A Primer on Performing Systematic Reviews and Meta-analyses

Craig A. Umscheid[1,2,3,4,5]

[1]Center for Evidence-based Practice, [2]Department of Medicine, [3]Center for Clinical Epidemiology and Biostatistics, [4]Leonard Davis Institute of Health Economics, and [5]Institute for Translational Medicine and Therapeutics, University of Pennsylvania, Philadelphia

The number of systematic reviews published in the peer-reviewed literature has increased dramatically in the last decade, and for good reason. They have become an essential resource for clinicians who want unbiased and current answers for their clinical questions; researchers and funders who want to identify the most critical evidence gaps for study; payers and administrators who want to make coverage, formulary, and purchasing decisions; and policymakers who want to develop quality measures and clinical guidelines. Targeted to beginners interested in conducting their own systematic reviews and users of systematic reviews looking for a brief introduction, this primer (1) highlights the differences between review types; (2) outlines the major steps in performing a systematic review; and (3) offers a set of resources to help authors perform and report valid and actionable systematic reviews.

Keywords. systematic review; meta-analysis; literature searching; heterogeneity; reporting bias.

The number of systematic reviews and meta-analyses published in the peer-reviewed literature has increased dramatically in the last decade, and for good reason [1]. They have become an essential resource for clinicians who want unbiased and up-to-date answers for their clinical questions; researchers and funders who want to identify the most critical evidence gaps for study [2]; payers and administrators who want to make coverage, formulary, and purchasing decisions [3–5]; and policy makers who want to develop quality measures and clinical practice guidelines [5–7]. As such, their impact can be potent. This makes it paramount that their conduct and interpretation are consistent with the most current and valid standards. This primer describes the differences between review types; outlines the major steps in performing a systematic review; and offers a set of resources for the appraisal, reporting, and performance of systematic reviews. It is written for beginners interested in writing their own systematic reviews and users of systematic reviews who are looking for a brief introduction to the process.

DEFINITIONS

A *narrative review* is the current term used to describe traditional reviews authored by recognized experts in a field. It is the most common method of summarizing a field. However, because such reviews lack systematic methods to identify, appraise, and synthesize information, they have a higher risk of bias than systematic reviews, as there is potential for authors to selectively include or exclude studies to support a position.

Unlike a narrative review, a *systematic review* is guided by key questions and a protocol for conduct, much like other scientific studies, thus mitigating bias. Having a systematic approach to answer a key question also allows systematic reviewers to identify critical evidence gaps, which researchers and funders can use to prioritize

Received 1 November 2012; accepted 4 May 2013; electronically published 22 May 2013.

Presented in part: Advanced Epidemiologic Methods Track, Society for Healthcare Epidemiology of America Meeting, Jacksonville, Florida, 14 April 2012.

Correspondence: Craig A. Umscheid, MD, MSCE, FACP, University of Pennsylvania Perelman School of Medicine, Penn Medicine Center for Evidence-based Practice, 3535 Market St, Ste 50, Philadelphia, PA 19104 (craig.umscheid@uphs.upenn.edu).

Clinical Infectious Diseases 2013;57(5):725–34

© The Author 2013. Published by Oxford University Press on behalf of the Infectious Diseases Society of America. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com.

DOI: 10.1093/cid/cit333

Downloaded from https://academic.oup.com/cid/article-abstract/57/5/725/311245 by Dana Medical Library, University of Vermont user on 26 June 2019

their research agendas [2]. Furthermore, a well-reported systematic review facilitates replication or updating of the review by others. Some systematic reviews also include a *meta-analysis*, which statistically pools the results of individual studies to produce a single estimate of effect.

Beyond addressing the limitations of narrative reviews, systematic reviews can also address the limitations of individual studies. Although well-designed and -conducted studies can provide unbiased information, there is always a risk of false-positive results, a risk of false-negative results for underpowered studies, and the potential for poor generalizability for studies performed in relatively limited and homogenous patient populations. By reviewing multiple studies performed in varying patient populations using slightly different interventions and outcomes, systematic reviews and meta-analyses often provide more precise estimates of effect, reduce the risk of false-positive and false-negative results, improve generalizability of findings, and allow for the exploration of differences in findings that exist between studies [1].

Other less commonly encountered review types are described in Table 1, and include the *integrative review* [8] and *individual patient-level meta-analyses* [9–11].

Table 1. Descriptions of Less Common Types of Reviews

Review Type	Description
Integrative review	Integrative reviews are the broadest type of review methodology available and include descriptive, experimental, and theoretical data in an attempt to more fully understand a topic of interest. They can be used to define concepts, analyze methodological approaches, and review theories and clinical evidence, and as a result are well suited to comprehensively address what are often complex issues arising from the nursing field [8].
Individual patient-level meta-analyses	Individual patient-level meta-analyses are similar to the more typical study-level meta-analyses described in the text, but instead of statistically combining study results at the level of the individual study, they statistically combine results at the level of the individual patients in the individual studies. This requires the meta-analyst to obtain individual patient-level data from each of the lead authors for the studies eligible for the meta-analysis. Although most evidence suggests that meta-estimates resulting from the 2 meta-analytic approaches are similar, the primary benefit of the patient-level meta-analysis is the ability to adjust for potential confounders of treatment effect as well as perform subgroup analyses to identify those in whom the intervention may be more effective. The major drawback to patient-level meta-analyses is the increase in time and cost required to obtain patient-level data, as well as the potential inability to obtain all patient-level data sets, resulting in a potentially biased subset of studies to analyze [9–11].

STEPS FOR PERFORMING OR APPRAISING A SYSTEMATIC REVIEW

Define the Question

The most critical part of a systematic review is asking the right question. If a systematic review is conducted in a methodologically flawless manner, but the clinical question(s) addressed are of little consequence to patients or providers, the review will have marginal clinical utility [12]. Questions asked in a systematic review most commonly address the effectiveness or safety of a therapeutic intervention, or the performance of a diagnostic test. Arriving at the correct question(s) requires the input and engagement of a multidisciplinary group of stakeholders, which often includes providers from the relevant specialties (usually >1), patients, payers, and policy makers (eg, hospital administrators in systematic reviews performed for local policy [3, 4], or specialty societies for reviews performed at the national or international level [13]) [14–17]. The acronym PICO (population, intervention, comparator, outcomes) helps clarify the critical elements of the key question(s) (Table 2) [18, 19].

Establish Eligibility Criteria

Once the question is defined, one must establish criteria to systematically include or exclude studies from the review. Limiting included studies by study design is one common approach (eg, only including randomized controlled trials [RCTs] if one is focused on the efficacy of an intervention; RCTs, cohort, and case-control studies if one is interested in the real-world effectiveness or safety of an intervention; or cross-sectional studies if one is interested in the characteristics of a diagnostic test). Other popular eligibility criteria include publication year and a minimum period of follow-up for a given outcome. As an example, if one is interested in the impact of a particular type of antimicrobial-impregnated prosthetic joint on joint infections at 1 year, one might exclude older studies examining obsolete versions of the device, as well as studies limited to analyzing the impact of the device on joint infections during the postoperative hospital stay [20].

One might also limit the search to English-language studies only, or by whether or not the study was published in a peer-reviewed journal; however, such exclusions may increase the risk of reporting bias. Such reporting biases most often occur when studies published in the English-language or the

Downloaded from https://academic.oup.com/cid/article-abstract/57/5/725/311245 by Dana Medical Library, University of Vermont user on 26 June 2019

Table 2. Description of "PICO" Elements With an Example [19]

PICO Element	Description	Example [19]
Population	What is the age range, sex, race, diagnosis, or disease severity of the patient population of interest? For example, one might be interested in all patients with heart failure, or only those with a particular severity class who have been recently discharged from the hospital.	Adult patients admitted to medical and surgical intensive care units
Intervention	What is the specific drug, device, test, or clinical practice of interest? Are there particular doses, frequencies, delivery routes, or treatment settings of interest? One might be interested in one particular treatment, or a combination of interventions that may be used in a test-and-treat strategy.	Antimicrobial therapy guided by the use of serum procalcitonin levels
Comparator	Is the comparator a placebo, standard care, or an active comparator?	Antimicrobial therapy without procalcitonin guidance
Outcomes	Are the outcomes of interest process measures, surrogate outcomes, clinical outcomes, and/or cost?	Mortality, duration of antimicrobial therapy, intensive care unit, and hospital length of stay

Abbreviation: PICO, population, intervention, comparator, outcomes.

peer-reviewed literature are more likely to be "positive" than those that are not [21–25]. Here, a "positive" study is defined as one where the findings are statistically significant and support the intervention of interest. Reporting biases such as "language" or "publication" bias often result in more favorable findings in a systematic review than would have otherwise occurred if non-English-language or unpublished studies were included. Conversely, including unpublished studies in a review might reduce the validity of the review, because studies that are not published in the peer-reviewed literature but are instead in abstract form are often preliminary work, the results of which can change significantly with further analysis and peer review [26]. Furthermore, including non-English-language studies might increase the time and cost of a review, which might not be feasible if the review is a rapid one to inform local hospital policy, and excluding them may not ultimately impact the final results of the review [27, 28].

When considering eligibility criteria, one must consider how sensitive or specific the search needs to be. If one is performing a review to inform a national guideline, the search should maximize sensitivity. Thus, there may be fewer eligibility criteria. If one is performing a rapid review to inform policy at a local institution, one might choose to perform a search with more specificity with more narrow eligibility criteria [29].

Search the Literature

After the key question(s) and eligibility criteria are finalized, one needs to devise a search strategy, and decide what resources to search. Librarians can be extremely helpful in translating key question(s) and eligibility criteria into search strategies using the most appropriate free text and structured terms (such as Medical Subject Heading [MeSH] terms in PubMed [30–33], or Emtree terms in Embase). They can also help combine these

Table 3. Examples of Search Filters for Identifying Various Study Designs in OVID Medline

Systematic Reviews [34]	Randomized Controlled Trials [38]	Observational Studies [37]	Diagnostic Studies [36]
(Medline or systematic review).tw or meta analysis.pt	1. randomized controlled trial.pt. 2. controlled clinical trial.pt. 3. randomized.ab. 4. placebo.ab. 5. clinical trials as topic.sh. 6. randomly.ab. 7. trial.ti. 8. or/1–7 9. exp animals/ not humans.sh. 10. 8 not 9	1. epidemiologic studies/ 2. exp case control studies/ 3. exp cohort studies/ 4. case control.tw. 5. (cohort adj (study or studies)).tw. 6. cohort analy$.tw. 7. (follow up adj (study or studies)).tw. 8. (observational adj (study or studies)).tw. 9. longitudinal.tw. 10. retrospective.tw. 11. cross sectional.tw. 12. cross-sectional studies/ 13. or/1–12	1. exp sensitivity and specificity/ 2. specificity.tw. 3. false negative.tw. 4. accuracy.tw. 5. screening.tw. 6. or/1–5

Abbreviations: ab, abstract; adj, adjacent; exp, explode; pt, publication type; sh, medical subject heading subheadings; ti, title; tw, text words.

Downloaded from https://academic.oup.com/cid/article-abstract/57/5/725/311245 by Dana Medical Library, University of Vermont user on 26 June 2019

terms into concepts, and link these concepts using the appropriate Boolean operators (ie, AND, OR, and NOT). Last, they can help one use validated search filters (or "hedges") to restrict one's search to particular study designs (Table 3) [34–38], and help one select the electronic literature databases most relevant to one's key question(s) (Table 4) [39]. For all of these reasons, the Institute of Medicine (IOM), in its recent report on standards for systematic reviews, recommends including librarians in the conduct of systematic reviews [31].

Besides searching electronic literature databases, one should also consider searching reference lists of those studies selected for inclusion as well as conference proceedings. In addition, trial registries such as ClinicalTrials.gov can help one identify studies not published in the peer-reviewed literature. Moreover, hand searching of high-value journals can help identify studies that are poorly indexed. Although this may seem impracticable, this can be high yield if one's question is particularly narrow such that there are only a few key journals in the field.

During the search process, one should share the search results with experts in the field to ensure the search has not missed any relevant studies. One might also seek the input of pharmaceutical or device manufacturers, particularly when trial registries include relevant ongoing or completed trials performed by the manufacturer that have not yet been published.

To maximize the yield of one's searches, one should review a small random sample of hits generated from one's search for each of the databases one is considering using, such that one can distinguish those databases with the most relevant studies. When performing a rapid review, this can help one focus their search on the 1 or 2 highest-yield databases and make the process more efficient.

After running the searches, one should combine the hits from the various databases used into 1 reference manager and remove duplicate publications. One will then need to decide whether to have 1 independent review of titles and abstracts to select full text to review more closely, or 2 independent reviews of titles and abstracts, with resolution of disagreements by consensus or a third independent review. This decision will rest on the resources available to conduct the review (ie, staff and funds), as well as the timeline of the review (eg, weeks vs months). As recommended by current standards, one should ideally have >1 independent review for each of the steps, as having multiple independent reviews helps to ensure eligibility criteria are applied in an unbiased way [31, 38]. If one needs to balance rigor of the review with efficiency of the process, one might choose to have 1 independent review of titles and abstracts to determine which full text to review (the default for the reviewer here would be to include any study that he or she is uncertain about), and then 2 independent reviews of the full text for inclusion in the review. For rapid reviews performed and used in local settings, one might choose to have 1 independent review for title, abstract, and full text screening; however, if using a single reviewer, the reviewer needs to be an experienced methodologist.

Table 4. Databases to Consider Searching in a Systematic Review of the Medical Literature

Database	Description
Medline	The premier database of medical research.
	Should generally be used in every systematic review.
	PubMed is the free interface , and OVID Medline is an interface commonly supported by institutional subscriptions that allows for more complex searching
	The list of references generated for any given search run in Medline should be similar regardless of the interface used; the main exception is that articles awaiting indexing by librarians at the National Library of Medicine will be included in PubMed as "in process," but not in the standard OVID Medline interface. To capture these articles in the search, one must use either PubMed or the OVID PreMedline database in addition to the standard OVID Medline.
Embase	A large European database similar in scope and content to Medline, but up to 70% of citations in Embase are not in Medline.
	Particularly good resource for pharmacoepidemiology or pharmacoeconomics topics.
	Many systematic reviews search both Medline and Embase, although some evidence suggests that the incremental value of searching Embase in the context of a Medline search may be limited [39].
Cochrane Library	Commonly used in systematic reviews, particularly for its CENTRAL database, which is a compilation of controlled clinical trials meticulously maintained by the many Cochrane Review Group.
CINAHL	A valuable database when questions involve the fields of nursing or quality and safety, or when searching for qualitative studies.
PsycINFO	Helpful when addressing questions in the area of mental health.

Data Extraction

After the studies are selected for inclusion, one needs to set priorities for data to abstract and avoid the temptation to extract everything from the identified studies. To help one prioritize data elements to extract, anticipate the structure and content of the evidence tables in the review. The most common data elements to extract include study author, year, design, population, setting, sample size, definitions, and results of clinically relevant outcomes, and information needed to assess the risk of bias of the individual study, such as randomization procedures [40]. To ensure the data are extracted accurately, quality assurance procedures should be established. Ideally, a review would have >1 individual independently extracting the data and then resolving discrepancies. If this is not possible, then extraction by a single reviewer should be followed by another review to check the accuracy of the data extracted.

Downloaded from https://academic.oup.com/cid/article-abstract/57/5/725/311245 by Dana Medical Library, University of Vermont user on 26 June 2019

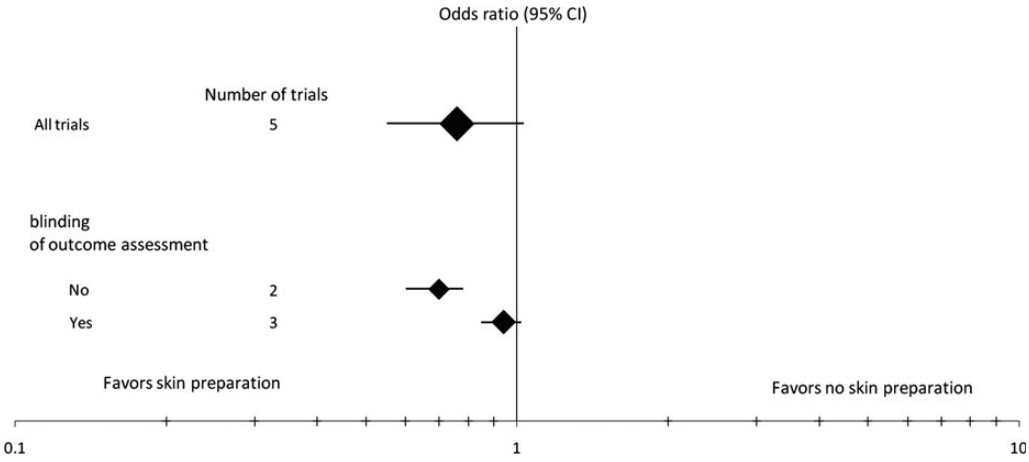

Figure 1. Performing subgroup analyses to determine the effect of study risk of bias on meta-analysis results. In a hypothetical meta-analysis of 5 randomized controlled trials (RCTs) examining the impact of a skin preparation on surgical site infections, the meta-estimate of the RCTs suggests that the skin preparation reduces superficial surgical site infections. However, when the 3 studies that meet the risk of bias criteria of blinding the outcome assessor to the use of the skin preparation are meta-analyzed, there is no difference in superficial surgical site infection for those who had the skin preparation applied. In this example, the subgroup analysis of those studies meeting the risk of bias criteria suggests that the findings of the meta-analysis are dependent on the risk of bias of the studies. The conclusions of the review should thus be tempered accordingly. Abbreviation: CI, confidence interval.

Evaluate the Risk of Bias of the Individual Studies

A study's risk of bias is a measure of its internal validity. Inclusion of biased studies in systematic reviews leads to erroneous findings and conclusions [41]. Hence, the risk of bias of the studies included in the review needs to be evaluated. Using established scales tailored to specific study designs is one approach to evaluating risk of bias. The Jadad scale for evaluating risk of bias of RCTs is validated and commonly used [40], but other RCT scales exist as well [42]. There are also scales to judge the risk of bias of non-RCTs [43, 44] and diagnostic studies [45]. The benefit of such scales is that they can produce one summary score representing a study's risk of bias. However, most of these scales are limited in that they have not been validated, they include different items even if they purport to measure risk of bias in the same study types, and there are differences in how items are weighted across scales. In addition, many study designs do not have established scales to judge their risk of bias. Because of these limitations, the Cochrane Collaboration and other authorities favor the use of individual items to distinguish studies with higher risks of bias from those with lower risks of bias. These items are not combined to produce a summary score but are used individually. They can be selected from items included in previously constructed scales or can be developed by the reviewer to address the specific question of interest. For example, if one were assessing the risk of bias of observational study designs, one might decide to note whether the study was prospective, or whether it adjusted for a particularly important confounder. These individual criterion might differentiate those studies at higher risk of bias from those at lower risk.

After deciding on an approach to evaluate a study's risk of bias, one then has to decide how the evaluation will be used. One could use a study's risk of bias as an inclusion/exclusion criteria, or to weight the studies in a meta-analysis, giving more weight to those studies with lower risk of bias. Most commonly, however, risk of bias is used to perform subgroup or sensitivity analyses, where studies with the lowest risk of bias are grouped together and analyzed, and the results compared to the analysis of all included studies to determine if the results are sensitive to risk of bias. For example, in a hypothetical meta-analysis of 5 RCTs examining the impact of a novel skin preparation on surgical site infections, the meta-estimate of the RCTs suggests that the skin preparation reduces superficial surgical site infections. However, when one meta-analyzes the 3 studies that meet the risk of bias criterion of blinding the outcome assessor to the use of the skin preparation, one finds no differences in the incidence of superficial surgical site infections between the groups. (Figure 1) In this example, the subgroup analysis of those studies meeting the criterion suggests that the findings of the meta-analysis are dependent on study risk of bias. The conclusions of the review should thus be tempered accordingly. Of note, in this example, an individual criterion was used to stratify studies by their risk of bias, rather than a specific scale. Evidence suggests that the use of risk of bias scales for this purpose can produce dramatically different results depending on the scale use, so using such scales for this purpose is generally not recommended [46].

Downloaded from https://academic.oup.com/cid/article-abstract/57/5/725/311245 by Dana Medical Library, University of Vermont user on 26 June 2019

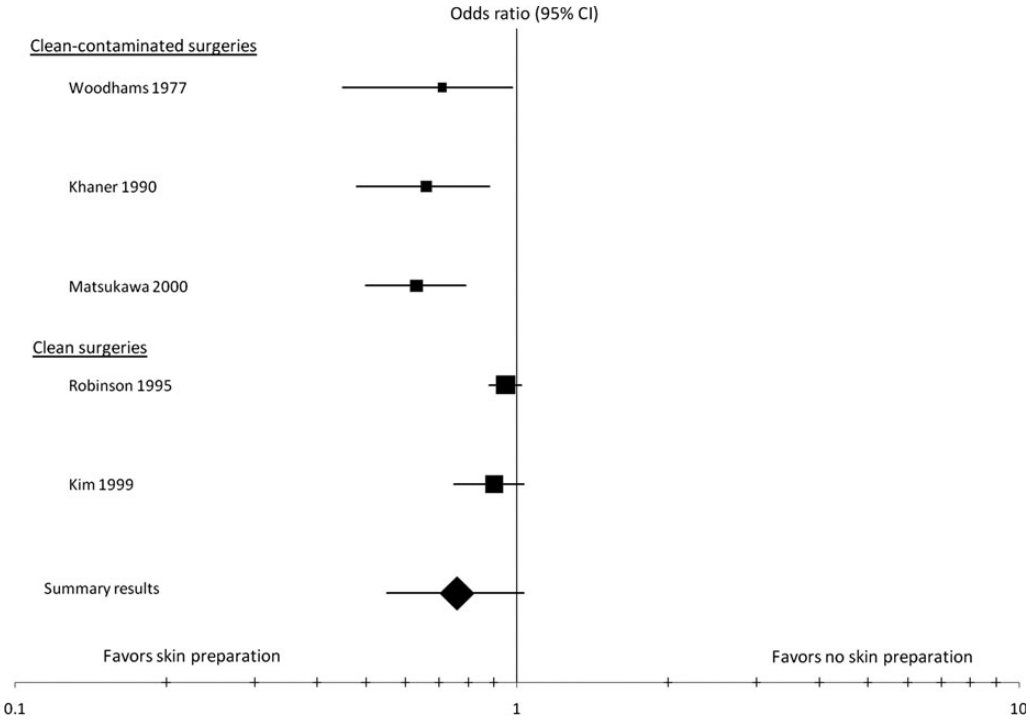

Figure 2. A forest plot demonstrating heterogeneity. Figure 2 is a typical "forest plot" depicting a meta-analysis of 5 randomized controlled trials (RCTs), where each row and each box represent an individual study, the size of each box represents the weight of that study, the horizontal line through the box represents the 95% confidence interval (CI) of the effect size from the study, and the diamond at the bottom of the figure represents the meta-estimate resulting from statistically combining the results of the individual studies. The lateral tips of the diamond represent the 95% CI of the meta-estimate. In this meta-analysis, 3 of the studies show a statistically significant reduction in superficial surgical site infections using the skin preparation under examination, whereas 2 of the studies do not. The 2 studies that show no difference in surgical site infections between those who received the preparation and those who did not are studies in those with clean surgeries, whereas the 3 studies that did show a difference examined populations undergoing clean-contaminated surgeries. These differences suggest that the effect of the skin preparation may be different based on the patients' baseline risk of surgical site infection. In other words, the preparation may only make a difference when the risk of infection exceeds some minimum threshold. If one were to meta-analyze these 5 studies without recognizing and exploring the heterogeneity between the individual study findings, the important interaction between the effectiveness of the preparation and the patients' risk of infection may not have been identified. Abbreviation: CI, confidence interval.

Data Synthesis

One can synthesize data qualitatively through the use of written evidence summaries and evidence tables or quantitatively through meta-analytic techniques. If the evidence is of sufficient homogeneity and quantity such that a meta-analysis can provide additional information beyond what is otherwise available, then one should strongly consider such an analysis. Meta-analysis may be most helpful in the situation where there are a number of small "negative" studies that are similar in terms of the populations, interventions, and outcomes that they examine, and combining the studies can increase one's power to find a difference if it exists. Meta-analyses can also be helpful when one wants to produce a more precise estimate of effect by combining the findings from individual studies. In addition, meta-analysis can help determine whether findings across studies are statistically different from one another, and subgroup analyses can help one understand the causes of those differences, otherwise known as heterogeneity between studies.

To perform a meta-analysis, one must first estimate a summary measure and its variance for each of the individual studies to be included in the meta-analysis. One then must weight each study according to its sample size, and then statistically combine the results of each study to obtain a weighted average, as opposed to a simple average [47].

Figure 2 is a typical forest plot depicting a meta-analysis of 5 RCTs, where each row and each box represents an individual study, the size of each box represents the weight of that study, the horizontal line through the box represents the 95% confidence interval (CI) of the effect size from the study, and the diamond at the bottom of the figure represents the meta-estimate resulting from statistically combining the results of the individual studies. The lateral tips of the diamond represent the

Downloaded from https://academic.oup.com/cid/article-abstract/57/5/725/311245 by Dana Medical Library, University of Vermont user on 26 June 2019

Table 5. Topics in Meta-analyses

Topic	Description
Model types	An important detail to note in any meta-analysis is the model used to perform the analysis. *Fixed-effects models* assume there is one underlying effect of the intervention under study, that all of the individual studies are measuring this effect, and that any differences between the results of individual studies are a consequence of sampling variation. Conversely, *random-effects models* assume that differences between the results of individual studies reflect a distribution of effects of the intervention under study. They often, but not always [48], offer a more conservative pooled estimate (ie, wider confidence intervals). Thus, if there is significant heterogeneity in a meta-analysis, a random-effects model should be used to perform the analysis [47]. More importantly, one should explore potential reasons for the observed heterogeneity, as described in the text.
Heterogeneity	Heterogeneity is typically estimated by the *Q test* (ie, the χ^2 test), and measures whether study-specific estimates are statistically different from one another. The test cannot determine which of the studies is different, or how many of the studies are different. The main limitations of the *Q* test is that it is underpowered when there are only a small number of studies included in the meta-analysis (ie, it will suggest no statistically significant heterogeneity when it in truth exists), and it has excessive power when there are a large number of studies included in the meta-analysis (ie, the test will suggest significant heterogeneity when it does not in truth exist). Given that the more common scenario in a meta-analysis is to have too few studies, and that the *Q* test is underpowered to detect significant heterogeneity in this scenario, the *P* value threshold is often relaxed to .10 when assessing the statistical significance of the *Q* test.
	More recently, the *I^2 test* was developed to account for these limitations [55]. The I^2 test is essentially the *Q* test adjusted for the number of studies included in the meta-analysis. It measures the percentage of variation across studies that is due to heterogeneity between studies and not chance, and is measured on a scale of 0 to 100%. An I^2 value >50% is considered to be moderate to high heterogeneity, but some consider values <50% to still be statistically significant heterogeneity.
Meta-regression	Meta-regression can help one examine the association between specific study characteristics and the meta-analysis results. Meta-regression is similar to standard regression techniques used to examine the association between individual patient characteristics and individual study outcomes, but occurs at the level of the study rather than the level of the patient [54].
Reporting bias	Given that it can often be challenging to visually detect asymmetry in funnel plots created to assess reporting bias (particularly if there are only a few studies included in your analysis) [58], statistical tests such as the *Egger and Begg test* can also be used to identify these biases [21]. Furthermore, statistical methods such as the *trim and fill technique* can help one model the impact of smaller, less positive studies on the meta-estimate to determine how sensitive the findings of the review might be to reporting bias. If adding smaller, less positive studies using modeling techniques does not significantly change the meta-estimate, your review findings may be robust to any reporting bias resulting from your search strategy.

95% CI of the meta-estimate. Meta-analyses can be performed using fixed-effects or random-effects models (Table 5) [47, 48].

One can quantitatively synthesize the results of RCTs, as well as non-RCTs. See the referenced publications to learn more about the quantitative synthesis of observational studies [49] and diagnostic studies [50–52], including studies of genetic tests [53].

Exploration of Heterogeneity

If heterogeneity exists in a meta-analysis, one must explore it and not ignore it. Exploration of heterogeneity is a critical objective of meta-analyses. Differences in study designs, populations, interventions, comparators, outcome definitions, and the conduct of studies can lead to differing results between studies. An examination of these differences using subgroup analyses or the technique of meta-regression [54] (Table 5) can help one understand the conditions most likely to yield positive or negative effects from an intervention (Figure 2). Importantly, significant clinical heterogeneity can exist between studies without significant statistical heterogeneity, particularly as statistical tests of heterogeneity are often underpowered (Table 5) [55]. Thus, if there are real differences between studies in the variables listed above, it is prudent to examine the effect of these differences on the meta-estimate, regardless of whether there is significant statistical heterogeneity present [56].

Exploration of Reporting Bias

To help identify reporting bias, one can plot the sizes of the studies included in the review on the y-axis against the effect sizes of the studies on the x-axis. This type of plot is called a "funnel plot," because the shape of the plot should represent an inverted funnel (Figure 3*A*) [21, 57, 58]. If there is a "missing quadrant" in the base of the funnel, this suggests that smaller, less positive studies that are theoretically in existence may not have been identified in the search, thus biasing the results in favor of the intervention of interest (Figure 3*B*).

ADDITIONAL RESOURCES FOR THE PERFORMANCE, REPORTING, AND APPRAISAL OF SYSTEMATIC REVIEWS

There are a number of standards for conducting, reporting, and judging the risk of bias of systematic reviews and meta-analyses,

Downloaded from https://academic.oup.com/cid/article-abstract/57/5/725/311245 by Dana Medical Library, University of Vermont user on 26 June 2019

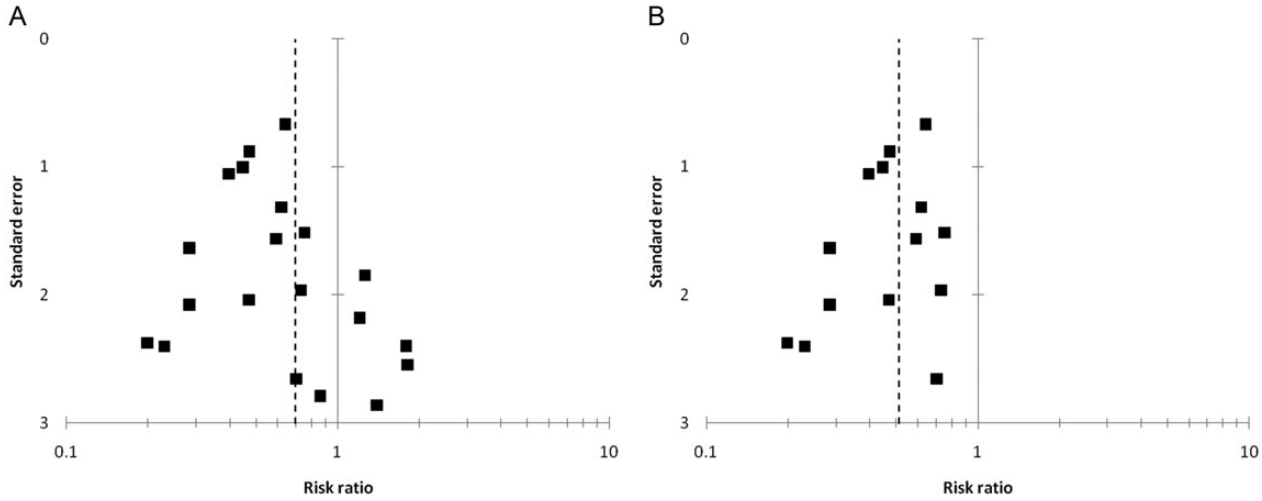

Figure 3. Assessing reporting bias using funnel plots. *A*, No reporting bias. *B*, Reporting bias. To help identify reporting bias, one can plot the sizes of the studies included in the review on the y-axis against the effect sizes of the studies on the x-axis. Study size is commonly represented by the variance or standard error of the study estimate. Because estimates from smaller studies will have larger variances or standard errors, they will be located lower on the y-axis given the axes on the typical plot. Ideally, these smaller studies will evenly distribute around the vertical line representing the meta-estimate resulting from statistically combining all of the studies (in this plot, risk ratio [RR] = 0.70). The larger studies will be at the top of the plot, and should huddle closer to the vertical line than the smaller studies because they are larger and contribute more weight to the overall meta-estimate. This type of plot is called a "funnel plot" because the shape of the plot should represent an inverted funnel (*A*). If there is a "missing quadrant" in the base of the funnel, this suggests that smaller, less positive studies that are theoretically in existence may not have been identified in the search, thus biasing the results in favor of the intervention of interest. Note that the absence of the 6 smaller, less positive studies in this plot changes the meta-estimate from RR = 0.70 to RR = 0.50 (*B*).

the most important of which are the recent IOM report for conducting [31], the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement for reporting [59], and the AMSTAR checklist [60] for measuring risk of bias. It is best to review these publications prior to starting a systematic review to ensure one considers the most appropriate methods and documents them accordingly during the conduct of the review. More detailed resources [35, 38, 61, 62] and textbooks [63] exist for those involved in writing reviews for established organizations, such as the Cochrane Collaboration [38, 62], or the Agency for Healthcare Research and Quality Evidence-based Practice Centers [13, 35, 61]. Software is also freely available from these organizations to help individuals perform many of the steps of a systematic review, including citation screening [64], data extraction [65, 66], and the meta-analysis itself [66, 67].

CONCLUSIONS

Performing a systematic review and meta-analysis is a scientifically rigorous process that allows one to identify and synthesize the available evidence addressing a key question. Systematic reviews can be both qualitative and quantitative (ie, a meta-analysis). A meta-analysis involves statistical pooling of findings from individual studies; the exploration of heterogeneity in a meta-analysis is critical. Potential limitations of systematic reviews and meta-analyses include the risk of bias of identified studies (ie, "garbage in, garbage out"), the heterogeneity of included studies, and reporting biases resulting from one's search strategy. Standards exist that can help authors perform and report valid and actionable systematic reviews and meta-analyses.

Notes

Acknowledgments. The author would like to thank Rajender K. Agarwal, Matthew D. Mitchell, and David R. Goldmann for reviewing an earlier version of this manuscript, and for their assistance with tables and figures.

Financial support. C. A. U. has received funds from the Centers for Disease Control and Prevention to coauthor infection control guidelines, and is the Senior Associate Director of the ECRI Institute–Penn Medicine Agency for Healthcare Research and Quality (AHRQ) Evidence-based Practice Center, which coauthors comparative effectiveness reviews for AHRQ.

Potential conflicts of interest. Author certifies no potential conflicts of interest.

The author has submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. Egger M, Davey Smith G. Meta-analysis: potentials and promise. BMJ **1997**; 315:1371–4.

Downloaded from https://academic.oup.com/cid/article-abstract/57/5/725/311245 by Dana Medical Library, University of Vermont user on 26 June 2019

2. Chang SM, Carey T, Kato EU, Guise JM, Sanders GD. Identifying research needs for improving health care. Ann Intern Med **2012**; 156:439–45.
3. Umscheid CA, Williams K, Brennan PJ. Hospital-based comparative effectiveness centers: translating research into practice to improve the quality, safety and value of patient care. J Gen Intern Med **2010**; 25: 1352–5.
4. Luce BR, Brown RE. The use of technology assessment by hospitals, health maintenance organizations, and third-party payers in the United States. Int J Technol Assess Health Care **1995**; 11:79–92.
5. Pearson SD, Rawlins MD. Quality, innovation, and value for money: NICE and the British National Health Service. JAMA **2005**; 294: 2618–22.
6. Umscheid CA, Agarwal RK, Brennan PJ. Updating the guideline development methodology of the Healthcare Infection Control Practices Advisory Committee (HICPAC). Am J Infect Control **2010**; 38:264–73.
7. Harris RP, Helfand M, Woolf SH, et al. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med **2001**; 20(3 suppl):21–35.
8. Whittemore R, Knafl K. The integrative review: updated methodology. J Adv Nurs **2005**; 52:546–53.
9. Stewart LA, Parmar MK. Meta-analysis of the literature or of individual patient data: is there a difference? Lancet **1993**; 341:418–22.
10. Steinberg KK, Smith SJ, Stroup DF, et al. Comparison of effect estimates from a meta-analysis of summary data from published studies and from a meta-analysis using individual patient data for ovarian cancer studies. Am J Epidemiol **1997**; 145:917–25.
11. Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. Stat Med **2002**; 21:371–87.
12. Umscheid CA. Maximizing the clinical utility of comparative effectiveness research. Clin Pharmacol Ther **2010**; 88:876–9.
13. Atkins D, Fink K, Slutsky J. Better information for better health care: the Evidence-based Practice Center program and the Agency for Healthcare Research and Quality. Ann Intern Med **2005**; 142(12 Pt 2): 1035–41.
14. Whitlock EP, Lopez SA, Chang S, Helfand M, Eder M, Floyd N. AHRQ series paper 3: identifying, selecting, and refining topics for comparative effectiveness systematic reviews: AHRQ and the effective health-care program. J Clin Epidemiol **2010**; 63:491–501.
15. Segal JB. Chapter 3: choosing the important outcomes for a systematic review of a medical test. J Gen Intern Med **2012**; 27(suppl 1):S20–7.
16. Samson D, Schoelles KM. Chapter 2: medical tests guidance (2) developing the topic and structuring systematic reviews of medical tests: utility of PICOTS, analytic frameworks, decision trees, and other frameworks. J Gen Intern Med **2012**; 27(suppl 1):S11–9.
17. Deverka PA, Lavallee DC, Desai PJ, et al. Stakeholder participation in comparative effectiveness research: defining a framework for effective engagement. J Comp Eff Res **2012**; 1:181–94.
18. Schardt C, Adams MB, Owens T, Keitz S, Fontelo P. Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC Med Inform Decis Mak **2007**; 7:16.
19. Agarwal R, Schwartz DN. Procalcitonin to guide duration of antimicrobial therapy in intensive care units: a systematic review. Clin Infect Dis **2011**; 53:379–87.
20. Chou R, Helfand M. Challenges in systematic reviews that assess treatment harms. Ann Intern Med **2005**; 142(12 Pt 2):1090–9.
21. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. J Clin Epidemiol **2000**; 53:1119–29.
22. Ioannidis JPA. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. JAMA **1998**; 279:281–6.
23. Egger M, Zellweger-Zähner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomized controlled trials in English and German. Lancet **1997**; 350:326–9.
24. Scherer RW, Langenberg P, von Elm E. Full publication of results initially presented in abstracts. Cochrane Database Syst Rev **2007**; MR000005.
25. Egger M, Davey Smith G. Bias in location and selection of studies. BMJ **1998**; 316:61–6.
26. Dundar Y, Dodd S, Williamson P, Dickson R, Walley T. Case study of the comparison of data from conference abstracts and full-text articles in health technology assessment of rapidly evolving technologies: does it make a difference? Int J Technol Assess Health Care **2006**; 22:288–94.
27. Juni P, Holenstein F, Sterne J, Bartlett C, Egger M. Direction and impact of language bias in meta-analyses of controlled trials: empirical study. Int J Epidemiol **2002**; 31:115–23.
28. Moher D, Pham B, Klassen TP, et al. What contributions do languages other than English make on the results of meta-analyses? J Clin Epidemiol **2000**; 53:964–72.
29. Ganann R, Ciliska D, Thomas H. Expediting systematic reviews: methods and implications of rapid reviews. Implement Sci **2010**; 5:56.
30. Shultz M. Mapping of medical acronyms and initialisms to Medical Subject Headings (MeSH) across selected systems. J Med Libr Assoc **2006**; 94:410–4.
31. Institute of Medicine. Finding what works in health care: standards for systematic reviews. **2011**. Available at: http://www.iom.edu/Reports/2011/Finding-What-Works-in-Health-Care-Standards-for-Systematic-Reviews.aspx. Accessed 1 November 2012.
32. Sood A, Erwin PJ, Ebbert JO. Using advanced search tools on PubMed for citation retrieval. Mayo Clin Proc **2004**; 79:1295–9.
33. Ebbert JO, Dupras DM, Erwin PJ. Searching the medical literature using PubMed: a tutorial. Mayo Clin Proc **2003**; 78:87–91.
34. Health Information Research Unit, McMaster University. Search filters for MEDLINE in Ovid Syntax and the PubMed translation. **2012**. Available at: http://hiru.mcmaster.ca/hiru/HIRU_Hedges_MEDLINE_Strategies.aspx. Accessed 1 November 2012.
35. Slutsky J, Atkins D, Chang S, Sharp BA. AHRQ series paper 1: comparing medical interventions: AHRQ and the effective health-care program. J Clin Epidemiol **2010**; 63:481–3.
36. Relevo R. Chapter 4: effective search strategies for systematic reviews of medical tests. J Gen Intern Med **2012**; 27(suppl 1):S28–32.
37. Relevo R, Balshem H. Finding evidence for comparing medical interventions: AHRQ and the Effective Health Care Program. J Clin Epidemiol **2011**; 64:1168–77.
38. The Cochrane Collaboration. Cochrane handbook for systematic reviews of interventions. **2011**. Available at: http://www.cochrane.org/training/cochrane-handbook. Accessed 1 November 2012.
39. Sampson M, Barrowman NJ, Moher D, et al. Should meta-analysts search Embase in addition to Medline? J Clin Epidemiol **2003**; 56:943–55.
40. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary?. Control Clin Trials **1996**; 17:1–12.
41. Schulz KF CI, Hayes RJ, Altman D. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA **1995**; 273:408–12.
42. Moher D, Jadad AR, Nichol G, Penman M, Tigwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. Controlled Clin Trials **1995**; 16: 62–73.
43. Sanderson S, Tatt ID, Higgins JPT. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol **2007**; 36:666–76.
44. The Ottawa Hospital Research Institute. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. **2011**. Available at: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp. Accessed 1 November 2012.
45. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med **2011**; 155:529–36.
46. Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trial for meta-analysis. JAMA **1999**; 282:1054–60.

Downloaded from https://academic.oup.com/cid/article-abstract/57/5/725/311245 by Dana Medical Library, University of Vermont user on 26 June 2019

47. Egger M, Davey Smith G, Phillips AN. Meta-analysis: principles and procedures. BMJ **1997**; 315:1533–7.
48. Poole C, Greenland S. Random-effects meta-analyses are not always conservative. Am J Epidemiol **1999**; 150:469–75.
49. Egger M, Schneider M, Davey Smith G. Spurious precision? Meta-analysis of observational studies. BMJ **1998**; 316:140–4.
50. 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 **2012**; 27(suppl 1):S56–66.
51. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol **2005**; 58:982–90.
52. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med **2001**; 20:2865–84.
53. Sun F, Bruening W, Erinoff E, Schoelles KM. Addressing challenges in genetic test evaluation: evaluation frameworks and assessment of analytic validity. **2012**. Avialable at: http://www.ncbi.nlm.nih.gov/books/NBK56750/. Accessed 29 May 2013.
54. 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**; 64:1187–97.
55. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ **2003**; 327:557–60.
56. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. BMJ **1994**; 309:1351–5.
57. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ **1997**; 315:629–34.
58. Terrin N, Schmid CH, Lau J. In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. J Clin Epidemiol **2005**; 58:894–901.
59. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. J Clin Epidemiol **2009**; 62:1006–12.
60. Shea BJ, Grimshaw JM, Wells GA, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol **2007**; 7:10.
61. Smetana GW, Umscheid CA, Chang S, Matchar DB. Methods guide for authors of systematic reviews of medical tests: a collaboration between the Agency for Healthcare Research and Quality (AHRQ) and the Journal of General Internal Medicine. J Gen Intern Med **2012**; 27(suppl 1):S1–3.
62. Cochrane Collaboration. Cochrane handbook for systematic reviews of diagnostic test accuracy. **2012**. Available at: http://srdta.cochrane.org/handbook-dta-reviews. Accessed 1 November 2012.
63. Egger M, Davey Smith G, Altman DG, eds. Systematic reviews in health care: meta-analysis in context. 2nd ed. London: BMJ Publishing Group, **2001**.
64. Center for Clinical Evidence Synthesis. Abstrackr. Available at: http://tuftscaes.org/citation_screening/. Accessed 19 April 2013.
65. Agency for Healthcare Research and Quality . Systematic Review Data Repository. **2013**. Available at: http://srdr.ahrq.gov/. Accessed 19 April 2013.
66. Cochrane Collaboration. RevMan. **2012**. Available at: http://ims.cochrane.org/revman. Accessed 19 April 2013.
67. OpenMeta[Analyst], **2012**. Available at: http://www.cebm.brown.edu/open_meta. Accessed 19 April 2013.

Downloaded from https://academic.oup.com/cid/article-abstract/57/5/725/311245 by Dana Medical Library, University of Vermont user on 26 June 2019