Introduction to Systematic Reviews

Table of Contents

- Module A: Overview of Systematic Reviews
- Module B: Evaluating Systematic Reviews
- Module C: Steps for Conducting a Systematic Review
  - Step 1: Assembling the team
  - Step 2: Develop the protocol or work plan
  - Step 3: Question/topic refinement
  - Step 4: Systematic and comprehensive searches for evidence
  - Step 5: Inclusion/exclusion rules
  - Step 6: Critical appraisal of relevant literature
  - Step 7: Data abstraction
  - Step 8: Data Synthesis
  - Step 9: Communication of results

Authors:
Elizabeth O'Connor, Ph.D.
Evelyn Whitlock, M.D., M.P.H.
Bonnie Spring, Ph.D., ABPP

Module A: Overview of Systematic Reviews

What are Systematic Reviews

(SRs)?

Simply, they are literature reviews that use rigorous, systematic, and transparent methods to minimize bias in the results. Transparent means that the decisions are clearly documented. Bias is a systematic distortion of an estimated effect, and can result from problematic decisions made at almost any point in the review process. We will discuss bias in more detail when we go over the steps to creating a systematic review.

They are a central link between evidence and health-related decision making. SRs provide the decision-maker with best available evidence. This evidence, in combination with clinical or field expertise and the client's values, characteristics, and circumstances, are necessary ingredients for making good decisions (Eden, 2008).

How do they differ from traditional narrative reviews?

- They have clear, explicit objectives with clearly stated inclusion criteria for studies to be selected (providing transparency)
- They use systematic searching methods that reduce the risk of selective sampling of studies, which may support preconceived conclusions (reduces risk of bias)
- They use consistent evaluation of available information such as outcomes and study quality (reduces risk of bias)
- They give the readers more information about decisions that were made along the way, which allows the readers to assess the quality of the review more directly (increases transparency)
- They may be able to provide greater precision in estimates of effect, especially if meta-analysis is involved (increases accuracy)
- They set the stage for updates as more data is published (because methods are transparent)

Where do they fit in evidence-based behavioral practice (EBBP)?

- SRs are an important source of evidence for decision-making
- The first three steps in the EBBP process are covered in the process of completing an SR: Ask, Acquire, and Appraise
- They provide a short-cut for EBBP practitioners who can use SRs to cover the first three steps of the EBBP process and then focus their efforts on the subsequent steps of Apply and Analyze &

Adjust	

What are they used for?

To help groups and individuals make decisions to improve people's health. Examples include:

- **Recommendations and guidelines**
  [e.g., United States Preventive Services Task Force (USPSTF), National Institute for Health and Clinical Excellence (NICE).] Should smokers routinely be advised to use quit smoking medications? Should primary care providers routinely screen patients for depression? What should be the components of an intervention to help overweight children manage their weight?
- **Benefit design, coverage and policy decisions**
  [e.g., Centers for Medicare and Medicaid Services (CMS), Drug Effectiveness Review Project (DERP), UK National Health Service (NHS).] Should we cover the use of medication to quit smoking? Should we reimburse visits with breast feeding specialists in mothers with babies?
- **Public Policy**
  Would it improve the health of our community if we increase funding for mass transit and bike facilities?
- **Performance measures**
  [e.g., Assessing Care of Vulnerable Elderly (ACOVE).] If a patient receives a new prescription for an antidepressant, what frequency and duration of followup constitute good quality care?
- **Research agendas**
  [e.g., National Institutes for Health (NIH), Agency of Healthcare Research and Quality (AHRQ).] What are the gaps in the evidence related to treatment of anxiety in children?
- **Individual patient care**
  Should I advise this client to use behavioral treatment in addition to medication to help her quit smoking?
- **Patient decisions**
  Should I try hypnosis to help me get over my fear of flying?

Example: Single-session Critical Incident Debriefing (CID)

Here is an example of the role of systematic reviews in the evolution of a body of literature. We will cover the methods used in the systematic review in some detail so you can begin to learn about SR methods. For a more detailed overview about how to conduct an SR, please read Module 3 of this tutorial, "Steps for Conducting a Systematic Review."

- **Background**

  Psychological interventions were developed in the 1980s to help people cope with traumatic events in their immediate aftermath and to prevent the onset of post-traumatic stress disorder (PTSD). Debriefing has been used by organizations such as the Red Cross and emergency services agencies to help workers who have been in traumatic disaster situations, and some psychotherapists have specialized in offering such services to individuals and organizations involved in traumatic events. One such approach involves single-session critical incident debriefing.

- **First effort at summarizing the evidence base**

  A 1995 editorial (Raphael et al, 1995) identified five studies on the topic, none involving randomized comparisons:

  - Results were mixed: some studies favored debriefing, some studies showed no differences, some studies reported that those participating in debriefing interventions did more poorly than those who weren't
  - No information on how the studies were found
  - No information on the decision process behind whether or not a study would be discussed in the review
  - Concluded that data was insufficient to determine if Critical Incident Debriefing was helpful or harmful; more research was needed, particularly randomized trials
  - This reviewer may have used very rigorous, unbiased methods for identifying and evaluating studies for inclusion, but we simply cannot tell

- **First systematic review appeared in 1998 (Rose & Bisson, 1998)**

  Aim: "To perform a systematic review of available research literature regarding the effectiveness of [single-session psychological debriefing] in preventing psychological sequelae following traumatic events (p 698)."

  Searching process: Electronically searched seven publication databases, contacted experts, hand-searched contents of a relevant journal, requested information from a research organization specializing in trauma, searched

references of relevant articles, and reviewed conference papers. Provided years searched and search terms used.

Inclusion/exclusion rules:

- Study participants experienced a traumatic event meeting Criteria A of DSM-III-R for PTSD
- Random assignment
- Clear inclusion/exclusion criteria for participants
- Outcomes were assessed using reliable and valid methods
- Ages 16+
- A structured or semi-structured intervention was used
- These criteria caused all five of the studies in the previously described non-systematic review to be excluded due to lack of randomization

- **First systematic review (cont'd)**

Critical appraisal of studies:

- Some quality criteria were embedded inclusion/exclusion rules, such as requiring the use of reliable and valid assessment methods, a structured or semi-structured intervention, and clear inclusion/exclusion rules for participants
- Narrative descriptions highlighted study quality issues

Data abstraction: Not described

Data synthesis: Qualitative

- Created summary table showing six included studies and key characteristics of the studies
- Results describe limits of the studies' generalizability and flaws that limit confidence in the studies' results
- Conclusions: Critical Incident Debriefing is most likely not effective and its routine use cannot be supported

- **The same author conducted a more updated review four years later (Rose et al, 2002)**
  - Similar methods
  - Included nine additional studies published in the intervening years
  - Conclusion: "There is no evidence that single session individual psychological debriefing is useful treatment for the prevention of post traumatic stress disorder after traumatic incidents. Compulsory debriefing of victims of trauma should cease. A more appropriate response could involve a 'screen and treat' model." They also report that single session critical debriefing may be harmful, though good quality data on harms were lacking. The World Health Organization (WHO) and British National Health Service each developed guidelines that advise

against the use of single-session CID (IASC, 2007; NICE, 2005).

Positive results of the evidence-based practice movement

- High quality evidence-based treatment guidelines (e.g., NICE guidances, Department of Veteran's Affairs clinical practice guidelines, clinical decision-support systems)
- High quality evidence-based recommendations (e.g., U.S. Guide to Community Preventive Services, U.S. Preventive Services Task Force)
- More synthesized evidence with explicit attention to quality
- Clearer standards of research reporting [e.g., CONSORT for RCTs, QUORUM for SRs (Moher et al, 1999), STROBE for observational studies]
- Trial registries (e.g., clinicaltrials.gov)
- Quality improvement initiatives in healthcare

Future directions

Dissemination and Translation:

For evidence-based practice to become viable to the average practitioner, it is essential to establish systems that provide ready access to regularly-updated synthesized evidence and decision support. Translation of research into practice can't become an everyday affair without such infrastructure. There are important and emerging issues related to translation of evidence-based interventions into practice, and a growing body of literature spells out the key issues. (Brownson, 2006; Kerner, 2005)

Module B: Evaluating Systematic Reviews

Quality indicators

Is there a clear review question or aim?

The question or questions the review is addressing should be clearly stated.

Was the literature search strategy stated?

The review should list the search terms, databases searched, years searched, and other identification strategies used.

Were there explicit inclusion/exclusion criteria reported relating to selection of the primary studies?

The inclusion/exclusion criteria should cover study design, populations, interventions, and outcomes of interest.

Selection Bias

Details should be reported relating to the process of decision-making (i.e., how many reviewers were involved, whether the studies were examined independently, and how disagreements between reviewers were resolved). Some measure of interrater reliability (e.g., kappa, percent agreement) is important if only a sample of abstracts or studies were dual-reviewed. Additionally, is there any evidence that the process was biased or inadequately implemented?

Is there evidence of a substantial effort to search for all relevant research?

In addition to details of the search terms and databases, descriptions of hand-searching, attempts to identify unpublished material, and any contact with authors, industry, and research institutes should be provided. The appropriateness of the database(s) searched by the authors should also be considered (e.g., if only MEDLINE is searched for a review looking at health education, then it is unlikely that all relevant studies will have been located).

Is there a transparent system to evaluate the quality of individual studies?

A systematic assessment of the quality of primary studies should include an explanation of the criteria used (e.g., method of randomization, whether outcome assessment was blinded, whether analysis was on an intention-to-treat basis). The process of the assessment should be explained (i.e., the number of reviewers who assessed each study, whether the assessment was independent, and how discrepancies between reviewers were resolved).

Is sufficient detail of the individual studies presented?

The review should demonstrate that the studies included are suitable to answer the question posed and that a judgement on the appropriateness of the authors' conclusions can be made. If a review includes a table giving information on the design and results of the individual studies, or includes a narrative description of the studies within the text, this criterion is usually fulfilled. If relevant, the tables or text should include information on study design, sample size in each study group, patient characteristics, description of interventions, settings, outcome measures, followup, drop-out rate (withdrawals), effectiveness, results, and adverse events.

Are the primary studies summarized appropriately?

The authors should attempt to synthesize the results from individual studies. In all cases, there should be a narrative summary of results, which may or may not be accompanied by a quantitative summary (meta-analysis). For reviews that use a meta-analysis, heterogeneity between studies should be assessed using statistical techniques. If heterogeneity is present, the possible reasons (including chance) should be investigated. In addition, the individual evaluations should be weighted in some way (e.g., according to sample size, or inverse of the variance) so that studies that are considered to provide the most reliable data have greater impact on the summary statistic.

Is there a transparent system to evaluate the quality of body of evidence? Were the authors' conclusions supported by the evidence they presented?

Optimally, a reviewer will also appraise the body of literature as a whole, pointing out strengths and weaknesses, and offering an assessment of their confidence in the conclusions.

What was the funding source and role of funder?

Further considerations

There are no clear guidelines for determining whether the methods used are adequate or not. After evaluating the quality indicators, the reader must make a judgment as to their confidence in the validity of the conclusions.

How well do their methods apply to your question? Sometimes good quality reviews on the same topic can still come to different conclusions because of differences in their methods, usually in their inclusion/exclusion criteria.

Module C: Steps for Conducting a Systematic Review

1. Assembling the team

SRs cannot be completed by a single person. They are always a team effort. Important areas of expertise to cover include:

- Content experts
- SR methods experts
- Statistician
- Medical librarian
- Reference management

- **Content experts**

  It is important to have either one or more team members or an active consultant to provide expertise in the area covered by the review. Input is usually needed from practitioners and researchers representing a variety of perspectives.

- **SR methods experts**

  One or more persons with expertise in the methods of conducting SRs is needed. This person may be responsible for developing the procedures and documentation standards for the review. A SR methods expert may also be a content expert, but more than one investigator-level reviewer is necessary, since some steps in the process require dual review or data checking that requires expertise in research and statistical methodology.

- **Statistician**

  If meta-analysis is to be considered, access to a statistician with experience in meta-analysis is needed.

- **Medical librarian**

  Database searching requires specialized knowledge that general research training does not provide. Preferably, the librarian searcher has experience with the extensive searching and documentation procedures that are a part of a systematic review.

- **Reference management**

  Someone must be responsible for maintaining the database of references. Most SRs involve thousands of abstracts, and the use of software to manage the references is necessary. This person must be able to track which abstracts have been reviewed and their disposition (e.g., included or excluded, reason for exclusion).

2. Develop the protocol or work plan

The following steps will walk through items that should be specified ahead of time in as much detail as possible.

3. Question/topic refinement

(Akin to the ASK step in Evidence-Based Behavioral Practice)

- Formulate a clearly defined answerable question or series of questions, identifying the **P**opulation, **I**ntervention, **C**omparison condition, and **O**utcome(s) of interest
- Example: "Do psychological treatments for PTSD improve PTSD symptomatology in adults, compared with control conditions (e.g., usual care or waiting list) or alternate psychological treatment?"
- If there are multiple, related questions, reviewers often develop an analytic framework, which visually shows the logical pathways underlying your questions. Sometimes reviewers need a higher level diagram and multiple logical pathways that show more detail contained in each high level box. An example of a higher level diagram can be found on page 84 of this report. To see an example of a simple analytic framework with four related questions, click here

4. Systematic and comprehensive searches for evidence

(Akin to the ACQUIRE step in Evidence-Based Behavioral Practice)

- Consultation with a medical librarian is essential. There is a great deal to know about what databases would be appropriate, how search terms are used in each database, and the logic of searching
- It is very important to keep good records of the search strategy (exact search strings used for which databases, covering what time periods)

4. Systematic and comprehensive

searches for evidence

Identify relevant databases.

- **MEDLINE**
  A general medical database, accessed through service providers such as Ovid or PubMed. The most widely used database, but a substantial proportion of journals are NOT indexed in MEDLINE, and use of MEDLINE alone is usually insufficient.
  http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed
- **PsycINFO**
  Covers the international literature in psychology and related behavioral and social sciences. Includes book chapters and dissertations as well as journal articles.
  http://www.apa.org/psycinfo
- **Cumulative Index of Nursing and Allied Health Literature (CINAHL)**
  Includes nursing and allied health disciplines such as occupational therapy, emergency services, and social services.
  http://www.ebscohost.com/cinahl
- **Embase**
  A pharmacological and biomedical database which includes the international literature.
  http://www.embase.com
- **BIOSIS**
  Covers the biological and biomedical sciences. Journal articles comprise the majority of the content, but also included are meeting and conference reports, books and patents.
  http://scientific.thomsonreuters.com/products/bp
- **Health Services/Technology Assessment (HSTAT)**
  Includes full text of documents such as clinical guidelines and the publications of the Agency for Healthcare Research and Quality (AHRQ) and the National Institutes of Health (NIH).
  http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=hstat
- **Cochrane Central Register of Controlled Trials (CCRCT)**
  Part of the Cochrane Library, includes details of published articles taken from bibliographic databases (notably MEDLINE and EMBASE), and other published and unpublished sources.
  http://apps1.jhsph.edu/cochrane/central.htm
- **Cochrane Database of Systematic Reviews (CDSR)**
  Contains systematic reviews and protocols. Reviews are updated regularly.
  http://www.cochrane.org
- **Database of Abstracts of Reviews of Effects (DARE)**
  A database of critically appraised systematic reviews, published by the Centre for Reviews and Dissemination. Also part of the Cochrane Library.
  http://www.crd.york.ac.uk/crdweb
- **Campbell Collaboration**

A library of systematic reviews in the areas of education, criminal justice and social welfare.
http://www.campbellcollaboration.org

4. Systematic and comprehensive searches for evidence

Identify and pilot test appropriate search terms.

Some groups have developed and tested search strings for different study designs (e.g., controlled trials, diagnostic studies, etc.). Click on the **Resources** button for further resources.

MeSH is the controlled vocabulary thesaurus used for indexing articles in MEDLINE. Queries using MeSH vocabulary permit a higher level of specificity, but because articles may not always be indexed correctly or appropriate MeSH terms may not exist, it is important to also use textword searching.

4. Systematic and comprehensive searches for evidence

Specify years for searching.

- Early years may or may not be relevant, depending upon differences between current and historical interventions, larger cultural/societal context, or other factors

Decide whether (and if so, how extensively) non-peer-reviewed literature will be pursued.

Determine which languages will be included.

Hand-search reference lists of relevant literature reviews and articles, and contents of selected journals.

Consult content experts to make sure important articles have not been missed.

4. Systematic and comprehensive searches for evidence

Potential biases that can occur at the searching phase:

Publication bias:

Unpublished trials have an average of 7% smaller effects than published trials in medical fields (Egger et al, 2003), but there is a large range across individual reviews and medical specialties. For example, in psychiatric reviews, unpublished trials showed substantially smaller effects than published trials, but there was little difference between published and unpublished trials in the oncology literature.

Database bias:

A substantial number of journals are not indexed in MEDLINE. Compared to trials that are not indexed in MEDLINE, those that are showed roughly similar rates of appropriate blinding (at treatment assignment and assessment) and had an average of 6% larger effects. (Egger et al, 2003)

Language bias:

Non-English language studies on average have smaller Ns, are less likely to have used appropriate blinding, and are more likely to have statistically significant effects. (Egger et al, 2003)

5. Inclusion/exclusion rules

These must be developed carefully and thoughtfully. Inclusion/exclusion rules have a

big impact on the generalizability of the results. Reviewers should always consider the larger goals of the review to help make decisions about where to draw the inclusion/exclusion line. If the inclusion rules are very broad it will be difficult to summarize the body of literature and the evidence will likely need to be broken into more homogeneous subgroups of studies to discuss the results. Reviewers must consider their resources and how broadly a topic can realistically be covered and still maintain rigorous, systematic methods.

5. Inclusion/exclusion rules

Population

There are a number of population-related inclusion/exclusion rules to consider.

- **Age ranges:**
  - All ages?
  - Adults only?
  - Both pre-retirement age and older adults?
- **Risk status:**

  Treatment outcomes may differ, and special treatment approaches may be needed, for some high risk groups. For example, adolescents who have already experienced an episode of depression or who have a depressed parent may have some special needs, compared with the youth experiencing their first episode of depression. Will studies that are limited to depressed youth with elevated risk of depression due to previous episodes of depression be included? Or a depressed parent? Or youth who are not currently depressed but are at risk of future episodes because they or their parents have had episodes of depression in the past?

- **Presence of co-morbidities:**

  Certain co-morbidities may also influence intervention approaches and affect outcomes. For example, in treatment for alcohol dependence or abuse, co-existing anxiety disorder may require concurrent treatment addressing anxiety issues in order to effectively reduce alcohol use. Or, clients with psychotic disorders may have special treatment needs in order to maintain sobriety. In a review of substance abuse treatment, will studies limited to people with co-existing anxiety disorders be included? Psychotic disorders?

- **Range of severity:**

  For example, in a review of treatment of eating disorders, will studies that require DSM diagnoses of an eating disorder and studies that merely require a person to score above a cut-point on an eating disorder screening instrument both be included? Will studies that include participants with all levels of eating disorder symptomatology be included? Studies to prevent eating disorders in asymptomatic children?

5. Inclusion/exclusion rules

Intervention

What are the critical treatment elements that must be present in order to be included in the review? For example, in a review of the effectiveness of cognitive behavioral (CBT) approaches in treating obsessive-compulsive disorder, will the use of both cognitive and behavioral elements be required? The presence of a treatment manual or specific training in the treatment approach? Evidence that the researchers examined the fidelity of the treatment delivered and provided supervision or quality control to ensure high fidelity? Are there treatment components that would cause a trial to be excluded? For example, will studies that involve adjunctive use of selective serotonin reuptake inhibitors (SSRIs) be excluded? Studies that compare SSRIs alone with SSRIs plus CBT?

5. Inclusion/exclusion rules

Comparator

It is important to clearly define allowable control conditions. Will trials comparing the effectiveness of two active treatment approaches be included? If a minimal treatment control group is required, what will be considered “minimal”?

5. Inclusion/exclusion rules

Outcomes

There are a number of outcomes-related inclusion/exclusion rules to consider.

- **Primary outcomes specified a priori:**

  Primary review outcomes should be specified a priori. Additional outcomes may be abstracted, but reporting of other outcomes should usually be done in terms of generating hypotheses for future reviews to test. Reviewers should be very cautious about making too much of non-primary outcomes, as these are very vulnerable to outcome reporting bias. (I.e., likely to have been reported when they were statistically significant and not reported when they were not significant.)

- **Three general categories of outcomes**

  Health outcomes: These are outcomes in which the person subjectively perceives an improvement in their health or quality of life. Examples include mortality, presence of a diagnosable disorder, subjective effects (e.g, feelings of hopelessness or sadness), and quality of life indicators. Smoking cessation, abstinence from other substances of abuse, and weight loss are also usually considered health outcomes, even though there may not be immediate measurable health impacts, because these outcomes are strongly related to future health and they can have a profound effect on a person's quality of life.

  Intermediate outcomes: These are outcomes where a change can be measured, but the person may not subjectively feel any improvement in their health. Examples are laboratory or clinical measures of health such as lipid levels or blood pressure, and behavioral outcomes such as increased numbers of fruits and vegetables eaten or increased number of pleasurable events initiated.

  Harms: The relative benefits of an intervention cannot be determined without an assessment of the likelihood and seriousness of harms.

- **Measurability:**

  Choose outcomes that can be reliably and validly measured. It is preferable to use validated instruments, particularly for self-report measures of subjective well-being.

- **Measurement interval:**

Determine if there should be a minimal follow-up. For some interventions, it may be relatively common to see an immediate effect, but if the effect does not persist, then the participants may not have really benefited. Also, some disorders can be episodic in nature (such as depression); be sure the outcomes are timed so that the improvements are most likely due to the treatment rather than the passage of time. In some cases, the reviewer may be interested in both short-term improvement and maintenance of improvement, so multiple followup windows are of interest.

- **Risk of bias:**

  Outcome reporting bias (i.e., the selective reporting of some outcomes and not others, depending on whether they were statistically significant or not). To combat this, reviewers should consider only including trials in which outcomes that are relevant to your review are reported as primary outcomes. The NIH clinical trials registry (clinicaltrials.gov) requires researchers to state the aims of their trials. This can be a useful tool to systematic reviewers when the primary aim of the study is not clear from the published write-up.

5. Inclusion/exclusion rules

Study design

Must be considered very carefully. It is difficult for a study to have very high internal validity and very high generalizability. Internal validity is always important, but the relative weight of generalizability varies somewhat depending on the question. For example, if the goal is to determine whether a treatment approach can be effective, generalizability is less of a concern. If the goal is to determine if a treatment approach is generally effective in normal clinical conditions, however, the generalizablity of the research methods to normal clinical conditions is extremely important.

Randomized Controlled Trials (RCTs) generally have the best internal validity, but they may not be available, they can have limited generalizability, and they may only report only very short-term outcomes.

Controlled Clinical Trials are trials in which the researcher assigns participants to treatment groups on a non-random basis. For example, if a researcher is working with a city to implement a community-wide intervention, he or she may assign another city in the same region of comparable size and with similar socioeconomic indices to act as a control. Often CCTs are less desirable than RCTs because they lack randomization, but a good quality CCT may provide better evidence of treatment efficacy than fair-

quality RCT. For example, an RCT that does not adequately control for baseline differences in the treatment and control groups does not provide as good evidence as a good quality CCT in which baseline differences were assessed and controlled for if any were found.

Observational studies may be the best design for answering some questions, such as those related to rare events (common when looking at harms) incidence, prevalence, and prognosis.

6. Critical appraisal of relevant literature

(This and the remaining steps are akin to the APPRAISE step of Evidence-Based Behavioral Practice.)

Rate the individual studies on the degree to which the results are true and free of bias:

- Numerous scales and checklists have been developed to rate study quality. Click the **Resources** button for some examples and further resources
- It is important to use a checklist or instrument with explicit criteria, either one that has already been published or one developed specifically for the review being undertaken
- Different rating schemes are needed for different study designs

6. Critical appraisal of relevant literature

There are different common quality elements for different study designs.

- **RCTs, common quality elements:**
    - Adequate randomization procedures, including a truly random process for

assigning participants to treatment groups
    - Adequate allocation concealment (i.e., ensuring that study personnel do not know which treatment group the potential participant will be assigned to)
    - Baseline comparability between groups verified, or where groups are not comparable, steps are taken to compensate for inequality (e.g., statistically controlling for age if the groups differed in average age)
    - Clear definition of the interventions
    - Outcomes and other measures assessed using reliable and valid methods
    - Identical assessment methods used in all treatment groups
    - Blinding of outcomes/followup assessment
    - Minimal loss to followup
    - Similar loss to followup in all treatment groups
    - Analyzing participants in the treatment group to which they were assigned
    - Appropriate data analysis methods, including appropriate handling of missing data
    - Funding/sponsorship by disinterested party
- **CCTs and controlled quasi-experimental studies, common quality elements:**
    - Consideration of potential confounders in constructing the groups
    - Adjustment for potential confounders in data analysis
    - Others cited under RCTs
- **Cohort studies:**
    - Groups are selected from source populations that are comparable
    - Similar rates of recruitment, refusal, and attrition in the two groups
    - The likelihood that some eligible participants might have the outcome at the time of enrollment is assessed and taken into account in the analysis
    - Outcomes are clearly defined and assessed blind to exposure status, or careful quality control of assessment where blinding is not possible
    - Main potential confounders are identified and taken into account in the design and analysis
- **Case-control studies:**
    - It is clearly established that controls are non-cases
    - The same inclusion/exclusion criteria are used for both cases and controls
    - Others cited under cohort studies
- **Other types of studies:**

  Other types of studies may include economic evaluations, studies of diagnostic accuracy of screening tests, other studies of instrument development or evaluation.

6. Critical appraisal of relevant literature

Make a judgment on how much the quality issues reduce confidence in the results. Minor flaws are the norm. Many reviewers exclude studies with major ("fatal") flaws. Some include those studies, but conduct meta-analyses with and without the methodologically poor studies. If results are inconsistent, then the analysis excluding the poor studies is usually more valid.

Quality assessment is subject to bias because it is easy to be influenced by another reviewer's judgment. Therefore, use multiple independent raters and decide on a method of resolving differences (e.g., third rater, consensus of larger research team).

7. Data abstraction

Data abstraction—identifying pre-specified data elements from individual studies and entering the data into a table or database.

Identify elements to be abstracted. These will vary by study design somewhat. Often reviewers are faced with tremendous heterogeneity in the included studies, along many dimensions. It is important to systematically capture this heterogeneity, since factors related to population, intervention, and design may have a big impact on effect size. Reviews that involve a very focused, narrow question and little heterogeneity in the included studies may not need to abstract as much detail about the individual studies.

7. Data abstraction

Commonly abstracted elements include:

- Study reference
- Study characteristics (e.g., design, N randomized, setting, country, recruitment source, stated aim of study)

- Participant characteristics (e.g., age range, mean age, sex, race/ethnicity, socioeconomic status, presence of selected co-morbidities)
- CONSORT-type numbers (e.g., number approached, completed screening, completed baseline assessment, eligible, refused, randomized)
- Inclusion/Exclusion criteria (list all)
- Description of the intervention and control conditions (e.g., general approach, treatment components, number of sessions, length of sessions, duration of intervention, group vs. individual)
- Outcomes. Determine allowable outcomes a priori. You may or may not choose to abstract other beneficial outcomes that are not among prespecified outcomes. If you do, these should only be used for hypothesis generation because of the high risk of outcomes reporting bias. Abstract reported statistics (include standard deviations and confidence intervals) and the measurement interval. It is useful to determine a priori what will count as short-term outcome (0-3 months? 3-6 months? 3-12 months?) and what will be considered maintenance or long-term outcomes (6-12 months? 12 or more months? 24 or more months?). Reviewers should report the Ns associated with the outcomes reported, as they will likely be different from the N randomized. These analysis-specific Ns may be needed for meta-analysis
- Comments. It is useful to make a note summarizing methodological limitations and generalizability of the study to the specific research question

7. Data abstraction

Data should be carefully checked by a second reviewer and differences reconciled by discussion or by a third reviewer. Sometimes statistics will be calculated by the reviewer, either for meta-analysis purposes or to facilitate the presentation of the data (e.g., converting all event rates to the rate per 10,000). These calculations must be carefully checked. Gotzsche et al (2007) found that 37% of standardized mean difference calculations in 27 meta-analyses had data abstraction or calculation errors on at least one of two trials randomly selected quality assessment.

8. Data Synthesis

In this step you are attempting to answer questions such as:

- Is there an intervention effect?
- How large is the effect? Is it clinically meaningful? For example, a trial may report that 62% of participants in the intervention group show symptom improvement compared to only 43% in the control group, which sounds impressive. But, the same trial may report that the intervention group improved only an average of 1.5 points (on a 30-point scale) more on a symptom severity scale than the control group. This may or may not be clinically meaningful.
- How confident are you that the effect is due to the intervention of interest?
- How consistent is the effect across studies?
- Are there factors that increase or decrease the likelihood of seeing an effect?

8. Data Synthesis

Qualitative Synthesis

Often, the studies are so heterogeneous that they cannot be combined statistically. Even if heterogeneity is not an issue, you usually cannot combine all included studies; it is rare for all studies to provide comparable outcome data. If a meta-analysis is conducted, it is still important to explain how studies that are not included in the meta-analysis support or oppose the meta-analysis findings.

8. Data Synthesis

Quantitative Synthesis (meta-analysis)

The reviewer must decide if the level of heterogeneity among the studies is low enough to justify combining them in a meta-analysis. There are no clear decision rules for this; it is simply a judgment that you as the reviewer must make. Ask yourself: "Would an average of these studies be meaningful?"

This is another area where consultation with an expert is essential. It is easy to run a

meta-analysis and get a result, but there are many factors that determine whether your results are valid. For example, statistical, methodological, and clinical heterogeneity (which is the norm), the presence of rare events, and missing data all present difficulties that can have multiple solutions, but that must be handled appropriately.

We will provide a brief overview of meta-analysis. More detailed information can be found in free on-line materials developed by the Cochrane Collaboration, including the Cochrane Handbook (Higgins & Green 2008, available at http://www.cochrane-handbook.org) and the Cochrane library online open learning material (http://www.cochrane-net.org/openlearning).

Step 1: Creating the input data file. This involves deciding on one or a small number outcomes and follow-up times of interest, choosing which statistic to analyze (mean, median, odds ratios, relative risks, etc.), choosing the N to report. Be careful about reporting outcomes at vastly different times. For example, 6-week and 12-month outcomes in same analysis is suspect.

Step 2: Calculating/estimating data that is not reported. For example, a study may report a standard error, but you may need the standard deviation for meta-analysis. See the Cochrane handbook section 7.7 for some useful formulae.

Step 3: Determining whether a random or fixed effects model is preferable (see Cochrane open learning material Module 12).

Step 4: Analyze the data. In addition to running the basic meta-analyses of interest, it is also important to examine the data for publication bias. A funnel plot is commonly used for this, but other methods exist as well. (see Cochrane handbook section 10.4.1).

Exploring sources of variability/heterogeneity. Can study characteristics be identified that have an impact on the effect size? Two commonly used approaches are subgroup analysis and meta-regression.

In **subgroup analysis**, data are analyzed in a subset of the studies that share a characteristic of interest, and the effect sizes are reviewed to see if they remain consistent or if they vary substantially for different subgroups. Sensitivity analysis is essentially the same: data are re-analyzed using a subset of the studies, testing how much the results change when different inclusion/exclusion rules are applied. For example, if the data were limited to RCTs and CCTs were dropped, how would the results change?

Meta-regression is similar to general linear (or logistic) model regression, except that the unit of analysis is a study. Multiple variables can be entered into the model, representing different study characteristics of interest.

Examples of sources of variability that may be explored:

- Study quality (overall, or specific indicators)
- Time to followup
- Study characteristics (setting, recruitment source, intervention approach)
- Participant characteristics (age, presence of comorbidities)
- Publication characteristics (source, peer-reviewed vs. non-peer reviewed, English language vs. other language)

Number Needed to Treat (NNT). If a reviewer has an estimate of the difference in the risk of an outcome between treated and untreated participants (or between two treatment options), he or she may estimate the Number Needed to Treat (NNT) to benefit (or harm) one person.

The estimate of the difference in risk (the probability of the outcome) can come from a single good quality study or from the results of a meta-analysis.

NNT is calculated by taking the reciprocal of the estimated difference in risk.

Example: If 65% of the treated participants had a good outcome and 45% of the untreated participants had a good outcome, the risk difference is 20 or 0.20. The NNT to benefit would be 1/0.20 or 5. So, for every five clients treated with treatment X, one will improve more than if treatment X had not been used.

Limits to meta-analysis:

- Decisions such as how to handle heterogeneity and whether or not to combine the data in a meta-analysis can have a major impact on a review, but there are no clear "rules" for these decisions. If the methods are transparent and sufficient detail of individual studies is provided, however, readers can judge for themselves whether they agree with the reviewers' decisions
- Sometimes a single large trial will contradict a meta-analysis, which is better evidence? (Answer: It depends on the quality of the trial and the meta-analysis)
- Garbage in-garbage out: is it useful to summarize a group of poorly done studies? Meta-analysis cannot magically produce good data from poor quality studies. An approach to handling this criticism is to have a quality threshold for inclusion
- If the effect is so small that we need to combine numerous studies to find it, is that really a valuable treatment approach? This criticism highlights the importance of discussing the size and clinical relevance of an effect, rather than only its statistical significance

There are numerous resources for getting started with meta-analysis, in print and on-line media.

Assessing the quality of the body of literature

It is also important to assess the overall quality of the entire body of literature addressing each key question, and the degree to which each piece in the analytic framework or logic model are supported. The Cochrane collaboration has adopted an approach called the "GRADE" method, and is detailed in Chapters 11 & 12 of the Reviewer's Manual (http://www.cochrane-handbook.org).

9. Communication of results

Once you've found the studies, abstracted the data, and analyzed the data, it is time to write up the results. Here are some things to cover when writing up the results:

- Summarize the trial characteristics: population, care setting, intervention, comparison groups, outcomes reported, and quality of the studies
- Describe the effect (how big is the benefit? is it clinically significant?)
- Summarize information on harms (how common and serious are the harms?)
- Describe the generalizability of the studies, or their applicability to your specific question

Why go to all the trouble of using these methods?

- To ensure that your review is a rigorous, unbiased, comprehensive, transparent, reproducible review of all the relevant evidence for a question
- Decision-makers have different priorities and goals in mind. Clear methods help the decision-maker determine how well this evidence applies to their specific situation
- Systematic, transparent methods facilitate the articulation of the "best evidence" to answer a particular question. Readers may disagree with your decisions, but they can clearly see your approach and possibly re-analyze the

data you present to fit their own needs better

Controversies and challenges of systematic reviews

Insufficient evidence. Reviewers are often faced with a body of literature with few or no good quality RCTs, the studies that do exist having numerous methodological flaws, comparison conditions that make it difficult to determine absolute effectiveness of an intervention, or so much heterogeneity in outcomes, methods, and intervention components that few or no general conclusions can be drawn that are useful to clinical care. However, this is a flaw in the literature, not in SR methods. SRs can provide an important function of showing where the evidence gaps exist and methodological problems with a body of literature.

With limited time and resources it is impossible comprehensively to cover everything on a topic that is relevant, or the review will be out of date by the time it is completed.

Judgments are required to limit the scope, and the decisions can affect the results of the review and may introduce bias (but at least the biases would be clear).