recommendation Into Testable and Nontestable Parts (Step 2)

It is important to distinguish between two types of statements found in methodological recommendations: those that can be tested in a practical manner, and those that cannot.

Testable Statements

Statements are testable when 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 compare empirically the six alternative choices listed in Table 2 in a sizable number of systematic reviews. In fact, human screening in one pass has been compared with double independent screening, and computer-assisted screening has been compared with human-only screening in limited settings. Similarly, for R2 it is practical to compare the alternative choices in Table 3 according to their statistical performance in simulation studies.

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.

Nontestable Statements

By contrast, R3 is different in a subtle but important way. R3 describes an attribute of the generative model (the model that describes the data generation process) in a meta-analysis. Thus, R3 describes a belief about the (unobserved and unknowable) process by which diagnostic accuracy studies are generated. The alternatives considered in Table 4 include random effects modeling, equal ("fixed") effect modeling, and choosing between the random and equal effects on the basis of model fit to data, i.e., on the basis of a heterogeneity test. R3 favors random effects modeling; while R3 can be motivated, it is not empirically testable. This is because, *in general, data do not suffice to determine the underlying statistical model.*[d] In particular, model fit cannot be used to infer the correct model, an intuition that has been conveyed in standard meta-analysis texts by advising to choose between equal ("fixed") and random effects a priori, and not on the basis of heterogeneity testing (Cochran's *Q* can be considered a measure of model fit).

So how does one choose the underlying true model in a meta-analysis? In the meta-analysis case, and for the alternatives in Table 4, we have some intuition about the attributes of study generating processes. With trivial (and exceedingly uncommon) exceptions, the equal ("fixed") effect scenario, which postulates that all studies have the same true effect and differ only because

[d] The exception is when one is aware of, and can measure and control for all existing confounders, and knows the time order of events: this is by design possible in a single randomized trial, but not in general –see reference 31, Robins JM, Scheines R, Spirtes P, et al. Uniform consistency in causal inference. Biometrica. 2003;90 (3):491-515, on the nonexistence of uniform consistency in causal inference.

of sampling error, is implausible: studies in a meta-analysis are unlikely to estimate a single true effect because they differ in terms of populations, interventions, comparators and outcomes, and in the details of their design and execution. It is therefore reasonable to choose random effects modeling over equal ("fixed") effect modeling.

Those making recommendation R3 apparently forgo the choice of data-driven model selection, presumably because they are comfortable in specifying the plausible model a priori. This is relatively straightforward for the background context in Table 4. In some sense, choosing between only two simple models is a "luxury" one has in the simple meta-analysis case. In other statistical problems (perhaps problems posed in a more nuanced way than R3 and), an a priori choice may be less straightforward, and many may reasonably choose the pragmatic approach of guiding model choice on the basis of model fit.[e]

Others[f] *may add more alternative choices* in Table 4, reflecting different understandings of the plausible data generation process, or a different stances to model selection altogether. In the judgment of those making the recommendation, R3 is acceptable in the given background context: peers can legitimately disagree about the acceptability of nontestable statements.

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

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, consider a composite hypothetical recommendation, within the background context in Table 5:

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

R4 can be decomposed into a nontestable statement:

R4.1: Use (i) a random effects model for the meta-analysis, and (ii) the odds ratio as the analysis metric, and a testable statement:

R4.2: Given R4.1, use a generalized random effects linear model (random effects logistic regression fit by maximum likelihood) to obtain the summary odds ratio.

The first part (i) of R4.1 is similar to R3, and has been discussed already. Part (ii) specifies using the odds ratio instead of the risk ratio, the risk difference, or other, less commonly considered metrics such as the difference in entropy or the Kullback-Leibler divergence. The choice of the metric is **arbitrary but motivated**. It is arbitrary, because the alternative options included only the small set of metrics traditionally used in clinical research. It is motivated, because it coheres with other methods choices that are common in clinical research, with statistical and epidemiological theory, and with common practices. Further—and this may be important to those making the recommendation—the log odds ratio is symmetric whereas the log risk ratio is not; and its range is the whole real axis, whereas the range of the risk difference is

[e] Relying on model fit has implications on the interpretation and contextualization of findings.

[f] A recommendation-making body other than the "hypothetical" one that made R3.

from -1 to +1. In all, 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—one can motivate its selection to communicate why it is deemed reasonable, but cannot prove its validity.

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. The following sections pertain to a real recommendation statement that is probably better handled after it has been decomposed into testable and nontestable parts.

Table 5. Background context for a composite recommendation statement constructed for exposition. R4 "Use a generalized mixed effects linear model (random effects logistic regression fit by maximum likelihood) 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 with dichotomous outcomes (>5% event rate per arm)
3	Alternative choices	Homogeneous effect
		Peto method (odds ratio metric)
		Mantel-Haenszel method (various metrics)
		Inverse-variance method (normal within study variance; various metrics)
		Logistic regression (binomial within study variance; odds ratio metric)
		Random effects
		Noniterative estimation of heterogeneity (inverse-variance, DerSimonian Laird; various metrics)
		Iterative estimation of heterogeneity (REML, normal-normal model; various metrics)
		Random effect logistic regression (ML, binomial-normal model; odds ratio metric)
4	Perspective	[Same as R1]
5	Measures to optimize	Maximize the generalizability of SR findings
		Statistical bias; coverage probability; mean squared error; maximize feasibility of implementation; maximize SR/meta-analysis credibility.

ML = maximum likelihood; REML = restricted maximum likelihood; SR = systematic review
Note: 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-1980s as a means to assess the likelihood of publication bias. In brief, publication bias is present when studies with statistically significant results are more likely to be published than those with statistically nonsignificant results. This phenomenon biases summaries of published data. The motivation for funnel plots is simple. If all studies are generated from a single underlying unimodal distribution,[g] 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

[g] See later in this section for a more explicit description.

the presence of publication bias (and assuming the generative story described below). The original paper emphasized that this test was valid in the equal effect setting. A seminal paper in the *British Medical Journal* popularized a method for testing for funnel plot asymmetry. Its authors were very careful to explain that publication bias is only one of many explanations of funnel plot asymmetry, but many others have chosen to ignore this advice. It is now common in the medical literature that authors use funnel plot asymmetry as a test for publication bias under either equal or random effect models.

The misuse of funnel plots in the literature has spurred an academic exchange with very little cross-talk between those who do not accept that funnel plot asymmetry is particularly informative on the likelihood of publication bias, and others who opted to focus on developing improved statistical tests and other methods for assessing funnel plot asymmetry. The parallel monologues continued in a methodological discussion in the Cochrane Collaboration, where the discussants agreed on technical aspects of testing for asymmetry, 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.

Nevertheless, to demonstrate how one might put these considerations into 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", 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.

R5 is vague, in that Table 6 does not state a generative model that includes publication bias. To identify the publication bias from the pattern of published studies, one has to state assumptions about (the details of) **the process that generates studies** and the **process that censors study publication** (or their complete reporting). With that in mind, R5 can be decomposed into a nontestable statement (embellished using rather typical assumptions):

R5.1 The generative model for publication bias comprises (i) study effects generated according to an equal effects model (i.e., the between study variance is 0), and (ii) a selection process, where the probability that study i is published is a function of $z_i = d_i / s_i$, with d_i the observed effect (log odds ratio) and s_i^2 its sampling variance;

and a testable statement:

R5.2 Given R5.1, regress the effect size d_i versus its precision$1/s_i$ and deduce that the funnel plot is asymmetric if the regression slope is different than 0.

The first statement (R5.1) is nontestable in that it describes an explicit model for processes that are complex and unobservable. (Other explicit versions of R5.1 are possible, and any of them, including this one, would suffice for the sake of exposition.) R5.1 implies that funnel plot asymmetry is the manifestation of publication bias. Given R5.1, funnel plot asymmetry detected by following R5.2 should be interpreted as (i) publication bias, (ii) a statistical artifact (e.g., asymmetry induced by the correlation between the log-odds ratio and its precision), or (iii) as a chance finding (type I error). **Only symmetric funnel plots are generated in the universe of R5.1, to be rendered asymmetric through the censoring/selection process of publication bias.** For example, R5.1 does not provide for a bimodal study generating distribution, i.e., for an intervention with different effects across two population strata. In such a case funnel plot asymmetry might arise without publication bias, if studies were designed efficiently: smaller (less precise) studies might be designed in the larger effect stratum, and larger (more precise) studies in the smaller effect stratum (example drawn from reference). In all, however, R5.1 is not empirically testable in a practical manner.

By contrast, R5.2 is testable in a practical manner, e.g., through simulation studies comparing alternative statistical tests (Table 7). The aforementioned consensus publication succeeds in focusing on testable statements, and avoids specifics on nontestable ones (said publication does not make explicit distinctions between testable and nontestable statements).

Table 6. Background context for recommendation R5.1 (describing a generative model for publication bias—see text)

#	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	Study effects are generated under an equal effects model. It is also assumed that the probability that studies are selected for publication as a function on the ratio $z_i = d_i / s_i$, where d_i is the treatment effect and s_i the sampling variance of study i. [Other generative models, e.g. models where the data generation process follows a random effect model, or where the probability that a study is selected for publication can be a function of d_i and s_i separately, rather than their ratio z_i.]
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

Note: This is the first part of the decomposition of a recommendation made by the MOOSE group on using funnel plots for assessing the presence of publication bias (see text). We (the authors) have little insight into the deliberations of MOOSE and thus make a good faith effort to reconstruct a plausible context. Other contexts can be constructed – but the current should suffice for exposition.

Table 7. Background context for recommendation R5.2 ("Given R5.1, regress the effect size d_i versus its precision $1/s_i$ and deduce that the funnel plot is asymmetric if the regression slope is different than 0")

#	Element of the Background Context	Description
1	Setting	[Same as R5.1 – see Table 6]
2	Problem	Assessing the likelihood of publication bias in a meta-analysis of at least 10 studies.
3	Alternative choices	Infer about the likelihood of publication bias based on funnel plot asymmetry
		Use a regression test of effect size versus its precision
		Use the trim-and-fill method
		Use visual inspection of contour enhanced funnel plots
		Infer on the likelihood of publication bias using selection models
		Search clinical trial registries and other sources and contact manufacturers to identify unpublished data
		[… another 4 options, combination of (c) with each of (a.i), (a.ii), (a.iii) or (b)]
4	Perspective	[Same as R5.1]
5	Measures to optimize	Minimize the likelihood of excluding the presence of publication bias

Note: This is the second part of the decomposition of a recommendation made by the MOOSE group on using funnel plots for assessing the presence of publication bias (see text). We (the authors) have little insight in to the deliberations of MOOSE and thus make a good-faith effort to reconstruct a plausible context. Other contexts can be constructed, but the current should suffice for exposition.

Describe Statements in Four Dimensions (Step 3)

The next step is to evaluate recommendations with respect to four dimensions: (1) **face-validity** (for nontestable statements only) or **scientific rigor and (adequacy of) evidentiary basis** (for testable statements only), (2) **feasibility of implementing** the recommendation (for all statements), (3) **expected practical impact of implementing** the recommendation (for all statement), and (4) **congruence with context-specific requirements** (for all statements). It is unclear whether qualitative descriptions are sufficient, or whether they should be further summarized by an explicit score for each dimension. To avoid encouraging an algorithmic and uncritical application of this framework we favor qualitative descriptions.

Face-Validity (for Nontestable Statements Only); and Scientific Rigor or Adequacy of Evidentiary Basis (for Testable Statements Only)

For **nontestable statements**, or nontestable parts of composite recommendations, those making the methodological recommendation should provide a justification. The justification should help readers understand the viewpoint of the nontestable recommendation, and explain why it has face validity. For example, those making the recommendation R3 (using random effects models for diagnostic test meta-analysis) may motivate their choice based on their assessment that a random effect model is a more plausible data generation model than an equal ("fixed") effect one. R4.1 (the nontestable part of R4) is similar to R3, and may be motivated using an analogous justification. R5 (and its decomposition into R5.1 and R5.2) was constructed for exposition, based on the MOOSE group statement. We (the authors) have little insight into the deliberations of MOOSE and thus make a good faith effort to reconstruct a plausible context and justification for it. For the record, we (the authors) do not agree with R5.1 (and thus with R5): We believe that R5.1 is probably too simplistic to be useful for identifying publication bias. Others can disagree. Being explicit can help identify the points of disagreement.

For **testable statements**, this dimension informs the likelihood that the statement holds, by evaluating the arguments or data that support it. As is the case for clinical practice guidance where current thinking favors different study designs for different tasks such as measuring the effectiveness of interventions, 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 to prefer for commonly encountered types of testable recommendations.

Categories of Supporting Information for Testable Statements

For description, we distinguish four categories of supporting information: (1) mathematical and technical arguments including statistical simulations; (2) empirical data of large scale; (3) case studies; and (4) 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 a technical nature. For example, it is easy to show that in a pairwise meta-analysis of intervention studies, naïve pooling (where one first combines proportions of events 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 with other linear weighting schemes (e.g., by sample size, or by inverse-standard error).

Statistical 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 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. Judging whether a simulation study is comprehensive or convincing requires content expertise and solid understanding of the methods being examined and of the background context of the methodological recommendation made. Almost invariably, this requires statistical expertise.

Empirical data of large scale[h] should be identified by means of methodology systematic reviews, such as the ones conducted by the Cochrane Methodology Review Group to inform the methods guidance provided by the Cochrane collaboration. Empirical data 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. Examples include a randomized comparison of blinding versus not blinding of data extractors to the names and affiliations of primary study authors, which showed

[h] "Large scale" means empirical data that include more than a couple of anecdotal examples. Of course, the exact number of examples included in these data will differ.

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). Not directly relevant to meta-analysis, randomized trials have been used to examine the effects of blinding vs. not blinding reviewers to author identities in journal peer-review, and having eponymous versus anonymous peer-reviewers. An ingenious study attempted to compare alternative study designs by randomizing participants to receive random versus nonrandom assignments, to empirically assess whether analysis of observational data can yield accurate answers.

Much more common are observational studies, such as those informing 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 (1,346 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. Finally, empirical data in the form of surveys of the literature can document current practices and inform on their prevalence, thereby improving the description of the background context, or even the assessment of the expected impact of a recommendation (as discussed in a later section). Simply describing prevalent practices almost never constitutes sufficient support for methods recommendations.

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 the paper by Berkey et al. that introduced a multivariate random effects meta-analysis method), or anecdotes that demonstrate the catastrophic implications or the infeasibility and inconsistency of a practice that should be avoided. Examples include documenting a large volume of unpublished trials of antipsychotics, that were highly contradictory to published data; and documenting the extreme fragility of study rankings and inconsistency of subgroup analyses produced by comparing 25 published scales rating the methodological quality of trials in a meta-analysis. Documentation of an undesirable event is valuable information, but it is seldom sufficient on its own to support methodological guidance. It is generally important to consider how often the undesirable event is likely to arise. Information from case studies is generally not sufficient for describing practices, documenting associations, or supporting the use of a methodology.

Finally, **expert opinion**,[i] as a summary of the experts' experience and intuition, can be invoked to support methodological recommendations when no other information is available or when it is necessary to interpolate or extrapolate from existing data. We believe that much of the existing guidance is based on expert opinion.

On the Adequacy of the Supporting Information

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 8 summarizes the evidentiary basis of R2. An analytical answer would give the strongest possible support 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

[i] An expert is imprecisely defined here as someone with an understanding of the nuances of the matter at hand, and one who has a broad overview of the field.

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.

We are not aware of empirical studies comparing all four methods in a large number of meta-analyses including mostly studies with rare events. In any event, an empirical comparison is inadequate to answer such questions as which method minimizes 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, case reports and expert opinion are much less useful in the presence of mathematical proof, simulation or empirical studies. They can demonstrate disagreement between methods, but cannot identify the better performing one.

One might perceive that the evidentiary base of R2 is adequate (mature), 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 8). One might have been more reserved in the absence of well-performed simulation studies, because of the inadequate evidence provided even by large-scale empirical studies. As discussed in a later section, recommendations that are based on an adequate evidence base do not automatically rise to mandatory items or minimum standards.

Table 8. Description of the evidentiary basis of R2 (“Use the Peto or the Mantel-Haenszel method for meta-analysis of rare events”)

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. In sum, the Peto or the Mantel-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 about the concordance of alternative choices in real life situations, but cannot inform about which choices are closer to the truth.
Case studies (proof of concept studies)	Isolated examples (nonexhaustive list references).	These are examples that alternative methods can yield different results.
Expert opinion	An example is reference.	Generic or irrelevant (nonmethodological) comments; no consideration of simulation data or empirical evidence.

Note: The contents of this table are not based on a systematic review of the methods literature; they are provided for exposition, and in good faith.

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 9). 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; in another limited assessment, single screening by a human was estimated to miss between 0 and 24 percent of finally eligible studies compared with more comprehensive screening (estimation based on a capture-recapture approach).

One might suggest that the evidentiary base of R1 is not robust, because no large-scale empirical studies exist (i.e., no supporting information of the appropriate type is available). One cannot use a mathematical or technical argument for R1, because measures to be optimized in the background context in Table 2 require specific empirical data.

Table 9. Description of the evidentiary basis of R1 ("Use two or more members of the review team, working independently, to screen and select studies")

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.
Case studies (proof of concept studies)	In an analysis of four screeners in a large systematic review, the estimated proportion of eligible citations missed by a single reviewer was between 0 and 24%. Computer-assisted screening has been compared with single or duplicate human screening in a small number of examples.	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.	No references are provided.

IOM = Institute of Medicine; MECIR = Methodological Expectations of Cochrane Intervention Reviews
Note: The contents of this Table are not based on a systematic review of the methods literature; they are provided for exposition, and in good faith.

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, of course, be more challenging to implement for groups that do not have such expertise or resources. For example, Cochrane and EPC guidance recommend the bivariate model with binomial likelihood, or the hierarchical summary ROC (HSROC) model 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 advanced meta-analysis. The same resource levels apply to 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 computer-assisted processes will not be widely feasible. By contrast, the recommended method in R2 for analyzing meta-analyses of studies with rare events is implemented in many widely available software tools, and requires only a very basic understanding of statistical concepts. Thus, R2 is probably quite feasible.

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 with substantial organizational and other support. One might deem that implementing R1 for (most if not all) publicly funded reviews in health care is probably feasible. However, this recommendation may not be feasible as a minimum standard for systematic reviews that do not have robust funding.

Expected Practical Impact of Implementation

The practical impact of following versus not following 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 statistical methods. 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 allow for between-study heterogeneity. However, in a large empirical evaluation of more than 300 meta-analyses of diagnostic test accuracy, estimates and confidence intervals for the primary metrics of sensitivity and specificity rarely differed substantially when comparing bivariate and separate univariate analyses. Comparing univariate versus multivariate meta-analysis for estimating marginal treatment effects also failed to find much benefit in the more complex multivariate approach for the major parameters of interest. However, both studies did suggest that the multivariate analyses might be beneficial in evaluating secondary parameters such as linear combinations of sensitivity and specificity. Thus, the choice between more advanced and simpler methods depends also on the measures of 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 other options for redundant screening in Table 2 is not obvious. For R2, it is unclear how often conclusions would change using alternative methods.[j]

Congruence With Context-Specific Requirements

This dimension captures requirements 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.

For example, R1 is probably 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 about R2.

Opine on Whether a Recommendation Constitutes a Mandatory Item or a Desirable (but not Mandatory) Item (Step 4)

While users should preferably follow well-developed methodological guidance whenever possible, certain items may be more important in a resource limited setting. Those making the

[j] This assessment can vary across background contexts.

recommendations may wish to convey that some recommendations represent **truly minimal standards** or **truly mandatory items**, in that they are perceived essential for producing high quality systematic reviews and meta-analyses. Failure to meet such minimal standards or mandatory items would raise concerns about the validity and comprehensiveness of the review's findings and conclusions. Other recommendations may be judged to represent **desirable—**but not mandatory practices. For this category, most peers would agree that following the recommendation would improve the systematic review or meta-analysis, and that not following the recommendation unlikely render the review unusable.

Assuming that one would never recommend generally harmful practices, there potentially remains a third category of recommendations whose utility or necessity remains unclear. We believe that this third category should be empty because **one should only make recommendations that one deems clearly beneficial** given the background context at hand. 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, e.g., along the lines of performing sensitivity analyses to assess robustness of results.

Those making methodological recommendations can also make comments or statements about the insufficiency of information to make a recommendation. This is analogous to the "insufficient evidence" statements ("I" statements) that the U.S. Preventive Services Task Force makes when the information is not sufficient to support a recommendation for or against an action.

A careful and balanced discussion of the aforementioned dimensions within the background context of the recommendation may help clarify the rationale for the recommendation, and may help deduce whether the recommendation is a minimum standard/mandatory item or a best practice/highly desirable item.

R1 ("use two or more members of the review team, working independently, to screen and select studies") was considered a minimum standard in the IOM report, and a mandatory item in MECIR. Table 10 summarizes ones thoughts on R1. R1, may be 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. One might prefer 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) are 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).

Many peers would probably agree that R2 is a minimum standard or mandatory item, presumably because it has an adequate evidentiary support, it is feasible and congruent with the requirements of scientific rigor and credibility, and despite the fact that its practical impact is not clear. 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 already, it has face validity. It is also very feasible and easy to implement. It is also likely that it can have an impact on the conclusions of systematic reviews, and it is in agreement with the desired characteristics of the background context.

Table 10. 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 o or the Mantel-Haenszel method rather than inverse variance weighting for meta-analysis of rare events	Use random effects models for diagnostic test meta-analysis
Is the recommendation testable?	Yes	Yes	No*
Adequacy of evidentiary basis or scientific rigor (testable statement) / face validity (nontestable statement)	See Table 9	See Table 8	A random rather than equal ("fixed") effect model is more plausible. Choosing models based on data is a needless compromise if a plausible generative story can be posed.
Feasibility of implementation	No obstacles to feasibility identified	No obstacles to feasibility identified	No obstacles to feasibility identified
Expected impact of implementation	Unclear	Unclear	May result in some practical changes in conclusions because uncertainty is more fully modeled compared with equal ("fixed") effects models
Congruence with context-specific requirements	High	High	High
Overall confidence that the recommendation is a mandatory item			
IOM	High	NA	NA
MECIR	High	NA	NA

IOM = Institute of Medicine; MECIR = Methodological Expectations of Cochrane Intervention Reviews; NA = not applicable (not included in IOM or MECIR guidance)

Discussion

We have proposed a framework for evaluating methodological recommendations, and for organizing and describing the logic behind them. We started by defining the background context of the recommendations. 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 recommendations can be justified, but their validity cannot be demonstrated by means of a "test". Testable recommendations 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 requirements of the background context. Considering these four dimensions, one might opine if a recommendation is a minimum standard or a mandatory item, a desirable (but not mandatory) item, or something in between.

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 such stringency 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 the suggested approaches may not be optimal.

We are ambivalent about the need for explicit scoring of testable statements on their four dimensions. Explicit scoring of individual dimensions is as subjective as rating the overall strength of the recommendation. It is conceivable that the same recommendation could receive different descriptions in some dimensions (e.g., feasibility of implementation, congruence with context-specific requirements), given different background contexts. On the other hand, explicit scoring could be used to abstract the logic for deducing whether the recommendation is a mandatory item or a desirable but not mandatory practice. We lean in favor of avoiding explicit scoring of the individual dimensions.

The proposed framework calls for a structured set of subjective judgments. Extensive as they may be, the subjectivity in rating the dimensions of testable statements, and the fuzziness of the dimensions themselves are probably much less extensive than the subjectivity and fuzziness of recommendations that do not have a defined background context, are composites of testable and nontestable statements, are in want of an organized categorization of their supporting information, or 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 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), 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,[k] which is similar to standard #C39 in MECIR) to illustrate our framework. We would probably agree with IOM and MECIR that this recommendation should be followed if the only alternative is single screening, but point out that as more options become available (e.g., computer-aided screening by humans), this recommendation might change.

In practice, an additional complication arises: When many methodological recommendations are made that tax resources and demand expertise, adhering to all of them can be a rather expensive proposition, even within well-funded programs. Deciding which recommendations are most important to adhere to and which are not is very difficult. In theory, those making the recommendations might try to rank them by relative merit, given the expected resource constraints, by following the same process as we discussed here for single recommendations. In reality this is impractical. For example, the feasibility of implementation is *consumer-dependent*, e.g., some systematic review teams have extensive expertise, others less so. It is also *domain- and topic-dependent*, in that adhering to all recommendations is feasible for reviews of small enough scope. Analogous considerations apply to the impact of the implementation. Fine tuning the relative merit of a large number of recommendations is thus too much to ask of those providing general methodological guidance, e.g., the IOM or Cochrane's MECIR. By contrast, very explicit prioritizations can be meaningful in narrow contexts, e.g., when developing internal guidance for a specific program or team. In the end, systematic reviewers will prioritize the methodological recommendations they can adhere to when they do a systematic review, given their operational abilities and budget.

Deeper epistemological and philosophical discussions can be had with respect to the validity or "truth" of methodological recommendation statements. The device employed in this work, namely distinguishing testable from nontestable statements, is utilized without explicit (at least without conscious or intended) reference to philosophical theories of truth. Djulbegovic et al. posit that Evidence-based Medicine as a (still evolving) structure for optimizing clinical practice that draws on several major philosophical theories of scientific evidence, but is in want of a rigorous epistemological stance. The proposed framework for communicating the strength of methodological recommendations is analogously (non)nuanced epistemologically: our designation of (empirically) nontestable statements draws from the coherence theory of truth, according to which a statement is considered true if it squares with other statements. The origins of nontestable statements can be traced, because they are developed within a historical body of knowledge in which the arguments strive to be internally consistent. So, ideally each (empirically) nontestable statement should be referred to the body of knowledge that led to its inclusion in the recommendations—a daunting but not impossible task. By contrast, our (empirically) testable statements reflect the correspondence theory of truth, in which arguments correspond with reality, and "truth" is based on the correspondence of ideas with facts.

The feasibility and ease of applying the proposed framework itself is not clear. 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 so will this proposal.

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

References

1. Anonymous. 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, et al. Methodological standards for the conduct of new Cochrane intervention reviews. 2011; 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. 2011;64 (11):1187-1197. PMID: 21477993.

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. 2012;27 Suppl 1 (S56-66). PMID: 22648676.

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. 2000;283 (15):2008-2012. PMID: 10789670.

7. Lau J, Ioannidis JP, Terrin N, et al. The case of the misleading funnel plot. BMJ. 2006;333 (7568):597-600. PMID: 16974018.

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. 2005;58 (9):894-901. PMID: 16085192.

9. Terrin N, Schmid CH, Lau J, et al. Adjusting for publication bias in the presence of heterogeneity. Statistics in medicine. 2003;22 (13):2113-2126. PMID: 12820277.

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. 2000;53 (11):1119-1129. PMID: 11106885.

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. 2005;58 (9):882-893. PMID: 16085191.

12. 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;22553887. PMID: N/A.

13. Juni P, Witschi A, Bloch R, et al. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA : the journal of the American Medical Association. 1999;282 (11):1054-1060. PMID: 10493204.

14. Greenland S. Invited commentary: a critical look at some popular meta-analytic methods. American journal of epidemiology. 1994;140 (3):290-296. PMID: 8030632.

15. Greenland S. Quality scores are useless and potentially misleading. American journal of epidemiology. 1994;140 (3):301-302. PMID: 8030632.

16. Bradburn MJ, Deeks JJ, Berlin JA, et al. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. Statistics in medicine. 2007;26 (1):53-77. PMID: 16596572.

17. Rucker G, Schwarzer G, Carpenter J, et al. Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. Statistics in medicine. 2009;28 (5):721-738. PMID: 19072749.

18. 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. 2004;23 (9):1351-1375. PMID: 15116347.

19. Helfand M, Balshem H. AHRQ series paper 2: principles for developing guidance: AHRQ and the effective health-care program. Journal of clinical epidemiology. 2010;63 (5):484-490. PMID: 19716268.

20. 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. 2009;7 (10):1060-1096. PMID: 19930975.

21. 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. 2010;8 (2):240-262. PMID: 20141680.

22. Mohler J, Bahnson RR, Boston B, et al. NCCN clinical practice guidelines in oncology: prostate cancer. Journal of the National Comprehensive Cancer Network : JNCCN. 2010;8 (2):162-200. PMID: 20141676.

23. 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. 2010;8 Suppl 3 (S8-S15). PMID: 20697133.

24. 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. 2010;8 Suppl 3 (S26-28. PMID: 20697128.

25. 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. 2012;22481134.

26. Wallace BC, Trikalinos TA, Lau J, et al. Semi-automated screening of biomedical citations for systematic reviews. BMC bioinformatics. 2010;11 (55). PMID: 20102628.

27. 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. 2004;8 (36):iii-iv, ix-xi, 1-158. PMID: 15361314.

28. 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. 1999;12 (6):477-485. PMID: 10612366.

29. Keeney RL, Raiffa H. Decisions with multiple objectives: preferences and value trade-offs. New York: Cambridge University Press; 1993

30. Edwards P, Clarke M, DiGuiseppi C, et al. Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records. Statistics in medicine. 2002;21 (11):1635-1640. PMID: 12111924.

31. Robins JM, Scheines R, Spirtes P, et al. Uniform consistency in causal inference. Biometrica. 2003;90 (3):491-515. PMID: N/A.

32. Egger M, Davey Smith G, Schneider M, et al. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315 (7109):629-634. PMID: 9310563.

33. Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. Journal of clinical epidemiology. 2001;54 (10):1046-1055. PMID: 11576817.

34. 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. 2007;176 (8):1091-1096. PMID: 17420491.

35. Rucker G, Schwarzer G, Carpenter J. Arcsine test for publication bias in meta-analyses with binary outcomes. Statistics in medicine. 2008;27 (5):746-763. PMID: 17592831.

36. 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. 2006;25 (20):3443-3457. PMID: 16345038.

37. Moreno SG, Sutton AJ, Thompson JR, et al. A generalized weighting regression-derived meta-analysis estimator robust to small-study effects and heterogeneity. Statistics in medicine. 2012;31 (14):1407-1417. PMID: 22351645.

38. Peters JL, Sutton AJ, Jones DR, et al. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. Journal of clinical epidemiology. 2008;61 (10):991-996. PMID: 18538991.

39. 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. PMID: 21784880.

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

41. Burton A, Altman DG, Royston P, et al. The design of simulation studies in medical statistics. Statistics in medicine. 2006;25 (24):4279-4292. PMID: 16947139.

42. Hopewell S, Clarke M, Lefebvre C, et al. Handsearching versus electronic searching to identify reports of randomized trials. Cochrane Database Syst Rev. 2007;2):MR000001. PMID: 17443625.

43. Hopewell S, McDonald S, Clarke M, et al. Grey literature in meta-analyses of randomized trials of health care interventions. Cochrane Database Syst Rev. 2007;2):MR000010. PMID: 17443631.

44. Young T, Hopewell S. Methods for obtaining unpublished data. Cochrane Database Syst Rev. 2011;11):MR000027. PMID: 22071866.

45. Dwan K, Altman DG, Cresswell L, et al. Comparison of protocols and registry entries to published reports for randomised controlled trials. Cochrane Database Syst Rev. 2011;1):MR000031. PMID: 21249714.

46. Odgaard-Jensen J, Vist GE, Timmer A, et al. Randomisation to protect against selection bias in healthcare trials. Cochrane Database Syst Rev. 2011;4):MR000012. PMID: 21491415.

47. Welch V, Tugwell P, Petticrew M, et al. How effects on health equity are assessed in systematic reviews of interventions. Cochrane Database Syst Rev. 2010;12):MR000028. PMID: 21154402.

48. Berlin JA. Does blinding of readers affect the results of meta-analyses? University of Pennsylvania Meta-analysis Blinding Study Group. Lancet. 1997;350 (9072):185-186. PMID: 9250191.

49. 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. 1998;280 (3):237-240. PMID: 9676667.

50. McNutt RA, Evans AT, Fletcher RH, et al. The effects of blinding on the quality of peer review. A randomized trial. JAMA : the journal of the American Medical Association. 1990;263 (10):1371-1376. PMID: 2304216.

51. van Rooyen S, Godlee F, Evans S, et al. Effect of blinding and unmasking on the quality of peer review: a randomized trial. JAMA : the journal of the American Medical Association. 1998;280 (3):234-237. PMID: 9676666.

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

53. 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. 2008;336 (7644):601-605. PMID: 18316340.

54. 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). PMID: 21401947.

55. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Statistics in medicine. 2008;27 (12):2037-2049. PMID: 18038446.

56. Berkey CS, Hoaglin DC, Antczak-Bouckoms A, et al. Meta-analysis of multiple outcomes by regression with random effects. Statistics in medicine. 1998;17 (22):2537-2550. PMID: 9839346.

57. 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. PMID: 22448149.

58. 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. PMID: 18375649.

59. Hernandez AV, Walker E, Ioannidis JP, et al. Challenges in meta-analysis of randomized clinical trials for rare harmful cardiovascular events: the case of rosiglitazone. American heart journal. 2008;156 (1):23-30. PMID: 18585493.

60. 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). PMID: 17244367.

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

62. Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. Journal of clinical epidemiology. 2006;59 (12):1331-1332; author reply 1332-1333. PMID: 17098577.

63. 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. 2004;57 (9):925-932. PMID: 15504635.

64. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Statistics in medicine. 2001;20 (19):2865-2884. PMID: 11568945.

65. Wallace BC, Dahabreh IJ, Trikalinos TA, et al. Closing the Gap between Methodologists and End-Users: R as a Computational Back-End. J Stat Softw. 2012;49 (5):1-15. PMID: N/A.

66. Wallace BC, Small K, Brodley CE, et al. 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. PMID: N/A.

67. Wallace BC, Schmid CH, Lau J, et al. Meta-Analyst: software for meta-analysis of binary, continuous and diagnostic data. BMC medical research methodology. 2009;9 (80. PMID: 19961608.

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

69. 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. 2008;61 (11):1095-1103. PMID: 19208372.

70. Trikalinos TA, Olkin I. Meta-analysis of effect sizes reported at multiple time points: A multivariate approach. Clin Trials. 2012 Oct;9(5):610-20; PMID: 22872546.

71. Trikalinos TA, Olkin I. A method for the meta-analysis of mutually exclusive binary outcomes. Statistics in medicine. 2008;27 (21):4279-4300. PMID: 18416445.

72. Petitti DB, Teutsch SM, Barton MB, et al. Update on the methods of the U.S. Preventive Services Task Force: insufficient evidence. Annals of internal medicine. 2009;150 (3):199-205. PMID: 19189910.

73. Anonymous. Methodological standards and patient-centeredness in comparative effectiveness research: the PCORI perspective. JAMA : the journal of the American Medical Association. 2012;307 (15):1636-1640. PMID: 22511692.

74. Djulbegovic B, Guyatt GH, Ashcroft RE. Epistemologic inquiries in evidence-based medicine. Cancer control : journal of the Moffitt Cancer Center. 2009;16 (2):158-168. PMID: 19337202.

75. Young JO. The Coherence Theory of Truth In: Zalta EN, ed. The Stanford Encyclopedia of Philosophy. (Summer 2013) ed. Stanford, CA: The Metaphysics Research Lab, Stanford University; 2013. Accessed June 05, 2013.

76. David M. The Correspondence Theory of Truth In: Zalta EN, ed. The Stanford Encyclopedia of Philosophy. (Summer 2013) ed. Stanford, CA: The Metaphysics Research Lab, Stanford University; 2013. Accessed June 05, 2013.