How to Write a Systematic Review: A Step-by-Step Guide

[1]Sarah M. Yannascoli, MD
[1]Mara L. Schenker, MD
[1]James L. Carey, MD, MPH
[1]Jaimo Ahn, MD, PhD
[1,2]Keith D. Baldwin MD, MSPT, MPH

[1]Department of Orthopaedic Surgery, University of Pennsylvania, Philadelphia, PA

[2]Division of Orthopaedic Surgery, Children's Hospital of Philadelphia, Philadelphia, PA

Corresponding author:
Keith D. Baldwin, MD, MSPT, MPH
Children's Hospital of Philadelphia
Assistant Professor of Orthopaedic Surgery
University of Pennsylvania
34th Street and Civic Center Boulevard
Philadelphia, PA 19104
keith.baldwin@uphs.upenn.edu

Introduction

A systematic review attempts to comprehensively and reproducibly collect, appraise, and synthesize all available empirical evidence that meets pre-defined criteria in order to answer a research question. The quantitative combination and statistical synthesis of the systematically-collected data is what defines a meta-analysis. Here, we first attempt to delineate the basic steps for conducting a systematic review: initial planning, conducting the search, data extraction, and quality analysis. We then outline the fundamental steps for assessing the appropriateness of meta-analytic technique for your review and an explanation of statistical tools available for data analysis and presentation. An academic discussion regarding the strengths and weaknesses of systematic review methodology is beyond the scope of this guide, as are detailed instructions regarding statistical analysis.

Initial Planning

When initiating a systematic review, it is important to plan ahead and anticipate problems. By maintaining a clear study focus from the beginning, identifying a well-defined research question, outlining strict inclusion and exclusion criteria, and understanding the eventual contribution of your work to the existing literature, you can effectively minimize reviewer bias and streamline the review process.

Defining a Research Question

An appropriate, focused research question is based on an extensive *a priori* literature review to understand the scope of evidence available on your topic. It is often helpful to write down your question first, then to conduct a literature review to determine whether your question has already been answered, can be answered, or is irrelevant and would pose an insignificant contribution. The PICO mnemonic (Population, Intervention, Comparison, Outcome) is a commonly used tool to help delineate a clearly defined, clinically-based question for your systematic review. In detail, the mnemonic refers to the following:

1. Population: Define your subject group. Think about the age, sex, race and other patient characteristics, as well as relevant co-morbidities, pathology, and outcomes.
2. Intervention: Consider the prognostic factor or exposure (includes intervention) of interest.
3. Comparison: Repeat steps 1 and 2 for the group to whom you will compare your initially defined population and intervention (note: this step does not apply to all questions).
4. Outcome: The item you hope to accomplish, measure, or define.

For example:

1. P: Are adults with open fractures
2. I: who undergo operative irrigation and debridement
3. C: after a delay of greater than six hours from the time of injury at an
4. O: increased risk of developing osteomyelitis, soft tissue infection, and fracture non-union?

In the process of outlining a study question, it is clear that many critical terms within the question stem will need to be defined and characterized further. It is important that these terms are evaluated and discussed amongst your collaborators at the initial stages of the project, so as to eliminate potential confusion moving forward. For instance, in the above question, terms that require strict definitions are age (include pediatric patients?), open fractures (gunshot injuries excluded? only long bones?), osteomyelitis (culture-positive patients only?), soft tissue infection (those treated with antibiotics? or those who required an additional surgery?), and non-union (how long from initial injury?). With your question and these terms in mind, the future identification of relevant and appropriate literature will be easier.

Study Justification

An initial literature review is required so that you can justify the significance of your work. Your study may intend to do one or more of the following: 1) clarify strengths or weaknesses of existing literature; 2) summarize large amounts of literature; 3) resolve conflicts; 4) evaluate the need for a large clinical trial; 5) increase the statistical power of smaller studies;

or 6) improve study generalizability. Bear in mind that the purpose of a systematic review is to not only collect all the relevant literature in an unbiased fashion, but to extract data presented in these articles in order to provide readers with a succinct *synthesis* of available evidence. As a general guide, you should easily find—on a broad non-systematic search—numerous papers that are relevant but may be excluded. Look carefully to see if your work has been previously published. If the most recent review is more than a few years old and the topic remains relevant, your contribution may still be of value. Finally, check the PROSPERO site (http://www.crd.york.ac.uk/NIHR_PROSPERO/) to see if others are working to answer the same questions. If not, consider registering your study.

Literature Search

To execute a well-designed study there are two requirements: 1) an organized team including a statistician, an expert in the field, and at least two individuals to oversee each section of the review process; and 2) a detailed study protocol. For the latter, consideration will need to be given to specific search terms, inclusion and exclusion criteria, databases to be searched, and eventual data which will need to be collected and reported. Finally, persons experienced in conducting a search, a medical librarian, or both may offer guidance as you proceed.

Selecting Search Terms

Selecting the appropriate terminology is what guides the entire search, and thus is of crucial importance. Consider alternate terms, historical terminology, and even common misspellings. List these terms in the protocol before starting the search. Each search term (or terms) will need to be queried in each database that is used. A detailed search may be produced with the assistance of an experienced librarian who can "customize" a search in order to limit the number of extraneous hits. In cases with extremely specific questions this may be appropriate. Additionally, the librarian may be able to assist you in obtaining rare journal articles or texts which may constitute part of your review. It is important to obtain the services of someone with expertise in systematic reviews to minimize the prospect of bias infused by terms which are too restrictive.

Inclusion and Exclusion Criteria

Strict criteria are necessary to determine the appropriate articles for inclusion. Some of these criteria will depend on your specific question (i.e. exclude gunshot injuries as a mechanism of open fracture). General criteria applicable to any systematic review are: level of evidence, language, and animal or human subjects. First, choose the level of evidence included for your particular study. This will depend on the existing literature and the overarching aim of your research. For topics that are well-represented in the literature with the aim to synthesize available evidence, it is common to include articles with high levels of evidence only (Levels I and II). For topics that are less well-characterized with the aim to justify a larger clinical trial, inclusion of all levels of evidence may be warranted. It is important to remember that the quality of a systematic review is defined by the lowest quality of the included studies. Next, decide whether the resources are available to include articles published in other languages. The inclusion of English-language articles only will certainly introduce bias, but is often necessary when resources are not available for translation.

Databases

Multiple information sources will need to be searched to perform a comprehensive systematic review. Medline includes articles published since 1966 and is freely available via PubMed. EMBASE includes articles published since 1974 and requires a personal or university subscription to access. Surprisingly, *there is only a 34% overlap* of journals included in these two databases.[1] Therefore, using a single database alone is insufficient, with reports that only 30 to 80% of randomized controlled trials will be identified with Medline database search alone.[2] Additionally, the Cochrane Controlled Trials Register does not overlap with the previous two databases and will need to be individually searched. If you are conducting a review that is related to medical education, quality control, bioengineering, etc., there are a number of additional databases for alternate fields including education-focused, nursing, or engineering literature that may require the use of additional resources. A further search of the "grey" literature may produce additional references. For example, sites like opengrey.eu may prove fruitful, especially if the topic is unusual or if a large publication bias is found. Keep in mind that Medline and EMBASE articles are more likely to be well vetted, but also more vulnerable to publication bias.

Data Organization

A key aspect to conducting and writing a systematic review is reporting your exact methods for data collection. The most recent guidelines on conducting and reporting systematic reviews are the PRISMA statement (Preferred Reporting Items for Systematic reviews and Meta Analyses).[3]These guidelines facilitate the reporting of appropriate information (Figure 1).

Conducting and Reviewing the Search

Once a justified study question and detailed study protocol are in place, the systematic review process can proceed. First, accounts must be created with each database (Medline, EMBASE, Cochrane) in order to save searches that may need to be retrieved at a later time. Terms must be entered into the search field only once, and the date the search is conducted must be recorded. If a search is conducted, and subsequently re-typed into the database *de novo* one week later, there may have been additional articles published or uploaded within that week. It is better to record the date and report this than to constantly re-do the search. Type search terms into the database (remember to use filters as defined by your study protocol). Export references to a reference-managing program that allows for efficient identification and exclusion of duplicate entries.

Once the references have been recorded, collected, and duplicates excluded (record this number too), the first-pass review may begin. In this stage, the reviewers (minimum of two), should read through each study title and exclude clearly irrelevant studies. If either reviewer feels that the study may be of value, it is included for further analysis. A second-pass review is then conducted where the abstracts of included titles are analyzed further. Eventually, articles still included must undergo full-text review. Once this is complete, the bibliographies of each article also need to be systematically reviewed for further relevant articles. This process again necessitates a first-pass review (exclusion by title), a second-pass review (exclusion by abstract), and a third-pass review (exclusion by full-text), as was conducted for the primary search. Any additional articles found to meet all inclusion criteria will again need a systematic bibliography review until no further articles are identified. Once this last step is complete, it is useful to provide a measure of inter-rater agreement in order to determine how robust the initial search words were. Be sure to record the number of studies searched and excluded at each stage of the process. Review the flowchart in Figure 1 frequently as a reminder of data which needs to be recorded and reported.

Data Extraction

The data extraction component of a systematic review is driven by a well-organized spreadsheet. The spreadsheet should be carefully piloted on a few select studies before incorporating it into the entire review. The structure of the data collection form will vary between different systematic

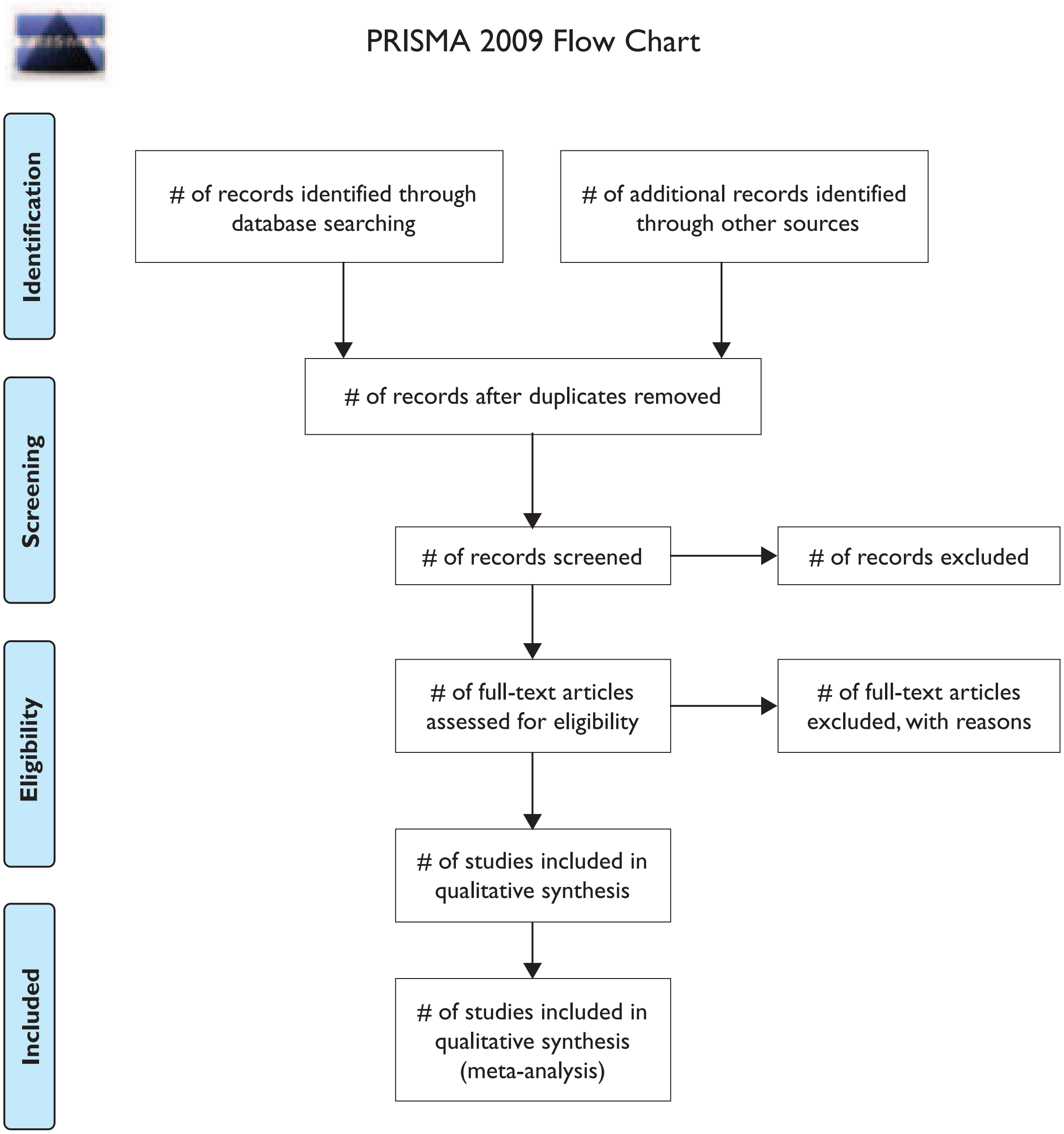
PRISMA 2009 Flow Chart
Identification
Screening
Eligibility
Included
# of records identified through database searching
# of additional records identified through other sources
# of records after duplicates removed
# of records screened
# of records excluded
# of full-text articles assessed for eligibility
# of full-text articles excluded, with reasons
# of studies included in qualitative synthesis
# of studies included in qualitative synthesis (meta-analysis)

Figure 1. PRISMA 2009 Flow Diagram. A flowchart outlining the information required for reporting in systematic reviews according to the PRISMA guidelines.[3]

reviews, thus depending on the systematic review, more specific data collection items may need to be extracted for full appropriate review. We recommend beginning with a more detailed spreadsheet to avoid having to return to the primary articles after the initial data extraction. It is important to remember that the data extraction should be performed by two independent reviewers and any differences need to be reconciled by mutual agreement.

Quality Analysis

A key step in a systematic review is the critical appraisal of the included studies. An assessment of "study quality" is a bit nebulous, but at a minimum, an assessment of the internal (i.e. minimization of study methodological error and bias) and external (i.e. generalizability to other populations) validity of all the studies included in the systematic review is necessary. Several potential threats to the validity of the studies need to be assessed in a reproducible manner, and include description bias, selection bias, measurement bias, analytic bias, and interpretation bias.

- Description bias: is the intervention well described?
    - *Did the authors report antibiotic timing and dosing, in addition to the operative debridement times?*
    - *Is the fracture population adequately described?*
- Selection bias: did the authors describe the screening criteria for study eligibility?
    - *Did the authors describe why some open fractures were excluded from the study or transferred to another facility?*
- Measurement bias: was the exposure (i.e. open fracture classification) and outcome measures (i.e. infection diagnosis) valid and reliable?
    - *Did the authors report reliability for the classification of the open fractures?*
    - *Did they use quantitative cultures or subjective clinical examination findings for the definition of "infection"?*
- Analytic bias: did the authors conduct an appropriate analysis by conducting statistical testing, controlling for repeated measures, etc.?
    - *Did the authors account for severity of injury in their statistical analysis?*
    - *Did they report any statistics or just observations?*
- Interpretation bias: did the authors correct for controllable confounders?
    - *Was there adequate follow-up of the patients with open fractures?*

Several quality scales and checklists have been reported,[1] including many that are available for randomized trials. Quality measures for non-randomized studies are variable, and none have been developed specifically for use in orthopaedic trials. Many of the available scales are able to generate overall quality scores. However, overall scores may not provide adequate information regarding the strengths and weaknesses of the individual studies, and may be misleading, by providing a high overall summary score in spite of a single critical methodological flaw. Therefore, many researchers prefer to use checklist of necessary elements to quality appraisal. The items on the checklist can be presented in a qualitative manner in the systematic review.[4] A minimum of two independent reviewers should assess the quality of the studies. Differences can be reconciled by mutual agreement or by a third reviewer.

Meta-Analysis

Prior to embarking on a meta-analysis, it is important to determine whether or not the data are appropriate for meta-analytic methods. The term "meta-analysis" is most commonly associated with the summation of randomized controlled trials. Every meta-analysis implies that a systematic review has been done, but not every systematic review is amenable to meta-analysis. True meta-analyses are somewhat uncommon in orthopaedic surgery relative to the number of systematic reviews. Recently, many authors have begun applying meta-analytic techniques to observational comparative studies. Care must be taken to consider selection bias when the groups are not similar, as well as reporting bias if it is unclear whether the entire sample was used in some of the parent studies. For these reasons, sensitivity analyses have shown that meta-analysis of different levels of evidence may produce disparate results.

Publication Bias

Every meta-analysis should include an assessment of publication bias. Publication bias, or "the file drawer effect," is the tendency for articles to get published based on the magnitude and direction of the results. As such, small studies that demonstrate a difference are more likely to get published than large studies. This type of publication bias may be assessed in several ways. Two of the most common means of assessing publication bias are funnel plots and the Egger's intercept.

A funnel plot should be symmetric (Figure 2). This indicates that the size of the studies did not correlate with the effect size of the outcome measure of interest. If the scatter plot here shows dots which fall outside the confidence ranges, then publication bias can be considered a possibility (i.e. small studies showing a greater effect size).[5] This implies that small studies with a smaller or negative effect size may exist but were never published.

Egger's intercept is a quantitative method to identify asymmetry in a funnel plot. Mathematically it is equal to the Y intercept of a line produced by a regression of the normalized effect size (estimate divided by standard error) by precision of that estimate (1/SE).[6] This produces a recognized p-value that can be interpreted as the chance that this funnel plot would have been produced by a random set of studies by chance.

Study Heterogeneity

Differences in study populations, methods, and in the case of surgery, surgical techniques and follow-up, can all have a profound influence on effect size. Unless techniques are relatively standardized, it is often useful to assume that heterogeneity is present between studies in surgical trials and

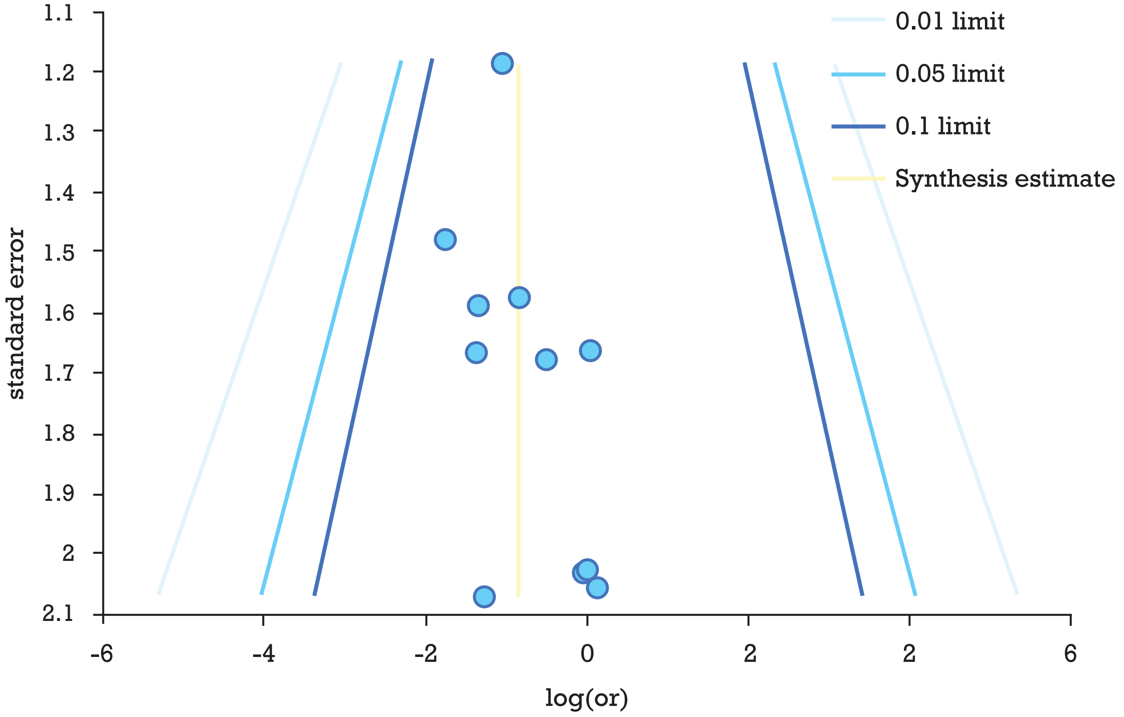

Figure 2. Funnel Plot. A symmetric representation or a funnel plotwith all studies (blue dots) remaining within the demonstrated confidence intervals (blue lines).This indicates that smaller studies do not demonstrate a greater effect (i.e. less likelihood of publication bias).

observational studies, simply by virtue of practice variation between surgeons. This does not mean that studies cannot be summed to produce a meaningful result. Common practice methodology is variable, and therefore a meta-analysis can be the means for increasing external validity.

In order to assess statistical heterogeneity, two methods are generally employed: The Cochran's Q statistic, and the I^2 range. Cochran's Q attempts to detect if greater heterogeneity exists in the effect sizes than can be accounted for by sampling error. This statistic has been criticized because it has poor power to detect heterogeneity when a small number of trials are present, and too much power when a large number of trials are present. The I^2 range is somewhat more descriptive in the sense that it describes the range of possible heterogeneity by the confidence interval. This confidence interval can be interpreted as the range of potential heterogeneity. It is acceptable to assume homogeneity if the I^2 range includes 0%. It is our practice, however, to assume heterogeneity to be conservative for the aforementioned rationale regarding surgical trials if the range is large (i.e. 0 to 50%). This increases the risk of a type II error, but decreases the rate of a type I error.

Fixed or Random Effects

Selection of an appropriate model is very important in meta-analysis. Fixed effects models assume that all the variance in the data comes from variance within studies. As such, an underlying assumption of such models are that variance between studies is negligible. Selection of a fixed effects model implies that heterogeneity statistics have been performed and show minimal inter-study statistical heterogeneity. It also demonstrates that populations and methods between studies are of sufficient similarity that the assumption of negligible inter-study variability can be safely made. This model assumes that there is one true effect, and the studies all estimate this effect. All differences from this effect are then assumed to be the result of sampling error. In surgical trials, this may not be believable, because surgical techniques are difficult to standardize.

A random effects model on the other hand, assumes that the studies performed represent a random sampling of the effect size which varies and has a mean and a 95% confidence interval. The underlying assumption is that the effect sizes follow a normal distribution. A random effects model is more conservative in the setting of study heterogeneity. This is because a random effects model allows for inter-study variability. The nature of many orthopaedic studies is that there are differences in populations and methods because the vast majority of studies are observational. An argument can be made that the fixed effects assumption is always faulty in this setting. Additionally, in the setting of minimal inter-study variance, a random effects model will approach the results of a fixed effects model. In random effects models, standard errors are larger and confidence intervals wider than in fixed effects models, hence researchers are less likely to reject the null hypothesis.

Although using a random effects model helps to account for inter-study variability, the statistical test itself does not

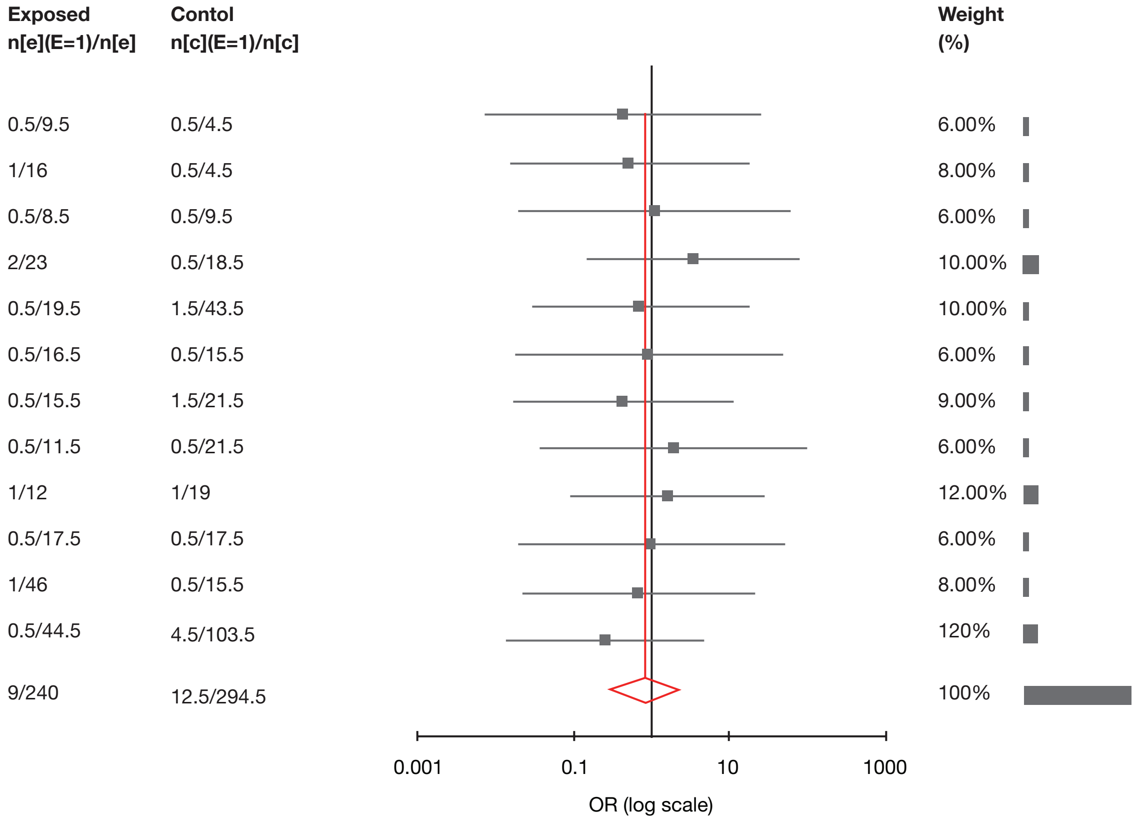

Figure 3. Forest Plot. A graphical representation of the distribution of effect sizes of the parent studies. The summary effect size is provided with confidence intervals is represented by the red diamond.

eliminate this heterogeneity. Rather, it adds variance to the summary effect proportional to variability. In other words, simply using a random effects model does not indicate that the statistical methods can in some way overcome the problem of heterogeneity. If the studies identified are clearly heterogeneous, a summary estimate should not be calculated.[7]

Summary Effect Sizes

After a model has been chosen, summary effect sizes can be generated. The most common graphical representation for summary effect sizes is the forest plot, illustrating the distribution of effect sizes of the parent studies (individual squares with horizontal bars to correspond to each study).A summary effect size is provided (vertical line) with confidence intervals (diamond at the bottom of the vertical line). The midline represents an odds ratio of 1 or "no difference" (Figure 3). Consideration for "zero-event" studies is given by adding 0.5 to each cell. This provides for an addition of 0.5/0.5 to each ratio or 1 (so nothing is added, but an odds ratio can be calculated), thus allowing for usage of zero-event studies and making the model more robust by increasing the n of available studies.[8]

Conclusion

Conducting a systematic review and incorporating meta-analytic statistical techniques takes a great deal of planning, cooperation among team members, time, and sincere effort to conduct a thorough analysis of all available empirical evidence. Adherence to guidelines and strict reporting of search methodology are essential. Most importantly, it is essential to remember that the quality of a systematic review and/or meta-analysis cannot exceed the quality of the individual studies included in the analysis.

References

1. **Wright RW, Brand RA, Dunn W, et al.** How to write a systematic review. *Clin Orthop Rel Res* 2007;455:23-9.

2. **Egger M, Smith GD, Altman DG (eds).** *Systematic Reviews in Healthcare: A Meta-Analysis in Context. 2nd Edition.* London: BMJ books, 2001.

3. **Moher D, Liberati A, Tetzlaff J, et al.** Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Int J Surg* 2010; 8:658.

4. **Zaza S, Wright-De Aguero LK, Briss PA, et al.** Data collection instrument and procedure for systematic reviews in the Guide to Community Preventive Services. *Am J Prev Med* 2000;18(S1):44-74.

5. **Light RJ, Pillemer DB.** *Summing up: the Science of Reviewing Research.* Cambridge, Massachusetts: Harvard University Press, 1984.

6. **Egger M, Davey Smith G, Schneider M, et al.** Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315:629-34.

7. **Hulley HB, Cummings SR, Browner WS, et al.** *Designing Clinical Research. 3rd Edition.* Philadelphia, Pennsylvania: Lippincott Williams and Wilkins, 2007.

8. **Higgins JPT, Green S (eds.)** Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org. Accessed 2/12/2013; chapter 16.9.2.