A Primer on Performing Systematic Reviews and Meta-analyses

Craig A. Umscheid1,2,3,4,5

1Center for Evidence-based Practice, 2Department of Medicine, 3Center for Clinical Epidemiology and Biostatistics, 4Leonard Davis Institute of Health Economics, and 5Institute 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.

Received 1 November 2012; accepted 4 May 2013.
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
© 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
gaps, which researchers and funders can use to prioritize 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].

**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

<table>
<thead>
<tr>
<th>Table 1. Descriptions of Less Common Types of Reviews</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Review Type</strong></td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>Integrative review</td>
</tr>
<tr>
<td>Individual patient-level meta-analyses</td>
</tr>
</tbody>
</table>
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 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.

| 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 | randomized controlled trial.pt | epidemiologic studies/ | exp sensitivity and specificity/ |
| 1. | controlled clinical trial.pt | exp case control studies/ | specificity.tw. |
| 2. | randomized.ab | exp cohort studies/ | false negative.tw. |
| 3. | placebo.ab | case control.tw | accuracy.tw. |
| 4. | clinical trials as topic.sh | cohort analy$.tw | screening.tw. |
| 5. | randomly.ab | (cohort adj (study or studies)).tw | or/1–5 |
| 6. | trial.ti | cohort analy$.tw | |
| 7. | or/1–7 | (follow up adj (study or studies)).tw | |
| 8. | exp animals/ not humans.sh | (observational adj (study or studies)).tw | |
| 9. | 8 not 9 | longitudinal.tw | |
| 10. | 8 not 9 | retrospective.tw | |
| 11. | cross sectional.tw | cross-sectional studies/ | |
| 12. | or/1–12 | 

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

### 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.

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

<table>
<thead>
<tr>
<th>Database</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medline</td>
<td>The premier database of medical research. Should generally be used in every systematic review.</td>
</tr>
<tr>
<td>PubMed</td>
<td>The free interface, and OVID Medline is an interface commonly supported by institutional subscriptions that allows for more complex searching</td>
</tr>
<tr>
<td>Embase</td>
<td>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 pharmacoconomics topics.</td>
</tr>
<tr>
<td>Cochrane Library</td>
<td>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.</td>
</tr>
<tr>
<td>CINAHL</td>
<td>A valuable database when questions involve the fields of nursing or quality and safety, or when searching for qualitative studies.</td>
</tr>
<tr>
<td>PsycINFO</td>
<td>Helpful when addressing questions in the area of mental health.</td>
</tr>
</tbody>
</table>
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 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].
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 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.

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.
Heterogeneity

Heterogeneity is typically estimated by the \( Q \) test and diagnostic studies [50, 53]. About the quantitative synthesis of observational studies [49], 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 3A) [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 3B).

### 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,
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.** Authors 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**


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.


